Universitat de València
Doctorado en Teledetección

VNIVERSITAT DE VALÈNCIA

Low-Dimensional Representations of Earth System Processes

July 2020

Author:
Guido Kraemer

Advisors:
Miguel Mahecha
Markus Reichstein
Gustau Camps-Valls

VNIVERSITAT DE VALÈNCIA Facultat de Física

Escuela de Doctorado

Miguel Mahecha, doctor en Geografía por la Universidad de Bayreuth. Catedrático de la Universidad de Leipzig, Alemania.

Markus Reichstein, doctor en Geografía por la Universidad de Bayreuth. Director del departamento de "Biogeochemical Integration" en el Instituto Max Planck para Biogeoquimica en Jena, Alemania.

Gustau Camps-Valls, doctor en Física por la Universitat de València. Catedrático de la Universitat de València.

Hacen constar que:

Guido Kraemer ha realizdo bajo nuestra dirección el trabajo titulado *"Low-Dimensional Representations of Earth System Processes"*, que se presenta en esta memoria para optar al grado de Doctor por la Universitat de València.

Y para que así conste a los efectos oportunos, firmamos el presente certificado.

València, a _____

_____ _____

Miguel Mahecha Markus Reichstein

_____

Gustau Camps-Valls

Tesis Doctoral:	**Low-Dimensional Representations of Earth System Processes**
Autor:	*Guido Kraemer*
Directores:	*Miguel Mahecha* *Markus Reichstein* *Gustau Camps-Valls*

El tribunal nombrado para juzgar la tesis doctoral arriba citada, compuesto por los señores/as:

Presidente/a: _____

Vocal: _____

Secretario/a: _____

Acuerda otorgarle la calificación de _____

Y para que así conste a los efectos oportunos, firmamos el presente certificado.

València, a _____

Acknowledgments

The research presented in this Thesis benefited from the efforts of many people who I want to mention here. Most importantly I would like to thank my supervisors, Miguel Mahecha, Gustau Camps-Valls, and Markus Reichstein, without their help and guidance this work have been impossible. I thank Miguel for offering me a job as a student research assistant and later a position as a doctoral candidate, he also was the person that gave me most guidance, I also want to thank him for all the positive energy that provided the basis for a great research environment and for the freedom to pursue my own ideas. The vision for this research project came from Miguel and Markus, therefore I want to thank them for trusting me with it, as well as for countless very productive discussions. I also want to thank Gustau for his hospitality at his research group in Valencia and for providing lots of methodological input as well as positive energy and enthusiasm.

A special thank goes to the members of the Empirical Inference of the Earth System group members at the Max Planck Institute for Biogeochemisty in Jena who provided lots of help and discussions. This work benefited especially from the work of Fabian Gans, whose efforts simplified much of data analysis that went into this Thesis. I thank Jacob Nelson and Andrew Durso for proofreading some of the manuscripts. I also want to thank the members of the Image and Signal Processing Group at the University of Valencia for their time and input, especially Valero Laparra for the helpful discussions.

Finally I want to thank my family: My wife for all the support provided during my time as a doctoral candidate and before, her support made everything much easier and allowed me to focus more on my research. My parents for supporting me thoughout my studies, and my father for proofreading some of the manuscripts. Without you I would not be where I am today.

Contents

Abstract

In times of global change, we must closely monitor the state of our planet in order to understand gradual or abrupt changes early on. In fact, each of the Earth's subsystems—i.e. the biosphere, atmosphere, hydrosphere, cryosphere, and anthroposphere—can be analyzed from a multitude of data streams. However, since it is very hard to jointly interpret multiple monitoring data streams in parallel, one often aims for some summarizing indicator. Climate indices, for example, summarize the state of atmospheric circulation in a region, e.g. the Multivariate ENSO (El Ñino-Southern Oscillation) Index. Indicator approaches have been used extensively to describe socioeconomic data too, and a range of indices have been proposed to synthesize and interpret this information. For instance the "Human Development Index" (HDI) by the United Nations Development Programme (UNDP, 2016) was designed to capture specific aspects of development.

"Dimensionality reduction" (DR) is a widely used approach to find low dimensional and interpretable representations of data that are natively embedded in high-dimensional spaces. Here, we propose a robust method to create indicators using dimensionality reduction to better represent the terrestrial biosphere and the global socioeconomic system. We aim to explore the performance of the approach and to interpret the resulting indicators.

For biosphere indicators, the concept was tested using 12 explanatory variables representing the biophysical states of ecosystems and land-atmosphere water, energy, and carbon fluxes. We find that two indicators account for 73% of the variance of the state of the biosphere in space and time. While the first indicator summarizes productivity patterns, the second indicator summarizes variables representing water and energy availability. Anomalies in the indicators clearly identify extreme events, such as the Amazon droughts (2005 and 2010) and the Russian heatwave (2010), they also allow us to interpret the impacts of these events. The indicators also reveal changes in the seasonal cycle, e.g. increasing seasonal amplitudes of productivity in agricultural areas and in arctic regions.

We also apply the method on the "World Development Indicators" (WDIs; The World Bank, 2018a), a database with more than 1500 variables, to track the socioeconomic development at a country level. The aim was to extract the core dimensions of development in a highly efficient way,

using a method of nonlinear dimensionality reduction. We find that over 90% of variance in the WDIs can be represented by five uncorrelated and nonlinear dimensions. The first dimension (explaining 74%) represents the state of education, health, income, infrastructure, trade, population, and pollution. The second dimension (explaining 10%) differentiates countries by gender ratios, labor market, and energy production patterns. Overall, we find that the data contained in the WDIs are highly nonlinear therefore requiring nonlinear methods to extract the main patterns of development. Globally, most countries show rather consistent temporal trends towards wealthier and aging societies. Deviations from the long-term trajectories are detected with our approach during warfare, environmental disasters, or fundamental political changes.

In general we find that the indicator approach is able to extract general patterns from complex databases and that it can be applied to databases of varying characteristics. We also find that indicators are can different kinds of changes occurring in the system, such as extreme events, permanent changes or trends. Therefore it is a useful tool for general monitoring and exploratory data analysis. The approach is flexible and can be applied to complex datasets, such as large data, nonlinear data, as well as data with many missing values.

Resumen

La Tierra es un sistema muy complejo, dinámico, e interconectado. Su estudio requiere del uso de técnicas de monitorización y procesado avanzado de datos. En el escenario actual de cambio climático se hace si cabe más necesaria y urgente la monitorización del estado del planeta mediante el seguimiento y estimación de variables climáticas esenciales (ECV por sus siglas en inglés). Pero tal vez más importante que la propia *estimación* de las ECVs como diagnóstico del estado del planeta, resulta de gran interés *entender* los mecanismos y procesos subyacentes, los cambios graduales o abruptos que se producen, y las interrelaciones en el 'sistema Tierra'. De hecho, cada uno de los subsistemas de la Tierra, es decir, la biosfera, la atmósfera, hidrosfera, criosfera, y antroposfera, pueden ser analizadas desde una multitud de flujos de datos. Todas estas esferas están relacionadas, y no se puede entender una sin las otras. Sin embargo, dado que es muy difícil interpretar conjuntamente múltiples variables en paralelo, se busca resumir el sistema a través de pocos indicadores. Por ejemplo, los índices de vegetación se emplean abundantemente en la literatura de teledetección para resumir el estado de bosques y cultivos. Asimismo, los índices climáticos, resumen el estado de la circulación atmosférica en una región. Cuando hablamos de la antroposfera, se han empleado una gran multitud de índices para describir los aspectos socio-económicos, y se han propuesto muchos índices para sintetizar e interpretar esta información. Por ejemplo, el "Índice de Desarrollo Humano" (Human Development Index, HDI; UNDP, 2016) fue diseñado para captar aspectos específicos del desarrollo. Sin embargo, una cuestión pendiente es si el HDI y los indicadores relacionados capturan el desarrollo en su totalidad. Aunque estos enfoques también se utilizan en otros campos de la ciencia, rara vez se utilizan para describir la dinámica de la superficie terrestre.

En esta Tesis Doctoral, proponemos un método robusto para crear indicadores para el sector terrestre la biosfera y el sistema socioeconómico mundial utilizando técnicas de aprendizaje estadístico conocidas como 'reducción de la dimensionalidad' (dimensionality reduction, DR). Aplicaremos estos métodos sobre grandes cantidades de datos globales de alta dimensionalidad: tanto variables esenciales climáticas como variables socioeconómicas. Nuestro objetivo final es resumir el contenido informativo en un subconjunto de componentes esenciales (es decir, unos indicadores

multidimensionales). Para ello, exploraremos el rendimiento de distintas técnicas e indicadores lineales y no lineales, evaluaremos su poder de compresión y estudiaremos e interpretaremos esos ejes principales que definen el subespacio acoplado de biosfera-antroposfera.

Para los indicadores biosféricos, el concepto se evaluó utilizando 12 variables explicativas que representan los estados biofísicos de los ecosistemas y los flujos tierra-atmósfera de agua, energía y carbono. Encontramos que dos indicadores representan el 73 % de la varianza del estado de la biosfera en el espacio y en el tiempo. Mientras que el primer indicador resume los patrones de productividad, el segundo indicador resume las variables que representan la disponibilidad de agua y energía. Las anomalías en los indicadores identifican claramente los eventos extremos, como sequías en la Amazonía (2005 y 2010) o la ola de calor en Rusia (2010), también nos permiten interpretar los impactos de estos eventos. Los indicadores también revelan cambios en el ciclo estacional, por ejemplo, un aumento de amplitudes estacionales y de la productividad en las zonas agrícolas y en las regiones árticas.

En cuanto a los indicadores socioeconómicos, empleamos los "Indicatores de Desarrollo Mundial" (World Development Indicators, WDI) publicado por el Banco Mundial, una base de datos con más de 1500 variables (The World Bank, 2018a) para hacer un seguimiento de la situación socioeconómica desarrollo a nivel de país. La intención aquí es extraer las dimensiones centrales del desarrollo de una manera altamente eficiente y no lineal. Encontramos que más del 90 % de la varianza de 621 WDI puede ser representada por únicamente cinco dimensiones no correlacionadas y no lineales. La primera dimensión (que explica el 74 %) representa el estado de la educación, la salud, ingresos, infraestructura, comercio, población y contaminación. La segunda dimensión (que explica el 10 %) diferencia a los países por proporcion de géneros, mercado laboral y patrones de producción de energía. En general, encontramos que los datos contenidos en el WDI son altamente no lineales por lo que se requieren métodos no lineales para extraer los principales patrones de desarrollo. A nivel mundial, la mayoría de los países muestra tendencias temporales bastante consistentes hacia sociedades más prósperas y envejecidas. Las desviaciones de las tendencias generales de una trayectoria se detectan con la metodología propuesta durante guerras, desastres ambientales o cambios políticos fundamentales.

Las implicaciones de este trabajo son abundantes. Resumir la ingente cantidad de información y variables de monitorización del sistema Tierra en las mínimas componentes explicativas de las distintas esferas resulta esencial para comprender, adaptarse y mitigar los efectos de los cambios climáticos antropogénicos.

Chapter 1

Introduction

Content

1.1 Changes in the Earth System

Human activity causes unprecedented changes to the Earth, especially the biosphere. The total impact of anthropogenic climate change is far from being understood yet but is already enough to have the magnitude of a mass extinction event (Ripple et al., 2017; Ceballos and Ehrlich, 2018; IPBES, 2019) and will only increase in the future (IPCC, 2019). These impacts can manifest in different ways:

1. As slow phase shifts, e.g. increasing greenhouse gas concentrations, changes in mean temperatures, increased nitrogen deposition, rising sea levels, desertification, large scale ecosystem greening, pollutants accumulating in ecosystems, and ocean acidification.

2. Rapid, but permanent, changes to the ecosystem state, e.g. human made land use change (Khanna et al., 2017), extractive exploitation, ecosystem tipping points due to changing weather patterns and climate change (Lenton et al., 2008).

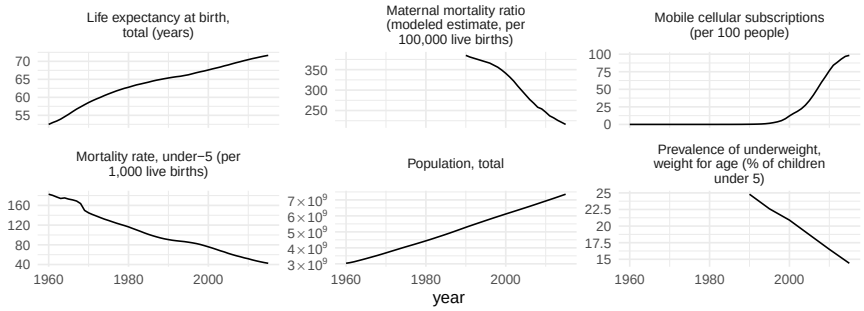

Figure 1.1: Better lives for all human beings have increased life expectancy and decreased mortality (source: The World Bank, 2018a).

3. Extreme events, abrupt but reversible changes, are becoming more frequent due to climate change, e.g. tropical cyclones (Easterling et al., 2000), heat waves and droughts, extreme rainfall events.

Not all occurring changes are bad; in the last decades humanity has made huge progress toward better lives for all human beings: Poverty and hunger have been reduced, despite an increasing population. Today's medicine can prevent or cure many illnesses that used to be fatal and has increased life expectancy dramatically, especially the mortality of mothers and infants has dropped significantly. Resolving many of these basic problems of humanity has caused an enormous increase in human population (see fig. 1.1) causing even more pressure on the Earth's ecosystems (UNDP, 2016; IPCC, 2019).

As the pressure of humanity on ecosystems increases, so does the need for tools to not only monitor the changes happening in ecosystems but also to monitor economic development. The monitoring tools should not only be able to detect a single type of impact but a broad range of the changes that can occur. Therefore, we require monitoring tools that are flexible enough to detect impacts in different types of systems, e.g. the socioeconomic systems *and* the biosphere, as well as different kinds of impact, e.g. slowly occurring trends, sudden extreme events and sudden changes in ecosystem state.

1.2 How Do We Observe the Earth?

Another positive change is the increasing amount of data collected by humanity that gives us a more complete picture of the state of the Earth

than ever before: New satellite missions have been launched, more ground based data are being collected and Earth system models produce more and more data (Overpeck et al., 2011) also data is being digitized by citizens (See et al., 2016). On the socioeconomic side, there are many initiatives to collect new data (The World Bank, 2018a) and reprocess existing data (e.g. Smits and Permanyer, 2019). From a standpoint of data we are better equipped than ever to monitor the Earth.

Earth sciences divide the earth into subsystems, usually referred to as "spheres" because of their shape. There is a large number of these spheres, such as the "atmosphere", which refers to the gaseous layer between the surface of the earth and space. In this work we look at the "biosphere", which constitutes all organic life on earth and the "anthroposphere", which is the human equivalent of the biosphere and constitutes everything that is made by or modified by humans (Bonan, 2015).

1.2.1 Biosphere

When analyzing global biosphere data, a number of different kinds of products are being used. At first the observation of vegetation by satellites was done using simple derived products, such as the Normalized Difference Vegetation Index (NDVI; Becker and Choudhury, 1988). Over time sensors became more specialized and sophisticated and the resulting products much better at measuring specific properties of ecosystems, such as fluorescence (Ryu et al., 2019) and canopy structure (Mathieu et al., 2013).

In addition to satellite observations, large networks to collect ground based information have been created (e.g. FLUXNET; Baldocchi, 2020), measuring properties of ecosystems such as carbon fluxes on the ground. Together with satellite observations, these measurements can be upscaled to create global datasets estimating ecosystem functions which otherwise would be impossible to observe with satellites (Bodesheim et al., 2018).

When direct observations of the system and underlying driver processes are not possible, we resort to simulations. Earth system models can be used to model properties which we cannot observe directly (Smith et al., 2001). These models typically use known physical and physiological processes to model the behavior of vegetation. Examples of processes that are hard to observe directly, are processes happening below ground, e.g. root-zone soil moisture (Martens et al., 2017) is a property derived from such a process, or empirical relations, which contain parameters such as respiration coefficients (van't Hoff, 1898).

Monitoring of biospheric data using advanced methods has come quite far. Current monitoring systems can usually detect a single type of change,

be it extremes (temporary deviations from the mean seasonal cycle, see e.g. Flach et al. 2017), breakpoints (permanent, abrupt changes to the mean seasonal cycle, see e.g. Verbesselt et al. 2010) or trends (slowly and steadily accumulating changes; Murthy and Bagchi, 2018). Many methods operate only on single variables (e.g. Alexander et al. 2006; Zhou et al. 2011) but observing a single variable may often not be enough to observe an extreme (Zscheischler et al., 2014) and therefore multiple covariates have to be observed at once (Flach et al., 2017).

Satellite data and model data are being produced in quantities that it is called "a deluge of Earth system data" (Reichstein et al., 2019). Together with the amount of data, computational power has also increased significantly allowing the development of ever more complex machine learning approaches. Interpretability is one of the main challenges when dealing with complex models (Runge et al., 2015; Chalupka et al., 2017; Montavon et al., 2018) and is a necessity if we want to use machine learning to further our understanding of the Earth system (Reichstein et al., 2019).

1.2.2 *Anthroposphere*

The Earth is a coupled system and society as one of the interacting spheres plays a key role (both a cause and effect) if we want to understand the Earth system as a whole. The observation of the socioeconomic development of countries is very different from the observations of natural processes. Only very little data can be gathered from satellite observations (these data come from integrating satellite data with national and subnational statistics or point observations and include population density, demographics, population counts, and Gross Domestic Product (GDP), but also wealth, health, education and development indicators; CIESIN, 2018; Yetman et al., 2010; Smits and Permanyer, 2019). On a global scale data are usually collected at a country level and most variables are only available at a yearly resolution.

In contrast to biospheric variables, there are many more variables describing the different facets of development; these variables are collected in the "World Development Indicators" database (The World Bank, 2018a). The monitoring of social development is much more difficult, mainly because no canonical measure exists and the meaning of "development" has changed significantly over time.

Originally the term development was purely economical and its main measure was the economic growth of a country, taking GDP as the basic indicator. Since the 1960s the concept of development was expanded, and economic growth started to be seen as only one of the aspects of development (Stanton, 2007), therefore development started to be measured using composite indicators, i.e. abstract magnitudes integrating over several

variables representing the desired property of a system. In 1990, the Human Development Index (HDI, last updated version: UNDP, 2019) was created by the United Nations Development Programme (UNDP) integrating health (in the form of life expectancy), education (by school enrollment and literacy rates), and standard of living (by per capita income) into a single indicator. The HDI was a major milestone in the adoption of composite indicators but also attracted wide criticism as "conceptually weak and empirically unsound" (Srinivasan, 1994). Despite the criticism, the general concept of creating indicators was widely adopted, to the point that today we have hundreds of composite indicators (Parris and Kates, 2003; Shaker, 2018; Ghislandi et al., 2018), each one providing improvements over previous ones (e.g. inequality adjusted variants of the HDI) or different specialized aspects of development, such as gender inequality (UNDP, 2016) or the ecological footprint (Wackernagel et al., 1999).

The mathematical expressions to create composite indicators are usually decided by experts, therefore they are often criticized as being subjective (Shaker, 2018). To address this criticism, multivariate methods, mostly Principal Component Analysis (PCA), started being used to create indicators from a pre-selected set of variables which jointly represent the desired properties of the final indicator (OEDC, 2008). This approach has also been criticized on the grounds of underrepresenting important indicators, not being robust to outliers, ignoring the polarization of indicators and being difficult to align with handmade indicators (Mazziotta and Pareto, 2019).

In Chapter 4 (Kraemer et al., 2020b) we followed a purely data driven approach based on modern machine learning techniques of dimensionality reduction. We explore the development space, assigning meaning to the resulting indicators *after* the creation of the indicators. In this analysis it became clear that the underlying development data is highly nonlinear and therefore a Principal Component Analysis is not enough to adequately represent the data in few dimensions. One of the main difficulties of analyzing these data was the large fraction of missing values which causes difficulties applying standard methods of nonlinear dimensionality reduction.

1.3 Indicator Approaches

There are two words describing the concept in the English language: *Index* with plural *indices* (the other plural, *indexes*, is not commonly used) and *indicator* with the regular plural *indicators*. In general, natural sciences use the word index and the plural indices, e.g. in Multivariate ENSO Index, Leaf Area Index, Normalized Difference Vegetation Index (cf. fig. 1.2). In

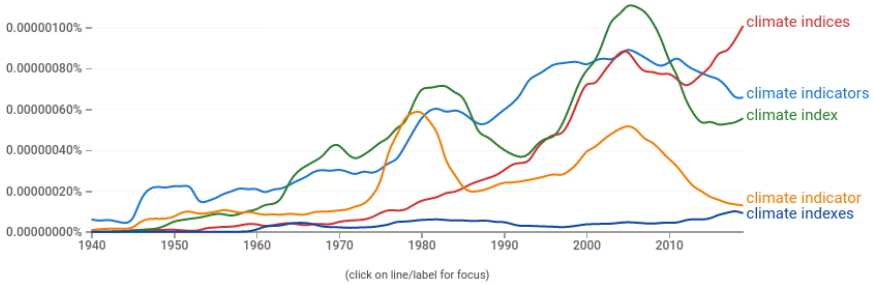

Figure 1.2: The use of terms referring to climate indices and indicators over time in the Google Books data base. We see that "index" and its associated plural "indices" are more commonly used than "indicator". The other indicator of "index", "indexes", is hardly used. Source: Google (2020)

social sciences a composite index generally is the function of a number of indicators: The World Development Indicators is a database of mostly non-composite indicators, while the Human Development Index is a composite index. But the line between the two is very blurry: One component of the Human Development Index is a "life expectancy index", which itself is a linear transform of the indicator "life expectancy at birth" and therefore not composite (UNDP, 2019). To create a mega-index, Shaker (2018) uses 31 indicators, which all are composite indices in itself and 26 have the word "Index" in their name, while 2 have the word "Indicator" in their name. Because of this large overlap between these concepts, in this thesis the words "indicator" and "index" will be used interchangeably.

Remote sensing is the acquisition of information about an object from afar. Remote sensing can be used to estimate vegetation properties. Because the reflective properties of different surfaces are known, we can combine different bands parametrically to calculate vegetation indices (VI) that try to model physical properties; this has been done extensively (Camps-Valls et al., 2011). Examples for such indicators include the Leaf Area Index (LAI), or Fractional Vegetation Cover (FVC), leaf Chlorophyll content (*Chl*) (Whittaker and Marks, 1975), and the NDVI (Rouse et al., 1973).

Climate indicators usually try to describe phenomena which are important to the global circulation. For example there are a variety of indicators describing the El Niño Southern Oscillation (ENSO), the most important coupled ocean-atmosphere phenomenon (Wolter and Timlin, 2011b). One way to create such indicators consists in using the first principal component

of fields of variables, such as sea surface temperature or sea level pressure over a region of the equatorial Pacific. These indicators use dimensionality reduction to reduce space (and variables if there is more than one) and keep the time dimension. Other important climate indicators calculated in a similar fashion include the North Atlantic Oscillation index, the Arctic Oscillation index, and the Antarctic Oscillation index.

Usually vegetation and climate classifications separate reasonably well in the vector space (climate space) spanned by temperature and precipitation (Köppen and Geiger, 1954; Kottek et al., 2006; Papagiannopoulou et al., 2018). Because one of the drivers for vegetation is climate, it is to be expected that variables representing vegetation allow a similar degree of separation using two components. Although this is not a direct application of dimensionality reduction, clustering and dimensionality reduction are similar in that both reduce input features: In the case of dimensionality reduction, the output is a number of continuous features while in the case of clustering the output consists in a number of discrete classes or homogeneous groups. Spatial classifications reduce over variables and time (in the form of the mean seasonal cycle) and keep only the spatial dimensions (in the form of a class per spatial pixel).

When observing many data streams, there will be redundancies in the data, e.g. different measures for GDP may be adjusted to inflation, to money exchange rates, to living costs, etc. Other measures that will be correlated with GDP are measures of poverty, measures for infrastructure, among others. All of these measures will covary strongly and even though they measure different aspects (which all are important in their own right) combined, these data will still contain large redundancies. The question is: *What are the redundancies and what are the independent dimensions?*

The natural way to address this question comes from multivariate statistics: Dimensionality reduction describes a family of multivariate methods that find alternative representations of data by constructing linear, or non-linear, combinations of the original variables so that important properties are maintained in as few dimensions as possible.

It should be noted that in Earth sciences there are a number of ways to reduce the dimensionality of data. We usually encounter a grid with dimensions space, time, and sometimes variables, the dimension space comes either in the form of a latitude and longitude grid or discrete spatial units, such as ecosystems or countries. Dimensionality reduction methods take a matrix and reduces the number of rows in the matrix while maintaining the number of columns, therefore we have to matricize the higher dimensional tensor by combining axes. There is only a limited number of ways this can be done with a tensor of order 3 or 4 and only a few of these combinations have been explored, cf. tab. 1.1. Here we

Table 1.1: Overview of ways to reduce the dimensionality of Earth observation data. While the System State Indicator and Empirical Orthogonal Functions return continuous components, climate classifications return discrete classes.

Method	Reduces over	Keeps
Climate classification	Time, variables	Space
Empirical Orthogonal Functions (space)	Time	Space
Empirical Orthogonal Functions (time)	Space	Time
System State Indicator	Variables	Time, space

propose a different way of matricizing the data by combining the space and time dimensions while reducing only over the variables.

1.4 Dimensionality Reduction and the System State Indicator

In this Thesis we propose a System State Indicator (SSI), a method that tracks the state of the elements of a complex and multivariate system over time and is explicit in space. In order to achieve this we apply dimensionality reduction in a distinct way to previous approaches. The SSI allows us to monitor and detect different kinds of events on any variable. A trajectory tracks the position of a spatial observation unit (a country or pixel) over time in an abstract space of reduced dimensionality which represents the data in high-dimensional space of observed variables faithfully.

Dimensionality reduction is a uniquely suited tool to create the monitoring indicators described above. If we observe a single object, be it a spatio-temporal pixel or a country over time with enough data streams, there will inevitably be redundancies in the data. These redundancies will cause the data to not fill the data space uniformly, but the observations will live on a manifold of lower dimensionality than the original space. Dimensionality reduction tries to find low-dimensional embedding of this data. We can now represent this manifold in a space of lower dimensionality, ideally of the same dimensionality as the manifold itself and describe the position of our object inside the manifold. This allows us to represent the position of the object inside our system faithfully in a low-dimensional space and therefore is optimally suited for the indicator approach presented in this Thesis.

There is a number of issues to consider, which complicates the creation of an SSI. The simplest method of dimensionality reduction is PCA which results in a linear transformation of the data. The simplicity comes at the advantage that PCA is relatively fast to learn and simple to apply, but it cannot deal with nonlinear relations. More complex methods are computationally much more expensive to train, e.g. often an eigenvalue decomposition of an $n \times n$ matrix, where n is the number of observations, is necessary or an expensive optimization has to be performed. Nonlinear methods also often require tuning several parameters, which requires retraining and makes finding an adequate model even more difficult.

Within linear methods, PCA is the canonical method for dimensionality reduction, but when it comes to nonlinear methods, there is no standard method and therefore the researcher has to choose a method from a large pool of existing methods (or develop a new method). There is a number of other difficulties when picking a nonlinear method for dimensionality reduction, which is why we created the **dimRed** package in the R language to aid the investigator with choosing the right method (Chapter 2; Kraemer et al., 2018).

Most importantly, there is no canonical measure to compare the goodness of fit of different methods for dimensionality reduction (we revise some methods to measure quality in Section 2.3) which makes the comparison of methods very difficult. The training of many nonlinear methods relies on non-convex optimization and therefore solutions may not be stable and a successful training may require several attempts. Other limitations include the lack of readily available and well tested implementations for methods. Often the publication of a method is not accompanied by an implementation that is easy to use by other people and therefore replication has to be accompanied by a reimplementation. Another important factor for the application on real world data is the ability of the method to deal with missing data because real world observations usually contain missing values, e.g. in Chapter 4 we implement an extension of Isomap (Tenenbaum et al., 2000) to cope with the sparseness of the input data.

If we have two models of dimensionality reduction that can represent the same amount of information of the original data in the same amount of dimensions, and one of the models is simpler than the other (e.g. PCA) then, following the principle of Ockham's Razor, we should choose the simpler model. This is the case in the analysis presented in Chapter 3 where PCA resulted to be sufficient for reducing the dimensionality of the dataset. PCA also provides a number of other benefits which are discussed in Chapter 3.

When choosing a nonlinear model, there are certain considerations to be made. As we are assuming that the data lies on a manifold of low

dimensionality inside the feature space, we want to use a method that preserves the intrinsic geometry of the manifold. A method that just preserves local neighborhoods (e.g. t-SNE) or large distances (e.g. PCA) but otherwise does not maintain the general structure of the manifold may not be a good choice. In Chapter 4 we show that Isomap can be a good choice to create indicators, as it tries to find an embedding of the manifold by unfolding it and preserving its internal Euclidean structure.

1.5 Objectives

The overarching goal of the Thesis can be stated as:

> *"Learn the intrinsic dimensionality of the biosphere and anthroposphere from data using advanced machine learning techniques."*

To attain this goal, we have defined a set of specific objectives:

1. *Find the dimensionality of the system.* Here we ask the question: How many dimensions are necessary to accurately describe the system?
2. *Find the dominant dimensions of the covariates describing the spheres of the Earth system and analyze the characteristics of the resulting components.* We apply methods of dimensionality reduction to real world global datasets and analyze and interpret the resulting components by looking at the covariates encoded into the components. This helps us to understand the most important dimensions of the system.
3. *Find global patterns using the resulting indicators.* We analyze how the objects are distributed in the space of reduced dimensionality. We see which patterns can be found and give an additional way to characterize the system.
4. *Use the resulting trajectories to characterize the observed objects.* Each object (spatial pixel or country) is described by time series of resulting indicators, just as the observed objects are described by time series of covariates. We analyze the trajectories of the observed objects in reduced space in terms of their relative positions, their direction and the encoded information in the time series of indicators to characterize properties of the global system and the observed objects.
5. *Find the changes and extremes described by the indicators.* We analyze how extreme events and other important changes are encoded in the

time series of indicators and how these changes reflect changes in a local ecosystem or a country.

1.6 Thesis Organization

Chapter 2 (Kraemer et al., 2018) discusses the proper way to apply dimensionality reduction to real world data and how to choose the right method, providing the basis for the creation of the indicators.

Chapters 3 (Kraemer et al., 2020a) and 4 (Kraemer et al., 2020b) contain the applications of the indicator framework onto real world data. In Chapter 3 we applied the SSI method to the biospheric variables of the Earth System Data Cube (Mahecha et al., 2019) to explore the global biosphere. In Chapter 4, we applied the SSI method on the World Development Indicators (The World Bank, 2018a), a key data source for global development data to explore development space.

Chapter 5 contains the concluding remarks and perspectives of the Thesis, as well as the achievements of the author reached during the doctoral studies.

Chapter 2

Unifying Dimensionality Reduction Methods

Content

This chapter is based on the following publication:

Kraemer, G., Reichstein, M., and Mahecha, M. D. (2018). dimRed and coRanking—Unifying Dimensionality Reduction in R. *The R Journal*, 10(1), 342–358. doi:10.32614/RJ-2018-039

Abstract

"Dimensionality reduction" (DR) is a widely used approach to find low dimensional and interpretable representations of data that are natively embedded in high-dimensional spaces. DR can be realized by a plethora of methods with different properties, objectives and hence (dis)advantages, resulting low-dimensional data embeddings are often difficult to compare with objective criteria. Here, we introduce the **dimRed** and **coRanking** packages for the R language. These open source software packages enable users to easily access multiple classical and advanced DR methods using a common interface. The packages also provide quality indicators for the embeddings and easy visualization of high-dimensional data. **coRanking** provides the functionality for assessing DR methods in the co-ranking matrix framework. In tandem, these packages allow for uncovering complex structures high dimensional data. Currently 15 DR methods are available in the package, some of which were not previously available to R users. Here, we outline the **dimRed** and **coRanking** packages and make the implemented methods understandable to the interested reader.

2.1 Introduction

Dimensionality Reduction (DR) essentially aims to find low dimensional representations of data while preserving their key properties. Many methods exist in literature, optimizing different criteria: maximizing the variance or the statistical independence of the projected data, minimizing the reconstruction error under different constraints, or optimizing for different error metrics, just to name a few. Either way choosing an inadequate method may imply that much of the underlying structure remains undiscovered. Often the structures of interest in a dataset can be well represented by fewer dimensions than existing in the original data. Data compression of this kind has the additional benefit of making the encoded information better conceivable to our brains for further analysis tasks like classification of regression problems.

There are a number of software packages that provide collections of methods: In Python there is scikit-learn (Pedregosa et al., 2011) which

contains a module for DR, in Julia we currently find ManifoldLearning.jl for nonlinear and MultivariateStats.jl for linear DR methods. There are several toolboxes for DR implemented in Matlab (Van Der Maaten et al., 2009; Arenas-Garcia et al., 2013) and the Shogun toolbox (Sonnenburg et al., 2010) implements a variety of methods for dimensionality reduction in C++ and offers bindings for a many common high level languages (including R, but the installation is everything but simple, i.e. there is no CRAN package). However, there is no comprehensive package for R and none of the former mentioned software packages provides means to consistently compare the quality of methods for DR.

For many applications it can be difficult to objectively find the right method or parameterization for the DR task. This chapter presents **dimRed** and **coRanking**, both are software package in the popular programming language R (R Core Team, 2016) and provide a standardized interface to dimensionality reduction methods and quality metrics for embeddings using the S4 class system making the packages both easy to use and to extend.

The goal is to enable researchers who may not necessarily be experts in DR to apply the methods in their own work to get and objectively identify suitable methods. This chapter aims to an overview of the methods collected in the package and how to use the packages.

The notation in this Thesis is as follows (unless specified otherwise in the text): The total dataset of observations is the matrix $\mathbf{X} = [\mathbf{x}_1 | \dots | \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, containing the observations $\mathbf{x}_i \in \mathbb{R}^d$. The observations may be centered and standardized to unit variance. A DR method then embeds each observation $\mathbf{x}_i$ onto $\mathbf{y}_i \in \mathbb{R}^p$, a vector containing the corresponding values of the indicators. The dataset of resulting indicators is $\mathbf{Y} = [\mathbf{y}_1 | \dots | \mathbf{y}_n] \in \mathbb{R}^{p \times n}$, ideally we expect $p \ll d$.

Some methods provide an explicit mapping $f(\mathbf{x}_i) = \mathbf{y}_i$ and some even offer an inverse mapping $f^{-1}(\mathbf{y}_i) = \hat{\mathbf{x}}_i$, such that one can reconstruct a (usually approximate) sample from the low-dimensional representation. For some methods pairwise distances between points are needed, we set $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ and $\hat{d}_{ij} = d(\mathbf{y}_i, \mathbf{y}_j)$, where d is some appropriate distance function and the the corresponding distance matrices are $\mathbf{D} = [d_{ij}] \in \mathbb{R}^{n \times n}$ and $\widehat{\mathbf{D}} = [\hat{d}_{ij}] \in \mathbb{R}^{n \times n}$.

When referring to `functions` in the **dimRed** package or base R simply the function name is mentioned, functions from other packages are referenced with `package::function`.

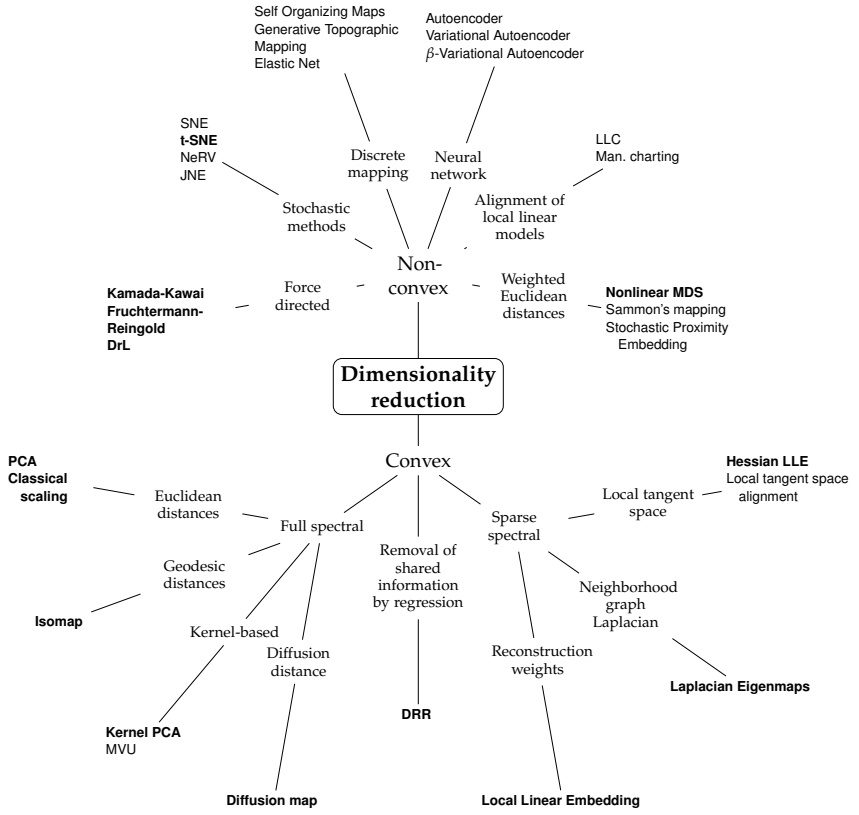

Figure 2.1: Classification of dimensionality reduction methods, methods in bold face are implemented in **dimRed**. Modified from Van Der Maaten et al. (2009).

2.2 *Dimensionality Reduction Methods*

In the following section we do not aim for an exhaustive explanation to every method in **dimRed** but rather to provide a general idea on how the methods work.

In all methods, parameters have to be optimized or decisions have to be made, even if it is just about the preprocessing steps of data. **dimRed** tries to make the optimization process for parameters as easy as possible, but if possible the parameter space should be narrowed down using prior knowledge. Often decisions can be made based on theoretical knowledge,

e.g. sometimes an analysis requires data to be kept in their original scales and sometimes this is exactly what has to be avoided (e.g. when comparing different physical units). Sometimes decisions based on the experience of others can be made, e.g. the Gaussian kernel is probably the most universal kernel and therefore should be tested first if there is a choice.

All methods presented here have the embedding dimensionality, q, as a parameter (or `ndim` as a parameter for `embed`). For methods based on eigenvector decomposition, the result generally does not depend on the number of dimensions, i.e. the first dimension will be the same, no matter if we decide to calculate only two dimensions or more. If more dimensions are added, more information is maintained, the first dimension is the most important and higher dimensions are successively less important. This means, that a method based on eigenvalue decomposition only has to be run once if one wishes to compare the embedding in different dimensions. In optimization based methods that use gradient descend this is generally not the case, the number of dimensions has to be chosen a priori, an embedding of 2 and 3 dimensions may vary significantly, and there is no ordered importance of dimensions. This means that comparing dimensions of gradient descend based methods is computationally much more expensive.

We try to give the computational complexity of the methods but because of the actual implementation computation times may differ largely. R is an interpreted language, so all parts of an algorithm that are implemented in R often will tend to be slow compared to methods shiped with efficient implementations in a compiled language. Methods where most of the computing time is spent for eigenvalue decomposition do have very efficient implementations because R uses optimized linear algebra libraries, although eigenvalue decomposition itself does not scale very well in naive implementations ($\mathcal{O}(n^3)$).

2.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is the most basic technique for reducing dimensions and dates back to Pearson (1901). PCA finds a linear projection ($\mathbf{V}$) of the high-dimensional space into a low-dimensional space $\mathbf{Y} = \mathbf{VX}$, maintaining maximum variance of the data. It is based on solving the following eigenvalue problem:

$$(\mathbf{Q} - \lambda_k \mathbf{I}) v_k = 0 \tag{2.1}$$

where $\mathbf{Q} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ is the covariance matrix, λ_k and $\mathbf{v}_k$ are the k-th eigenvalue and eigenvector, and $\mathbf{I}$ is the identity matrix. The equation has

several solutions for different values of λ_k (leaving aside the trivial solution $\mathbf{v}_k = 0$). PCA can be efficiently applied to large datasets, because it computationally scales $\mathcal{O}(nd^2 + d^3)$, i.e. it scales linearly with the number of samples and R uses specialized linear algebra libraries for such kind of computations.

PCA is a rotation around the origin and there exist a forward and inverse mapping. PCA may suffer from a scale problem, i.e. when one variable dominates the variance simply because it is in a higher scale, to remedy this the data can be scaled to zero mean and unit variance, it depends on the use case if this is necessary or desired.

Base R implements PCA in the functions `prcomp` and `princomp`; but several other implementations exist i.e. **pcaMethods** from Bioconductor which implements versions of PCA that can deal with missing data. The **dimRed** package wraps around `prcomp`.

2.2.2 Kernel Principal Component Analysis

Kernel Principal Component Analysis (kPCA) extends PCA to deal with nonlinear dependencies among variables. The idea behind kPCA is to map the data into a very high-dimensional feature space using a possibly nonlinear function ϕ and to perform a PCA in feature space. Some mathematical tricks are used for efficient computation.

If the rows of $\mathbf{X}$ are centered around 0, then the principal components can also be computed from the inner product matrix $\mathbf{K} = \mathbf{X}^T\mathbf{X}$. Due to this way of calculating a PCA, we do not need to explicitly map all points into feature space and do the calculations there, it is enough to obtain the inner product matrix or kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ of the mapped points (Schölkopf et al., 1998). This is called the "kernel trick".

Here is an example calculating the kernel matrix using a Gaussian kernel,

$$k_{ij} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) = \kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \tag{2.2}$$

where σ is a length scale parameter accounting for the width of the kernel. The other trick used is known as the "representer theorem" and the interested reader is referred to Schölkopf et al. (2001).

kPCA is very flexible and there exist many kernels for special purposes, the most common kernel function is the Gaussian kernel (eq. 2.2). The flexibility comes at the price that the method has to be finely tuned for the dataset because some parameter combinations are simply unsuitable for certain data. The method is not suitable for very large datasets, because memory scales with $\mathcal{O}(n^2)$ and computation time with $\mathcal{O}(n^3)$.

Diffusion Maps, Isomap, Locally Linear Embedding and some other techniques can be seen as special cases of kPCA (Ham et al., 2004) and an out-of-sample extension using the Nystöm formula can be applied (Bengio et al., 2004). This can also yield applications for bigger data, where an embedding is trained with a sub-sample of all data and then the data is embedded using the Nyström formula.

Kernel PCA in R is implemented using the function `kernlab::kpca`, and supports a number of kernels and user defined functions, for details see `kernlab::kpca`.

The **dimRed** package wraps around `kernlab::kpca` but additionally provides forward and inverse (Bakir et al., 2004) methods which can be used to fit out-of sample data or to visualize the transformation of the data space.

2.2.3 Classical Scaling

What today is called classical scaling (cMDS) was first introduced by Torgerson (1952) and uses an eigenvalue decomposition of a transformed distance matrix to find an embedding that maintains the distances between observations. The method works because of the same reason that kPCA works, i.e. classical scaling can be seen as a kPCA with the linear kernel, $\kappa(x_i, x_j) = \mathbf{x}_i^T \mathbf{x}_j$. A matrix of squared Euclidean distances can be transformed into an inner product matrix using double centering[1] and therefore yields the same result as a PCA. Classical scaling is conceptually more general than PCA in that arbitrary distance matrices can be used, i.e. the method does not even need the original coordinates just a distance matrix $\mathbf{D}$. Then it tries to find and embedding $\mathbf{Y}$ so that $\hat{d}_{ij}$ is as similar to d_{ij} as possible.

The disadvantage is that is computationally much more demanding, i.e. an eigenvalue decomposition of a $n \times n$ matrix has to be computed which requires $\mathcal{O}(n^2)$ memory and $\mathcal{O}(n^3)$ computation time, while PCA requires only the eigenvalue decomposition of a $d \times d$ matrix and usually $n \gg d$. R implements classical scaling in the `cmdscale` function.

The **dimRed** package wraps around `cmdscale` and allows the specification of arbitrary distance functions for calculating the distance matrix. There is also a method to calculate the embedding of new points.

[1] $X^T X = -\frac{1}{2}\mathbf{H}[d_{ij}^2]\mathbf{H}$, where $\mathbf{H} = [\delta_{ij} - \frac{1}{n}]$

2.2.4 Isomap

As Classical Scaling can deal with arbitrarily defined distances, Tenenbaum et al. (2000) suggested to approximate the structure of the manifold by using geodesic distances. In practice, a graph is created by either keeping only the connections between every point and its k nearest neighbors to produce a k-nearest neighbor graph (k-NNG), or simply by keeping all distances smaller than a value ε producing an ε-neighborhood graph (ε-NNG). Geodesic distances are obtained by recording the distance on the graph and classical scaling is used to find an embedding in fewer dimensions. This leads to an "unfolding" of possibly convoluted structures (see fig. 2.3).

Isomap's computational cost is dominated by the eigenvalue decomposition and therefore scales with $\mathcal{O}(n^3)$. Other related techniques can use more efficient algorithms because the distance matrix becomes sparse due to a different preprocessing.

In R Isomap is implemented in the function `vegan::isomap` and the calculation of geodesic distances in `vegan::isomapdist`. The **dimRed** package uses its own implementation which is faster mainly due to using a KD-tree for the nearest neighbor search (from the **RANN** package) and a faster implementation for the shortest path search in the k-NNG (from the **igraph** package). The implementation in **dimRed** also includes a forward method that can be used to embed a subset of data points and then use these points to approximate an embedding for the remaining points, this technique is generally referred to as landmark Isomap (de Silva and Tenenbaum, 2004).

2.2.5 Locally Linear Embedding

Points that lie on a manifold in a high-dimensional space can be reconstructed through linear combinations of their neighborhoods. If the manifold is well sampled and the neighborhood lies on a locally linear patch, these reconstruction weights are the same in the high-dimensional space and the low-dimensional space. Locally Linear Embedding (LLE; Roweis and Saul, 2000) is a technique that constructs a weight matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ with elements w_{ij} so that

$$\sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{n} w_{ij} \mathbf{x}_j \right\|^2 \tag{2.3}$$

is minimized under the constraints that $w_{ij} = 0$ if x_j does not belong to the neighborhood and that $\sum_{j=1}^{n} w_{ij} = 1$. Finally the embedding is made

in such a way that the following cost function is minimized for $\mathbf{Y}$,

$$\sum_{i=1}^{n} \left\| \mathbf{y}_i - \sum_{j=1}^{n} w_{ij} \mathbf{y}_j \right\|^2, \tag{2.4}$$

which can be solved using the eigenvalue decomposition of the matrix

$$\mathbf{M} = \mathbf{I}_n - \mathbf{W} - \mathbf{W}^T + \mathbf{W}^T \mathbf{W}, \tag{2.5}$$

where $\mathbf{I}_n$ is the identity matrix. The bottom $d + 1$ eigenvectors of $\mathbf{M}$ are computed and the bottom eigenvector is discarded, the other eigenvectors represent our embedding.

Conceptually, the method is similar to Isomap but it is computationally much nicer because $\mathbf{M}$ is sparse and there exist efficient solvers. In R LLE is implemented by the function `lle::lle`, which unfortunately does not make use of the sparsity. The manifold must be well sampled and the neighborhood size must be chosen appropriately for LLE to give good results.

2.2.6 Laplacian Eigenmaps

Laplacian Eigenmaps were originally developed under the name spectral clustering to separate non-convex clusters. Later they were also used for graph embedding and DR (Belkin and Niyogi, 2003).

A number of variants have been proposed. First a graph is constructed, usually from a distance matrix, the graph can be made sparse by keeping only the k nearest neighbors, or by specifying an ε neighborhood. Then a similarity matrix $\mathbf{W}$ is calculated by using a Gaussian kernel (see eq. 2.2), if $c = 2\sigma^2 = \infty$, then all distances are treated equally, the smaller c the more emphasis is given to differences in distance. The degree of vertex i is $d_i = \sum_{j=1}^{n} w_{ij}$ and the degree matrix is the matrix $\mathbf{D}$ with the entries d_i in the diagonal. Then we can form the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$ and there are several ways how to proceed, an overview can be found in von Luxburg (2007).

The **dimRed** package implements the algorithm from Belkin and Niyogi (2003). Analogously to LLE, Laplacian Eigenmaps avoids computational complexity by creating a sparse matrix and not having to estimate the distances between all pairs of points. Then the eigenvectors corresponding to the lowest eigenvalues larger than 0 of either the matrix $\mathbf{L}$ or the symmetric normalized Laplacian $\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$ are computed and form the embedding.

2.2.7 Diffusion Maps

Diffusion Maps (Coifman and Lafon, 2006) takes a distance matrix as input and calculates the transition probability matrix P of a diffusion process between the points to approximate the manifold. Then the embedding is done by an eigenvalue decompositon of P to calculate the coordinates of the embedding. The algorithm for calculating Diffusion Maps shares some elements with the way Laplacian Eigenmaps are calculated. Diffusion Map calculate the transition probability on the graph after t time steps and do the embedding on this probability matrix.

The idea is to simulate a diffusion process between the nodes of the graph, which is more robust to short-circuiting than the k-NNG from Isomap (see bottom right fig. 2.3). Diffusion maps in R are accessible via the `diffusionMap::diffuse()` function. Additional points can be approximated into an existing embedding using the Nyström formula (Bengio et al., 2004). The implementation in **dimRed** is based on the `diffusionMap::diffuse` function.

2.2.8 non-Metric Dimensional Scaling

While Classical Scaling and derived methods (see Section 2.2.3) use eigenvector decomposition to embed the data in such a way that the given distances are maintained, non-Metric Dimensional Scaling (nMDS, Kruskal, 1964a,b) uses gradient based optimization methods to reach the same goal. Therefore a Stress function,

$$S = \sqrt{\frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}} \tag{2.6}$$

is used, and the algorithm tries to embed $\mathbf{y}_i$ in such a way that the order of the d_{ij} is the same as the order of the $\hat{d}_{ij}$. Because optimization methods can fit a wide variety of problems, there are very loose limits set to the form of the error or stress function. For instance Mahecha et al. (2007c) found that nMDS using geodesic distances can be almost as powerful as Isomap for embedding biodiversity patterns. Because of the flexibility of nMDS, there is a whole package in R devoted to Multidimensional Scaling, `smacof` (de Leeuw and Mair, 2009).

R implements nMDS by `MASS::isoMDS` and `vegan::monoMDS`, related methods include Sammons Mapping which con be found as `MASS::sammon`. The **dimRed** package wraps around `vegan::monoMDS`.

2.2.9 Force Directed Methods

The data $\mathbf{X}$ can be considered as a graph with weighted edges, where the weights are the distances between points. Force directed algorithms see the edges of the graphs as springs or the result of electric charges of the nodes that results in an attractive or repulsive force between the nodes, the algorithms then try to minimize the overall energy of the graph,

$$E = \sum_{i<j} k_{ij}(d_{ij} - \hat{d}_{ij})^2, \tag{2.7}$$

where k_{ij} is the spring constant for the spring connecting points i and j.

Because graph embedding algorithms are gradient based and optimization is non-convex, they tend to suffer from long running times compared to eigendecomposition based methods and may get stuck in local optima. This is why a number of methods that try to deal with some of the shortcomings have been developed, e.g. the Kamada-Kawai (Kamada and Kawai, 1989), the Fruchtermann-Reingold (Fruchterman and Reingold, 1991), or the DrL (Martin et al., 2007) algorithms.

There are a number of graph embedding algorithms included in the `igraph` package. They can be accessed using the `igraph::layout_with_*` function family. The **dimRed** package only wraps the three algorithms mentioned above. The **igraph** package contains many more algorithms which are not interesting for dimensionality reduction.

2.2.10 t-SNE

Stochastic Neighbor Embedding (SNE; Hinton and Roweis, 2003) is a technique that minimizes the Kullback-Leibler divergence of scaled similarities of the points i and j in high-dimensional space, p_{ij}, and low dimensional space, q_{ij}:

$$KL(\mathbf{P}\|\mathbf{Q}) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{2.8}$$

SNE uses a Gaussian kernel (see eq. 2.2) to compute similarities in high- and low-dimensional space. t-Distributed Stochastic Neighborhood Embedding (t-SNE; van der Maaten and Hinton, 2008) improves on SNE by using a t-Distribution as a kernel in low-dimensional space. Because of the heavy-tailed t-distribution, t-SNE maintains local neighborhoods of the data better and penalizes wrong embeddings of dissimilar points. This property makes it especially suitable to represent clustered data and complex structures in few dimensions.

t-SNE has one parameter, perplexity, to tune which determines the neighborhood size of the kernels used.

The general runtime of t-SNE is $\mathcal{O}(n^2)$, but an efficient implementation using tree search algorithms that scales $\mathcal{O}(n \log n)$ exists and can be found in the `Rtsne` package in R. The t-SNE implementation in **dimRed** wraps around the **Rtsne** package.

There exist a number of derived techniques for dimensionality reduction, e.g. NeRV (Venna et al., 2010), and JNE (Lee et al., 2013), that improve results but there do not exist packages implementing them on CRAN yet.

2.2.11 Independent Component Analysis

Independent Component Analysis (ICA) interpretes the data $\mathbf{X}$ as a mixture of independent signals, e.g. a number of sound sources recorded by several microphones and tries to "un-mix" them and find the original signals in the recorded signals. ICA is a linear rotation of the data just as PCA but instead of capturing the maximum variance, it preserves statistically independent components. A signal matrix $\mathbf{S}$ and a mixing matrix $\mathbf{A}$ are estimated so that $\mathbf{X} = \mathbf{AS}$.

There are a number of algorithms for ICA, the most widely used is fastICA (Hyvarinen, 1999) because it provides a fast and robust way to estimate $\mathbf{A}$ and $\mathbf{S}$. FastICA maximizes a measure for nongaussianity called negentropy J (Comon, 1994), which is equivalent to minimizing mutual information between the resulting components. Negentropy J is defined as follows:

$$H(u) = -\int f(u) \log f(Y)\, \mathrm{d}u, \tag{2.9}$$

$$J(u) = H(u_{\text{gauss}}) - H(u), \tag{2.10}$$

where $u = (u_1, \ldots, u_n)^T$ is a random vector with density $f(\cdot)$ and u_{gauss} is a Gaussian random variable with the same covariance structure as u. FastICA uses a very efficient approximations to calculate negentropy. Because ICA can be translated into a simple linear projection, it is possible to project new data points and reconstruct embedded points.

There are a number of packages in R that implement algorithms for ICA, the **dimRed** package wraps around the `fastICA::fastICA()` function.

2.2.12 DRR

Dimensionality Reduction via Regression is a recent technique extending PCA (Laparra et al., 2015). Starting from a rotated (PCA) solution $\mathbf{X}' = \mathbf{VX}$,

it predicts redundant information from the remaining components using nonlinear regression.

$$\mathbf{y}_{i\cdot} = \mathbf{x}'_{i\cdot} - f_i(\mathbf{x}'_{1\cdot}, \mathbf{x}'_{2\cdot}, \ldots, \mathbf{x}'_{i-1\cdot}) \tag{2.11}$$

with $\mathbf{x}'_{i\cdot}$ and $\mathbf{y}_{i\cdot}$ being the loading of observations on the i-th axis, i.e. the rows of the matrices $\mathbf{X}'$ and $\mathbf{Y}$ respectively. In theory any kind of regression can be used to estimate f_i, the authors of the original paper choose Kernel Ridge Regression (KRR; Saunders et al., 1998) because it is a flexible nonlinear regression technique and computational optimizations for a fast calculation exist. DRR has another advantage over other techniques presented here, because it provides an exact forward and inverse function.

The use of KRR also has the advantage of making the method convex, here we list it under non-convex methods, because other types of regression may make it non-convex.

Mathematicaly, functions are limited to map one input to a single output point, therefore DRR reduces to PCA if manifolds are too complex. But it seems very useful for slightly curved manifolds. The initial rotation is important, because the result strongly depends on the order of dimensions in high-dimensional space.

DRR is implemented in the package **DRR**. The package provides methods to project new data and reconstruct embedded data.

2.3 Quality Criteria

The advantage of unsupervised learning is that one does not need to specify classes or a target variable for the data under scrutiny. Instead the chosen algorithm arranges the input data e.g. into clusters or a lower dimensional representation. In contrary to a supervised problem, there is no natural way to directly measure the quality of any output or to compare two methods by an objective measure like for instance modeling efficiency or classification error. The reason is that every method optimizes a different error function, and it would be unfair to compare t-SNE and PCA by means of either recovered variance or KL-Divergence. One fair measure would be the reconstruction error, i.e. reconstructing the original data from a limited number of dimensions, but as shown above not many methods provide forward and inverse mappings.

However, there are a series of independent estimators on the quality of a low-dimensional embedding. The **dimRed** package provides a number of quality measures which have been proposed in literature to measure performance of dimensionality reduction techniques.

2.3.1 Co-Ranking Matrix Based Measures

The co-ranking matrix (Lee et al., 2009) is a way to capture the changes in ordinal distance, just as before, let $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ be the distances between $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e. in high-dimensional space and $\hat{d}_{ij} = d(\mathbf{y}_i, \mathbf{y}_j)$ the distances in low dimensional space, then we can define the rank of $\mathbf{y}_j$ with respect to $\mathbf{y}_i$

$$\hat{r}_{ij} = |\{k : \hat{d}_{ik} < \hat{d}_{ij} \text{ or } (\hat{d}_{ik} = \hat{d}_{ij} \text{ and } 1 \leq k < j \leq n)\}|, \tag{2.12}$$

and analogously the rank in high-dimensional space as

$$r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq n)\}|, \tag{2.13}$$

where $|A|$ is the number of elements in a set A. This means that we simply replace the distances in a distance matrix column wise by their ranks. It also means that r_{ij} is an integer which indicates that $\mathbf{x}_i$ is the r_{ij}-th closest neighbor of $\mathbf{x}_j$ in the set $\mathbf{X}$.

The co-ranking matrix $\mathbf{Q}$ then has elements

$$q_{kl} = |\{(i, j) : \hat{r}_{ij} = k \text{ and } r_{ij} = l\}|, \tag{2.14}$$

which is the 2d-histogram of the ranks, i.e. q_{kl} is an integer which counts how many points of distance rank l became rank k. In a perfect DR, this matrix will only have non-zero entries in the diagonal, if most of the non-zero entries are in the lower triangle, then the DR collapsed far away points onto each other and if most of the non-zero entries are in the upper triangle, then the DR teared close points apart. For a detailed description of the properties of the co-ranking matrix the reader is referred to Lueks et al. (2011).

The functions `coRanking::coranking` and `coRanking::imageplot` can be used to calculate and visualize the co-ranking matrix. A good embedding should scatter the values around the diagonal of the matrix; if the values are in the lower triangle, then the embedding collapses the original structure causing far away points to be much closer, if the values are predominantly in the upper triangle the points from the original structure are torn apart. Nevertheless this method requires visual inspection of the matrix. For an automated assessment of quality, a scalar value that assigns a quality to an embedding is needed.

A number of metrics can be computed from the co-ranking matrix:

$$Q_{NX}(k) = \frac{1}{kn} \sum_{i=1}^{k} \sum_{j=1}^{k} q_{ij}, \tag{2.15}$$

which is the number of points that belong to the k-th nearest neighbors in both high- and low-dimensional space, normalized to give a maximum of 1 (Lee et al., 2009). This quantity can be adjusted for random embeddings, which gives the Local Continuity Meta Criterion (Chen and Buja, 2009):

$$\mathrm{LCMC}(k) = Q_{NX}(k) - \frac{k}{n-1} \tag{2.16}$$

The above measures still depend on k, but LCMC has a well defined maximum at $k_{\max}$. Two measures without parameters can be defined, then:

$$Q_{\text{local}} = \frac{1}{k_{\max}} \sum_{k=1}^{k_{\max}} Q_{NX}(k), \text{ and} \tag{2.17}$$

$$Q_{\text{global}} = \frac{1}{n - k_{\max}} \sum_{k=k_{\max}}^{n-1} Q_{NX}(k), \tag{2.18}$$

which measure the preservation of local and global distances respectively. The original authors advised using Q_{local} over Q_{global}, but this depends on the application.

LCMC(k) can be normalized to a maximum of 1, which yields the following measure for a quality embedding (Lee et al., 2013):

$$R_{NX}(k) = \frac{(n-1)Q_{NX}(k) - k}{n - 1 - k}, \tag{2.19}$$

where a value of 0 corresponds to a random embedding and a value of 1 to a perfect embedding into the k-ary neighborhood. To transform $R_{NX}(k)$ into a parameterless measure, the area under the curve can be used:

$$\mathrm{AUC}_{\ln k}\left(R_{NX}(k)\right) = \left(\sum_{k=1}^{n-2} R_{NX}(k)\right) \Bigg/ \left(\sum_{k=1}^{n-2} 1/k\right). \tag{2.20}$$

This measure is normalized to one and takes k at a log-scale, therefore it gives higher scores to methods that preserve local distances.

In R, the `coRanking::coranking` function calculates the co-ranking matrix. The **dimRed** package contains the functions `Q_local`, `Q_global`, `Q_NX`, `LCMC`, and `R_NX` to calculate the above quality measures and `AUC_lnK_R_NX`.

Calculating the co-ranking matrix is a relatively expensive operation because it requires sorting every row or the distance matrix twice and therefore scales with $\mathcal{O}(n^2 \log n)$. There is also a plotting function `plot_R_NX`, which plots the R_{NX} values with log-scaled K and adds the $\mathrm{AUC}_{\ln k}$ to the legend (see fig. 2.2).

There are a number of other measures that can be computed from a co-ranking matrix, we will not provide these measures here because in literature the by far most used measure derived from the co-ranking matrix is $R_{NX}(k)$ and the associated $\text{AUC}_{\ln k}(R_{NX}(k))$, see Lueks et al. (2011), Lee et al. (2009), or Babaee et al. (2013).

2.3.2 Cophenetic Correlation

An old measure originally developed to compare clustering methods in the field of phylogenetics is cophenetic correlation (Sokal and Rohlf, 1962). This method consists simply of the correlation between the upper or lower triangles of the distance matrices (in dendrograms they are called cophenetic matrices, hence the name) in high- and low-dimensional space. Additionally, the distance measure and correlation method can be varied. In the **dimRed** package this is implemented in the `cophenetic_correlation`.

Some studies use a measure called "residual variance" (Tenenbaum et al., 2000; Mahecha et al., 2007c), which is defined as

$$1 - r^2(\mathbf{D}, \widehat{\mathbf{D}}),$$

where r is the Pearson correlation and $\mathbf{D}$, $\widehat{\mathbf{D}}$ are the distances matrices consisting of elements d_{ij} and $\hat{d}_{ij}$, respectively.

2.3.3 Reconstruction Error

The fairest and most common way to assess the quality of a dimensionality reduction if the method provides an inverse mapping is the reconstruction error. The **dimRed** package includes a function to calculate the root mean squared error which is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d(\hat{\mathbf{x}}_i, \mathbf{x}_i)^2} \tag{2.21}$$

with $\hat{\mathbf{x}}_i = f^{-1}(\mathbf{y}_i)$, and f^{-1} being the function that maps an embedded value back to feature space. The **dimRed** package provides the `reconstruction_rmse` and `reconstruction_error` functions.

2.4 Test Datasets

There are a number of test datasets that are often used to showcase a dimensionality reduction technique. Common ones being the 3d S-curve

and the Swiss roll among others. These datasets have in common that they usually have three dimensions, and well defined manifolds. Real world examples usually have more dimensions and often are much noisier, the manifolds may not be well sampled and exhibit holes and large pieces may be missing also we cannot be sure if we can observe all the relevant variables.

The **dimRed** package implements a number of test datasets that are used in literature to benchmark methods with the function `loadDataSet`. For artificial datasets the number of points and the noise level can be adjusted, the function also returns the internal coordinates.

2.5 Examples

The comparison of different DR methods, choosing the right parameters for a method, and the inspection of the results is made very simple by **dimRed**. This section contains a number of examples to highlight the use of the package. The code to reproduce these figures can be found in Kraemer et al. (2018).

To compare methods of dimensionality reduction, first a test dataset is loaded using `loadDataSet`, then the `embed` function is applied for DR using `lapply` this is a one-liner and it is very simple to add more methods. For inspection **dimRed** provides methods for the `plot` function to visualize the resulting embedding (fig. 2.2 b and d), internal coordinates of the manifold are represented by color gradients. To visualize how well embeddings represent different neighborhood sizes, the function `plot_R_NX` is used on a list of embedding results (fig. 2.2 c).

```
## define which methods to apply
embed_methods <- c("Isomap", "PCA")
## load test dataset
data_set <- loadDataSet("3D S Curve", n = 1000)
## apply dimensionality reduction
data_emb <- lapply(embed_methods,
                   function(x) embed(data_set, x))
names(data_emb) <- embed_methods
## plot dataset, embeddings, and quality analysis
plot(data_set, type = "3vars")
lapply(data_emb, plot, type = "2vars")
plot_R_NX(data_emb)
```

Often the quality of an embedding strongly depends on the choice of

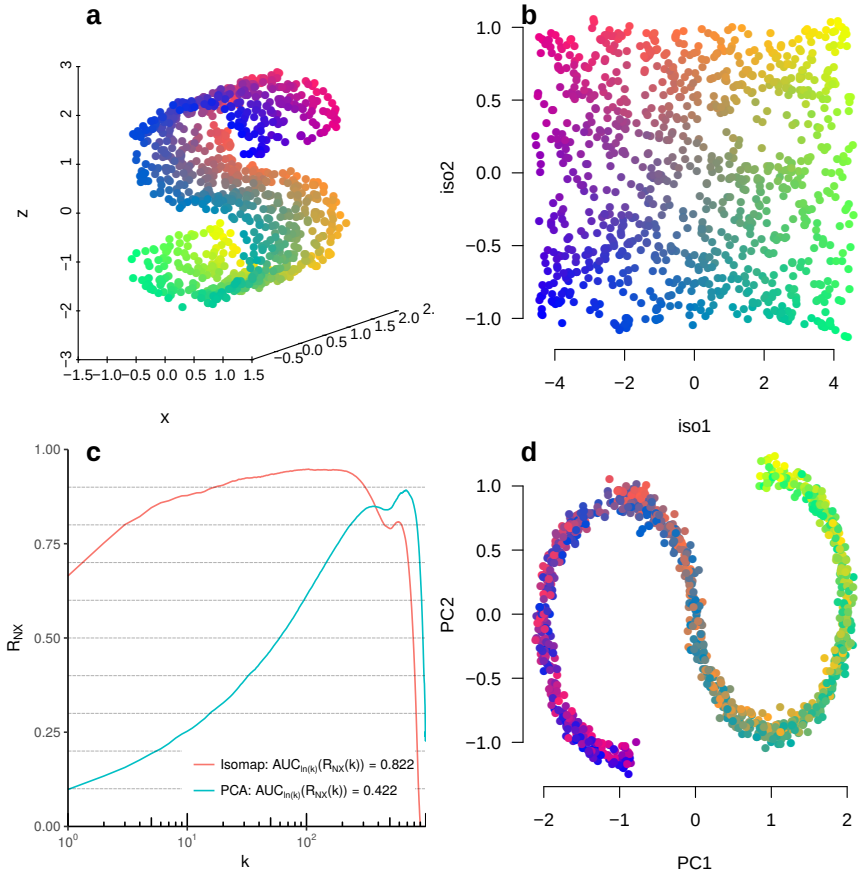

Figure 2.2: Comparing PCA and Isomap, (a) An S-shaped manifold, colors represent the internal coordinates of the manifold. (b) Isomap unfolds the S-shaped manifold. (d) PCA projects the data preserving the directions of maximum variance. (c) R_{NX} plotted against neighborhood sizes, Isomap is much better at preserving local distances and PCA is better at preserving global Euclidean distances. The numbers on the legend are the $\text{AUC}_{\ln K}$.

parameters; the interface of **dimRed** can be used to facilitate searching the parameter space.

Isomap (see Section 2.2.4) has one parameter k which determines the number of neighbors used to construct the kNN graph. If this number is too large, then Isomap will resemble an cMDS; if the number is too small, the resulting embedding contains holes. In fig. 2.3 we show how to estimate the optimal value $k_{\max}$, for k using the Q_{local} criterion.

It is also very easy to compare across methods and quality scores, fig. 2.4 compeares a number of quality indicators and methods for dimensionality reduction.

2.6 The *dimRed* Package

The **dimRed** package wraps DR methods readily implemented in R, implements some methods, and offers means to compare the quality of embeddings. The package is open source and available under the GPL3 license on github (`https://github.com/gdkrmr/dimRed`) and CRAN (`https://cran.r-project.org/package=dimRed`). **dimRed** provides a common interface and convenience functions for a variety of different DR methods so that it is made easier to use and compare different methods. An overview can be found in table 2.1.

Table 2.1: The main interface functions of the **dimRed** package.

Function	Description
`embed`	Embed data using a DR method.
`quality`	Calculate a quality score from the result of `embed`.
`plot`	Plot a `dimRedData` or `dimRedResult` object, colors the points automatically, for exploring the data.
`plot_R_NX`	For comparing the quality of various embeddings.
`dimRedMethodList`	Returns a character vector that contains all implemented DR methods.
`dimRedQualityList`	Returns a character vector that contains all implemented quality measures.

Internally, the package uses S4 classes but for normal usage the user does not need to have any knowledge on the inner workings of the S4 class system in R (s. table 2.2). The package contains simple conversion functions from and to standard R-objects like `data.frame` and `matrix`. The `dimRedData` class provides an container for the data to be processed. The

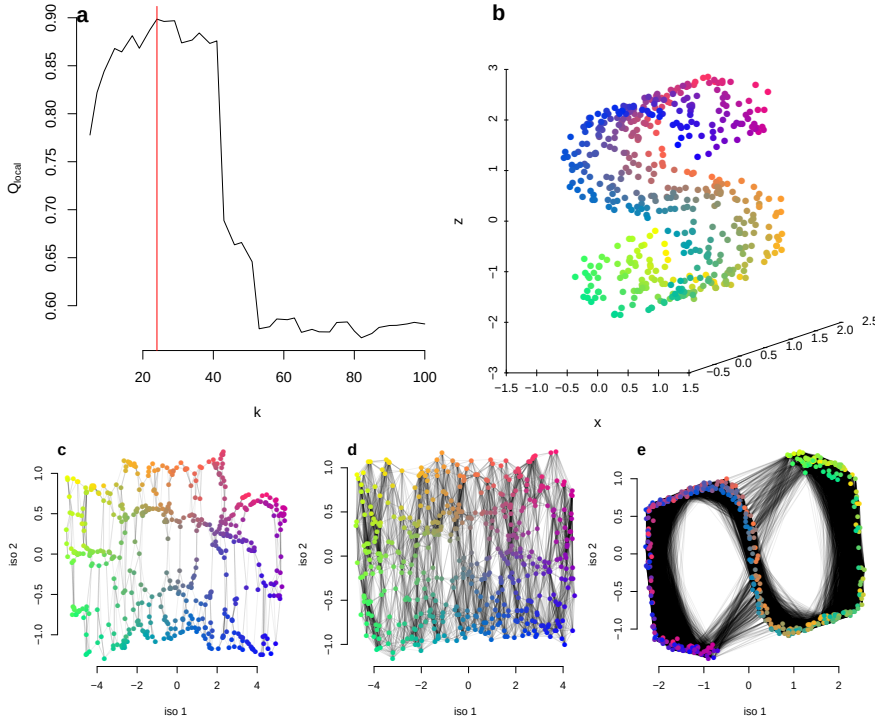

Figure 2.3: Using **dimRed** and the Q_{local} indicator to estimate a good value for the parameter k in Isomap. Top left: Q_{local} for different values of k, the vertical red line indicates the maximum $k_{\text{max}} = 24$. Top right: The original dataset, a 2 dimensional manifold bent in an S-shape in 3 dimensional space. Bottom row: Embeddings and k-NNG for different values of k. Left: $k = 5$, $Q_{\text{local}} = 0.78$. The value for k is too small resulting in holes in the embedding; the manifold itself is still unfolded correctly. Middle: $k = k_{\text{max}} = 24$, $Q_{\text{local}} = 0.90$. The best representation of the original manifold in two dimensions achievable with Isomap. Right: $k = 100$, $Q_{\text{local}} = 0.58$. k is too large, the kNN graph does not approximate the manifold any more.

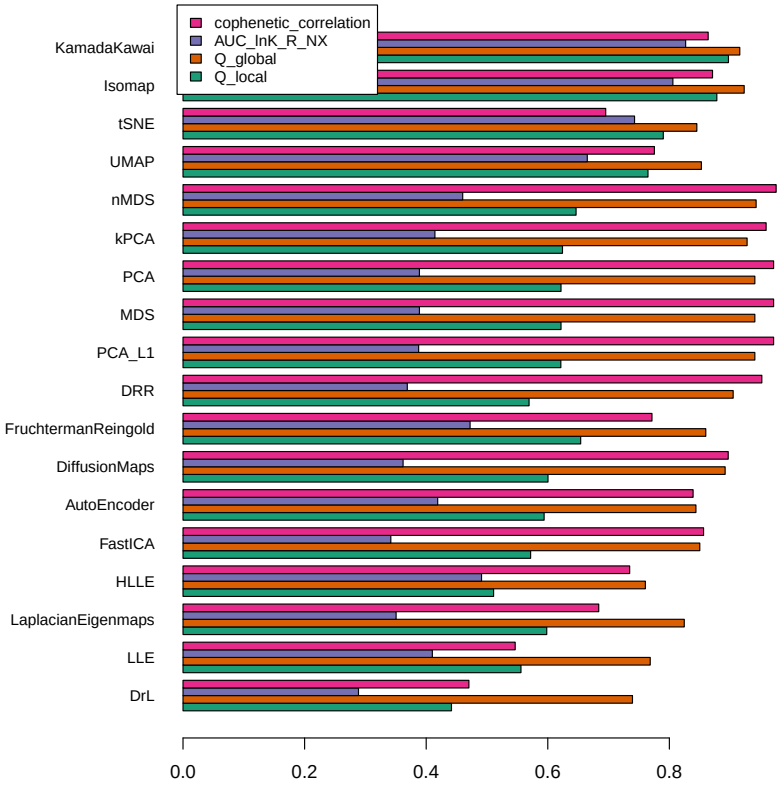

Figure 2.4: A visualization of the `quality_results` matrix. The methods are ordered by mean quality score. The reconstruction error was omitted, because a higher value means a worse embedding, while in the shown methods a higher score means a better embedding. Parameters for the methods were not tuned, therefore it should not be seen as a general quality assessment of methods.

slot[2] `data` contains a `matrix` with dimensions in columns and observations in rows, the slot `meta` may contain a `data.frame` with additional information, e.g. categories or other information of the data points.

Table 2.2: The S4 classes used in the **dimRed** package

Class Name	Function
`dimRedData`	Holds the data for a DR. Fed to `embed()`. `as.dimRedData()` for `data.frame`, `matrix`, and `formula` exist.
`dimRedMethod`	Virtual class, ancestor of all DR methods.
`dimRedResult`	The result of `embed()`, the embedded data.

Each embedding method is a class which inherits from `dimRedMethod` which means that it contains a function to generate `dimRedResult` objects and a `list` of standard parameters. The class `dimRedResult` contains the data in reduced dimensions, the original meta information along with the original data, and, if possible, functions for the forward and inverse mapping, see tab. 2.3

From a user-perspective, the central function of the package is `embed` which is called in the form `embed(data,method,...)`, `data` can take standard R objects like `data.frame`, `matrix`, or `formula` as input, which are automatically coerced to the internal S4 classes. The `method` is given as a character vector, all available methods can be listed by calling the function `dimRedMethodList`. Method specific parameters in `...`[3] can be given, if no method specific parameters are given, method specific defaults are chosen. `embed` returns an object of class `dimRedResult`

For comparing different embeddings, **dimRed** contains the function `quality` which relies on the output of `embed` and a string specifying the quality metric. This function returns a scalar quality score; a vector that contains the names of all quality functions is returned by calling `dimRedQualityList()`.

For easy visual examination the package contains `plot` functions for `dimRedData` and `dimRedResult` objects in order to plot high dimensional data like parallel plots and pairwise scatter plots. Automatic coloring of data points is done automatically if additional properties for the observations are provided.

[2] A slot of an S4 class in R can be accessed via the `@` symbol, i.e. `object@data`

[3] `...` in R refers to a variable number of function arguments, this is how method specific arguments are handled by the `embed` function.

Table 2.3: Methods in the **dimRed**, the code used in the `embed` function and projection and inverse projection functions provided.

Method	Code	$f(\mathbf{x}_i)$	$f^{-1}(\mathbf{y}_i)$
DRR	`DRR`	✓	✓
Diffusion Maps	`DiffusionMaps`	✓	✓
DrL	`DrL`	x	x
ICA	`FastICA`	✓	✓
Fruchterman-Reingold	`FruchtermanReingold`	x	x
Isomap	`Isomap`	✓	✓
Kamada-Kawai	`KamadaKawai`	x	x
LLE	`LLE`	x	x
Laplacian Eigenmaps	`LaplacianEigenmaps`	x	x
cMDS	`MDS`	✓	✓
NNMF	`NNMF`	✓	✓
PCA	`PCA`	✓	✓
kPCA	`kPCA`	✓	✓
nMDS	`nMDS`	x	x
t-SNE	`tSNE`	x	x

2.7 Conclusion

This chapter presented the **dimRed** and **coRanking** packages and provides a brief overview of the methods implemented therein. **dimRed** is written in the R language, which is one of the most popular languages for data analysis and is freely available through the built in package management system CRAN. It is object oriented and completely open source and therefore easily available and extensible. Although most of the DR methods already had implementations in R, **dimRed** adds some previously not implemented methods for dimensionality reduction, and **coRanking** adds methods for an independent quality control of DR methods to the R ecosystem. DR is a widely used technique but due to the lack of easily usable tools, choosing the right method is complex and depends on a variety of factors. The **dimRed** package aims to facilitate experimentation with different techniques, parameters, and quality measures so that choosing the right method becomes easier. **dimRed** wants to enable the user to objectively compare methods for dimensionality reduction that rely on very different conceptual approaches. It makes the life of the programmer easier, because all methods are aggregated in one place and there is a

single interface and standardized classes to access the functionality.

Chapter 3

Summarizing the State of the Terrestrial Biosphere in Few Dimensions

Content

This chapter is based on the following publication

Kraemer, G., Camps-Valls, G., Reichstein, M., and Mahecha, M. D. (2020). Summarizing the state of the terrestrial biosphere in few dimensions. *Biogeosciences*, 17(9), 2397–2424. doi:10.5194/bg-2019-307.

Abstract
In times of global change, we must closely monitor the state of the planet in order to understand the full complexity of these changes. In fact, each of the Earth's subsystems—i.e. the biosphere, atmosphere, hydrosphere, and cryosphere—can be analyzed from a multitude of data streams. However, since it is very hard to jointly interpret multiple monitoring data streams in parallel, one often aims for some summarizing indicator. Climate indices, for example, summarize the state of atmospheric circulation in a region. Although such approaches are also used in other fields of science, they are rarely used to describe land surface dynamics. Here, we propose a robust method to create global indicators for the terrestrial biosphere using principal component analysis based on a high-dimensional set of relevant global data streams. The concept was tested using 12 explanatory variables representing the biophysical state of ecosystems and land–atmosphere water, energy, and carbon fluxes. We find that three indicators account for 82% of the variance of the selected biosphere variables in space and time across the globe. While the first indicator summarizes productivity patterns, the second indicator summarizes variables representing water and energy availability. The third indicator represents mostly changes in surface albedo. Anomalies in the indicators clearly identify extreme events, such as the Amazon droughts (2005 and 2010) and the Russian heatwave (2010). The anomalies also allow us to interpret the impacts of these events. The indicators can also be used to detect and quantify changes in seasonal dynamics. Here we report, for instance, increasing seasonal amplitudes of productivity in agricultural areas and arctic regions. We defend that this generic approach has great potential for the analysis of land surface dynamics from observational or model data.

3.1 Introduction

Today, humanity faces the negative global impacts of land use and land cover change (Song et al., 2018), global warming (IPCC, 2014), and associated losses of biodiversity (IPBES, 2019; Díaz et al., 2019), to only mention the most prominent transformations. Over the past decades, new satellite missions (e.g., Berger et al., 2012; Schimel and Schneider, 2019), along with the continuous collection of ground based measurements (e.g., Wingate et al., 2015; Nasahara and Nagai, 2015; Baldocchi, 2020), and the integration of both (e.g., Papale et al., 2015; Babst et al., 2017; Jung et al., 2019) have increased our capacity to monitor the Earth's surface enormously. However, there are still large knowledge gaps limiting our capacity to monitor and understand the current transformations of the Earth system (Steffen et al.,

2015; Rosenfeld et al., 2019; Yan et al., 2019; Piao et al., 2020b).

Many recent changes due to increasing anthropogenic activity are manifested in long-term transformations. One prominent example is "global greening" that has been attributed to fertilization effects, temperature increases, and land use intensification (de Jong et al., 2011; Zhu et al., 2016; Piao et al., 2020a). It is also known that phenological patterns change in the wake of climate change (Schwartz, 1998; Parmesan, 2006). However, these phenological patterns vary regionally. In "cold" ecosystems one may find decreased seasonal amplitudes on primary production due to warmer winters (Stine et al., 2009). Elsewhere, seasonal amplitude increased in agricultural areas, for example, due to the so-called "green revolution" (Zeng et al., 2014; Chen et al., 2019). Another change in terrestrial land surface dynamics is induced by increasing frequencies and magnitudes of extreme events (Barriopedro et al., 2011; Reichstein et al., 2013). The consequences for land ecosystems have yet to be fully understood (Flach et al., 2018; Sippel et al., 2018) and require novel detection and attribution methods tailored to the problem (Flach et al., 2017; Mahecha et al., 2007c). While extreme events are typically only temporary deviations from a normal trajectory, ecosystems may change their qualitative state permanently, for example shift from grassland to shrubland. Such shifts or tipping points can be induced by changing environmental conditions or direct human influence, and they pose yet another problem that needs to be considered (Lenton et al., 2008). The question we address here is how to uncover and summarize changes in land surface dynamics in a consistent framework. The idea is to simultaneously take advantage of a large array of global data streams, without addressing each observed phenomenon in a specific domain only. We seek to develop a general approach to uncover changes in the land surface dynamics based on a very generic method.

The problem of identifying patterns of change in high-dimensional data streams is not new. Extracting the dominant features from high-dimensional observations is a well-known problem in many disciplines. One approach is to manually define indicators that are known to represent important properties such as the "Bowen ratio" (Bowen, 1926, find a more complete description of the concept in Section 3.3.3). Another one consists in using machine learning to extract unique, and ideally independent features from the data. In the climate sciences, for instance, it is common to summarize atmospheric states using empirical orthogonal functions (EOFs), also known as principal component analysis (PCA; Pearson, 1901). The rationale is that dimensionality reduction (see Chapter 2) only retains the main data features, which makes them more easily accessible for analysis. One of the most prominent examples is the description of the El Niño–Southern Oscillation (ENSO) dynamics in the multivariate ENSO

index (MEI; Wolter and Timlin, 2011b), an indicator describing the state of atmospheric and oceanic circulation patterns at a certain point in time. The MEI is a very successful index that can be easily interpreted and used in a variety of ways; most basically it provides a measure for the intensity and duration of the different quasi-cyclic ENSO events but it can also be associated with its characteristic impacts, e.g. seasonal warming, changes in seasonal temperatures and overall dryness in the Pacific Northwest of the United States (Abatzoglou et al., 2014), drought-related fires in the Brazilian Amazon (Aragão et al., 2018), and crop yield anomalies (Najafi et al., 2019).

In plant ecology, indicators based on dimensionality reduction methods are used to describe changes to species assemblages along unknown gradients (Legendre and Legendre, 1998; Mahecha et al., 2007c). The emerging gradients can be interpreted using additional environmental constraints, or based on internal plant community dynamics (van der Maaten et al., 2012). It is also common to compress satellite-based Earth observations via dimensionality reduction to get a notion of the underlying dynamics of terrestrial ecosystems. For instance, Ivits et al. (2014) showed that one can understand the impacts of droughts and heatwaves based on a compressed view of the relevant vegetation indices. In general, dimensionality reduction is the method of choice to compress high-dimensional observations in a few (ideally) independent components with little loss of information (Van Der Maaten et al., 2009, Chapter 2).

Understanding changes in land–atmosphere interactions is a complex problem, as all aforementioned patterns of change may occur and interact: land cover change may alter biophysical properties of the land surface such as (surface) albedo with consequences for the energy balance (Song et al., 2018). Long-term trends in temperature, water availability, or fertilization may affect each other and impact productivity patterns and biogeochemical processes (Zhu et al., 2016; Sitch et al., 2015). In fact, these land surface dynamics have implications for multiple dimensions and require monitoring of biophysical state variables such as leaf area index, albedo, etc., as well as associated land–atmosphere fluxes of carbon, water, and energy.

Here, we aim to summarize these high-dimensional surface dynamics and make them accessible for subsequent interpretations and analyses such as mean seasonal cycles (MSCs), anomalies, trend analyses, breakpoint analyses, and the characterization of ecosystems. Specifically, we seek a set of uncorrelated, yet comprehensive, state indicators. We want to have a set of very few indicators that represent the most dominant features of the above-described temporal ecosystem dynamics. These indicators should also be uncorrelated, so that one can study the system state by looking and interpreting each indicator independently. The approach should also give

an idea of the general complexity contained in the available data streams. If more than a single indicator is required to describe land surface dynamics accurately, then these indicators shall describe very different aspects. While one indicator may describe global patterns of change, others could be only relevant in certain regions, for certain types of ecosystems, or for specific types of impacts. The indicators shall have a number of desirable properties: (1) represent the overall state of observations comprising the system in space and time; (2) carry sufficient information to allow for reconstructing the original observations faithfully from these indicators; (3) be of lower dimensionality than the number of observed variables; and (4) allow intuitive interpretations.

In this work, we first introduce a method to create such indicators, and then we apply the method to a global set of variables describing the biosphere. Finally, to prove the effectiveness of the method, we interpret the resulting set of indicators and explore the information contained in the indicators by analyzing them in different ways and relating them to well-known phenomena.

3.2 Methods

3.2.1 Data

Table 3.1 gives an overview of the data streams used in this analysis (for a more detailed description see Appendix A). For an effective joint analysis of more than a single variable, the variables have to be harmonized and brought to a single grid in space and time. The Earth System Data Lab (ESDL; `www.earthsystemdatalab.net`, last accessed 27/04/2020; Mahecha et al., 2019) curates a comprehensive set of data streams to describe multiple facets of the terrestrial biosphere and the associated climate system. The data streams are harmonized as analysis-ready data on a common spatiotemporal grid (equirectangular grid 0.25° in space and 8 d in time, 2001–2011), forming a 4D hypercube, which we call a "data cube". The ESDL not only curates Earth system data, but also comes with a toolbox to analyze these data efficiently. For this study, we chose all available variables in the ESDL v1.0 (the most recent version available at the time of analysis), divided the available variables into meteorological and biospheric variables and discarded the atmospheric variables. We also discarded variables with distributions that are badly suited for a linear PCA (e.g. burnt area contains mostly zeros) and variables with too many missing values. The only dataset that was added post hoc was fAPAR which represents an important aspect of vegetation which was not available in the data cube at

Table 3.1: Variables used describing the biosphere. For a description of the variables, see Appendix A.

Variable	Details	Source
Black-sky albedo	Directional reflectance	Muller et al. (2011)
Evaporation	$[mm\ day^{-1}]$	Martens et al. (2017)
Evaporative stress	Modeled water stress	Martens et al. (2017)
fAPAR	fraction of absorbed photosynthetically active radiation	Disney et al. (2016)
Gross primary productivity (GPP)	$[gCm^{-2}day^{-1}]$	Tramontana et al. (2016); Jung et al. (2019)
Latent energy (LE)	$[Wm^{-2}]$	Tramontana et al. (2016); Jung et al. (2019)
Net ecosystem exchange (NEE)	$[gCm^{-2}day^{-1}]$	Tramontana et al. (2016); Jung et al. (2019)
Root-zone soil moisture	$[m^3m^{-3}]$	Martens et al. (2017)
Sensible heat (H)	$[Wm^{-2}]$	Tramontana et al. (2016); Jung et al. (2019)
Surface soil moisture	$[mm^3mm^{-3}]$	Martens et al. (2017)
Terrestrial ecosystem respiration (TER)	$[gCm^{-2}day^{-1}]$	Tramontana et al. (2016); Jung et al. (2019)
White-sky albedo	Diffuse reflectance	Muller et al. (2011)

the time of analysis (it is part of the most recent version of the data cube).

The datasets taken from Tramontana et al. (2016); Jung et al. (2019) are derived from flux tower measurements (Baldocchi, 2020). The flux towers are not equally distributed in the space spanned by climatic variables, i.e. there are many flux towers in temperate areas, but much less in tropic and arctic regions, which may lead to less accurate data in these regions. These datasets also exclude large arid areas such as the Sahara and Gobi deserts and parts of the Arabian Peninsula which may affect the resulting loadings of the PCA slightly.

In this study, each variable was normalized globally to zero mean and unit variance to account for the different units of the variables, i.e. trans-

form the variables to have standard deviations from the mean as the common unit. Because the area of the pixel changes with latitude in the equirectangular coordinate system used by the ESDL, the pixels were weighted according to the represented surface area. Only spatiotemporal pixels without any missing values were considered in the calculation of the covariance matrix.

3.2.2 Dimensionality Reduction with PCA

As a method for dimensionality reduction, we used a modified principal component analysis to summarize the information contained in the observed variables. PCA transforms the set of d centered and, in this case, standardized variables into a subset of p, $1 \leq p \leq d$, principal components (PCs). Each component is uncorrelated with the other components, while the first PCs explain the largest fraction of variance in the data.

The data streams consist of $d = 12$ observed variables at the same time and location. Each observation is defined in a d-dimensional space, $\mathbf{x}_i \in \mathbb{R}^d$, and we define the dataset by collecting all samples in the matrix $\mathbf{X} = [\mathbf{x}_1 | \cdots | \mathbf{x}_n] \in \mathbb{R}^{d \times n}$. The observations are repeated in space and time and lie on a grid of lat $\times$ lon $\times$ time. In our case, we have $n = |\text{lat}| \times |\text{lon}| \times |\text{time}| = 720 \times 1440 \times 506 = 524,620,800$ observations, where $|\cdot|$ denotes the cardinality of the dimension. Note that the actual number of observations was lower, $n = 106,360,156$, because we considered land points only and removed missing values.

The fundamental idea of PCA is to project the data to a space of lower dimensionality that preserves the covariance structure of the data. We treat time equal to space, this gives us the advantage, that we only have to calculate a single PCA and all observations are projected into the same space of reduced dimensionality, which makes them comparable. If we treated time differently from space we would have to compute a separate PCA for each time step and the resulting indicators would not be comparable because each would be projected into a different space.

The fundament of a PCA is the computation of a covariance matrix, $\mathbf{Q}$. When all variables are centered to global zero mean and normalized to unit variance, the covariance matrix can in principle be estimated as

$$\mathbf{Q} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T. \tag{3.1}$$

However, in our case the data cube lies on a regular 0.25° grid and estimating $\mathbf{Q}$ as above would lead to overestimating the influence of dynamics in relatively small pixels of high latitudes compared to lower latitudes where

each data point represents a larger areas. Hence, one needs a weighted approach to calculate the covariance matrix,

$$\mathbf{Q} = \frac{1}{w} \sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i^T, \tag{3.2}$$

where $w_i = \cos(\text{lat}_i)$ and lat_i is the latitude of observation i, $w = \sum_{i=1}^{n} w_i$ is the total weight, and n is the total number of observations. Equation (3.2) has the additional property that it can be computed sequentially on very big datasets, such as our Earth System Data Cube, by consecutively adding observations to an initial estimate.

Note that the actual calculation of the covariance matrix is even more complicated, because summing up many floating-point numbers one by one can lead to large inaccuracies due to precision issues of floating-point numbers and instabilities of the naive algorithm (Higham, 1993; the same holds for the implementations of the `sum` function in most software used for numerical computing). Here, we used the Julia package `WeightedOnlineStats.jl`[1] (implemented by the first author of this paper), which uses numerically stable algorithms for summation, higher-precision numbers, and a map-reduce scheme that further minimizes floating-point errors.

Based on this weighted and numerically stable covariance matrix, the PCA can be computed using an eigendecomposition of the covariance matrix,

$$\mathbf{Q} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \in \mathbb{R}^{d \times d}. \tag{3.3}$$

In this case, the covariance matrix $\mathbf{Q}$ is equal to the correlation matrix because we standardized the variables to unit variance. $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues, $\lambda_1, \ldots, \lambda_d$, in the diagonal in decreasing order and $\mathbf{V} \in \mathbb{R}^{d \times d}$, the matrix with the corresponding eigenvectors in columns. $\mathbf{V}$ can project observations, $\mathbf{x}_i$ (centered and standardized), onto the retained PCs,

$$\mathbf{y}_i = \mathbf{V}^T \mathbf{x}_i \in \mathbb{R}^d, \tag{3.4}$$

where $\mathbf{y}_i$ is the projection of the observation $\mathbf{x}_i$ onto the d PCs.

The canonical measure of the quality of a PCA is the fraction of explained variance by each component, σ_i^2, calculated as

$$\sigma_i^2 = \frac{\lambda_i}{\sum_{i=1}^{d} \lambda_i}. \tag{3.5}$$

[1] DOI: 10.5281/zenodo.3360311, repository: `https://github.com/gdkrmr/WeightedOnlineStats.jl/`

To get a more complete measure of the accuracy of the PCA, we used the "reconstruction error" in addition to the fraction of explained variance. PCA allows a simple projection of an observation onto the first p PCs and a consecutive reconstruction of the observations from this p-dimensional projection. This is achieved by

$$\mathbf{Y}_p = \mathbf{V}_p^T \mathbf{X} \in \mathbb{R}^{p \times n} \text{ and } \hat{\mathbf{X}}_p = \mathbf{V}_p \mathbf{Y}_p \in \mathbb{R}^{d \times n}, \tag{3.6}$$

where $\mathbf{Y}_p$ is the projection onto the first p PCs, $\mathbf{V}_p \in \mathbb{R}^{d \times p}$ the matrix with columns consisting of the eigenvectors belonging to the p largest eigenvalues, and $\mathbf{X}_p$ the observations reconstructed from the first p PCs.

The reconstruction error, $\mathbf{e}_i$, was calculated for every point, $\mathbf{x}_i$ in the space–time domain based on the reconstructions from the first p principal components:

$$\mathbf{e}_i = \mathbf{V}_p \mathbf{V}_p^T \mathbf{x}_i - \mathbf{x}_i \in \mathbb{R}^d. \tag{3.7}$$

As this error is explicit in space, time and variable, it allows for disentangling the contribution of each of these domains to the total error. This can be achieved by estimating the (weighed) mean square error,

$$\text{MSE} = \frac{1}{w} \sum_i w_i \mathbf{e}_i^2. \tag{3.8}$$

This approach can give a better insight into the compositions of the error than a single global error estimate based on the eigenvalues.

3.2.3 Pixel-Wise Analyses of Time Series

The principal components estimated as described above are ideally low-dimensional representations of the land surface dynamics that require further interpretation. These components have temporal dynamics that need to be understood in detail. One crucial question is how the dynamics of a system of interest deviate from its expected behaviour at some point in time. A classical approach is inspecting the "anomalies" of a time series, i.e., the deviation from the mean seasonal cycle at a certain day of year.

Another key description of such system dynamics are trends. We estimated trends of the indicators as well as of their seasonal amplitude using the Theil–Sen estimator (Theil, 1950; Sen, 1968). The advantage of the Theil–Sen estimator is its robustness to up to 29.3% of outliers[2], while

[2]We need a proportion, of at least $\frac{1}{2}$ of valid slopes. Let ε be the fraction of outliers. The fraction of slopes (combination of two points) that do not contain any outlier is $(1-\varepsilon)^2$ and must be larger or equal to $\frac{1}{2}$. This means $\varepsilon \leq 1 - \frac{1}{\sqrt{2}} \approx 29.3\%$ (Rousseeuw and Leroy, 1987, p. 67).

ordinary least-squares regression is highly sensitive to such values. The calculation of the estimator consists simply in computing the median of the slopes spanned by all possible pairs of points,

$$\text{slope}_{ij} = \frac{z_i - z_j}{t_i - t_j}, \tag{3.9}$$

where z_i is the value of the response variable at time step i and t_i the time at time step i. In our experiments, we computed the slopes separately per pixel and principal component with time as the predictor and the value of the principal component as the response variable.

To test the slopes for significance, we used the Mann–Kendall statistics (Mann, 1945; Kendall, 1970) and adjusted the resulting p values with the Benjamini–Hochberg method to control for the false discovery rate (Benjamini and Hochberg, 1995). Slopes with an adjusted $p < 0.05$ were deemed significant.

To identify disruptions in trajectories, breakpoint detection provides a good framework for analysis (Coppin et al., 2004; Tewkesbury et al., 2015; Gómez et al., 2016; Zhu, 2017). For the estimation of breakpoints, the generalized fluctuation test framework (Kuan and Hornik, 1995) was used to test for the presence of breakpoints. The framework uses recursive residuals (Brown et al., 1975)[3] such that a breakpoint is identified when the mean of the recursive residuals deviates from zero. We used the implementation in Zeileis et al. (2002). For practical reasons, here we only focus on the largest breakpoint.

The analysis of a different type of dynamic considers bivariate relations. In the context of oscillating signals it is particularly instructive to quantify their degree of phase shift and direction—even if both signals are not linearily related. A "hysteresis" would be such a pattern describing how the pathways $A \to B$ and $B \to A$ between states A and B differ (Beisner et al., 2003). We estimated hysteresis by calculating the area inside the polygon formed by the mean seasonal cycle of the combinations of two components.

$$\text{Area} = \frac{1}{2} \sum_{i=1}^{n} \mathbf{x}_i (\mathbf{y}_{i+1} - \mathbf{y}_{i-1}), \tag{3.10}$$

where $n = 46$, the number of time steps in a year, and $\mathbf{x}_i$ and $\mathbf{y}_i$ are the mean seasonal cycle of two PCs at time step i. The polygon is circular; i.e., the indices wrap around the edges of the polygon so that $x_0 = x_n$ and $x_{n+1} = x_1$. This formula gives the actual area inside the polygon

[3] Recursive residuals are a framework that gives a statistical test for changing regression parameters.

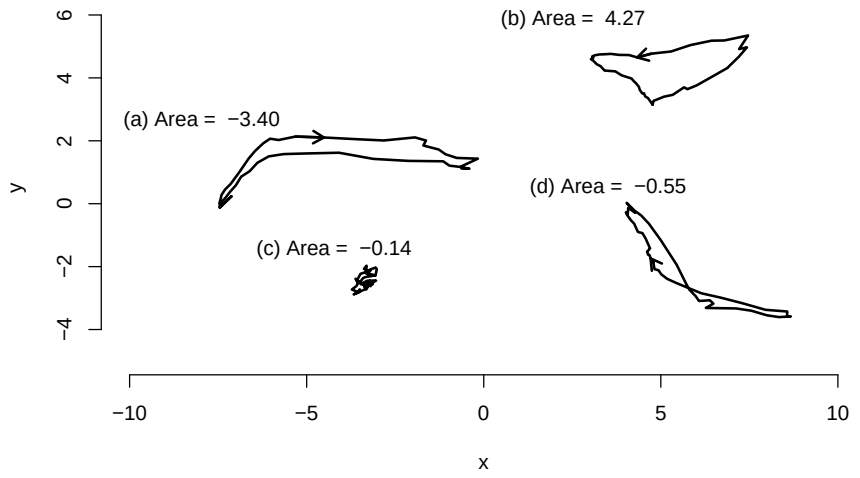

Figure 3.1: Example polygons and their areas, Eq. 3.10, the arrows indicate the directionality. (a) Clockwise polygon with a negative area. (b) Counterclockwise polygon with a positive area. (c) Chaotic polygon with a very low area. (d) Polygon with a single intersection and both a clockwise and counterclockwise portion. The clockwise portion is slightly larger than the counterclockwise portion; therefore the area is slightly negative.

only if it is non-self-intersecting and the vertices run counterclockwise. If the vertices run clockwise, the area is negative. If the polygon is shaped like an 8, the clockwise and counterclockwise parts will cancel each other (partially) out. Trajectories that have larger amplitudes will also tend to have larger areas as illustrated in fig. 3.1.

3.3 Results and Discussion

In the following, we first briefly present and discuss the quality of the global dimensionality reduction (Sect. 3.3.1) and interpret the individual components from an ecological point of view (Sect. 3.3.2). We summarize the global dynamics that we uncovered in the low-dimensional space (Sect. 3.3.3). We characterize the contained seasonal dynamics (Sect. 3.3.4), including spatial patterns of hysteresis (Sect. 3.3.5). We then describe global anomalies of the identified trajectories (Sect. 3.3.6), and discuss

the identified anomalies in depth based on local phenomena (Sect. 3.3.7). Finally, we present global trends and their breakpoints (Sect. 3.3.7).

3.3.1 Quality of the PCA

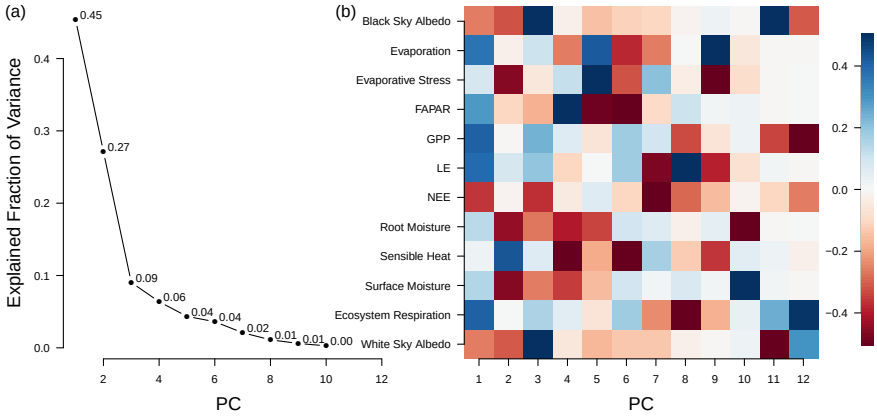

Figure 3.2: (a) Fraction of explained variance of the PCA by component. The knee at component three suggests that components four and higher do not contribute much to total variance. (b) Rotation matrix of the global PCA model (also-called *loadings*, eq. 3.4). The columns of the rotation matrix describe the linear combinations of the (centered and standardized) original variables that make up the principal components. PC_1 is dominated by primary-productivity-related variables, PC_2 by variables describing water availability, and PC_3 by variables describing albedo. Values of the rotation matrix are clamped to the range $[-0.5, 0.5]$, the actual range of the values is $[-0.73, 0.74]$, and $[-0.46, 0.54]$ for the first three components.

Figure 3.2a shows the explained fraction of variance (Eq. 3.5) for the global PCA based on the entire data cube. The two leading components explain 73% of the variance from the 12 variables; additional components contribute relatively little additional variance (PC_3 contributes 9%, and all subsequent PCs less than 7%) each. This results in a "knee" at component 3, which suggests that two indicators are sufficient to capture the major global dynamics of the terrestrial land surface, but we will also consider the third component in the following analyses (Cattell, 1966).

We estimated the reconstruction error sequentially up to the first three

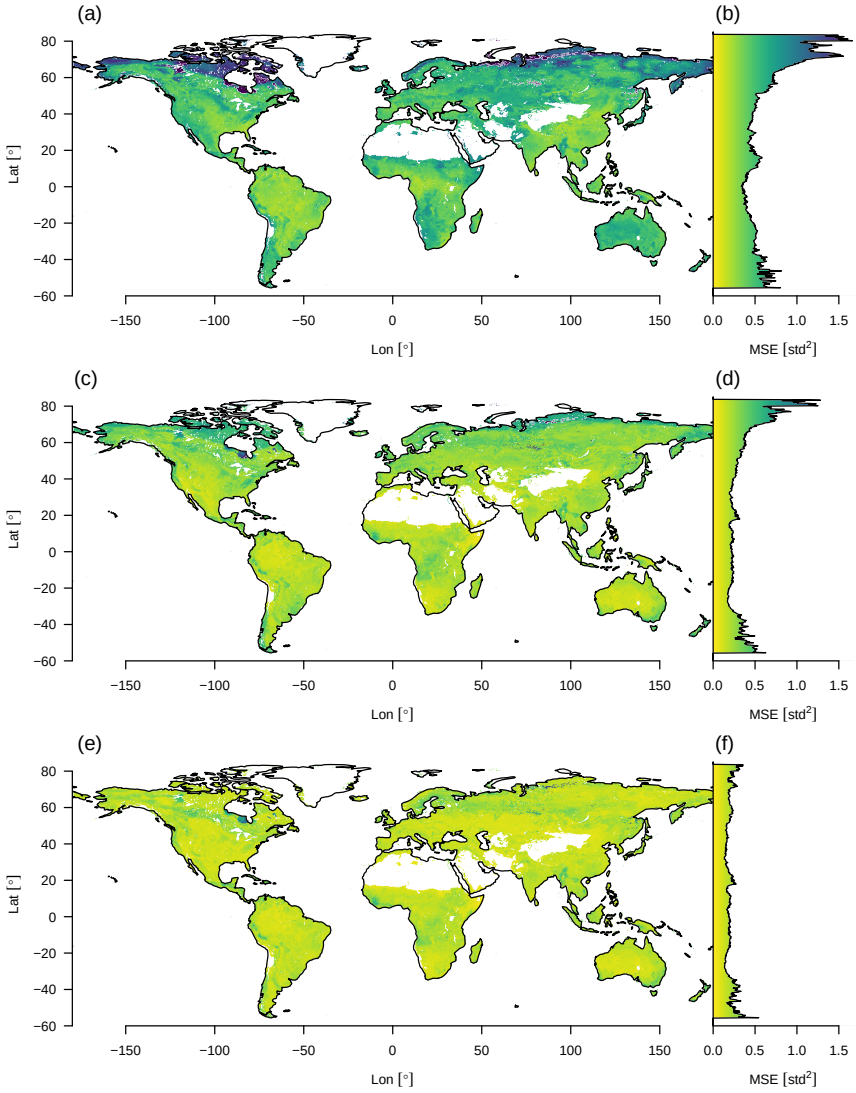

Figure 3.3: Reconstruction error of the data cube using varying numbers of principal components aggregated by the mean squared error. Reconstruction errors aggregated over all time steps and variables are shown in the left column: (a) using only the first component, (c) using the first two, (e) and using the first three. Corresponding right plots (b, d, f) show the mean reconstruction error aggregated by latitude.

principal components (fig. 3.3). Regions that do not fit the model well show a higher reconstruction error. Considering one component only, the highest reconstruction errors appear in high latitudes but decrease strongly with each additional component and nearly vanish if the third component is included.

3.3.2 Interpretation of the PCA

The first PC summarizes variables that are closely related to primary productivity (GPP, LE, NEE, fAPAR) and therefore are highly interrelated (see fig. 3.2b). The energy for photosynthesis comes from solar radiation, and fAPAR is an indicator for the fraction of light absorbed for photosynthesis. The available photosynthetic radiation is used by photosynthesis to fix CO_2 and to produce sugars that maintain the metabolism of the plant. The total uptake of CO_2 is reflected in GPP, which is also closely related to water consumption. The flow of water within the plant is not only essential to enable photosynthesis but also drives the transport of nutrients from the roots. The uplift of water in the plant is ultimately driven by transpiration—together with evaporation from soil surfaces one can obtain the integrated latent energy needed for the phase transition (LE). However, ecosystems also respire; CO_2 is produced by plants in energy-consuming processes as well as by the decomposition of dead organic materials via soil microbes and other heterotrophic organisms. This total respiration can be observed as terrestrial ecosystem respiration (TER). The difference between GPP and TER is the net ecosystem exchange (NEE) rate of CO_2 between ecosystems and atmosphere (Chapin et al., 2006). GPP and TER are also well represented in the first dimension (see fig. 3.2b).

The second component represents variables related to the surface hydrology of ecosystems (see fig. 3.2b). Surface moisture, evaporative stress, root-zone soil moisture, and sensible heat (H) are all essential indicators for the state of plant-available water. While surface moisture is a rather direct measure, evaporative stress is a modeled quantity summarizing the level of plant stress: a value of 0 means that there is no water available for transpiration, while a value of 1 means that transpiration equals the potential transpiration (Martens et al., 2017). Root-zone soil moisture is the moisture content of the soil at rooting depth. If this quantity is below the wilting point, there is no water available for uptake by the plants. Sensible heat is the exchange of energy by a change in temperature; if there is enough water available, then most of the surface available energy will dissipate via evaporation (latent heat), and with decreasing water availability more of the surface heat will be lost due to sensible heat.

We observe that the third component is most strongly related to albedo

(fig. 3.2b). Albedo describes the overall reflectiveness of a surface. Here we refer to broadband (400-3000nm) surface albedo; for an exact definition see Appendix A. Light surfaces, such as snow and sand, reflect most of the incoming radiation, while surfaces that have a high liquid water content or active vegetation absorb most of the incoming radiation. Local changes to albedo can be due to many causes, e.g. snowfall, vegetation greening and browning, or land use change.

The relation of PC_3 to productivity and hydrology is opposite to what we would expect from an albedo axis due to snow and ice in high latitudes. When water is liquid, albedo is negatively correlated with the productivity of the vegetation because vegetation uses radiation as an energy source, hence the negative correlation of albedo with PC_1. Given that liquid water also absorbs radiation we can observe a negative correlation of albedo with PC_2 (see fig. 3.2b). We observe that PC_1 and PC_2 are positively correlated with PC_3 on the positive portion of their axes (see fig. 3.4d and f), which means that the indicator representing albedo is positively correlated with primary productivity and moisture content due to the linearity of the method and the large increase in albedo on the negative portion of PC_1 and PC_2. Finally we can observe that PC_1 and PC_2 have a much higher reconstruction error in snow-covered regions, which is strongly improved by adding PC_3 (see fig. 3.3f). Therefore the third component should be regarded mostly as a binary variable that introduces snow cover, as the other information that is usually associated with albedo is already contained in the first two components.

3.3.3 Distribution of points in PCA space

The bivariate distribution of the first two principal components forms a "triangle" (gray background in fig. 3.4a). At the positive extreme of PC_1 we find one point of the triangle in which ecosystems have a high primary productivity (high values of GPP, fAPAR, LE, TER, and evaporation), mostly limited by radiation. On the lower end of the first principal component we find the other two points of the triangle describing two alternative states of low productivity: These can happen either when the second principal component coincides with temperature limitation (the negative extreme of the second principal component) as seen in the lower left corner of the distribution in fig. 3.4a and b or due to water limitation (positive extreme of the second principal component, the upper left corner in fig. 3.4a). This pattern reflects the two essential global limitations of GPP in terrestrial ecosystems (Anav et al., 2015).

Components 1 and 2 form a subspace in which most of the variability of ecosystems takes place. Component one describes productivity and

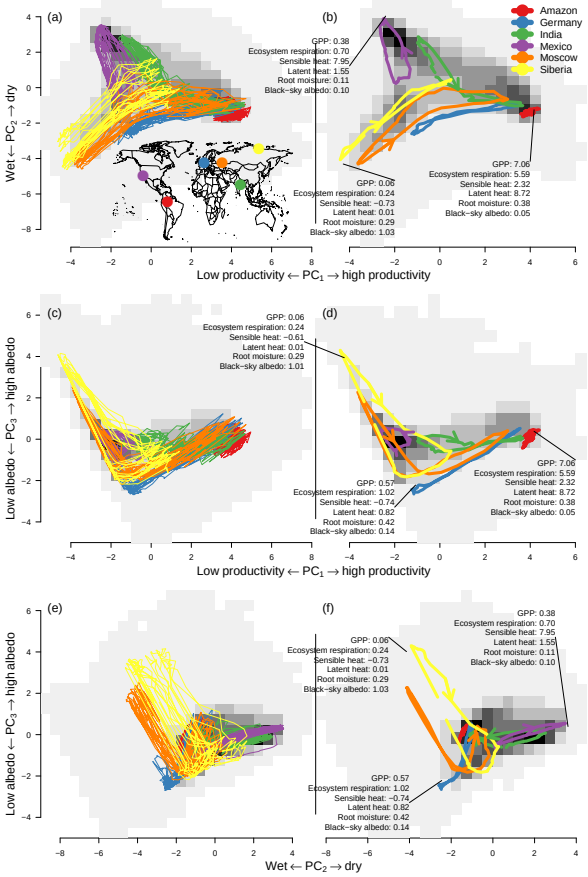

Figure 3.4: Trajectories of some points (colored lines) and the area-weighted density over principal components one and two (the gray background shading shows the density) for (left column) the raw trajectories and (right column) the mean seasonal cycle. The trajectories are shown in the space of PC_1–PC_2 (first row), PC_1–PC_3 (second row), and PC_2–PC_3 (third row). The trajectories were chosen to cover a large area in the space of the first two principal components. Some of the trajectories have an arrow indicating the direction. The numbers illustrate the value of some variables; for units see tab. 3.1. Description of the points is as follows.: Red: tropical rain forest, 2.625°S, 67.625°Wi; blue: maritime climate, 52.375°N, 7.375°E; green: monsoon climate, 22.375°N, 82.375°E; purple: subtropical, 34.875°, 117.625°W; orange: continental climate, 52.375°N, 44.875°E; yellow: arctic climate, 72.375°N, 119.875°E.

component two the limiting factors to productivity. Therefore, we can see that most ecosystems with high values on component one (a high productivity) are at the approximate center of component two. When ecosystems are found outside the center of component two, they have lower values on component one (lower productivity) because they are limited by water or temperature (see fig. 3.4b).

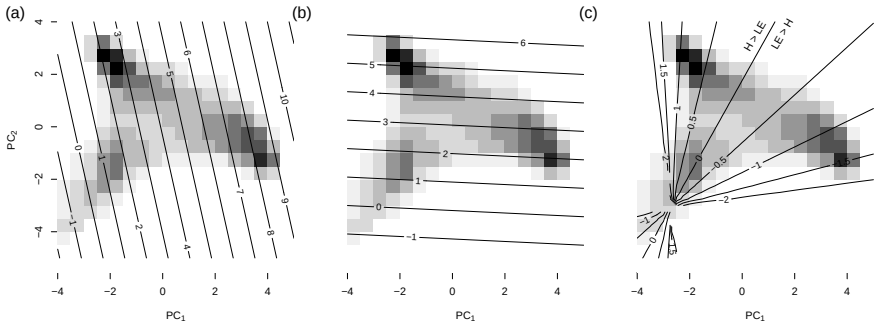

Figure 3.5: The background shading shows the distribution of the mean seasonal cycle of the spatial points (see fig. 3.4). The contour lines represent the reconstruction of the variables from the first two principal components. The reconstructed variables are (a) latent heat (LE), (b) sensible heat (H), and (c) $\log_{10}\left(\frac{\text{sensible heat}}{\text{latent heat}}\right)$, the $\log_{10}$ of the Bowen ratio. Note that the LE and H have been considered in the construction of the PCs and hence are a linear function of the PCs. The Bowen ratio, instead, was not considered here and clearly responds in a nonlinear form.

To further interpret the triangle we analyze how the Bowen ratio embeds in the space of the first two dimensions (see fig. 3.5). Energy fluxes from the surface into the atmosphere can represent either a transfer by conduction and convection (sensible heat) or evaporation (latent heat). Their ratio is the "Bowen ratio", $B = \frac{\mathrm{H}}{\mathrm{LE}}$ (Bowen, 1926). When water is available most of the available energy will be dissipated by evaporation, $B < 1$, resulting in a high latent heat flux. Otherwise, the transfer by latent heat will be low and most of the incoming energy has to be dissipated via sensible heat, $B > 1$. In higher latitudes, there is relatively limited incoming radiation and temperatures are low; therefore there is not much energy to be dissipated and both heat fluxes are low. A high sensible heat flux with respect to the available energy is an indicator of water limitation.

3.3.4 Seasonal Dynamics

The leading principal components represent most of the variability of the space spanned by the observed variables, summarizing the state of a spatiotemporal pixel efficiently. This means that the PCs track the state of a local ecosystem over time (fig. 3.4 left column) or, in the case of the mean seasonal cycle, time of the year (fig. 3.4 right columns). For a representation of the state of the first three components in time and space, see appendix fig. A.1.

A first inspection reveals a substantial overlap of seasonal cycles of very different regions of the world. We also see that very different ecosystems may reach very similar states in the course of the season, even though their seasonal dynamics are very different. For instance, a midlatitude pixel (blue trajectory in fig. 3.4) shows very similar characteristics to tropical forests during peak growing season. This indicates that an ecosystem of the midlatitudes can reach similar levels of productivity and water availability as a tropical rain forest (see also SI fig. A.2). Likewise, for the first two components, many high-latitude areas show similar characteristics to midlatitude areas during winter (low latent and sensible energy release as well as low GPP), and many dry areas such as deserts show similar characteristics to areas with a pronounced dry season, e.g. the Mediterranean.

Depending on their position on Earth, ecosystem states can shift from limitation to growth during the year (fig. 3.4b, e.g. Forkel et al., 2015). For example, the orange trajectory in fig. 3.4, an area close to Moscow, shifts from a temperature-limited state in winter to a state of very high productivity during summer. Other ecosystems remain in a single limitation state with only slight shifts, such as the red trajectory in fig. 3.4. In the corner of maximum productivity of the distribution, we find tropical forests characterized by a very low seasonality. We also observe that very different ecosystems can have very similar characteristics during their peak growing season; e.g. green (located in northeast India), blue (northwest Germany), and orange (located close to Moscow) trajectories have very similar characteristics during peak growing season compared to the red trajectory.

The third component shows a different picture. Due to a consistent winter snow cover in higher latitudes the albedo is much higher and the amplitude of the mean seasonal cycle is much larger than in other ecosystems. Other areas show comparatively little variance on the third component and their relation to productivity and moisture content is even positively correlated to the third component, which is the opposite of what is expected from an albedo axis.

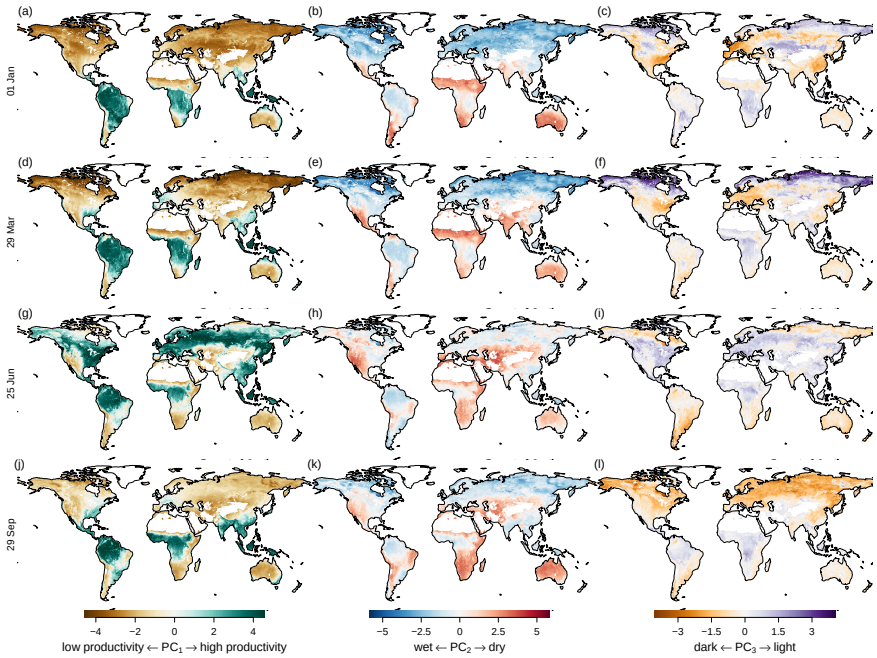

Figure 3.6: Mean seasonal cycle of the first three principal components (in columns) during the seasons (in rows). Left column: first principal component. Middle column: second principal component. Right column: third principal component. Rows from top to bottom: equally spaced intervals during the year. Values have been clamped to 0.7 times their range to increase contrast.

The global pattern of the first principal component follows the productivity cycles during summer and winter (fig. 3.6, left column) of the Northern Hemisphere, with positive values (high productivity, green) during summer and negative values (low productivity, brown) during winter. The tropics show high productivity all year. The global pattern shows the well-known green wave (Schwartz, 1994, 1998) because the first dimension integrates over all variables that correlate with plant productivity.

The second principal component (fig. 3.6, middle column) tracks water availability: red and light red areas indicate water deficiency, light blue areas excess water, and dark blue areas growth limitation due to cold. Areas which are temperature limited during winter but have a growing season during summer, such as boreal forests, change from dark blue in

winter to light blue during the growing season. Areas which have low productivity during a dry season change their coloring from red to light red during the growing season, e.g the northwest of Mexico and southwest of the United States.

The third principal component (fig. 3.6, right column) tracks surface reflectance. Therefore we can see the highest values in the arctic region during winter, and other areas vary much less in their reflectance throughout the year. Again, the third component shows a counterintuitive behavior in the midlatitudes, as it is positively correlated with productivity and therefore shows the opposite behavior of what would be expected from an indicator tracking albedo.

Although the principal components are globally uncorrelated, they covary locally (see fig. A.3). Ecosystems with a dry season have a negative covariance between PC_1 and PC_2, while ecosystems that cease productivity in winter have a positive covariance. Cold arid steppes and boreal climates show a negative covariance between PC_1 and PC_3. While other ecosystems that have a strong seasonal cycle show a positive correlation, many tropical ecosystems do not show a large covariance. A very similar picture is painted between the covariance of PC_2 and PC_3: boreal and steppe ecosystems show a negative covariance, while most other ecosystems show a more or less pronounced positive covariance, again depending on the strength of the seasonality.

Observing the mean seasonal cycle of the principal components gives us a tool to characterize ecosystems and may also serve as a basis for further analysis, such as a global comparison of ecosystems (Metzger et al., 2013; Mahecha et al., 2017).

3.3.5 *Hysteresis*

The alternative return path between ecosystem states forming the hysteresis loops (see Methods) arises from the ecosystem tracking seasonal changes in the environmental condition, e.g. summer–winter or dry–rainy seasons (fig. 3.4b). Hysteresis is a common occurrence in ecological systems (Folke et al., 2004; Blonder et al., 2017; Renner et al., 2019). For instance, a hysteresis loop can be found when plotting soil respiration against soil temperature (Tang et al., 2005). The sensitivity of soil respiration to soil temperature changes seasonally due to changing soil moisture and photosynthesis (by supplying carbon to the rhizosphere), producing a seasonally changing hysteresis effect (Gaumont-Guay et al., 2006; Richardson et al., 2006; Zhang et al., 2018). Biological variables also show a hysteresis effect in their relations with atmospheric variables; e.g. Mahecha et al. (2007a) found a hysteresis effect between seasonal NEE, temperature, and a num-

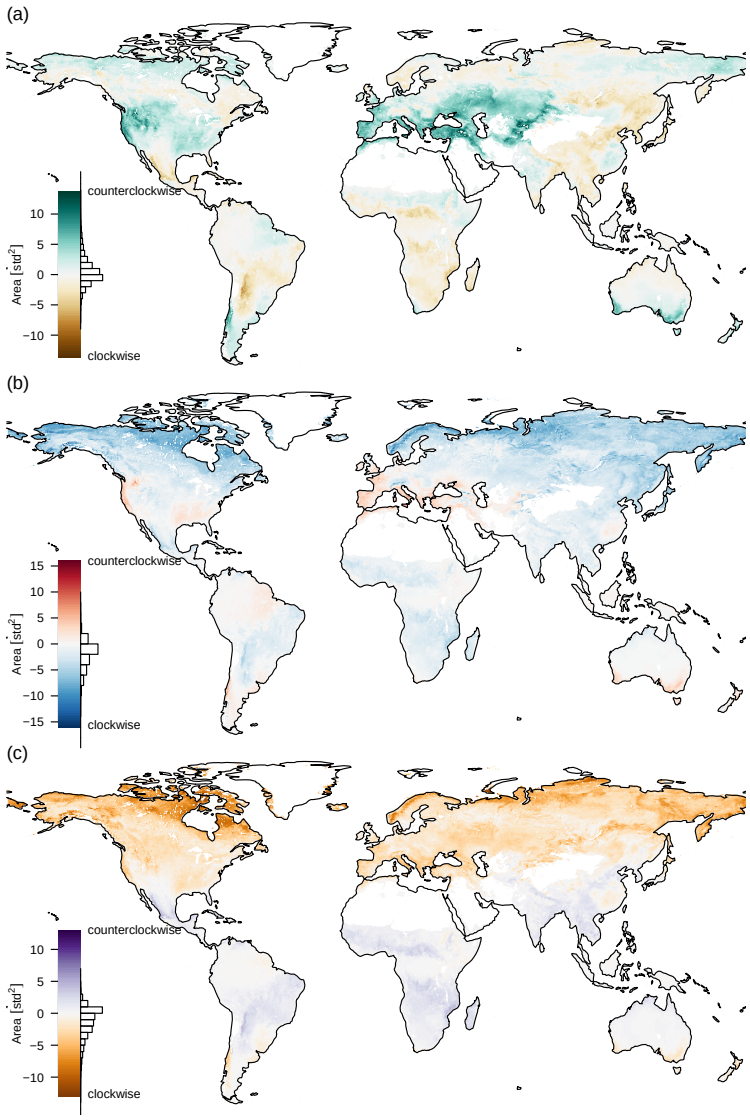

Figure 3.7: The area inside the mean seasonal cycles of (a) PC_1–PC_2, (b) PC_1–PC_3, and (c) PC_2–PC_3. The area is positive if the direction is counterclockwise and negative if the direction is clockwise. Most of the trajectories need a strong seasonal cycle to show a pronounced hysteresis effect. If the mean seasonal cycle intersects, the areas may cancel each other out, e.g. the green trajectory of 3.4b.

ber of other ecosystem and climate-related variables. Here we look at the mean seasonal cycles of pairs of indicators and the area they enclose.

The orange trajectory (area close to Moscow) in fig. 3.4b shows that the paths between maximum and minimum productivity can be very different, in contrast to the blue trajectory located in the northwest of Germany which also has a very pronounced yearly cycle but shows no such effect. Figure 3.4 also indicates that the area inside the mean seasonal cycles of PC_1–PC_2 and PC_1–PC_3 shows important characteristics while hysteresis in PC_2–PC_3 is a much less pronounced feature; i.e., we can only see a pronounced area inside the yellow curve in fig. 3.4f.

The trajectories that show a more pronounced counterclockwise hysteresis effect in PC_1–PC_2 (fig. 3.7a) are areas with a warm and temperate climate and partially those that have a snow climate with warm summers, i.e. areas that have pronounced growing, dry, and wet seasons and therefore shift their limitations more strongly during the year. That means the moisture reserves are depleted during growing season and therefore the return path has higher values on the second principal component (the climatic zones are taken from the Köppen–Geiger classification; Kottek et al., 2006). We can also see that areas with dry winters tend to have a clockwise hysteresis effect, e.g. many areas in East Asia. Due to the humid summers there is no increasing water limitation during the summer months which causes a decrease for PC_2 instead of an increase. Other areas with clockwise hysteresis can be found in winter dry areas in the Andes and the winter dry areas north and south of the African rain forests. Tropical rain forests do not show any hysteresis effect due to their low seasonality. In general we can say that the area inside the mean seasonal cycle trajectory of PC_1–PC_2 depends mostly on water availability in the growing and non-growing seasons, i.e., the contrast of wet summer and dry winter vs. dry summer and wet winter.

The hysteresis effect on PC_1–PC_3 (fig. 3.7b) shows a pronounced counterclockwise MSC trajectory mostly in warm temperate climates with dry summers, while it shows a clockwise MSC trajectory in most other areas; again tropical rain forests are an exception due to their low seasonality. The most pronounced clockwise MSC trajectories can be found in tundra climates in arctic latitudes, where we have a consistent winter snow cover and a very short growing period. A counterclockwise rotation can be found in summer dry areas, such as the Mediterranean and California, but also some more humid areas, such as the southeast United States and the southeast coast of Australia. In these areas we can find a decrease for PC_3 during the non-growing phase which probably corresponds to a drying out of the vegetation and soils.

The hysteresis effect on PC_2–PC_3 (fig. 3.7c) mostly depends on latitude.

There is a large counterclockwise effect in the very northern parts, due to the large amplitude of PC_3. The amplitude gets smaller further south until the rotation reverses in winter dry areas at the northern and southern extremes of the tropics and disappears at the equatorial humid rain forests.

We can see that the hysteresis of pairs of indicators represents large-scale properties of climatic zones. The enclosed area and the direction of the rotation provide interesting information. Hysteresis can provide information on the seasonal availability of water, seasonal dry periods or snowfall. With the method presented here, we can not observe intersecting trajectories, which would probably provide even more interesting insights (e.g. the green trajectory in fig. 3.4b).

3.3.6 Anomalies of the Trajectories

The deviation of the trajectories from their mean seasonal cycle should reveal anomalies, extreme events, and land cover changes. These anomalies have a directional component which makes them interpretable the same way the original PCs are. Therefore one can infer the state of the ecosystem during an anomaly. For instance the well-known Russian heatwave in summer 2010 (Flach et al., 2018) appears in fig. 3.8 as a dark brown spot in the southern part of the affected area, indicating lower productivity, and as a thin green line in the northern parts, indicating increased productivity. This confirms earlier reports in which only the southern agricultural ecosystems were negatively affected by the heatwave, while the northern predominantly forest ecosystems rather benefited from the heatwave in terms of primary productivity (Flach et al., 2018).

Another example of an extreme event that we find in the PCs is the very wet November rainy season of 2006 in the Horn of Africa after a very dry rainy season in the previous year. This event was reported to bring heavy rainfall and flooding events which caused an emergency for the local population but also increased ecosystem productivity (Nicholson, 2014). The rainfall event appears as green and blue spots in fig. 3.8b and c, preceded by the drought events which appear as red and brown spots.

Figure 3.8f and g also show the strong drought events in the Amazon, particularly the droughts of 2005 and 2010 (Doughty et al., 2015; Feldpausch et al., 2016) appear strongly north and south of the Amazon basin. The central Amazon basin does not show these strong events, because the observable response of the ecosystem was buffered due to the large water storage capacity in the central Amazon basin.

Another extreme event that can be seen is the extreme snow and cold event affecting central and south China in January 2008, causing the temporary displacement of 1.7 million people and economic losses of

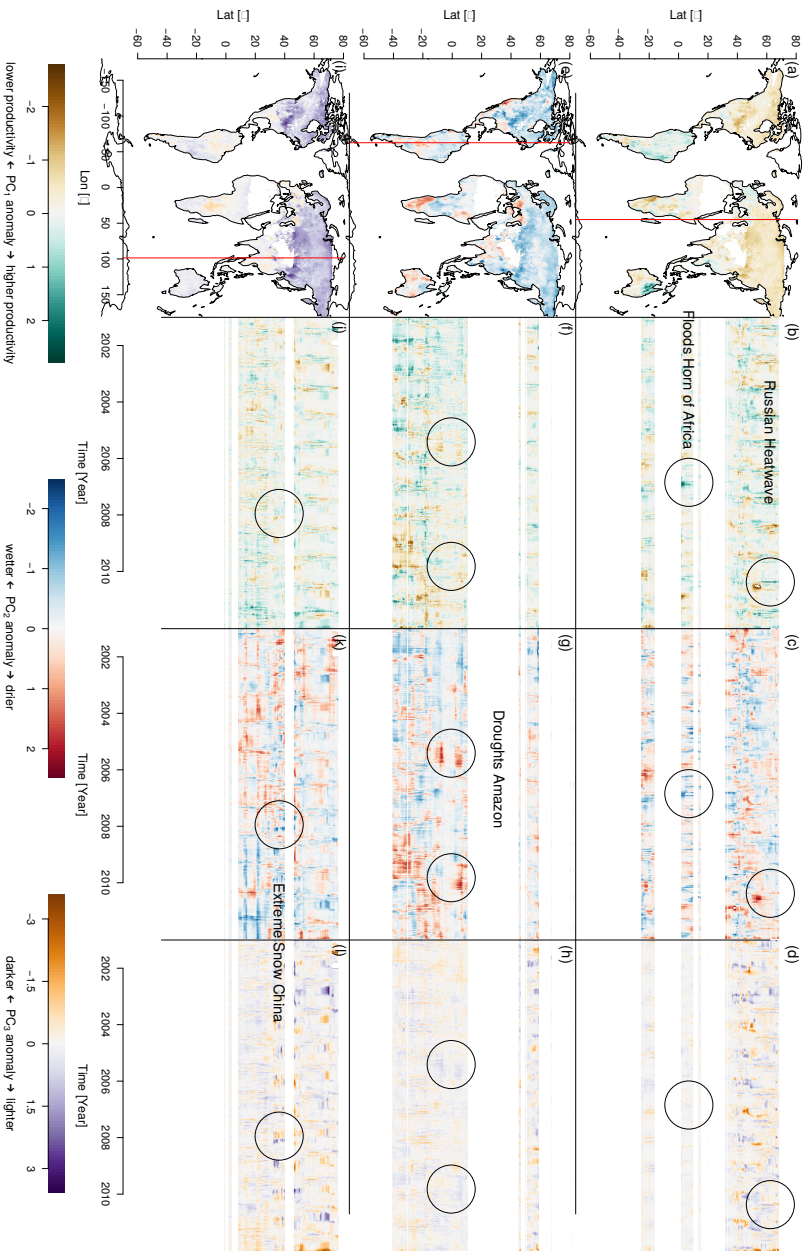

Figure 3.8: (Caption next page.)

Figure 3.8: (Previous page.) Anomalies of the first three principal components. The brown–green contrast shows the anomalies on PC_1, a relative low productivity or greening respectively. The blue–red contrast shows the anomalies on PC_2, a relative wetness or dryness respectively. The brown–purple contrast shows the anomaly on PC_3, a relative deviation in albedo. Panels (a), (e), and (i) are maps showing the anomalies on January 1, 2001, of PC_1–PC_3, respectively. Panels (b), (c), and (d) show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line (a). The effects of the floods on the Horn of Africa (2006) and the Russian heatwave (2010) are highlighted by circles. Panels (f), (g), and (h) show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in sub-figure (e). Strong droughts in the Amazon during 2005 and 2010 can be observed as large red spots on the fringes of the Amazon basin (highlighted by circles). Panels (j), (k), and (l) show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in (i) respectively. A strong snowfall event affecting central and southern China is marked as circles.

approximately US $ 21 billion (Hao et al., 2011). This event shows up clearly on PC_2 and PC_3 as cold and light anomalies respectively (see fig. 3.8k and f).

3.3.7 Single Trajectories

Exploring single temporal trajectories can give insight into past events that happened at a certain place, such as extreme events or permanent changes in ecosystems. The creation of trajectories is an old method used by ecologists, mostly on species assembly data of local communities, to observe how the composition changes over time (e.g. Legendre et al., 1984; Ardisson et al., 1990). In this context, we observe how the states of the ecosystems inside the grid cell shift over time, which comprises a much larger area than a local community but is probably also less sensitive to very localized impacts than a community-level analysis. One of the main differences of the method applied here from the classical ecological indicators is that the trajectories observed here are embedded into the space spanned by a single global PCA, and therefore we can compare a much broader range of ecosystems directly.

The seasonal amplitude of the trajectory in the Brazilian Amazon increases due to deforestation and crop growth cycles. Figure 3.9a shows an area in the Brazilian Amazon in Rondônia (9.5°S, 63.5°W) which was affected by large-scale land use change and deforestation. It can be seen that

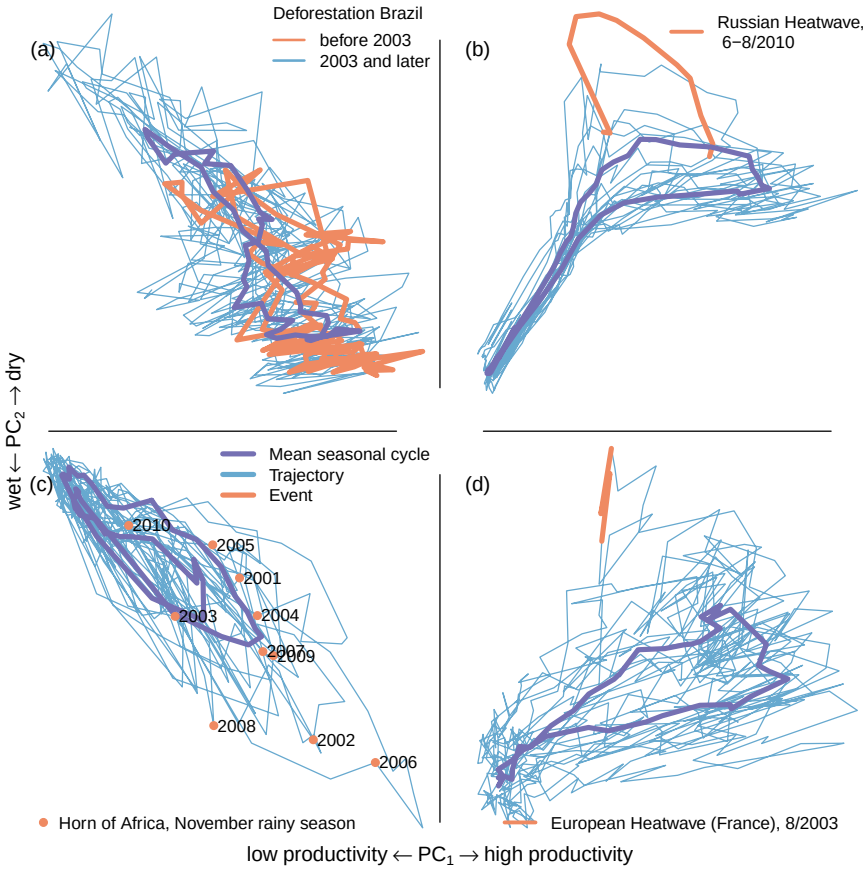

Figure 3.9: Trajectories of the first two principal components for single pixels. (a) Deforestation increases the seasonal amplitude of the first two PCs (Brazilian rain forest, 9.5°S 63.5°W). The red line shows the trajectory before 2003 and the blue line the trajectory 2003 and later, a strong increase in seasonal amplitude can be observed after 2003. (b) The heatwave is clearly visible in the trajectory (red, Russian heat wave, summer 2010, 56°N 45.5°E). (c) Rainfall in the short rainy season (November–December) influences agricultural yield and can cause flooding (extreme flooding after drought, 11/2006, 3°N 45.5°E). (d) The European heatwave in summer 2003 was one of the strongest on record (France, 47.2°N 3.8°E). The mean seasonal cycles of the trajectories are shown in purple.

the seasonal amplitude increases strongly after the beginning of 2003. This increased amplitude could be due to a decreased water storage capability and drying out of soils because of deforestation which in turn leads to a larger variability in ecosystem productivity. Therefore, during periods of no rain, large-scale deforestation can cause a shift in local-scale circulation patterns, causing lower local precipitation (Khanna et al., 2017). Another possible reason could be agriculture, i.e. crop growth and harvest cause an increased amplitude in the cycle of productivity. An analysis of the trajectory can point to the nature of the change, however finding the exact causes for the change requires a deeper analysis.

The 2010 Russian heatwave has a very clear signal in the trajectories. Figure 3.9b shows the deviation of the trajectory during the Russian heatwave (red line) in an area east of Moscow (56°N 45.5°E). In the southern grass- and croplands, the heatwave caused the productivity to drop significantly during summer due to a depletion of soil moisture. In the northern forested parts affected, the heatwave caused an increase in ecosystem productivity during spring due to higher temperatures combined with sufficient water availability. This shows the compound nature of this extreme event (see fig. 3.8a and Flach et al. 2018). The analysis of the trajectory points directly towards the different types of extremes and responses that happened in the biosphere during the heatwave.

Precipitation variability during the November rainy season in the Horn of Africa (3°N 45.5°E) can be seen in fig. 3.9c as red dots. The November rains have implications for food security because the second crop season depends on them. In 2006, the rainfall events were unusually strong and caused widespread flooding and disaster but also higher ecosystem productivity (see also fig. 3.8). This was especially devastating because it followed a long drought that caused crop failures. Note also the two rainy seasons in the mean seasonal cycle (purple line in fig. 3.9c).

The 2003 European heatwave is highlighted in the trajectories just like the 2010 Russian heatwave. Figure 3.9d shows the trajectory during the August 2003 heatwave in Europe (France, 47.2°N 3.8°E). The heatwave was unprecedented and caused large-scale environmental, health, and economic losses (Ciais et al., 2005; García-Herrera et al., 2010; Miralles et al., 2014). The 2010 heatwave was stronger than the 2003 heatwave but the strongest parts of the 2010 heatwave were in eastern Europe (see fig. 3.8), while the epicenter of the 2003 heatwave was located in France.

As we have seen here, observing single trajectories in reduced space can give us important insights into ecosystem states and changes that occur. While the trajectories can point us towards abnormal events, they can only be the starting points for deeper analysis to understand the details of such state changes.

3.3.8 Trends in Trajectories

The accumulation of CO_2 in the atmosphere should cause an increase in global productivity of plants due to CO_2 fertilization, while larger and more frequent droughts and other extremes may counteract this trend. Satellite observations and models have shown that during the last decades the world's ecosystems have greened up during growing seasons. This is explained by CO_2 fertilization, nitrogen deposition, climate change and land cover change (Zhu et al., 2016; Huang et al., 2018; Anav et al., 2015). Tropical forests especially showed strong greening trends.

General patterns of trends can be observed, such as a positive trend (higher productivity) on the first principal component in many arctic regions, see fig. 3.10. Many of these regions also show a wetness trend, with the notable exception of the western parts of Alaska, which have become drier. This is important, because wildfires play a major role in these ecosystems (Jolly et al., 2015; Foster et al., 2019). These changes are also accompanied by a decrease for PC_3 due to a loss in snow cover. A large-scale drying trend can also be observed across large parts of western Russia. Increasing productivity can also be observed for large parts of the Indian subcontinent and eastern Australia. Negative trends in the first component can also be observed: they are generally smaller and appear in regions around the Amazon and the Congo Basin, but also in parts of western Australia. The main difference from previous analyses on the observations presented here is that Zhu et al. (2016), for example, looked only at trends during the growing season while this analysis uses the entire time series to calculate the slope.

In the Amazon basin, we find a drying trend accompanied by a decrease in productivity and a slight increase in PC_3. In the Congo Basin, we find a wetting trend and an increasing productivity in the northern parts, while the southern part and woodland south of the Congo basin show a strong drying trend with decreased productivity. This is different to the findings of Zhou et al. (2014), who found a widespread browning of vegetation in the entire Congo Basin for the April–May–June seasons during the period 2000–2012. The findings of Zhou et al. (2014) are not reflected in our data, especially compared to the areas surrounding the Congo Basin. We can find only minor browning effects inside the basin and our findings are more in line with the global greening (Zhu et al., 2016), which shows a browning mostly outside the Congo Basin.

In eastern Australia we find a strong wetting and greening trends due to Australia having a "millennium drought" since the mid-1990s with peak year in 2002 and 2006 (Nicholls, 2004; Horridge et al., 2005; van Dijk et al., 2013) and extreme floods in 2010–2011 (Hendon et al., 2014).

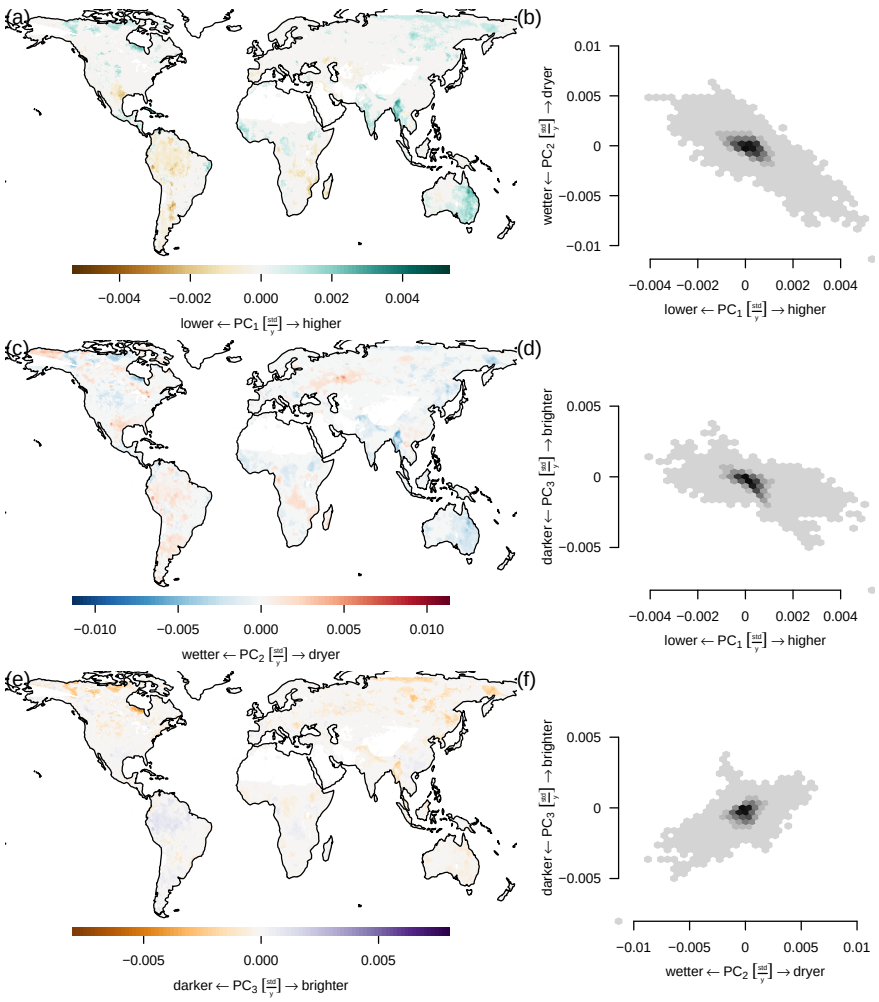

Figure 3.10: (a), (c), (e) Trends in PC_1–PC_3 respectively (2001–2011). (b), (d), (f) Bivariate distribution of trends. Trends were calculated using the Theil–Sen estimator. Panels (a), (c), and (e) show significant trends only ($p < 0.05$, Benjamini–Hochberg adjusted).

Large parts of the Indian subcontinent show a trend towards higher productivity and an overall wetter climate. The greening trend in India happens mostly over irrigated cropland. However browning trends over natural vegetation have been observed but do not emerge in our analysis (Sarmah et al., 2018). A very notable greening and wetting trend can be observed in Myanmar due to an increase in intense rainfall events and storms, although the central part experienced some strong droughts at the same time (Rao et al., 2013). In Myanmar we also find one of the strongest trends in PC_3 outside of the Arctic.

In large parts of the Arctic, a trend towards higher productivity can be observed. Vegetation models attribute this general increase in productivity to CO_2 fertilization and climate change. This also changes the characteristics of the seasonal cycles (Forkel et al., 2016). Stine et al. (2009) found a decreased seasonal amplitude of surface temperature over northern latitudes due to winter warming.

The seasonal amplitude of atmospheric CO_2 concentrations has been increasing due to climate change causing longer growing seasons and changing vegetation cover in northern ecosystems (Forkel et al., 2016; Graven et al., 2013; Keeling et al., 1996). Therefore we checked for trends in the seasonal amplitude, but because each time series only consists of 11 values (one amplitude per year), after adjusting the p values for false discovery rate, we could not find a significant slope. However, there were many significant slopes with the unadjusted p values; see the appendix, fig. A.4.

Another way to detect changes to the biosphere consists in the detection of breakpoints, which has been applied successfully to detect changes in global normalized difference vegetation index (NDVI) time series (de Jong et al., 2011; Forkel et al., 2013) or generally to detect changes in time series (Verbesselt et al., 2010). A proof-of-concept analysis can be found in fig. A.5. We hypothesize that applying this method to indicators instead of variables can detect a wider range of breakpoints analyzing a single time series.

3.3.9 Relations to Other PCA-type Analyses

One of the most popular applications of PCA in meteorology are EOFs, which typically apply PCA to single variables, i.e., on a dataset with the dimensions lat $\times$ lon $\times$ time, although EOFs can be calculated from multiple variables. EOFs can be calculated in S mode and T mode. If we matricize our data cube so that we have time in rows and lat $\times$ lon $\times$ variables in columns, then S mode PCA works on the correlation matrix of the combined variable and space dimension. In T mode, the PCA works

on the correlation matrix formed by the time dimension (Wilks, 2011). The PCA presented here works slightly differently: (1) We performed a different matricization (lat $\times$ lon $\times$ time in rows and variables in columns) and then (2) the PCA works on the correlation matrix formed by the variables. Therefore in this framework we could call this a V mode PCA.

Ecological analyses usually use PCA with matrices of the shape object $\times$ descriptors. When calculating the PCA on the correlation matrix formed by the objects, then it is called a Q mode analysis. When the PCA is applied to the correlation matrix formed by the variables, then it is called an R mode analysis (Legendre and Legendre, 1998). The PCA carried out in this study is closest to an R mode analysis. In the present case the descriptors are the various data streams and the objects are the spatiotemporal pixels. These modes have been defined by Cattell (1952), an extensive description of all modes can be found in Richman (1986).

Using PCA as a method for dimensionality reduction means that we are assuming linear relations among features. A nonlinear method could possibly be more efficient in reducing the number of variables, but would also have significant disadvantages. In particular: nonlinear methods typically require tuning specific parameters, objective criteria are often lacking, a proper weighting of observations is difficult, the methods are often not reversible, and it is harder to interpret the resulting indicators due to their nonlinear nature (see Chapter 2). The salient feature of PCA is that an inverse projection is well defined and allows for a deeper inspection of the errors, which is not the case for nonlinear methods which learn a highly flexible transformation that is hard to invert. Therefore interpretability of the transform in meaningful physical units in the input space is often not possible. In the machine-learning community, this problem is known as the "pre-imaging problem" (Mika et al., 1999; Arenas-Garcia et al., 2013) and is a matter of current research.

3.4 Conclusions

To monitor the complexity of the changes occurring in times of an increasing human impact on the environment, we used PCA to construct indicators from a large number of data streams that track ecosystem state in space and time on a global scale. We showed that a large part of the variability of the terrestrial biosphere can be summarized using three indicators. The first emerging indicator represents carbon exchange, the second indicator shows the availability of water in the ecosystem, while the third indicator mostly represents a binary variable that indicates the presence of snow cover. The distribution in the space of the first two prin-

cipal components reflects the general limitations of ecosystem productivity. Ecosystem production can be limited by either water or energy.

The first three indicators can detect many well-known phenomena without analyzing variables separately due to their compound nature. We showed that the indicators are capable of detecting seasonal hysteresis effects in ecosystems, as well as breakpoints, e.g. large-scale deforestation. The indicators can also track other changes to the seasonal cycle such as patterns of changes to the seasonal amplitudes and trends in ecosystems. Deviations from the mean seasonal cycle of the trajectories indicate extreme events such as the large-scale droughts in the Amazon during 2005 and 2010 and the Russian heatwave of 2010. The events are detected in a similar fashion as with classical multivariate anomaly detection methods but provide additional information on the underlying variables.

Using multivariate indicators (see. Chapter 2, we gain a high level overview of phenomena in ecosystems, and the method therefore provides an interesting tool for analyses where it is required to capture a wide range of phenomena which are not necessarily known a priori. Future research should consider nonlinearities, adding data streams describing other important biosphere variables (e.g. related to biodiversity and habitat quality), and including different subsystems, such as the atmosphere or the anthroposphere.

Chapter 4

The Low Dimensionality of Development

Content

This chapter is based on the following publication:

Kraemer, G., Reichstein, M., Camps-Valls, G., Smits, J., and Mahecha, M. D. (2020). The Low Dimensionality of Development. *Social Indicators Research*, . doi:10.1007/s11205-020-02349-0

Abstract
The World Bank routinely publishes over 1500 "World Development Indicators" to track the socioeconomic development at the country level. A range of indices has been proposed to interpret this information. For instance, the "Human Development Index" was designed to specifically capture development in terms of life expectancy, education, and standard of living. However, the general question which independent dimensions are essential to capture all aspects of development still remains open. Using a nonlinear dimensionality reduction approach we aim to extract the core dimensions of development in a highly efficient way. We find that more than 90% of variance in the WDIs can be represented by solely five uncorrelated dimensions. The first dimension, explaining 74% of variance, represents the state of education, health, income, infrastructure, trade, population, and pollution. Although this dimension resembles the HDI, it explains much more variance. The second dimension (explaining 10% of variance) differentiates countries by gender ratios, labor market, and energy production patterns. Here, we differentiate societal structures when comparing e.g. countries from the Middle-East to the Post-Soviet area. Our analysis confirms that most countries show rather consistent temporal trends towards wealthier and aging societies. We can also find deviations from the long-term trajectories during warfare, environmental disasters, or fundamental political changes. The data-driven nature of the extracted dimensions complements classical indicator approaches, allowing a broader exploration of global development space. The extracted independent dimensions represent different aspects of development that need to be considered when proposing new metric indices.

4.1 Introduction

During the last decades, humanity has achieved on average longer life spans, decreased child mortality, better access to health care and economic growth (UNDP, 2019). In emerging countries like China and India many people have escaped extreme poverty (less than 1.90 US$ per person per day) in the wake of persistent economic growth (UNDP, 2016). To measure development, a wide range of variables are routinely made available by the World Bank, describing multiple facets of societal conditions. These "World Development Indicators" (WDIs, revision of 3/5/2018; The World Bank, 2018a) have become a key data resource that today contains more than 1500 variables with annual values for most countries of the world.

A widely accepted method for assessing development consists in the construction of indicators (hereafter called "classical" indicators) based on

expert knowledge that allow ranking countries by their development status and tracking them over time. A multitude of such classical indicators have been developed over the past few decades (Parris and Kates, 2003; Shaker, 2018; Ghislandi et al., 2018), focusing on different aspects of development. For instance, the United Nations Development Programme (UNDP) uses a Multidimensional Poverty Index, a Gender Development Index, a Gender Inequality Index, amongst others for reporting on human development (UNDP, 2019). The UNDP's most prominent indicator is the Human Development Index (HDI), which is the geometric mean of indicators describing life expectancy, education, and income (UNDP, 2019). However, there are many other efforts to produce relevant indicators, such as the Genuine Progress Index (Kubiszewski et al., 2013), the Global Footprint and Biocapacity indicators (McRae et al., 2016), and the POLITY scores (Marshall and Elzinga-Marshall, 2017), to name just a few. These classical approaches are well suited for describing and communicating selected aspects of development, e.g. the HDI has been specifically developed "to shift the focus of development economics from national income accounting to people-centered policies" (UNDP, 2018).

An alternative to approach to constructing indices consists in using purely data-driven methods, such as PCA (Pearson, 1901) or "Factor Analysis" (FA; Spearman, 1904). As seen in Chapters 2 and 3, PCA linearly compresses a set of variables of interest. The resulting principal component or components represent the main dimensions of variability and can then be interpreted as an emerging indicator (OEDC, 2008). This approach has been used to create indicators of well-being from sets of co-varying variables (Mazziotta and Pareto, 2019). While PCA refers to a well defined method which tries to summarize the variance of an entire dataset, FA refers to a family of methods which assumes a multivariate linear model to explain the influences of a number of latent factors on observed variables. PCA and FA have been used extensively in the social sciences, e.g. to create indicators of well-being (Stanojević and Benčina, 2019) or to construct wealth indices (Filmer and Scott, 2012; Smits and Steendijk, 2015). An advantage of such data-driven methods is that they follow well defined mathematical behaviors and are not subjective, while there is no well established method for the creation of classical indicators (Shaker, 2018). A disadvantage of these methods is that they do not consider the polarity of the variables nor allow for expert based weighting (Mazziotta and Pareto, 2019). A detailed comparison between classical indicators and data driven indicators can be found in the Appendix tab. B.1.

The rationale for dimensionality reduction methods like PCA is that often the intrinsic dimension of a dataset is much lower than the number of variables describing it. In climate science, for example, a set of co-varying

variables observed over a region in the equatorial pacific can be compressed into the Multivariate ENSO Index (MEI, Wolter and Timlin, 1993, 2011a), to describe the state of the El Niño Southern Oscillation (ENSO)—the principal climate mode that determines e.g. food security in many regions of the world. In image vision, the number of main features from a set of images is much less than the number of pixels per image. For example Tenenbaum et al. (2000) shows that pictures taken from the same object at different angles have the viewing angle as the main feature of the set of images. These main features are called "intrinsic dimensions" because they are sufficient to describe the essential nature of the entire dataset, the number of such intrinsic dimensions is called the "intrinsic dimensionality" of the dataset (Bennett, 1969).

Development is a complex concept though, which is reflected in the large number of variables included in the WDI database. However, the large number of indicators let us expect substantial redundant information (Shaker, 2018; Rickels et al., 2016). This issue has also been discussed in the context of the Sustainable Development Goals (SDGs; The World Bank, 2018b). Since their introduction by the United Nations in 2015, the SDGs have become a widely accepted framework to guide policymakers. Today 17 SDGs address the issues of poverty, hunger, health, education, climate change, gender inequality, water, sanitation, energy, urbanization, environment and social justice. To monitor the SDGs, 169 specific targets have been developed which are measured using 232 different indicators included in the WDIs (The World Bank, 2018b; United Nations General Assembly, 2017a), leading to substantial interactions across and within the targets that need to be analyzed (Costanza et al., 2016). Hence, the question emerges how to extract the key information jointly contained in the WDIs that leads to a succinct, objective, and tangible picture of development.

In this paper, we aim to elucidate the most important dimensions of development contained in the WDI dataset, using a data-driven approach. Specifically, we aim to answer the question, how many independent indicators are necessary to summarize development space and what is their interpretation. We exploit the potential of nonlinear dimensionality reduction to identify dimensions that represent these (typically mutually dependent) variables, while preserving relevant properties of the underlying data. The rationale is that we expect strong interactions between the different WDIs which may not be linear.

Understanding what intrinsic dimensionality our current indicators of development have, could have important implications for policy makers. If the intrinsic dimensionality of development proves to be high, one would indeed need to track many indicators synchronously to understand the interplay of different aspects of development. On the contrary, in the case

of a low-dimensional development space, it would be sufficient to track either the emerging dimensions, or the closely related variables to monitor development across countries and time. In fact there is already substantial evidence that supports our hypothesis of a low-dimensional development space. For instance Pradhan et al. (2017) found strong correlations between all SDGs, suggesting that the intrinsic dimensionality of the SDGs is relatively low, but this has not been quantified yet.

This Chapter is divided into five sections. Section 4.2 presents a data-driven approach to extract nonlinear components from the WDI database, Section 4.3 presents the resulting dimensions, their interpretations, global distributions, trends and trajectories. Section 4.4 discusses the relation of the indicators produced by the method presented here with previous indicator approaches, and finally Section 4.5 gives some concluding remarks.

4.2 Data and Methods

4.2.1 Data

To understand the structure and dimensionality of development we rely on the WDI dataset, which is the primary World Bank collection of development indicators, compiled from officially-recognized international sources. The WDIs comprise a total of 1549 variables with yearly data between 1960 and 2016 for 217 countries. As such, it represents the most current, extensive, and accurate global development database available (The World Bank, 2018a).

Even though the WDI dataset is the most comprehensive set of development indicators available, it contains many missing values. Only for the most developed countries the dataset is (nearly) complete. For many other countries—particularly low and middle income countries—many indicators are partly or completely missing. This is problematic, as for most dimension reduction methods a dataset without missing observations is required. To make our analyses possible, we therefore had to select a subset of indicators, countries and years with few missing observations and to fill in the remaining missing observations using gapfilling techniques (see next section). To avoid arbitrariness of the subset selection, a scoring approach was used (see Section 4.2.2) and the 1000 subsets with the highest scores were selected. These 1000 subsets contained a total of 621 variables, 182 countries and the years ranging from 1990 to 2016. The subsets cover almost all categories of variables. The categories with their respective number of variables in the entire WDI dataset and the subsets are “Economic Policy & Debt” (120 out of 518), “Education” (73

out of 151), "Environment" (74 out of 138), "Financial Sector" (29 out of 54), "Gender" (1 out of 21), "Health" (123 out of 226), "Infrastructure" (19 out of 41), "Poverty" (0 out of 24), "Private Sector & Trade" (103 out of 168), "Public Sector" (31 out of 83), and "Social Protection & Labor" (48 out of 161). Jointly these subsets are representative for the original dataset while avoiding large gaps.

4.2.2 *Gapfilling*

The dimensionality reduction approach we have chosen (see Sect. 4.2.3) relies on a full matrix of distances between the different country–year data points. However, given the large amount of data gaps this global distance matrix cannot be computed directly. In the following, we develop an approach to find subsets of the WDI database which we can gapfill and use for estimating distances among data points.

In order to choose subsets of the WDI database covering a wide range of WDIs, countries, and years, but also having as few missing values as possible, the following method was applied: A series of subsets was created from the full WDI dataset using a combination of thresholds for the maximum fraction of missing values for the WDIs, f_v, and countries, f_c, as well as a starting year, y_{start}, and an ending year, y_{end}. We assigned a score to each of the resulting subsets by using a grid search over the parameters, $f_v, f_c \in (0.05, 0.15, \ldots, 0.65)$ and $y_{\text{start}}, y_{\text{end}} \in (1960, 1961, \ldots, 2017), y_{\text{start}} < y_{\text{end}}$. The size of this parameter space is 80997, each with a different combination of missing value thresholds and starting and ending year combinations. The $m = 1000$ subsets with the highest scores were finally chosen to build the global distance matrix. For an overview of the entire method, see fig. 4.1.

Each subset was created from the full WDI dataset by choosing consecutive years with starting year, y_{start}, and ending year, y_{end}, $y_{\text{start}} \leq y_{\text{end}}$; WDIs with a higher missing value fraction, p_v, than the corresponding threshold were dropped ($p_v > f_v$). Then, countries with higher missing value fractions, p_c, than the corresponding threshold were dropped as well ($p_c > f_c$). The number of remaining countries, n_c, and WDIs, n_v, was recorded and the resulting subsets were filtered to retain more observations (the number of countries times the number of years) than variables, leaving a total of 77610 subsets of the WDI for score calculation.

To account for different scales of the parameters, the values had to be rescaled, i.e. we calculated n'_v from n_v by scaling the values from subsets linearly to a minimum of 0 and a maximum of 1, analogously for n'_c, f'_c,

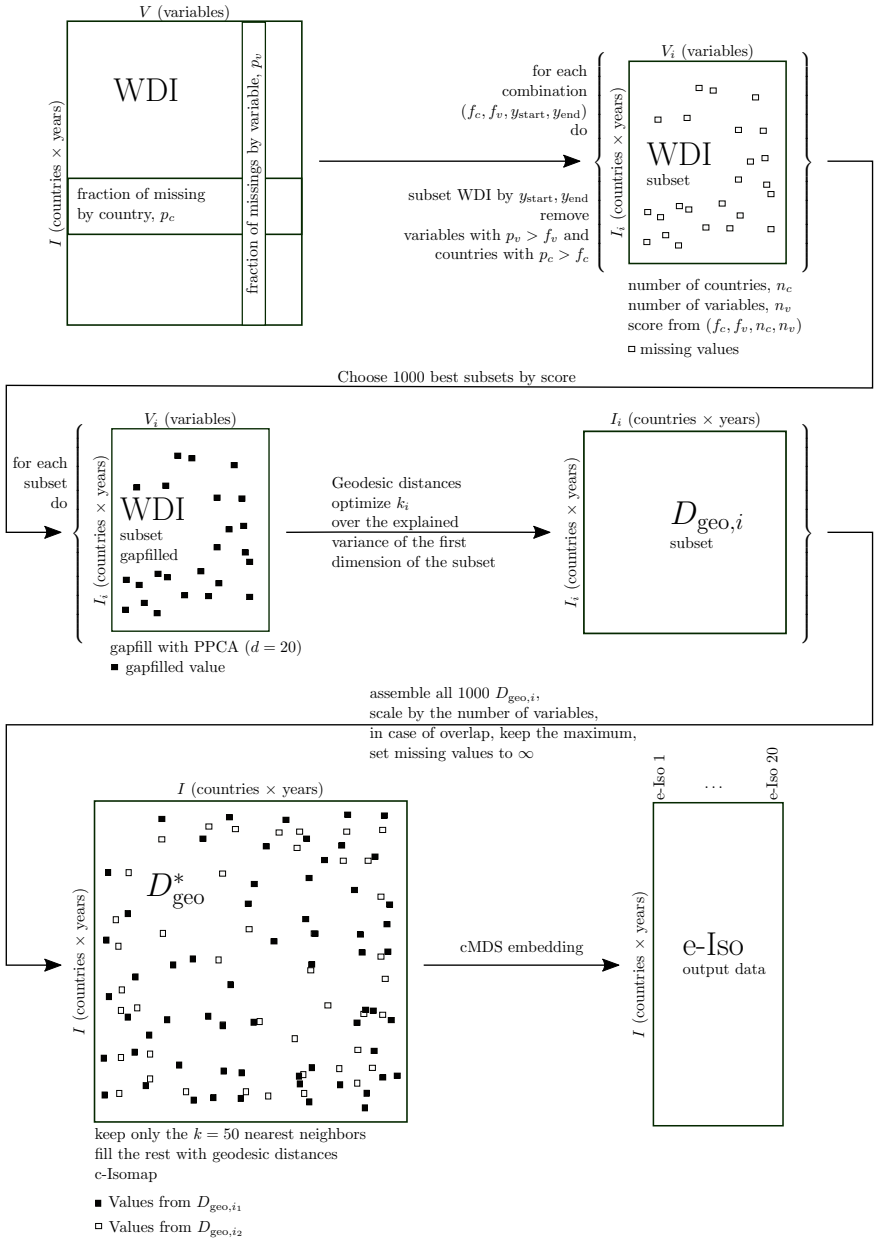

Figure 4.1: Schematic presentation of the ensemble Isomap (e-Iso) algorithm, for details see text.

and f'_v. The final score was then calculated as

$$\text{score} = \sqrt{n'_v n'_c} - \sqrt{f'_c f'_v}.$$

This score calculates the geometric means of the variables of interest. The geometric mean has the advantage over the arithmetic mean that it is very sensitive to single bad values. As we want to maximize the number of countries and WDIs chosen and have as few missing values possible, the final score is the difference between the geometric means. Finally the subsetted WDI data matrices with the 1000 highest scores were selected and a gapfilling procedure using Probabilistic PCA (Porta et al., 2005; Stacklies et al., 2007) was performed on the centered and standardized (z-transformed) variables using the leading 20 dimensions.

4.2.3 Dimensionality Reduction

Dimensionality reduction describes a family of multivariate methods that find alternative representations of data by constructing linear or, in our case, nonlinear combinations of the original variables so that important properties are maintained in as few dimensions as possible. A plethora of algorithms is currently available for dimensionality reduction, both linear and nonlinear (Arenas-Garcia et al., 2013; Van Der Maaten et al., 2009; Kraemer et al., 2018), but PCA is dominating in applied sciences because of ease of use and interpretation.

One method to find an embedding from a known distance matrix is "classical Scaling" (cMDS; Torgerson, 1952), this method is equivalent to PCA if the distance matrix is computed from the observations using Euclidean distance. cMDS finds coordinates in a reduced Euclidean space of dimension i minimizing

$$\|\tau(\mathbf{D}) - \tau(\mathbf{D}_i)\|_2,$$

where $\mathbf{D}$ is the matrix of Euclidean distances of observations and $\mathbf{D}_i$ the matrix of Euclidean distances of the embedded points. $\tau(\mathbf{D}) = -\frac{1}{2}\mathbf{HSH}$, is the "double centering operator", with $\mathbf{S} = [d_{ij}^2]$, $\mathbf{H} = [\delta_{ij} - \frac{1}{n}]$, and $\|\mathbf{X}\|_2 = \sqrt{\sum_{ij} x_{ij}^2}$ the L_2-norm. cMDS and therefore PCA tend to maintain the large scale gradients of the data and cannot cope with nonlinear relations between the covariates.

"Isometric Feature Mapping" (Isomap; Tenenbaum et al., 2000) extends cMDS, but instead of Euclidean distances, it preserves geodesic distances, i.e. the distances measured along a manifold of possibly lower dimensionality,

$$\|\tau(\mathbf{D}_{\text{geo}}) - \tau(\mathbf{D}_i)\|_2.$$

Specifically, Isomap uses geodesic distances, $\mathbf{D}_{\text{geo}} = [d_{\text{geo}}(\mathbf{x}_i, \mathbf{x}_j)]$, which are the distances between two points following a k-nearest neighbor graph of points sampled from the manifold.

Isomap is guaranteed to recover the structure of nonlinear manifolds whose intrinsic geometry is that of a convex region of Euclidean space (Tenenbaum et al., 2000). Isomap unfolds curved manifold which makes the method more efficient than PCA in reducing the number of necessary dimensions in the presence of nonlinearities.

To construct the geodesic distances, a graph is created by connecting each point to its k nearest neighbors and distances are measured along this graph. If the data samples the manifold well enough, then the distances along the graph will approximate the geodesic distances along the manifold. The value of k will determine the quality of the embedding and has to be tuned.

We applied Isomap on the 1000 previously generated subsets of the WDI database. To find the optimal value k of each subset, k_i, Isomap was calculated first with $k_i = 5$ and the residual variance for the embedding of the first component was calculated (see below). This process was repeated increasing the values of k_i by 5 in each step until there was no decrease in the residual variance for the first component any more (Mahecha et al., 2007b). In order to get an intuition of Isomap, we recommend the original publication of the Isomap method (Tenenbaum et al., 2000) which contains an excellent didactic explanation of the method.

4.2.4 Ensemble PCA and Ensemble Isometric Feature Mapping

An observation consists of a country name and year. To calculate a linear embedding (ensemble PCA) over the union of all countries, years and variables chosen before, we used a Probabilistic PCA ($d = 80$, where d is the number of dimensions used in the Probabilistic PCA) to gapfill all the observations and variables occurring in the subsets of the WDI dataset and applied a normal PCA to the gapfilled dataset. This was done to get a baseline for a linear embedding.

We developed "Ensemble Isometric Feature Mapping" (e-Isomap) to produce the final nonlinear embedding based on the different gapfilled subsets of data. E-Isomap combines $m = 1000$ geodesic distance matrices created from the subsets of the previous step and constructs a global ensemble geodesic distance matrix, D^*_{geo}, from the geodesic distance matrices of the m Isomaps.

Let the total set of observations be $I = \{1, \ldots, n\}$ (a country–year combination) and the observed variables $V = \{1, \ldots, r\}$ (the WDIs). We first

perform one Isomap $i \in \{1, \ldots, m\}$ per subset of I and V, I_i and V_i respectively, where $|V_i|$ is the number of variables for Isomap i. The geodesic distance matrix for Isomap i is $\mathbf{D}_{\text{geo},i} = (d_{\text{geo},i}(\mathbf{x}_j, \mathbf{x}_k))_{j,k}$ with $j, k \in I_i$. If a pair of observations $(\mathbf{x}_j, \mathbf{x}_k)$ does not occur in Isomap i, it is treated as a missing value. First the geodesic distance matrices are scaled element-wise to account for the different number of variables used,

$$d'_{\text{geo},i}(\mathbf{x}_j, \mathbf{x}_k) = d_{\text{geo},i}(\mathbf{x}_j, \mathbf{x}_k)\sqrt{\frac{|V|}{|V_i|}},$$

which are then combined into a single geodesic distance matrix $\mathbf{D}^*_{\text{geo}}$ by using the maximum distance value,

$$d^*_{\text{geo}}(\mathbf{x}_j, \mathbf{x}_k) = \max_i d'_{\text{geo},i}(\mathbf{x}_j, \mathbf{x}_k).$$

Missing values are ignored if all values are missing for a pair $(\mathbf{x}_j, \mathbf{x}_k)$ and they are treated as infinite distances. Taking the maximum avoids short-circuiting distances as long as there are few missing values. This provides an accurate approximation of the internal distances.

Finally the k nearest neighbor graph G is constructed from the distance matrix, and each edge $\{\mathbf{x}_i, \mathbf{x}_j\}$ is weighted by $\frac{|\mathbf{x}_i - \mathbf{x}_j|}{\sqrt{M(i)M(j)}}$, where $M(i)$ is the mean distance of $\mathbf{x}_i$ to its k nearest neighbors. This last step is called c-Isomap (Silva and Tenenbaum, 2003) and it contracts sparsely sampled regions of the manifold and expands densely sampled regions, the c-Isomap step proved to give a more evenly distributed embedding. Finally the geodesic distances are calculated on G and classical scaling is performed to find the final embeddings.

4.2.5 Quality Measurement of an Embedding and Influence of Variables

The quality for the embedding is estimated by calculating the residual variance (Tenenbaum et al., 2000) computed as

$$\text{residual variance}_i = 1 - r^2(\widehat{\mathbf{D}}, \mathbf{D}_i) = 1 - \text{explained variance}_i,$$

where $\mathbf{D}_i$ is the matrix of Euclidean distances of the first i embedded components and $\widehat{\mathbf{D}}$ is the distance matrix in the original space, Euclidean distances for PCA and geodesic distances for Isomap. Note that because $\mathbf{D}_i$ and $\widehat{\mathbf{D}}$ are symmetric, we only use one triangle for the calculation of the residual variance. This notion of explained variance is different from the

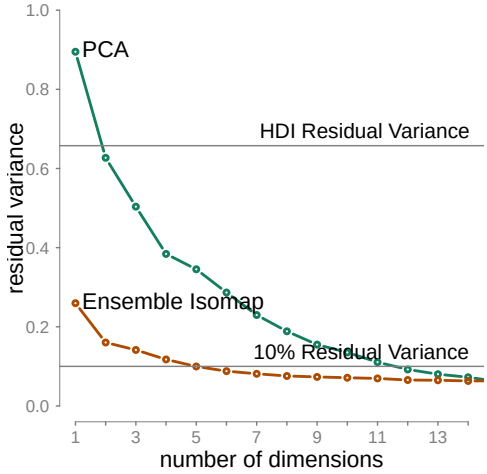

Figure 4.2: The residual variance for the first 14 components. The circled lines represent the residual variance of the Ensemble Isomap and the PCA. Isomap is much more efficient in compressing dimensionality of the data requiring only 5 components to describe more than 90% of the variance, while PCA requires 12 components to describe 90% of variance. The upper grey horizontal line represents the residual variance for the HDI (66%) and the lower one the 10% residual variance boundary.

one usually used for PCA, which is derived from the eigenvalue spectrum, but the measure used here has the advantage that it gives comparable results for arbitrary data such as the HDI and Isomap.

To assess the influence of single variables on the final e-Isomap dimensions, we calculated the distance correlation (dcor, Székely et al., 2007), which is a measure of dependence between variables that takes nonlinearities into account. Due to the strong nonlinearities in the dataset and the embedding method, a simple linear correlation would not have provided sufficient information about the relationships between variables and the embedding dimensions.

4.3 Results

4.3.1 Required Number of Dimensions

Our results suggest that the "development space" described by the WDI data is of low intrinsic dimensionality. Using e-Isomap we needed five dimensions only to explain 90% of the variance of global development (see fig. 4.2). The first dimension alone explains 74% of the variance in the WDI data; Dimension 2 explains 9.9% of the variance and dimensions 3 to 5 explain less than 3% of the variance each. Although the explained variance of dimensions 2–5 seems small compared to that of the first dimension, each of these dimensions still represents a distinct, well defined and highly significant aspect of development, as we will show later. Therefore the raw variances should not be used as the sole measure to discard dimensions.

The finding that such a high compression can be achieved with e-Isomap indicates that the WDIs are highly interdependent and that the underlying processes are highly nonlinear (see fig. 4.2). This is also confirmed by an analogous analysis using linear PCA which cannot compress the data with the same efficiency: the first PCA dimension only explains 10% of the variance, and 12 dimensions are required to express more than 90% of the variance. The cumulative explained variances for the first five e-Isomap dimensions are 74%, 84%, 86%, 88%, and 90%, which is much more than the respective PCA dimensions (10%, 37%, 50%, 61%, and 65%).

To understand if the HDI can compress the data in the same way, we compute the variance of the HDI using the same method. We find that the HDI captures 34% of the variance (see fig. 4.2), which is less than half of the variance captured by the first dimension extracted via nonlinear dimensionality reduction but more than three times the variance explained by the first PCA dimension. If the target is reducing the WDI data to a single dimension, the best performing method is e-Isomap, followed by the HDI, while PCA does not perform this task very well. In other words, the first e-Isomap dimension seems to be a more powerful summary of the WDI data than the HDI.

4.3.2 Intrinsic Dimensions of Development

Our results suggest that the dimensions resulting from the e-Isomap can be indeed interpreted analogously to traditional indicators of development. The main difference from classical indicators is that these dimensions emerge directly from the data. Hence, the interpretation of these indicators has to be achieved *a posteriori*. We also find that the relationship between the WDIs and the dimensions is highly nonlinear (see fig. 4.3) requiring the

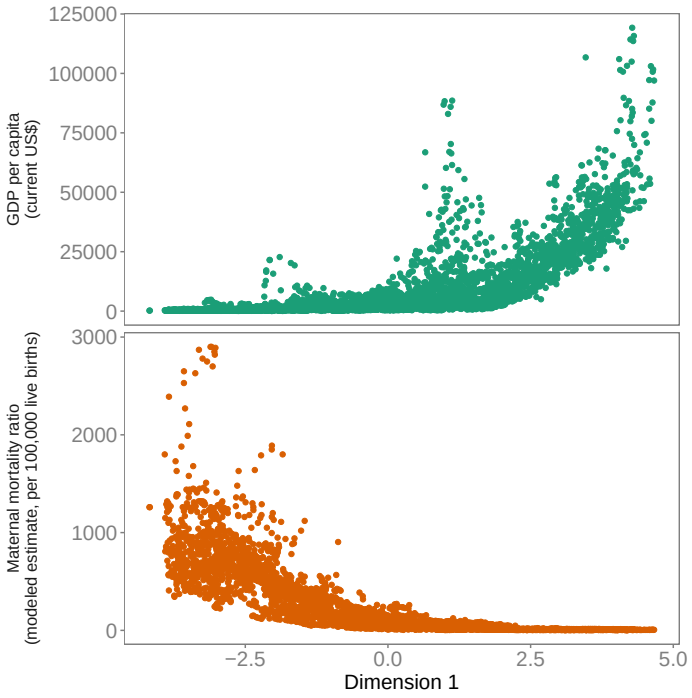

Figure 4.3: Illustrating the nonlinear relation between dimension 1 and GDP per capita and maternal mortality rates. Top: There is a positive correlation between GDP per capita and dimension 1. On the positive end of dimension 1 the per capita income increases strongly, while it increases very slowly on the negative side of dimension 1. Bottom: There is a negative correlation between the maternal mortality rate and dimension 1. The maternal mortality rate decreases strongly on the negative end but does not decrease any more on the positive end.

use of nonlinear measurements of correlation. Here we relate the extracted dimensions to the original data using distance correlation. See fig. 4.4, for a complete and interactive table in the supporting information[1].

We find that dimension 1 essentially represents progress in education, life expectancy, health, and relates to the population pyramid (see fig. 4.4). Additionally, dimension 1 is associated with infrastructure and income-related indicators. Other indicators that strongly correlate with this dimension are related to pollution and primary production and include tariffs

[1] http://bgc-jena.mpg.de/~gkraemer/consolidated_cor_table

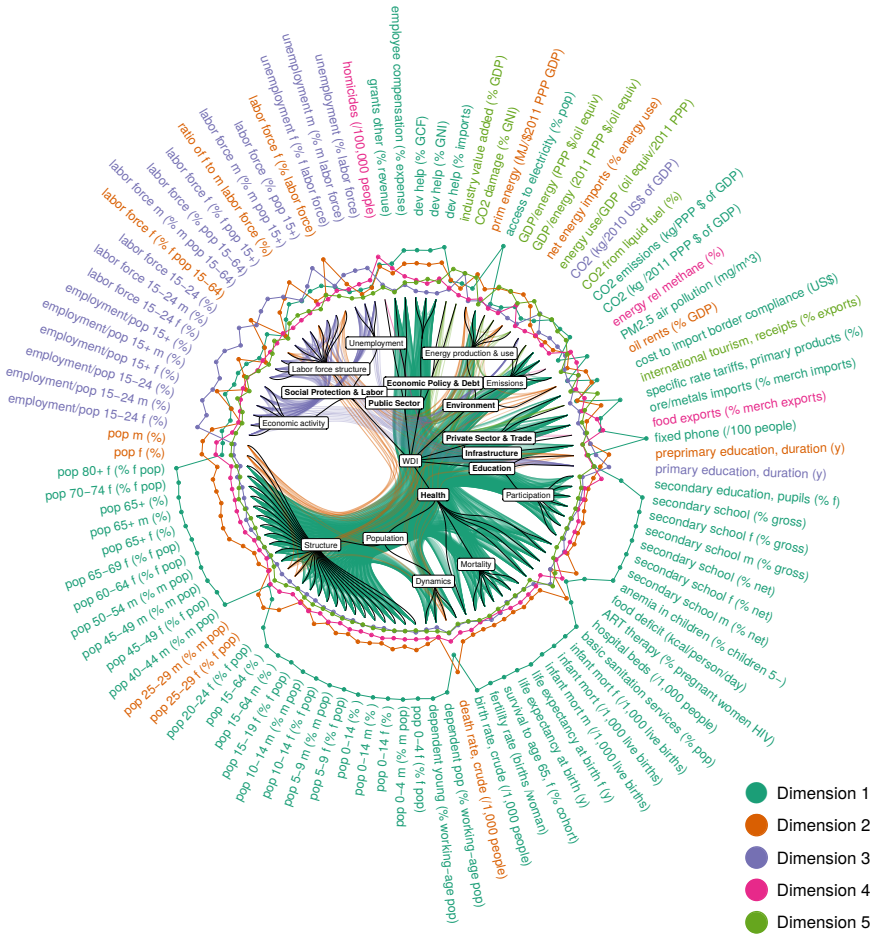

Figure 4.4: The importance of single WDIs to the dimensions. Outer circle: indicator names, colored by the maximum distance correlation dimension. Middle circle: distance correlation values with dimensions 1–5. Black center dendrogram: thematic ordering of the indicators by the World Bank. Colored lines in the center circle: connect all indicators that have maximum distance correlation with the same dimension, bundled along the center dendrogram for easier visualization. The figure only shows the WDIs that have the highest distance correlations with dimensions 1–5.

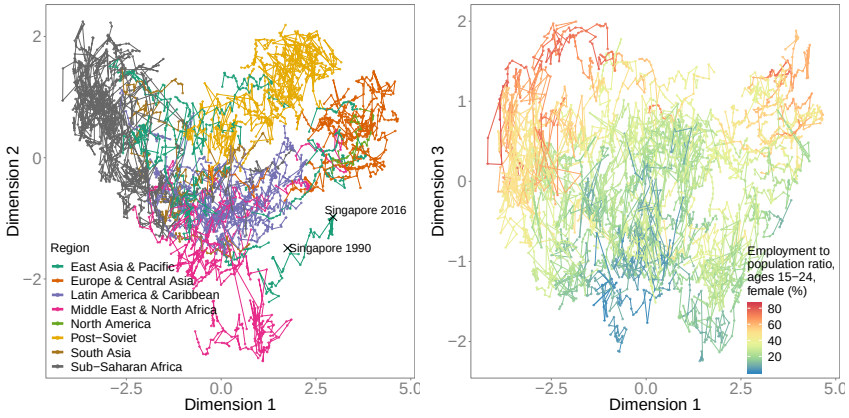

Figure 4.5: E-Isomap dimensions 1–3. Left: Dimension 1 and 2, colored by World Bank regions and former East Block and allies in Eastern Europe. Right: Dimensions 1 and 3 colored by Employment to population ratio, ages 15–24, female (%). Dimension 1 (the horizontal axis on both panes) is a general wealth gradient, on the far left side are poor countries, mostly classified as "Sub-Saharan Africa" while on the right side are the developed countries with most Western European countries on the far right. Dimension 2 (vertical axis on the left pane) spans mostly the percentage of female population and labor force participation of women. Dimension 3 (vertical axis on the right pane) spans employment ratios, employment ratios for women and labor force participation of young working age women. There is an interactive online version available (`http://bgc-jena.mpg.de/~gkraemer/consolidated_dimred/`).

and imports as well as trade, the climate impact of GDP (gross domestic product), and development aid received. Because dimension 1 embraces education, health, and life expectancy, it is conceptually similar to the HDI. In fact, dimension 1 has a strong nonlinear correlation with the HDI (dcor = 0.93, see Appendix fig. B.2), and can be interpreted as a measure of development *sensu* HDI, even though it includes much more than the aspects reflected by the HDI. We also find that the correlation is much lower for most sub-Saharan countries (Fig. S2).

Dimension 2 (9.9% of the variance) is strongly related to gender ratios in the general population and the labor market, as well as primary energy production and consumption and the fraction of 25–29 year old people. This dimension spans a gradient between the extremes of dimension 1 and former Soviet allied countries on one end, and rich mostly oil exporting

nations on the other end (see fig. 4.5). On the positive extreme on this axis are countries that have a very high participation of women in the labor market (e.g Mozambique has the highest participation of women in the labor force with around 55%, similar to countries like Lithuania with a rate of approx 50%) on the negative extreme we can find countries with a very low participation of women in the labor market: Rich countries like the United Arabian Emirates have a female labor force of around 12%, just as poorer countries like Yemen that has a participation rate of women of around 8%, and low death rates. Crude death rates also correlate well with this dimension and do not separate regions, e.g. Latvia in 1994 had a crude death rate of 16.6/1000 people, Denmark in 1993 a crude death rate of 12.1 per 1000 people, while similar crude death rates can be found in undeveloped countries (Democratic Republic of the Congo, 1996, 16.655 death per 1000 people; or Liberia, 2005, 12.128 deaths per 1000 people), on the low extreme we find mostly rich oil exporting nations (e.g. Qatar and the United Arabian Emirates with values around 1.5 deaths per 1000 people).

The third to fifth dimensions explain much less variance but are still important in that they account for variables not found in the first two dimensions: Dimension 3 (1.9% of the variance) is a labor market gradient representing descriptors like ratios of labor force, employment, and unemployment. Dimension 4 (2.4% of the variance) summarizes homicide rates, methane emissions and food exports. Dimension 5 (1.8% of the variance) represents the CO_2 impact of GDP, tourism and value added to products by industry.

4.3.3 Global Trends

Development is dynamic. Over time each country moves along a characteristic trajectory in development space. Along the first dimension, clear trends can be observed. Most countries have a positive slope (see fig. 4.6). Given that dimension 1 essentially spans a gradient between wealthy and poor countries, this reveals the overall global trend towards a wealthier world (Gapminder Foundation, 2018). Only a few countries have negative slopes. Comparing the slopes of "Sub-Saharan Africa" with the rest of the world reveals a widening gap in the development gradient *sensu* HDI. Dimensions 2 to 4 do not show such pronounced overall trends.

Dimension 2 shows positive trends in most of the "Western World" and North Africa and negative trends in most parts of Asia and Sub-Saharan Africa. The positive trends in the "Western World" countries are due to an increased participation of women in the labor market, declining death rates in countries with young populations, and climbing death rates in countries

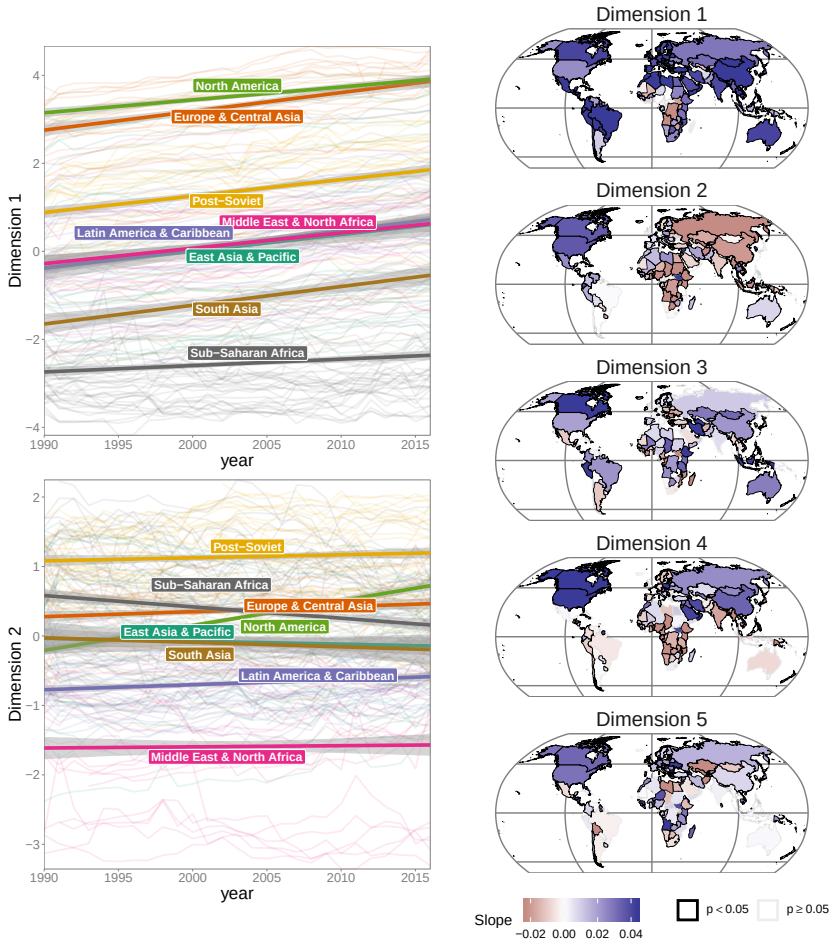

Figure 4.6: Trends in the first e-Isomap axis over time. Left: transparent lines are the trajectories of all countries over time (dimensions 1 and 2, other dimension see annex), colored by geographic regions, the straight lines are linear regressions over all data points of a region and the 95% confidence intervals over their coefficients. Right: World maps of slopes over time of dimension 1–5, the color represents the value of Sen's slope, countries with significant slopes have black borders.

with aging societies. Many developing countries in Sub-Saharan Africa and Asia show negative trends, which seems to be a common interaction between dimensions 1 and 2 on the far negative end of dimension 1.

Dimension 3 shows mostly employment/unemployment ratios, but there are no really strong general trends observable. We note that eastern and western Europe show fundamentally different trends, most of eastern Europe has predominantly negative trends, while in the rest of Europe there are few significant slopes reflecting the increase in unemployment in Eastern Europe. Other notable countries include Peru, Ethiopia, and Azerbaijan, where unemployment rates have strongly decreased; these countries show strong positive trends.

Dimension 4 shows energy-related methane emissions, which have increased in most parts of the northern hemisphere and decreased in most other parts of the world, as well as homicide rates, which have decreased in large parts of the world, but increased in parts of Latin America. The data on homicide rates in large parts of Africa are very sparse.

Dimension 5 shows tourism and the ecological impact of GDP. In general, more GDP is produced per unit of energy. This trend seems to be stronger in the Western World.

4.3.4 Trajectories

Changes in trajectories in development space are very likely to be a major disruption of a given development path. Some examples can be found in fig. 4.7. For example, the earthquake in Haiti in 2010 coincides with a major disruption in the trajectory. The financial crisis and the onset of austerity measures can be noted from a dent in 2008 in the trajectory of Greece. A few years after massive privatizations in Argentina the trajectory of Argentina changes drastically. Major disruptions in the trajectory of the United States happen during the burst of the dot-com bubble in 2000–2001 and the financial crisis in 2008. Attribution of changes to the trajectories to only these events can be challenging, and would require a formal causal framework (Pearl et al., 2016; Peters et al., 2017). For instance, in the case of the US, the changes in the trajectory could equally be attributed to changes in the presidency or to politics after 9/11/01. In the case of Argentina, it is not clear if the changes were caused by changes in politics during the Kirchner presidencies, problems that set in later after the privatizations, or a mixture of both, and remain of purely speculative nature.

In the overall view, some countries appear to change their centers of attraction recovered space of human development, e.g. Singapore in 1990 appears to be similar to the rich oil exporting Arab countries, but its trajectory suggests that it is currently gravitating towards most of the

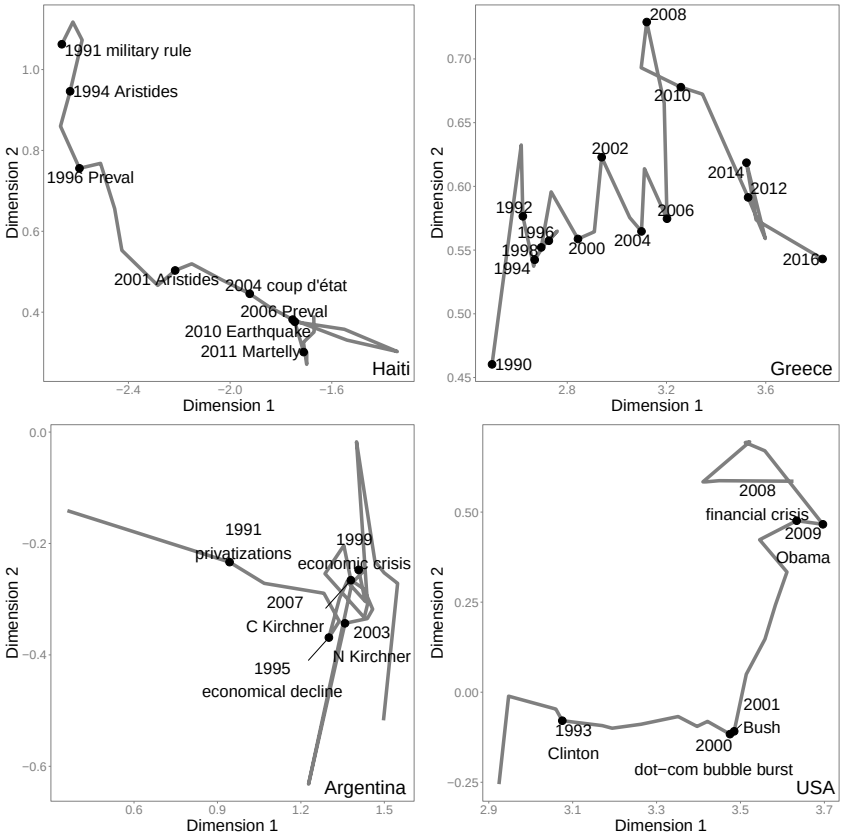

Figure 4.7: Example trajectories. Haiti: Large jumps in the trajectory and kinks are preceded by changes in government and natural disasters. Greece: Complete change of direction of the trajectory seem to appear during the financial crisis 2007–08. Argentina: The trajectory reflects important changes in economic policies. USA: The trajectory reflects the economic crises and changing presidencies.

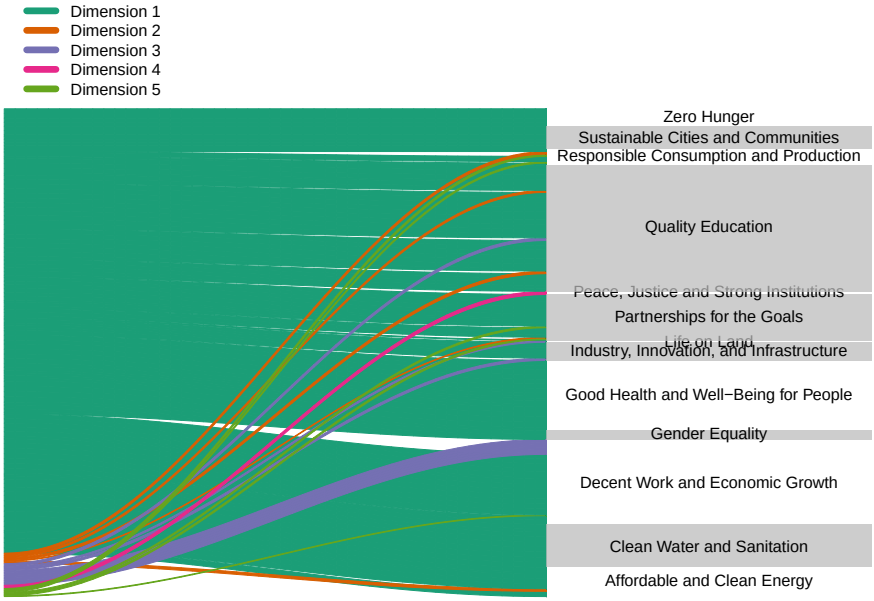

Figure 4.8: Showing the importance of the dimensions (color coded) for the SDGs. Dimension (left, unlabelled) are connected to the SDGs (right) through the corresponding WDIs (not shown, see text for details). The thickness of the connection reflects the distance correlation between the WDIs and the dimensions. See SI fig. 3 for a more detailed version of the figure.

wealthy European countries, see fig. 4.5. Countries that share similar history also seem to be close in the final dimensions, e.g. former Soviet countries, rich oil exporting nations, western European nations.

4.3.5 Sustainable Development

To understand the relevance of the emerging dimensions for the different SDGs, we again use distance correlation and the WDIs that the World Bank uses to track the SDGs (United Nations General Assembly, 2017b). We consider only the dimension with the maximum distance correlation to each WDI which is used to track an SDG. The results are shown in fig. 4.8.

As most goals are poverty related, they load most strongly on the first dimension. The goals "Decent Work and Economic Growth" and "Industry,

Innovation, and Infrastructure" also load on dimension 3, as this dimension describes the labor market. Dimension 2 describes educational and energy aspects and is related to "Affordable and Clean Energy" and "Quality Education". We found a relationship between dimension 4 and the SDG "Peace, Justice and Strong Institutions" due to the homicide rate indicator. Dimension 5 was important to the "Partnership SDG and Responsible Consumption and Production", due to relatedness of non-renewable energy sources and statistical reporting indicators.

Surprisingly, dimension two does not have any influence on the SDG "Achieve gender equality and empower all women and girls" despite describing aspects of gender equality. The reason for this may be that the SDG "Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all" is described by many of the variables loading on dimension two.

We can also see wich SDGs are well represented by the data (the height of the SDG in fig. 4.8) and which ones are not. For example, the best represented SDGs are describing traditional ideas of development, such as "Quality Education", "Decent Work and Economic Growth", "Good Health and Well-Being for People", while environmental SDGs such as "Life on Land", "Life Below Water", or "Climate Action" are not well or not at all represented.

4.4 Discussion

The assessment of development on the basis of a few key indicators has often proven very useful, but has also been controversial. As early as in the 1960s, GDP was recognized to be a very incomplete measure of development (Ram, 1982; McGillivray, 1991; Göpel, 2016). Later, a large number of indicator approaches emerged, each constructed to describe specific aspects of development (Parris and Kates, 2003; Shaker, 2018). The large number of measured variables and derived indicators that are used today to describe development could suggest that global development is a high-dimensional process requiring many indicators to describe it accurately. This perception contrasts with our finding that three quarters of the variability of the development space can be explained by only one dimension, and five dimensions recover 90% of variance. This indicates that the dimensionality of development is much lower than one would expect. Or, to put in other terms, the fact that many properties of development are highly correlated (Ghislandi et al., 2019) also means that one can summarize them efficiently in very few dimensions.

The notion that development is of low-dimensionality, however, by no

means implies that it is a "simple" process. In general it is well-known that low-dimensional spaces can still contain and depict very complex and unpredictable dynamics: Prominent examples are the logistic map (Verhulst, 1845, 1847), describing population dynamics in a space of a single dimension, or the Lorenz (1963) attractor in physics, describing hydrodynamic flow in a three dimensional space.

The question whether data-driven indicators as presented here can be an alternative to classical indicators has been widely discussed (Ram, 1982; OEDC, 2008; Gapminder Foundation, 2018). One argument in favour of such an approach is to overcome the lack of objectivity, which is a common criticism of classical indicators (Monni and Spaventa, 2013; Göpel, 2016). Consequently, PCA is increasingly used for the creation of wealth indicators (Filmer and Pritchett, 2001; Smits and Steendijk, 2015; Shaker, 2018), as well as other approaches to identify suitable variable weights (Seth and McGillivray, 2018). In our study we show that the PCA approach is less effective due to the strong nonlinear relations among the covariates present in the dataset.

Nonlinear data dimensionality reduction, however, makes the assessment of the identified dimensions difficult and hard to trace back to the underlying processes. Dimension 1, for example, includes both basic health and wealth variables. Figure 4.3 illustrates the reason for this. On the negative end of dimension 1, the maternal mortality rate is high and per capita income is low. When moving upwards along this dimension, first maternal mortality rates drop steeply, while the per capita income hardly changes. When moving towards the positive end of dimension 1, maternal mortality cannot decrease much further, as it is already close to zero, but the per capita income starts to increase strongly (fig. 4.3). Combining both effects, dimension 1 manages to incorporate wealth as well as mortality related variables into a single (nonlinear) indicator. Each indicator can have a strong influence on a subset of a dimension (e.g. maternal mortality rate on the negative side of dimension 1) and a very low impact on other subsets (e.g. maternal mortality rate on the positive end of dimension 1). Still, the fact that these factors co-vary in a way that we can represent them in a single dimension can guide the development of novel metric indices.

While dimension 1 allows for a relatively straightforward interpretation, we see in dimension 2 that there are more complex patterns to discuss. We find that Post-Soviet countries, Western European countries and Sub-Saharan African countries all lay on similar high coordinate values in dimension 2 (fig. 4.5). Looking at variables that correlate strongly with dimension 2, we find that the participation of women in the labor market can be similar for very different states of dimension 1. We probably also uncover certain socio-cultural divides: most countries classified as "Middle

East & North Africa" show a very low participation of women in the labor marked while in other parts of the world participation of women in the labor marked is much higher and does not depend on the geopolitical region of a country or its development status (see fig. 4.5). For example in many European countries 45–50% of the working population is female, the same or even less than in most Sub-Saharan countries. Another variable that is orthogonal to development are crude death rates, where a rich country like Germany can have very similar rates to many countries in central Africa. Death rates in the WDI database are not resolved by age groups, given the aging societies in the developed world and the very young societies in many African countries, the death rates affect mostly older age groups in countries with high values on dimension 1, while it affects many younger age groups in the African countries.

In general, data-driven approaches to index construction can be criticized for not taking the polarity, i.e. the "direction", into account (Mazziotta and Pareto, 2019). This means that it remains subject to a subsequent interpretation whether a high value of a principal component (or nonlinearly derived component) is a sign of a positive state in a certain domain or the opposite. The reason is that the underlying eigenvectors can be of arbitrary sign. However, we have shown (in fig. 4.4) that an interpretation is possible, and the analysis of trends and trajectories can remedy this issue. Collapsing many aspects of development into a single dimension, which in turn forms the main gradient along which countries move over time, essentially expresses (nonlinear) covariations that should not be studied in isolation. For example, higher employment rates and an increased per capita income often go hand in hand. Here we showed that these connections between the 621 measured variables are so strong that a single dimension suffices to represent 74% of the variance. In this sense, we also see our approach as an opportunity to generate novel hypotheses on development that can guide policy making e.g. towards achieving the SDGs.

A general criticism of machine learning approaches is that underlying data biases are propagated and exacerbated. For instance, if the training data contain biases against minority groups, e.g. gender or race, these groups will systematically be put in a disadvantage by the algorithm (Barocas and Selbst, 2016). Latest research tries to detect such biases (Obermeyer et al., 2019) and to avoid them during the training phase (Pérez-Suay et al., 2017). Therefore the implications of every machine learning based analysis have to be seen in the light of the dataset used for training. Here we summarize the WDI database, which represents the efforts of the World Bank to collect information on development at the global scale. The high variance explained by variables representing basic infrastructure, per capita income, and the population pyramid therefore

reflects the (historic) emphasis that has been given to these kinds of basic indicators. For instance financial accounting has been ubiquitous, there are large scale efforts to monitor infrastructure and poverty, and census data is globally available.

Our analysis does not reveal an "environmental axis", a component that is essential to sustainable development (Steffen et al., 2015). We can therefore also read our analysis as a gap analysis and conclude that future versions of the WDI database should put more emphasis on environmental data that are now widely available (Mahecha et al., 2020). Another essential component are inequalities (UNDP, 2019). While some aspects are recovered by our analysis, such as between country inequalities on dimension one and some aspects of gender inequality on dimension 2, others do not emerge, e.g. income inequalities inside a country.

The best represented SDGs are those related to traditional ideas of development, while "Life on Land", "Life Below Water" or "Climate Action" are not well or not at all represented. This shows a clear bias towards classical development data, and a lack of environmental data in the WDI database. The reasons for this lie in the topics that have been emphasized for development historically (Griggs et al., 2013).

An analysis like the present one can be informative for policy making in various ways. It reveals general constraints of the development manifold, i.e. which combinations of WDIs are possible, which trajectories in the development space have been observed and which ones not. In particular, the trajectories can inform policy makers regarding the general present and past position of a country in this space beyond a single metric like the HDI (or our dimension 1). This means that also the less obvious changes, e.g. the changes of post-Soviet countries along dimension 2, can be taken into consideration.

Focusing on these dimensions is not trivial. It allows to target a few orthogonal aspects of developments only, instead of screening hundreds of individual WDIs. Another way this analysis can guide policies is by seeing the results in the context of the dataset and pointing out weaknesses and underrepresented dimensions in the dataset, such as the environment and within-country inequalities. The key difference between our approach and the classical approaches is that we try to describe development space in its entirety, and hence the extracted components are neutral and agnostic to any societal or political agenda.

In particular the trajectory of single countries can yield essential information on important events for a country. The trajectories analyzed in this paper all showed changes that are obvious to the human eye, such as temporary deviations or changes in speed and directions. We could find connections for all of the observed changes in the trajectories of fig. 4.7

with important socioeconomic or environmental events, although we were not able to automatically detect changes in all trajectories due to the different characteristics of each change. Future research is needed, to better understand the anomalies in the extracted trajectories.

In our opinion, a main advantage of data-driven approaches compared to classical indicator approaches is that the number of necessary indicators emerges naturally and the resulting indicators represent orthogonal features. The main disadvantage is the loss of indicators that represent very specific aspects of the data. Obviously, dimensionality reduction can only summarize the available data which also means that data incompleteness, data errors, and reporting biases are inherited—as it is also the case for classical indicators. Still, the proposed approach can help in the planning of adding measures of development and testing their redundancy with respect to the existing indicators, simplifying e.g. reporting of complementary dimensions of development.

A general limitation of the data under scrutiny is their aggregation at the country level. This means that our analyses cannot account for the often large socioeconomic differences and developments *within* a country. Also localized disasters may not influence the trajectory of a large economy as a whole, e.g. a large hurricane causing damage in Florida will only have a very marginal influence on the trajectory of the United States. Today there are efforts to collect data on sub-national levels which would alleviate this problem, see e.g. Smits and Permanyer (2019). However these efforts are relatively recent and there are still not many variables available.

4.5 Conclusions

In this study we investigated the "World Development Indicators" from 1990 to 2016 using a method of nonlinear dimensionality reduction. Our study led to three key insights. Firstly, the WDI database is of very low intrinsic dimensionality: We found that the WDIs are strongly interconnected, but we also showed that these connections are highly nonlinear. This is the reason why linear indices based on PCA cannot compress the information on human development that efficiently, while our approach only needs five dimensions to represent 90% of the data variance. The first dimension partly resembles the HDI, but also reveals much more differentiated patters in low-income countries. The subsequent dimensions show orthogonal aspects such as the participation of women in the labor market and complex demographic dynamics. Quantifying such interactions uncovered by this approach can lead to new approaches to quantify different aspects of development. Exploring the meaning of the emerging dimensions allows

us to understand which aspects of development are underrepresented in current databases. The second insight is that development as described by the dimensional space, remains to be a highly complex process that involves strong nonlinear interactions. We have elaborated some of these aspects, but a more profound exploration of the five-dimensional development space is still needed. Clearly, our approach can only account for the information in the data and ignore any additional aspects such as environmental issues that are expected to be critical for sustainable development. The third insight is that single countries' trajectories in the low-dimensional space show abrupt changes that coincide with major environmental hazards or socioeconomic anomalies. As these changes in the trajectories can be of different nature, automatized detection is non-trivial and may require further causal explorations. Overall, our analysis gives new insights into the general structure of development which is of low dimensionality, but highly nonlinear and interconnected. Future work is needed to understand the observed trajectories in development space in more detail, as well as to exploit them for achieving the Sustainable Development Goals.

Chapter 5

Conclusions

Content

5.1 General Conclusions

To find the intrinsic dimensions of the biosphere and anthroposphere we used machine learning. Because the many different data streams that are being used to observe these systems we expected the data streams not to be independent and to find redundancies between these measurements. To quantify and explore these redundancies we used linear and nonlinear methods for dimensionality reduction. We found that the redundancies are substantial, e.g. we only needed 5 dimensions to represent the variance of 621 dimensions comprising our observations of the anthroposphere. Dimensionality reduction also helps us to gain a deeper insight into the main dimensions of the observed system. Trajectories help us to characterize objects, their changes over time and extremes in the space of reduced dimensionality. As we showed in Chapters 3 and 4, dimensionality reduction provides the ideal framework for the development of a data driven system state indicator as proposed in Section 1.4.

In Chapter 1 we introduced the concept of system state indicators to track the elements of a system over time and proposed dimensionality reduction

as an ideal tool for the creation of such indicators. Then, in Chapter 2, we revised many methods for dimensionality reduction, how to apply them and how to compare them (Kraemer et al., 2018). This provided the framework for the proper application of dimensionality reduction on real world data. Finally, in Chapters 3 (Kraemer et al., 2020a) and 4 (Kraemer et al., 2020b), we applied dimensionality reduction to create indicators from two data sources to prove the concept of the system state indicator.

The two indicators were created from datasets with very different characteristics. The biospheric data was comparatively large, with relatively few missing values, and did not contain a very large number of variables. The socioeconomic dataset was comparatively small, contained a very large number of variables and almost all observations were partially incomplete. This allowed us to gain insight into the challenges for the creation of indicators on real world datasets and their potential. Through indicators, we could gain insights into the general structure of the datasets and create trajectories for the observed objects, i.e. spatial pixels in the case of the biosphere and countries in the case of the anthroposphere. The System State Indicator approach showed a lot of promise for exploratory analysis of complex datasets, we could extract general patterns in the data, observe the effect of permanent changes and extreme events on trajectories, and in general gain a lot of insight into the functioning of the system.

In Chapter 3 (Kraemer et al., 2020a), we created a biosphere indicator to track the state of ecosystems globally. The dataset for the biosphere indicators is derived directly or indirectly from remote sensing products. Some of the variables use very simple radiative transfer models (e.g. albedo), others use complex models of biological processes to derive the products (e.g. root-zone soil moisture). Other products use local observations and upscale these observations using other satellite remote sensing products (e.g. GPP).

There was no need to use a nonlinear method in the creation of the biosphere indicators because a simple PCA resulted to be enough to represent the system in three dimensions. This linear method also made interpretability a lot simpler. Interpretation is one of the main goals in data analysis in general and Earth system science in particular (Reichstein et al., 2019). The first emerging indicator represented carbon exchange, while the second indicator showed the availability of water in the ecosystem. The first two indicators can detect many well-known phenomena, without analyzing each original variable separately, due to their compound nature. We showed that the indicators are capable of detecting seasonal hysteresis effects in ecosystems, as well as extremes and breakpoints. Using System State Indicators we gained a high level overview of phenomena in ecosystems and the method therefore provides an interesting tool for analyses

where it is required to capture a wide range of phenomena which are not necessarily known *a priori*.

In Chapter 4 (Kraemer et al., 2020b), we created indicators of development and showed, that a space comprised of 621 social indicators can be represented in very few dimensions with the most important dimension representing 74% of the total variance. The analysis gave new insights into the general structure of development. The findings suggest that development can be characterized in a space of much lower dimensionality than previously thought. The most important dimension indicating a development towards a wealthier world, but also showing a widening gap between "Sub-Saharan Africa" and the rest of the world. The country trajectories derived from the indicators were able to show important events and their distribution characterizes development space very well.

We strove to create interpretable indicators in order to make the resulting indicators useful for the respective audience. In the case of the biosphere indicators in Chapter 3 this meant using a Principal Component Analysis, because the matrix of loadings gives the linear relations between the original variables and the indicators. The use of PCA also has the advantage that it is an invertible transformation which allows for a deeper analysis of the errors, i.e. in the case of the biosphere indicators it enables us to measure how well an ecosystem is represented by the indicator or indicators. In the case of the socioeconomic indicators (Chapter 4) the data was too nonlinear for a linear method and therefore an extension of Isomap had to be used in order to achieve a good compression of the data. The use of a nonlinear method also meant that the interpretation of the resulting indicators was not straightforward and required a substantial amount of additional effort.

In the biosphere as well as in the anthroposphere, we were able to characterize the most important gradients of the system by analyzing the space of the resulting indicators. In the biosphere the most important gradients were ecosystem productivity, water availability and albedo. In the socioeconomic system, the most important axis resulted to be similar to the Human Development Index, other gradients included the age structure of the population, the labor market, and crude death rates.

In the case of the biosphere, the yearly cycle dominated the extent of the space occupied by the trajectories, e.g. a summer in Germany exhibits similar indicator scores to a tropical rainforest. The differences between pixels can be found mostly in their seasonal cycles, e.g. tropical rainforests show no large seasonal cycle, while the pixel in Germany showed a strong limitation by cold during winter. Contrary to this, trajectories in the anthroposphere were dominated by differences between countries and trajectories did not show a cyclic behavior, but a distribution along the development

manifold. Therefore, while the biosphere trajectories where dominated by their seasonal cycle and overlapped strongly, the anthropospheric trajectories where overlapping much less and showed relatively consistent trends in most cases. A general global trend towards a wealthier world could be observed, while changes in the biosphere manifest mostly as deviations from seasonal means but can also manifest as permanent changes. In both cases, we could find important events, such as extremes or permanent changes, by a visual analysis of the trajectory. Due to the diversity in the types of changes, e.g. short deviations, changes in direction, changes in "speed". An automatic detection of events will be a topic for future research.

In general we can conclude that the resulting indicators are very useful for exploratory analysis because the low-dimensional trajectories maintain their essential properties and can represent them in a space of low dimensionality that we can then explore visually. The indicators can be used for a general characterization of the observed objects, as well as the detection of changes, such as temporary extreme events, "permanent" changes, or long-term trends.

5.2 Outlook

Although we developed a robust framework for the creation of indicators, the application still requires a lot of care. The right method has to be chosen. In the nonlinear case but also in the linear case, there are many possible pitfalls in the application. Currently there are not many methods implemented that can be readily used for such analyses: either they are lacking the ability to be used with out of memory data, or they cannot be used with missing values. This limits the number of methods that can be tested easily.

The resulting trajectories help with the characterization of the dataset and finding events is very easy for the human observer. Because there are many different types of changes, an automatic detection is not trivial to implement therefore a general method to detect all kinds of changes will require more research. The long-term goal must be to jointly interpret biospheric and socioeconomic datasets. The main challenge with this task lies in the large differences between both datasets, especially the differences in spatial and temporal scale.

Another important challenge to overcome when combining both datasets are spurious correlations, e.g. there is a wealth gradient along a latitudinal gradient (rich north–poor south) as well as climatic gradients. When training a machine learning model, special caution has to be taken in order

for the model to not learn such biases present in the training data.

Combining socioeconomic datasets with biospheric and atmospheric datasets has the potential to reveal important insights into the interactions between these systems, e.g. the vulnerability of populations to extreme events, or the anthropogenic factors that determine the loss of biodiversity.

5.3 Achievements and Relevance

Papers directly related to this Thesis are attached in the Annex (page 159ff). The conclusions of this work have been presented in several publications as research papers. A visualization derived from Kraemer et al. (2020b) received a "special mention" by the United Nations Development Programme.

5.3.1 International Journal Papers

Kraemer, G., Camps-Valls, G., Reichstein, M., and Mahecha, M. D. (2020). Summarizing the state of the terrestrial biosphere in few dimensions. *Biogeosciences*, 17(9), 2397–2424. doi:10.5194/bg-2019-307.

Kraemer, G., Reichstein, M., Camps-Valls, G., Smits, J., and Mahecha, M. D. (2020). The Low Dimensionality of Development. *Social Indicators Research*, . doi:10.1007/s11205-020-02349-0

Kraemer, G., Reichstein, M., and Mahecha, M. D. (2018). dimRed and coRanking—Unifying Dimensionality Reduction in R. *The R Journal*, 10(1), 342–358. doi:10.32614/RJ-2018-039

Migliavacca, M., Musavi, T., Mahecha, M. D., Nelson J. A., Knauer, J., Baldocchi, D. D., Perez-Priego, O., Anderson, K., Bahn, M., Black, A. T., Blanken, P. D., Bonal, D. , Buchmann, N., Caldararu, S., Carrara, A., Cescatti, A., Chen, J., Cleverly, J., Cremonese, E., Desai, A. R., El-Madany, T. S., Filippa, G., Forkel, M., Galvagno, M., Gough, C. M., Göckede, M., Ibrom, A., Ikawa, H., Janssens, I., Jung, M., Kattge, J., Keenan, T. F., Knohl, A., Kobayashi, H., **Kraemer, G.**, Law, B. E., Liddell, M. J., Ma, X., Mammarella, I., Martini, D., MacFarlaine, C., Matteucci, G., Montagnani, L., Pabon-Moreno, D. E., Panigada, C., Papale, D., Pendall, E., Penuelas, J., Phillips, R. P., Reich, P. B., Rossini, M., Scott, R. L., Gebhardt, M. M., Stahl, C., Wohlfahrt, G., Wolf, S., Wright, I. J., Yakir, D., Zaehle, S., and Reichstein, M. (under review). The global spectrum of ecosystem function. Nature

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., **Kraemer, G.**, Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J. F., Dorigo, W., Estupinan-Suarez, L. M., Gutierrez-Velez, V. H., Gutwin, M., Jung, M., Londoño, M. C., Miralles, D. G., Papastefanou, P., and Reichstein, M. (2020). Earth system data cubes unravel global multivariate dynamics. Earth System Dynamics, 11(1), 201–-234. doi:10.5194/esd-11-201-2020

Sierra, C. A., Mahecha, M., Poveda, G., Álvarez-Dávila, E., Gutierrez-Velez, V. H., Reu, B., Feilhauer, H., Anáya, J., Armenteras, D., Benavides, A. M., Buendia, C., Duque, Á., Estupiñan-Suarez, L. M., González, C., Gonzalez-Caro, S., Jimenez, R., **Kraemer, G.**, Londoño, M. C., Orrego, S. A., Posada, J. M., Ruiz-Carrascal, and D., Skowronek, S. (2017). Monitoring ecological change during rapid socio-economic and political transitions: Colombian ecosystems in the post-conflict era. Environmental Science & Policy, 76, 40–49. doi:10.1016/j.envsci.2017.06.011

5.3.2 Other Publications

Adsuara, J. E., Pérez-Suay, A., Moreno-Martínez, Camps-Valls, G., **Kraemer, G.**, Reichstein, M., and Mahecha, M. D. (2020). Discovering Differential Equations from Earth Observation Data [abstract]. In: IEEE International Geoscience and Remote Sensing Symposium.

Adsuara, J. E., Pérez-Suay, A., Moreno-Martínez, A., Mateo-Sanchis, A., Piles, M., **Kraemer, G.**, Reichstein, M., Mahecha, M. D., and Camps-Valls, G. (2020). Learning ordinary differential equations from remote sensing data [Other]. oral. https://doi.org/10.5194/egusphere-egu2020-19620

Estupinan-Suarez, L. M., Brenning, A., Gans, F., **Kraemer, G.**, Sierra, C. A., and Mahecha, M. D. (2020). Capturing the influence of ENSO on land surface variables for Tropical South America [Other]. oral. https://doi.org/10.5194/egusphere-egu2020-4187

Mahecha, M. D., Guha-Sapir, D., Smits, J., Gans, F., and **Kraemer, G.** (2020). Chapter 13—Data challenges limit our global understanding of humanitarian disasters triggered by climate extremes. In J. Sillmann, S. Sippel, and S. Russo (Eds.), Climate Extremes and Their Implications for Impact and Risk Assessment (pp. 243–256). Elsevier. doi:10.1016/B978-0-12-814895-2.00013-6

5.3.3 Awards

The interactive visualization from Kraemer et al. (2020b) received an honorable mention in the 2019 Human Development Data Visualization Challenge[1].

5.3.4 Visits to National and International Research Centers

This Thesis made possible due to the collaboration of national and international researchers. The author elaborated this Thesis as a PhD student enrolled at the Universitat de València and employed as a PhD researcher at the Max Planck Insitute for Biogeochemistry in Jena and spent three months at the Image Processing Lab at the Universitat de València.

5.3.5 Related Projects and Acknowledgements

The outcomes of this work are relevant to the research carried out by the author and his colleagues at the Max Planck Institute for Biogeochemistry and the Image Processing Lab at the Universitat de València in the context of different research projects. A list of projects in which the author of this Thesis has collaborated follows:

- oBEF-across
- BACI: H2020 EU project under grant agreement No. 640176
- ESDL: A project by the European Space Agency.
- ILeaps
- CubeColombia
- ERC Consolidator Grant SEDAL

This Thesis was made possible thanks to the support of the following grants an projects:

- Earth System Data Lab (ESDL) `http://earthsystemdatalab.net`
- BACI the H2020 project BACI under grant agreement No. 640176
- ILeaps

[1] `http://hdr.undp.org/en/content/2019-human-development-data-visualization-challenge-winner-gender-inequality-visual-story`
`https://web.archive.org/web/20190712180204/http://hdr.undp.org/en/content/2019-human-development-data-visualization-challenge-winner-gender-inequality-visual-story`

Chapter 6

Resumen en Español

Contenido

6.1 Motivación

La actividad humana causa cambios sin precedentes en la Tierra, especialmente en la biosfera. Su impacto total está lejos de ser comprendido todavía, pero ya es suficientemente fuerte para causar un evento de extinción masiva (Ripple et al., 2017; Ceballos and Ehrlich, 2018; IPBES, 2019) y el impacto aumentará en el futuro.

A medida que aumenta la presión de la humanidad sobre los ecosistemas, también aumenta la necesidad de herramientas, no solo para monitorear los cambios que ocurren en los ecosistemas, sino también sobre el sistema y desarrollo económico. Los instrumentos de monitoreo no solo deben ser capaces de detectar un solo tipo de impacto sino en realidad una amplia gama de cambios del sistema que puedan ocurrir. Por lo tanto, necesitamos herramientas de vigilancia que sean lo suficientemente flexibles para detectar impactos en diferentes tipos de sistemas, por ejemplo, los sistemas socioeconómicos y la biosfera. Asimismo, estos sistemas deben ser capaces de detectar diferentes tipos de impacto como, por ejemplo, tendencias

lentas, eventos extremos repentinos y cambios abruptos en el estado de los ecosistemas.

6.2 Objetivos

El objetivo general de esta Tesis Doctoral es:

> *"Aprender la dimensionalidad intrínseca de la biosfera y la antroposfera a partir de datos usando técnicas avanzadas de aprendizaje de máquinas."*

Para alcanzarlo, hemos definido un conjunto de objetivos específicos:

1. *Buscar la dimensionalidad del sistema.* Aquí nos hacemos la siguiente pregunta: ¿Cuántas dimensiones son necesarias para describir con precisión el sistema?

2. *Encontrar las dimensiones dominantes de las variables que describen las esferas del sistema terrestre -biosfera- y analizar las características de las componentes.* En este caso aplicamos métodos de reducción de la dimensionalidad a conjuntos de datos globales del mundo real y analizamos e interpretamos las componentes resultantes mirando las variables codificadas en ellas. Esto nos ayuda a entender las dimensiones más importantes que describen el sistema.

3. *Buscar patrones globales usando los indicadores resultantes.* Analizamos cómo los objetos se distribuyen en el espacio de dimensionalidad reducida, y vemos qué patrones se pueden encontrar y cómo esto caracteriza el sistema.

4. *Utilizar las trayectorias resultantes para caracterizar los objetos.* Cada objeto (píxel espacial o país) se describe por series temporales de los indicadores resultantes de la misma forma que los objetos observados se describen por las series temporales de las variables. Analizamos las trayectorias de los objetos observados en el espacio reducido en términos de sus posiciones relativas, sus direcciones y la información codificada en la serie temporal de indicadores para caracterizar las propiedades del sistema global y de los objetos observados.

5. *Encontrar los cambios y extremos descritos por los indicadores.* Analizamos cómo se codifican los eventos extremos y otros cambios importantes en las series temporales de los indicadores y cómo estos cambios reflejan las alteraciones en un ecosistema o en un país.

6.3 Metodología

6.3.1 Enfoques de Indicadores

La teledetección es la adquisición de información sobre un objeto desde la distancia. La teledetección puede utilizarse para estimar las propiedades de la superficie terrestre, las aguas o la atmósfera. En el caso de la vegetación, por ejemplo, debido a que se conocen las propiedades reflectantes de diferentes superficies, podemos combinar diferentes bandas de forma paramétrica para calcular índices de vegetación (VI) que tratan de modelar propiedades físicas (Camps-Valls et al., 2011). Existe un elevado número de este tipo de índices e indicadores usados ampliamente en modelar parámetros como el índice de área foliar (LAI, por sus siglas en inglés), fracción de la cobertura vegetal (FVC) o el contenido de clorofila en hoja. El índice de vegetación más extendido es el basado en la diferencia normalizada entre los canales del rojo y el verde, el conocido NDVI (Rouse et al., 1973).

Los indicadores climáticos normalmente tratan de describir fenómenos que son importantes para la circulación global. Por ejemplo, hay una variedad de indicadores que describen la Oscilación del Sur de El Niño (ENSO), el más importante de los fenómenos de interacción océano–atmósfera (Wolter and Timlin, 2011b). Una forma de crear tales indicadores consiste en utilizar el primer componente principal de los campos de las variables físicas involucradas, como la temperatura de la superficie del mar o la presión del nivel del mar en una región del Pacífico ecuatorial. Estos indicadores utilizan la reducción de la dimensionalidad para comprimir el espacio de representación (y las variables si hay más de una) y mantener la dimensión temporal. Otros indicadores climáticos importantes se calculan de manera similar, por ejemplo la Oscilación del Atlántico Norte, la Oscilación del Ártico y la Oscilación Antártica.

Normalmente las clasificaciones de la vegetación y el clima se separan razonablemente bien en un el espacio climático abarcado por la temperatura y la precipitación (Köppen and Geiger, 1954; Kottek et al., 2006; Papagiannopoulou et al., 2018). Por lo tanto, es de esperar que las variables que representan la vegetación permitan al menos el mismo grado de separación utilizando dos componentes. Aunque esta no es una aplicación directa de la reducción de la dimensionalidad, la agrupación y la reducción de la dimensionalidad son similares en el sentido de que ambas reducen las características de entrada: En el caso de la reducción de la dimensionalidad, el resultado es una serie de características continuas, mientras que en el caso de la agrupación el resultado consiste en una serie de clases o grupos homogéneos. Las clasificaciones espaciales se reducen sobre las variables y tiempo (en forma del ciclo estacional medio) y se mantienen sólo las

dimensiones espaciales (en forma de una clase por píxel espacial).

Al observar muchos flujos de datos, habrá redundancias en los datos, por ejemplo, diferentes medidas para el PIB de un país pueden ajustarse a la inflación, al cambio de moneda a las tarifas, a los costes de vida, etc. Otras medidas que se correlacionarán con el PIB son medidas de pobreza y medidas que describan la infraestructura, entre otras. Todas estas medidas covariarán fuertemente y, aunque midan diferentes aspectos (que son todos importantes por derecho propio), estos datos reflejarán grandes redundancias. La cuestión es: *¿Cuáles son las redundancias? y ¿cuáles son las dimensiones independientes?*

La forma natural de abordar esta cuestión proviene del campo de la estadística multivariada. La reducción de la dimensionalidad describe una familia de métodos multivariados que encuentran representaciones alternativas de los datos construyendo combinaciones lineales o no lineales de las variables originales, de modo que las propiedades importantes de la señal original se mantinen en ese subespacio de menor dimensionalidad. Cabe señalar que existe una variedad de métodos que crean indicadores y reducen la dimensionalidad de los datos de diferentes maneras en las ciencias de la Tierra, que son diferentes del enfoque propuesto aquí.

6.3.2 Reducción de Dimensionalidad

En esta Tesis proponemos un Indicador de Estado del Sistema (SSI) que resume el estado de los elementos de un sistema complejo y multivariado a lo largo del tiempo y es explícito en el espacio. Para lograrlo, aplicamos la reducción de la dimensionalidad de una manera distinta a los enfoques estándar anteriores. El SSI nos permite monitorear y detectar diferentes tipos de eventos en cualquier variable. Una trayectoria sigue la posición, en el espacio reducido, de una unidad de observación espacial (en el espacio geográfico, un país o un píxel) a lo largo del tiempo. El espacio de dimensionalidad reducida es un espacio abstracto que mantiene propiedades importantes del espacio formado por todas las variables observadas.

La reducción de la dimensionalidad es una herramienta única y válida para crear indicadores descritos anteriormente. Si observamos un solo objeto, ya sea un píxel espacio-temporal o un país a lo largo del tiempo con suficientes flujos de datos, inevitablemente habrá redundancias en los datos. Estas redundancias causan que los datos no llenen el espacio de representación de manera uniforme, pero las observaciones vivirán en una variedad diferencial (*manifold*) de menor dimensionalidad que el espacio original. La reducción de la dimensionalidad trata pues de encontrar representaciones de baja dimensión de esta variedad. Podemos ahora representar esta variedad en un espacio de menor dimensionalidad,

idealmente de la misma dimensionalidad que la variedad misma y describir el estado actual de nuestro objeto dentro de la variedad. Esto nos permite representar fielmente el estado actual del objeto dentro de nuestro sistema en un espacio de baja dimensión, y por lo tanto que se adapte de manera óptima al enfoque de los indicadores presentados en esta tesis.

Hay una serie de cuestiones a considerar, lo que complica la creación de un SSI. El método más simple de reducción de la dimensionalidad es el de análisis en componentes principales (PCA), que resulta en un transformación lineal de los datos. La simplicidad viene con la ventaja de que PCA es relativamente rápido de calcular y simple de aplicar e interpretar, pero no puede tratar con relaciones no lineales. Los métodos más complejos no lineales son computacionalmente mucho más caros de entrenar, por ejemplo a menudo es necesario una descomposición de una matriz de tamaño $n \times n$ por valores propios o hay que realizar una costosa optimización. Los métodos no lineales también suelen requerir el ajuste de varios parámetros, lo que hace más difícil encontrar un modelo adecuado.

Dentro de los métodos lineales, el PCA es el método canónico para la reducción de la dimensionalidad, pero cuando se trata de métodos no lineales, no hay un método estándar y por lo tanto el investigador tiene que elegir de un gran conjunto de metodos preexistentes (o desarrollar un nuevo método). Hay una serie de dificultades adicionales a la hora de elegir un método no lineal para la reducción de la dimensionalidad. Es por esto que creamos el paquete **dimRed** en el lenguaje R para ayudar al investigador a elegir el método correcto (Capítulo 2; Kraemer et al., 2018).

La mayor dificultad consiste en que no hay una medida canónica para comparar la bondad de ajuste de diferentes métodos para la reducción de la dimensionalidad (revisamos algunos métodos para medir la calidad en sec. 2.3) lo que hace que la comparación de métodos sea muy difícil sino imposible. El entrenamiento de muchos métodos no lineales se basa en la optimización no convexa y, por lo tanto, las soluciones pueden no ser estables y un entrenamiento exitoso puede requerir varios intentos. Otras limitaciones son las implementaciones de métodos fácilmente disponibles y bien probados. A menudo la publicación de un método no va acompañada de una implementación que sea fácil de utilizar por otras personas y, por lo tanto, la replicación debe ir acompañada de una reimplementación. Otro factor importante para la aplicación en los datos del mundo real es la capacidad del método para tratar los datos perdidos porque las observaciones del mundo real suelen no estar completos o las series temporales contienen muestras perdidas debido al proceso de adquisición, distorsiones o malas medidas. Por ejemplo, en el capítulo 4 desarrollamos y aplicamos una extensión de Isomap (Tenenbaum et al., 2000) para hacer frente a la alta proporción de valores faltantes en los datos de entrada.

Si tenemos dos modelos de reducción de la dimensionalidad que pueden representar la misma cantidad de información de los datos originales en la misma cantidad de dimensiones, y uno de los modelos es más simple que el otro (por ejemplo, PCA) entonces, siguiendo el principio de la Navaja de Ockham, deberíamos elegir el modelo más simple. Este es el caso en el análisis presentado en el Capítulo 3 donde el PCA resultó ser suficiente para reducir la dimensionalidad del conjunto de datos. La PCA también proporciona varios otros beneficios que se examinan en Capítulo 3.

Al elegir un modelo no lineal, hay ciertas consideraciones que deben hacerse. Como estamos asumiendo que los datos se encuentran en una variedad de baja dimensionalidad en el interior el espacio original, queremos usar un método que preserve la geometría intrínseca del conjunto. Un método que sólo preserva las vecindades locales (por ejemplo *t*-SNE) o globales (por ejemplo, PCA), pero que por lo demás no mantiene la estructura general de la variedad puede no ser una buena elección. En el capítulo 4 mostramos que el Isomap puede ser una buena opción para crear indicadores, ya que intenta encontrar una representación de la variedad, desplegando y preservando su estructura Euclideana interna.

6.4 Resultados

6.4.1 Biosfera

En tiempos de cambio global, debemos vigilar de cerca el estado del planeta para entender la complejidad total de estos cambios. De hecho, cada uno de los subsistemas de la Tierra—es decir, la biosfera, la atmósfera, la hidrosfera y la criósfera—puede ser analizada a partir de una multitud de flujos de datos. Sin embargo, dado que es muy difícil interpretar conjuntamente las relaciones entre las distintas variables, resulta ser una práctica común desarrollar algún indicador que resuma estas relaciones. Los índices climáticos, por ejemplo, resumen el estado de la circulación atmosférica en un región. Aunque estos enfoques también se utilizan en otros campos de la ciencia, raras veces se utilizan para describir la dinámica de la superficie terrestre. Proponemos un método robusto para crear indicadores globales para la biosfera utilizando el análisis de componentes principales basado en un conjunto de datos de alta dimensionalidad a escala global. El concepto se probó utilizando 12 variables explicativas que representan el estado biofísico de los ecosistemas y el intercambio de agua, energía y carbono con la atmósfera. Encontramos que tres indicadores explican el 82 % de la variación de la variables de la biosfera seleccionadas en el espacio y el tiempo en todo el mundo. Mientras que el primer indicador resume los

patrones de productividad, el segundo indicador resume variables que representan el intercambio de agua y energía. El tercer indicador representa mayormente cambios en el albedo de la superficie. Las anomalías en los indicadores claramente identifican eventos extremos, como las sequías del Amazonas (2005 y 2010) y la ola de calor en Rusia (2010). Las anomalías también nos permiten interpretar los impactos de estos eventos. Además, estos indicadores también pueden utilizarse para detectar y cuantificar los cambios en la dinámica estacional. Identificamos, por ejemplo, aumentos de la dinámica estacional amplitud de la productividad en las zonas agrícolas y las regiones árticas. Encontramos que este enfoque genérico tiene un gran potencial para el análisis de la superficie terrestre dinámica a partir de datos de observación o de modelos.

6.4.2 *Antroposfera*

El Banco Mundial publica rutinariamente más de 1500 "Indicadores de Desarrollo Mundial" (WDIs) para seguir el desarrollo socioeconómico a nivel de país. PAra poder interpretar esta ingente cantidad de información, se han creado una serie de índices que la resumen. Por ejemplo, el "El Índice de Desarrollo Humano" (HDI) se diseñó para captar específicamente el desarrollo en términos de la esperanza de vida, la educación y el nivel de vida. Sin embargo, la cuestión sobre qué dimensiones esenciales, o independientes, son fundamentales para representar todos los aspectos del desarrollo sigue abierto. Usando una reducción de dimensionalidad no lineal extrajimos las dimensiones centrales del desarrollo de una manera eficiente. Encontramos que más del 90 % de la variación en los WDIs puede ser representada por sólo cinco dimensiones no correlacionadas. La primera dimensión, explicando el 74 % de la variación, representa el estado de la educación, la salud, los ingresos, infraestructura, comercio, población y contaminación. Aunque esta dimensión no lineal se asemeja al HDI, esta dimension explica mucha más variación. La segunda dimensión (que explica el 10 % de la variación) diferencia a los países por las relaciones de género, el trabajo y los patrones de producción de energía. Aquí, diferenciamos las estructuras de las sociedad: por ejemplo, los países de Oriente Medio con los post-soviéticos. Nuestro análisis confirma que la mayoría de los países muestran más bien tendencias temporales consistentes hacia sociedades más ricas y envejecidas. También podemos encontrar desviaciones de las trayectorias a largo plazo durante la guerra, los desastres, o cambios políticos fundamentales. Las características extraídas a partir de datos complementa los enfoques clásicos definidos mediante indicadores y permite una exploración más amplia del espacio de desarrollo mundial así como relaciones más complejas entre las variables involucradas. Las

dimensiones extraídas representan diferentes aspectos del desarrollo que deben ser considerados al proponer nuevos índices métricos.

6.5 Conclusiones

Para encontrar las dimensiones intrínsecas de la biosfera y la antroposfera utilizamos el aprendizaje estadístico (conocido actualmente por el término en inglés 'machine learning'). Por la multitud de los flujos de datos que se están usando para observar estos sistemas, esperamos que las variables observadas no sean necesariamente independientes y por tanto muestren una elevada redundancia entre estas medidas. Para cuantificar y explorar estas redundancias usamos métodos lineales y no lineales para la reducción de la dimensionalidad. Encontramos que las redundancias son sustanciales en los dos casos de estudio. Por ejemplo, sólo necesitábamos 5 dimensiones para representar las 621 dimensiones que comprenden nuestras observaciones de la antroposfera. La reducción de la dimensionalidad también nos ayuda a obtener una visión más profunda de la dimensiones del sistema observado. Las trayectorias nos ayudan a caracterizar los objetos, sus cambios a lo largo del tiempo y extremos en el espacio de dimensionalidad reducida. Como mostramos en los capítulos 3 y 4, la reducción de la dimensionalidad proporciona el marco ideal para la elaboración de un indicador de estado del sistema basado en datos como se ha propuesto en la sección 1.4.

En el capítulo 1, introdujimos el concepto de indicador de estado del sistema para seguir los elementos de un sistema a lo largo del tiempo, y propusimos la reducción de la dimensionalidad como herramienta ideal para la creación de tales indicadores. A continuación, en el capítulo 2, revisamos muchos métodos para la reducción de la dimensionalidad, cómo aplicarlos y cómo compararlos (Kraemer et al., 2018). Esto proporcionó el marco ideal para la aplicación e intercomparación adecuada de métodos de reducción de dimensionalidad en datos y problemas arbitrarios del mundo real. Finalmente, en los capítulos 3 (Kraemer et al., 2020a) y 4 (Kraemer et al., 2020b), aplicamos métodos de reducción de la dimensionalidad para crear indicadores de dos fuentes de datos como prueba de concepto del indicador de estado del sistema.

Los dos indicadores se crearon a partir de conjuntos de datos con muy diferentes características. Los datos de la biosfera eran comparativamente grandes, con relativamente pocos valores faltantes, y sin un número muy elevado de variables. El conjunto de datos socioeconómicos era comparativamente más pequeño, contenía un gran número de variables y casi todas las observaciones fueron parcialmente incompletas. Esto nos permi-

tió comprender los desafíos para la creación de indicadores en el mundo real en diferentes conjuntos de datos y su potencial práctico. A través de los indicadores, podemos obtener información sobre la estructura general de los conjuntos de datos y crear trayectorias para los objetos observados, es decir, píxeles espaciales en el caso de la biosfera y países en el caso del conjunto de datos socioeconómicos. El enfoque de los indicadores resultó muy prometedor para el análisis exploratorio de conjuntos de datos complejos, pudimos extraer patrones generales en los datos, observar el efecto de los cambios permanentes y los eventos extremos en las trayectorias, y en general obtener una gran cantidad de información sobre el funcionamiento del sistema.

En el capítulo 3 (Kraemer et al., 2020a), se creó un indicador de la biosfera para seguir el estado de los ecosistemas a escala mundial. El conjunto de datos de los indicadores de la biosfera se deriva directa o indirectamente de productos de teledetección. Algunas de las variables utilizan modelos de transferencia radiativa muy simples (por ejemplo, el albedo), otros utilizan modelos complejos de procesos biológicos para derivar los productos (por ejemplo la humedad del suelo en la zona de las raíces). Otros productos utilizan observaciones locales y amplían estas observaciones utilizando otros productos de teledetección derivados de satélites.

No hubo necesidad de usar un método no lineal en la creación de la biosfera porque un simple PCA resultó ser suficiente para representar el sistema en tres dimensiones. Este método lineal también simplificó mucho la interpretación. De hecho, la interpretabilidad es uno de los principales objetivos del análisis de datos en general y de las ciencias de las sistemas de la Tierra en particular. El primer indicador emergente representa el intercambio de carbono, mientras que el segundo muestra la disponibilidad de agua en los ecosistemas. Los dos primeros indicadores pueden detectar muchos fenómenos conocidos, sin analizar cada variable original por separado, debido a su naturaleza compuesta. Demostramos que los indicadores son capaces de detectar la histéresis estacional, efectos en los ecosistemas, así como en los extremos y puntos de ruptura. Los indicadores también pueden seguir otros cambios del ciclo estacional, así como de patrones de cambios en las amplitudes y tendencias estacionales de los ecosistemas. Usando indicadores compuestos obtenemos una visión general de alto nivel de los fenómenos en ecosistemas y, por lo tanto, el método proporciona una herramienta interesante para el análisis donde se requiera capturar una amplia gama de fenómenos que no son necesariamente conocidos a priori.

En el capítulo 4 (Kraemer et al., 2020b), creamos indicadores de desarrollo y se mostró, que un espacio compuesto por 621 los indicadores pueden representarse en muy pocas dimensiones con la más importantes represen-

tando el 74 % de la variación total. El análisis amplió nuestro conocimiento sobre la estructura general del desarrollo. De hecho, los hallazgos sugieren que el desarrollo puede caracterizarse en un espacio de dimensionalidad mucho menor que el que se pudiera pensar a priori. La dimensión más importante indica un desarrollo hacia un mundo más rico, pero también mostrando una brecha creciente entre el "África Subsahariana" y el resto del mundo. Las trayectorias de los países derivadas de los indicadores son capaces de mostrar los acontecimientos importantes y su distribución caracteriza muy bien el espacio de desarrollo.

Nos esforzamos en crear indicadores interpretables y en hacer que los indicadores resultantes sean útiles para las diferentes audiencias respectivamente. En el caso de los indicadores de la biosfera en el Capítulo 3 esto significaba usar un Análisis de Componentes Principales, porque es simple relacionar las variables originales con los indicadores resultantes mediante la matriz de vectores propios. El uso del PCA también tiene la ventaja de que es una transformación invertible que permite un análisis más profundo de los errores, es decir, en el caso de los indicadores de la biosfera, que permite medir lo bien que está representado un ecosistema por el indicador o indicadores. En el caso de los indicadores socioeconómicos (Capítulo 4) los datos eran demasiado no lineales para aplicar un PCA sin más, y por lo tanto se utilizó una extensión de Isomap para lograr una buena compresión de los datos. El uso de un método no lineal también significó que la interpretación de los indicadores resultantes no fuera sencilla, requiriendo un esfuerzo adicional. Tanto en la biosfera como en la antroposfera, fuimos capaces de caracterizar los gradientes más importantes del sistema analizando el espacio de la indicadores resultantes. En la biosfera los gradientes más importantes fueron la productividad del ecosistema, la disponibilidad del agua y el albedo. En el sistema socioeconómico, el eje más importante resultó ser similar al Índice de Desarrollo Humano, si bien otros gradientes incluían la estructura de edad de la población, el mercado laboral y las tasas de mortalidad.

En el caso de la biosfera, el ciclo anual dominó la extensión del espacio ocupado por las trayectorias. Por ejemplo, un verano en Alemania resultó similar al de selva tropical. Las diferencias entre los píxeles se encuentran principalmente en los ciclos estacionales: por ejemplo, las selvas tropicales no muestran un gran ciclo estacional, mientras que el píxel en Alemania mostró una fuerte limitación por el frío durante el invierno. Contrariamente a esto, las trayectorias en la antroposfera estaban dominadas por diferencias entre países y las trayectorias no mostraron un comportamiento cíclico, sino un distribución a lo largo de la variedad de desarrollo.

Mientras que las trayectorias de la biosfera estaban dominadas por su ciclo estacional y se superponían fuertemente, las trayectorias antroposféri-

cas se superponían mucho menos y mostraron tendencias relativamente consistentes en la mayoría de los casos. Se observó una tendencia global hacia un mundo más rico, mientras que los cambios en el biosfera se manifiestan principalmente como desviaciones de los ciclos estacionales medios, pero también pueden manifestarse como cambios "permanentes".

En ambos casos, hemos podido identificar eventos importantes, como extremos o cambios "permanentes" mediante un simple análisis visual de la trayectoria extraída. Esto ha dado lugar a la caracterización e identificación de una gran diversidad en los tipos de cambios, por ejemplo desviaciones cortas, cambios de dirección o, cambios en "velocidad". La detección automática de eventos será un tema de trabajo futuro.

En general podemos concluir que los indicadores resultantes son muy útiles para el análisis exploratorio porque las trayectorias de baja dimensión mantienen sus propiedades esenciales y se puede representar en un espacio de baja dimensionalidad que luego se puede explorar visualmente. También pueden ser usados para una caracterización de los objetos observados, así como la detección de cambios, como eventos extremos temporales, cambios permanentes o tendencias a largo plazo.

Bibliography

J. T. Abatzoglou, D. E. Rupp, and P. W. Mote. Seasonal Climate Variability and Change in the Pacific Northwest of the United States. *Journal of Climate*, 27(5):2125–2142, March 2014. ISSN 08948755. doi: 10.1175/JCLI-D-13-00218.1.

L. V. Alexander, X. Zhang, T. C. Peterson, J. Caesar, B. Gleason, A. M. G. Klein Tank, M. Haylock, D. Collins, B. Trewin, F. Rahimzadeh, A. Tagipour, K. Rupa Kumar, J. Revadekar, G. Griffiths, L. Vincent, D. B. Stephenson, J. Burn, E. Aguilar, M. Brunet, M. Taylor, M. New, P. Zhai, M. Rusticucci, and J. L. Vazquez-Aguirre. Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research*, 111(D5):D05109, 2006. ISSN 0148-0227. doi: 10.1029/2005JD006290.

A. Anav, P. Friedlingstein, C. Beer, P. Ciais, A. Harper, C. Jones, G. Murray-Tortarolo, D. Papale, N. C. Parazoo, P. Peylin, S. Piao, S. Sitch, N. Viovy, A. Wiltshire, and M. Zhao. Spatiotemporal patterns of terrestrial gross primary production: A review: GPP Spatiotemporal Patterns. *Reviews of Geophysics*, 53(3):785–818, September 2015. ISSN 87551209. doi: 10.1002/2015RG000483. URL `http://doi.wiley.com/10.1002/2015RG000483`.

L. E. O. C. Aragão, L. O. Anderson, M. G. Fonseca, T. M. Rosan, L. B. Vedovato, F. H. Wagner, C. V. J. Silva, C. H. L. Silva Junior, E. Arai, A. P. Aguiar, J. Barlow, E. Berenguer, M. N. Deeter, L. G. Domingues, L. Gatti, M. Gloor, Y. Malhi, J. A. Marengo, J. B. Miller, O. L. Phillips, and S. Saatchi. 21st Century drought-related fires counteract the decline of Amazon deforestation carbon emissions. *Nature Communications*, 9(1):536, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-017-02771-y.

P.-L. Ardisson, E. Bourget, and P. Legendre. Multivariate Approach to Study Species Assemblages at Large Spatiotemporal Scales: The Community Structure of the Epibenthic Fauna of the Estuary and Gulf of St. Lawrence. *Canadian Journal of Fisheries and Aquatic Sciences*, 47(7):1364–1377, July 1990. ISSN 0706-652X, 1205-7533. doi: 10.1139/f90-156.

J. Arenas-Garcia, K. B. Petersen, G. Camps-Valls, and L. K. Hansen. Kernel Multivariate Analysis Framework for Supervised Subspace Learning: A Tutorial on Linear and Kernel Multivariate Methods. *IEEE Signal Processing Magazine*, 30(4):16–29, July 2013. ISSN 1053-5888. doi: 10.1109/MSP.2013.2250591.

M. Babaee, M. Datcu, and G. Rigoll. Assessment of dimensionality reduction based on communication channel model; application to immersive information visualization. In *Big Data 2013*, pages 1–6. IEEE Xplore, 2013. doi: 10.1109/bigdata.2013.6691726.

F. Babst, B. Poulter, P. Bodesheim, M. D. Mahecha, and D. C. Frank. Improved tree-ring archives will support earth-system science. *Nature Ecol. Evolut*, 1:1–2, 2017.

G. H. Bakir, J. Weston, and P. B. Schölkopf. Learning to Find Pre-Images. In S. Thrun, L. K. Saul, and P. B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 449–456. MIT Press, 2004. doi: 10.1007/978-3-540-28649-3_31.

D. D. Baldocchi. How eddy covariance flux measurements have contributed to our understanding of Global Change Biology. *Global Change Biology*, 26(1):242–260, 2020. ISSN 1365-2486. doi: 10.1111/gcb.14807.

S. Barocas and A. D. Selbst. Big Data's Disparate Impact. *SSRN Electronic Journal*, 2016. ISSN 1556-5068. doi: 10.2139/ssrn.2477899.

D. Barriopedro, E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. García-Herrera. The Hot Summer of 2010: Redrawing the Temperature Record Map of Europe. *Science*, 332(6026):220–224, April 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1201224.

F. Becker and B. J. Choudhury. Relative sensitivity of normalized difference vegetation Index (NDVI) and microwave polarization difference Index (MPDI) for vegetation and desertification monitoring. *Remote Sensing of Environment*, 24(2):297–311, March 1988. ISSN 0034-4257. doi: 10.1016/0034-4257(88)90031-4.

B. Beisner, D. Haydon, and K. Cuddington. Alternative stable states in ecology. *Frontiers in Ecology and the Environment*, 1(7):376–382, September 2003. ISSN 1540-9295. doi: 10.1890/1540-9295(2003)001[0376:ASSIE]2.0.CO;2.

M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373, 2003. ISSN 08997667. doi: 10.1162/089976603321780317.

Y. Bengio, O. Delalleau, N. L. Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Learning Eigenfunctions Links Spectral Embedding and Kernel PCA. *Neural Computation*, 16(10):2197–2219, 2004. ISSN 0899-7667. doi: 10.1162/0899766041732396.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 0035-9246.

R. Bennett. The intrinsic dimensionality of signal collections. *IEEE Transactions on Information Theory*, 15(5):517–525, September 1969. ISSN 1557-9654. doi: 10.1109/TIT.1969.1054365.

M. Berger, J. Moreno, J. A. Johannessen, P. F. Levelt, and R. F. Hanssen. ESA's sentinel missions in support of Earth system science. *Remote Sensing of Environment*, 120:84–90, May 2012. ISSN 00344257. doi: 10.1016/j.rse.2011.07.023.

B. Blonder, D. E. Moulton, J. Blois, B. J. Enquist, B. J. Graae, M. Macias-Fauria, B. McGill, S. Nogué, A. Ordonez, B. Sandel, and J.-C. Svenning. Predictability in community dynamics. *Ecology Letters*, 20(3):293–306, 2017. ISSN 1461-0248. doi: 10.1111/ele.12736.

P. Bodesheim, M. Jung, F. Gans, M. D. Mahecha, and M. Reichstein. Upscaled diurnal cycles of land–atmosphere fluxes: A new global half-hourly data product. *Earth System Science Data*, 10(3):1327–1365, July 2018. ISSN 1866-3508. doi: 10.5194/essd-10-1327-2018.

G. B. Bonan. *Ecological Climatology: Concepts and Applications*. Cambridge University Press, New York, NY, USA, third edition edition, 2015. ISBN 978-1-107-04377-0 978-1-107-61905-0.

I. S. Bowen. The Ratio of Heat Losses by Conduction and by Evaporation from any Water Surface. *Physical Review*, 27(6):779–787, June 1926. doi: 10.1103/PhysRev.27.779.

R. L. Brown, J. Durbin, and J. M. Evans. Techniques for Testing the Constancy of Regression Relationships over Time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(2):149–192, 1975. ISSN 0035-9246.

G. Camps-Valls, D. Tuia, L. Gómez-Chova, S. Jiménez, and J. Malo. Remote Sensing Image Processing. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 5(1):1–192, December 2011. ISSN 1559-8136. doi: 10.2200/S00392ED1V01Y201107IVM012.

R. B. Cattell. *Factor Analysis an Introduction and Manual for the Psychologist and Social Scientist.* Harper, New York, 1952.

R. B. Cattell. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, 1(2):245–276, April 1966. ISSN 0027-3171, 1532-7906. doi: 10.1207/s15327906mbr0102_10.

G. Ceballos and P. R. Ehrlich. The misunderstood sixth mass extinction. *Science*, 360(6393):1080–1081, June 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aau0191.

K. Chalupka, F. Eberhardt, and P. Perona. Causal feature learning: An overview. *Behaviormetrika*, 44(1):137–164, January 2017. ISSN 1349-6964. doi: 10.1007/s41237-016-0008-2.

F. S. Chapin, G. M. Woodwell, J. T. Randerson, E. B. Rastetter, G. M. Lovett, D. D. Baldocchi, D. A. Clark, M. E. Harmon, D. S. Schimel, R. Valentini, C. Wirth, J. D. Aber, J. J. Cole, M. L. Goulden, J. W. Harden, M. Heimann, R. W. Howarth, P. A. Matson, A. D. McGuire, J. M. Melillo, H. A. Mooney, J. C. Neff, R. A. Houghton, M. L. Pace, M. G. Ryan, S. W. Running, O. E. Sala, W. H. Schlesinger, and E.-D. Schulze. Reconciling Carbon-cycle Concepts, Terminology, and Methods. *Ecosystems*, 9(7):1041–1050, November 2006. ISSN 1435-0629. doi: 10.1007/s10021-005-0105-7.

C. Chen, T. Park, X. Wang, S. Piao, B. Xu, R. K. Chaturvedi, R. Fuchs, V. Brovkin, P. Ciais, R. Fensholt, H. Tømmervik, G. Bala, Z. Zhu, R. R. Nemani, and R. B. Myneni. China and India lead in greening of the world through land-use management. *Nature Sustainability*, 2(2):122–129, February 2019. ISSN 2398-9629. doi: 10.1038/s41893-019-0220-7.

L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, 104(485):209–219, 2009. doi: 10.1198/jasa.2009.0111.

P. Ciais, M. Reichstein, N. Viovy, A. Granier, J. Ogée, V. Allard, M. Aubinet, N. Buchmann, C. Bernhofer, A. Carrara, F. Chevallier, N. D. Noblet, A. D. Friend, P. Friedlingstein, T. Grünwald, B. Heinesch, P. Keronen, A. Knohl, G. Krinner, D. Loustau, G. Manca, G. Matteucci, F. Miglietta,

J. M. Ourcival, D. Papale, K. Pilegaard, S. Rambal, G. Seufert, J. F. Soussana, M. J. Sanz, E. D. Schulze, T. Vesala, and R. Valentini. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, 437(7058):529, September 2005. ISSN 1476-4687. doi: 10.1038/nature03972.

CIESIN. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11. Technical report, NASA Socioeconomic Data and Applications Center (SEDAC) - Center For International Earth Science Information Network (CIESIN) - Columbia University, Palisades, NY, 2018.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. ISSN 10635203. doi: 10.1016/j.acha.2006.04.006.

P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, 1994. ISSN 01651684. doi: 10.1016/0165-1684(94)90029-9.

P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin. Review change detection methods in ecosystem monitoring: A review. *International Journal of Remote Sensing*, 25(9):1565–1596, May 2004. ISSN 0143-1161, 1366-5901. doi: 10.1080/0143116031000101675.

R. Costanza, L. Fioramonti, and I. Kubiszewski. The UN Sustainable Development Goals and the dynamics of well-being. *Frontiers in Ecology and the Environment*, 14(2):59–59, 2016. ISSN 1540-9309. doi: 10.1002/fee.1231.

R. de Jong, S. de Bruin, A. de Wit, M. E. Schaepman, and D. L. Dent. Analysis of monotonic greening and browning trends from global NDVI time-series. *Remote Sensing of Environment*, 115(2):692–702, February 2011. ISSN 0034-4257. doi: 10.1016/j.rse.2010.10.011.

J. de Leeuw and P. Mair. Multidimensional scaling using majorization: Smacof in r. *Journal of Statistical Software, Articles*, 31(3):1–30, 2009. ISSN 1548-7660. doi: 10.18637/jss.v031.i03. URL `https://www.jstatsoft.org/v031/i03`.

V. de Silva and J. B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford, 2004.

S. Díaz, J. Settele, E. S. Brondízio, H. T. Ngo, J. Agard, A. Arneth, P. Balvanera, K. A. Brauman, S. H. M. Butchart, K. M. A. Chan, L. A. Garibaldi, K. Ichii, J. Liu, S. M. Subramanian, G. F. Midgley, P. Miloslavich, Z. Molnár, D. Obura, A. Pfaff, S. Polasky, A. Purvis, J. Razzaque, B. Reyers, R. R. Chowdhury, Y.-J. Shin, I. Visseren-Hamakers, K. J. Willis, and C. N. Zayas. Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science*, 366(6471), December 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax3100.

M. Disney, J.-P. Muller, S. Kharbouche, T. Kaminski, M. Voßbeck, P. Lewis, and B. Pinty. A New Global fAPAR and LAI Dataset Derived from Optimal Albedo Estimates: Comparison with MODIS Products. *Remote Sensing*, 8(4):275, April 2016. doi: 10.3390/rs8040275.

C. E. Doughty, D. B. Metcalfe, C. a. J. Girardin, F. F. Amézquita, D. G. Cabrera, W. H. Huasco, J. E. Silva-Espejo, A. Araujo-Murakami, M. C. da Costa, W. Rocha, T. R. Feldpausch, A. L. M. Mendoza, A. C. L. da Costa, P. Meir, O. L. Phillips, and Y. Malhi. Drought impact on forest carbon dynamics and fluxes in Amazonia. *Nature*, 519(7541): 78–82, March 2015. ISSN 1476-4687. doi: 10.1038/nature14213. URL `https://www.nature.com/articles/nature14213`.

D. R. Easterling, G. A. Meehl, C. Parmesan, S. A. Changnon, T. R. Karl, and L. O. Mearns. Climate Extremes: Observations, Modeling, and Impacts. *Science*, 289(5487):2068–2074, September 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.289.5487.2068. URL `http://science.sciencemag.org/content/289/5487/2068`.

T. R. Feldpausch, O. L. Phillips, R. J. W. Brienen, E. Gloor, J. Lloyd, G. Lopez-Gonzalez, A. Monteagudo-Mendoza, Y. Malhi, A. Alarcón, E. A. Dávila, P. Alvarez-Loayza, A. Andrade, L. E. O. C. Aragao, L. Arroyo, G. A. A. C, T. R. Baker, C. Baraloto, J. Barroso, D. Bonal, W. Castro, V. Chama, J. Chave, T. F. Domingues, S. Fauset, N. Groot, E. H. Coronado, S. Laurance, W. F. Laurance, S. L. Lewis, J. C. Licona, B. S. Marimon, B. H. Marimon-Junior, C. M. Bautista, D. A. Neill, E. A. Oliveira, C. O. dos Santos, N. C. P. Camacho, G. Pardo-Molina, A. Prieto, C. A. Quesada, F. Ramírez, H. Ramírez-Angulo, M. Réjou-Méchain, A. Rudas, G. Saiz, R. P. Salomão, J. E. Silva-Espejo, M. Silveira, H. ter Steege, J. Stropp, J. Terborgh, R. Thomas-Caesar, G. M. F. van der Heijden, R. V. Martinez, E. Vilanova, and V. A. Vos. Amazon forest response to repeated droughts. *Global Biogeochemical Cycles*, 30(7):964–982, July 2016. ISSN 1944-9224. doi: 10.1002/2015GB005133.

D. Filmer and L. H. Pritchett. Estimating Wealth Effects Without Expenditure Data—Or Tears: An Application To Educational Enrollments In States Of India*. *Demography*, 38(1):115–132, February 2001. ISSN 0070-3370, 1533-7790. doi: 10.1353/dem.2001.0003. URL https://link.springer.com/article/10.1353/dem.2001.0003.

D. Filmer and K. Scott. Assessing Asset Indices. *Demography*, 49(1):359–392, February 2012. ISSN 1533-7790. doi: 10.1007/s13524-011-0077-5.

M. Flach, F. Gans, A. Brenning, J. Denzler, M. Reichstein, E. Rodner, S. Bathiany, P. Bodesheim, Y. Guanche, S. Sippel, and M. D. Mahecha. Multivariate anomaly detection for Earth observations: A comparison of algorithms and feature extraction techniques. *Earth System Dynamics*, 8 (3):677–696, August 2017. ISSN 2190-4987. doi: 10.5194/esd-8-677-2017.

M. Flach, S. Sippel, F. Gans, A. Bastos, A. Brenning, M. Reichstein, and M. D. Mahecha. Contrasting biosphere responses to hydrometeorological extremes: revisiting the 2010 western Russian heatwave. *Biogeosciences*, 15(20):6067–6085, October 2018. ISSN 1726-4170. doi: https://doi.org/10.5194/bg-15-6067-2018. URL https://www.biogeosciences.net/15/6067/2018/.

C. Folke, S. Carpenter, B. Walker, M. Scheffer, T. Elmqvist, L. Gunderson, and C. Holling. Regime Shifts, Resilience, and Biodiversity in Ecosystem Management. *Annual Review of Ecology, Evolution, and Systematics*, 35(1): 557–581, November 2004. ISSN 1543-592X. doi: 10.1146/annurev.ecolsys.35.021103.105711.

M. Forkel, N. Carvalhais, C. Rodenbeck, R. Keeling, M. Heimann, K. Thonicke, S. Zaehle, and M. Reichstein. Enhanced seasonal CO2 exchange caused by amplified plant productivity in northern ecosystems. *Science*, 351(6274):696–699, February 2016. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aac4971.

M. Forkel, N. Carvalhais, J. Verbesselt, M. Mahecha, C. Neigh, M. Reichstein, M. Forkel, N. Carvalhais, J. Verbesselt, M. D. Mahecha, C. S. R. Neigh, and M. Reichstein. Trend Change Detection in NDVI Time Series: Effects of Inter-Annual Variability and Methodology. *Remote Sensing*, 5 (5):2113–2144, May 2013. doi: 10.3390/rs5052113.

M. Forkel, M. Migliavacca, K. Thonicke, M. Reichstein, S. Schaphoff, U. Weber, and N. Carvalhais. Codominant water control on global

interannual variability and trends in land surface phenology and greenness. *Global Change Biology*, 21(9):3414–3435, 2015. ISSN 1365-2486. doi: 10.1111/gcb.12950.

A. C. Foster, A. H. Armstrong, J. K. Shuman, H. H. Shugart, B. M. Rogers, M. C. Mack, S. J. Goetz, and K. J. Ranson. Importance of tree- and species-level interactions with wildfire, climate, and soils in interior Alaska: Implications for forest change under a warming climate. *Ecological Modelling*, 409:108765, October 2019. ISSN 03043800. doi: 10.1016/j.ecolmodel.2019.108765.

T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Softw: Pract. Exper.*, 21(11):1129–1164, 1991. ISSN 1097-024X. doi: 10.1002/spe.4380211102.

Gapminder Foundation. Gapminder: Unveiling the beauty of statistics for a fact based world view., 2018. URL `https://www.gapminder.org/`.

R. García-Herrera, J. Díaz, R. M. Trigo, J. Luterbacher, and E. M. Fischer. A Review of the European Summer Heat Wave of 2003. *Critical Reviews in Environmental Science and Technology*, 40(4):267–306, March 2010. ISSN 1064-3389, 1547-6537. doi: 10.1080/10643380802238137. URL `http://www.tandfonline.com/doi/abs/10.1080/10643380802238137`.

D. Gaumont-Guay, T. A. Black, T. J. Griffis, A. G. Barr, R. S. Jassal, and Z. Nesic. Interpreting the dependence of soil respiration on soil temperature and water content in a boreal aspen stand. *Agricultural and Forest Meteorology*, 140(1):220–235, November 2006. ISSN 0168-1923. doi: 10.1016/j.agrformet.2006.08.003.

S. Ghislandi, W. C. Sanderson, and S. Scherbov. A Simple Measure of Human Development: The Human Life Indicator. *Population and Development Review*, November 2018. ISSN 1728-4457. doi: 10.1111/padr.12205.

S. Ghislandi, W. C. Sanderson, and S. Scherbov. A simple measure of human development: The human life indicator. *Population and development review*, 45(1):219, 2019.

C. Gómez, J. C. White, and M. A. Wulder. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:55–72, June 2016. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2016.03.008.

Google. Google Books Ngram Viewer, 2020. URL https://books.google.com/ngrams/graph?content=climate+indicators%2Cclimate+indices%2Cclimate+index%2Cclimate+indicator%2Cclimate+indexes&year_start=1940&year_end=2019&corpus=26&smoothing=3.

M. Göpel. *The Great Mindshift*, volume 2 of *The Anthropocene: Politik—Economics—Society—Science*. Springer International Publishing, Cham, 2016. ISBN 978-3-319-43765-1 978-3-319-43766-8. doi: 10.1007/978-3-319-43766-8. URL http://link.springer.com/10.1007/978-3-319-43766-8.

H. D. Graven, R. F. Keeling, S. C. Piper, P. K. Patra, B. B. Stephens, S. C. Wofsy, L. R. Welp, C. Sweeney, P. P. Tans, J. J. Kelley, B. C. Daube, E. A. Kort, G. W. Santoni, and J. D. Bent. Enhanced Seasonal Exchange of CO2 by Northern Ecosystems Since 1960. *Science*, 341(6150):1085–1089, September 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1239207.

D. Griggs, M. Stafford-Smith, O. Gaffney, J. Rockström, M. C. Öhman, P. Shyamsundar, W. Steffen, G. Glaser, N. Kanie, and I. Noble. Sustainable development goals for people and planet. *Nature*, 495(7441):305–307, March 2013. ISSN 1476-4687. doi: 10.1038/495305a.

J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 47, New York, NY, USA, July 2004. Association for Computing Machinery. ISBN 978-1-58113-838-2. doi: 10.1145/1015330.1015417.

Z. Hao, J. Zheng, Q. Ge, and W. Wang. Historical analogues of the 2008 extreme snow event over Central and Southern China. *Climate Research*, 50(2):161–170, December 2011. ISSN 0936-577X, 1616-1572. doi: 10.3354/cr01052.

H. H. Hendon, E.-P. Lim, J. M. Arblaster, and D. L. T. Anderson. Causes and predictability of the record wet east Australian spring 2010. *Climate Dynamics*, 42(5):1155–1174, March 2014. ISSN 1432-0894. doi: 10.1007/s00382-013-1700-5.

N. J. Higham. The Accuracy of Floating Point Summation. *SIAM J. Scientific Computing*, 14:783–799, 1993. doi: 10.1137/0914050.

G. E. Hinton and S. T. Roweis. Stochastic Neighbor Embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information*

Processing Systems 15, pages 857–864. MIT Press, 2003. URL http://papers.nips.cc/paper/2276-stochastic-neighbor-embedding.pdf.

M. Horridge, J. Madden, and G. Wittwer. The impact of the 2002–2003 drought on Australia. *Journal of Policy Modeling*, 27(3):285–308, April 2005. ISSN 0161-8938. doi: 10.1016/j.jpolmod.2005.01.008.

K. Huang, J. Xia, Y. Wang, A. Ahlström, J. Chen, R. B. Cook, E. Cui, Y. Fang, J. B. Fisher, D. N. Huntzinger, Z. Li, A. M. Michalak, Y. Qiao, K. Schaefer, C. Schwalm, J. Wang, Y. Wei, X. Xu, L. Yan, C. Bian, and Y. Luo. Enhanced peak growth of global vegetation and its key mechanisms. *Nature Ecology & Evolution*, 2(12):1897–1905, December 2018. ISSN 2397-334X. doi: 10.1038/s41559-018-0714-0.

A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999. ISSN 1045-9227. doi: 10.1109/72.761722.

IPBES. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Summary for policymakers, IPBES, May 2019.

IPCC. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Technical report, IPCC, Geneva, Swizerland, 2014.

IPCC. IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse gas fluxes in Terrestrial Ecosystems Summary for Policymakers Approved Draft. Technical report, IPCC, August 2019.

E. Ivits, S. Horion, R. Fensholt, and M. Cherlet. Drought footprint on European ecosystems between 1999 and 2010 assessed by remotely sensed vegetation phenology and productivity. *Global Change Biology*, 20 (2):581–593, 2014. ISSN 1365-2486. doi: 10.1111/gcb.12393.

W. M. Jolly, M. A. Cochrane, P. H. Freeborn, Z. A. Holden, T. J. Brown, G. J. Williamson, and D. M. J. S. Bowman. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, 6 (1):7537, November 2015. ISSN 2041-1723. doi: 10.1038/ncomms8537.

M. Jung, S. Koirala, U. Weber, K. Ichii, F. Gans, G. Camps-Valls, D. Papale, C. Schwalm, G. Tramontana, and M. Reichstein. The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Scientific Data*, 6(1):1–14, May 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0076-8.

T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989. ISSN 0020-0190. doi: 10.1016/0020-0190(89)90102-6.

C. D. Keeling, J. F. S. Chin, and T. P. Whorf. Increased activity of northern vegetation inferred from atmospheric CO 2 measurements. *Nature*, 382 (6587):146, July 1996. ISSN 1476-4687. doi: 10.1038/382146a0.

M. G. Kendall. *Rank Correlation Methods.* Griffin, London, 1970. ISBN 0-85264-199-0 978-0-85264-199-6.

J. Khanna, D. Medvigy, S. Fueglistaler, and R. Walko. Regional dry-season climate changes due to three decades of Amazonian deforestation. *Nature Climate Change*, 7(3):200–204, March 2017. ISSN 1758-6798. doi: 10.1038/nclimate3226. URL `https://www.nature.com/articles/nclimate3226`.

M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel. World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, 15(3):259–263, July 2006. ISSN 0941-2948. doi: 10.1127/0941-2948/2006/0130. URL `http://www.schweizerbart.de/papers/metz/detail/15/55034/World_Map_of_the_Koppen_Geiger_climate_classificat?af=crossref`.

G. Kraemer, M. Reichstein, and M. D. Mahecha. dimRed and coRanking - Unifying Dimensionality Reduction in R. *The R Journal*, 10(1):342–358, 2018. ISSN 2073-4859. doi: 10.32614/RJ-2018-039.

G. Kraemer, G. Camps-Valls, M. Reichstein, and M. D. Mahecha. Summarizing the state of the terrestrial biosphere in few dimensions. *Biogeosciences*, 17(9):2397–2424, May 2020a. ISSN 1726-4170. doi: 10.5194/bg-17-2397-2020.

G. Kraemer, M. Reichstein, G. Camps-Valls, J. Smits, and M. D. Mahecha. The Low Dimensionality of Development. *Social Indicators Research*, May 2020b. ISSN 1573-0921. doi: 10.1007/s11205-020-02349-0.

J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964a. ISSN 0033-3123, 1860-0980. doi: 10.1007/bf02289565.

J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964b. ISSN 0033-3123, 1860-0980. doi: 10.1007/bf02289694.

C.-M. Kuan and K. Hornik. The generalized fluctuation test: A unifying view. *Econometric Reviews*, 14(2):135–161, January 1995. ISSN 0747-4938, 1532-4168. doi: 10.1080/07474939508800311.

I. Kubiszewski, R. Costanza, C. Franco, P. Lawn, J. Talberth, T. Jackson, and C. Aylmer. Beyond GDP: Measuring and achieving global genuine progress. *Ecological Economics*, 93:57–68, September 2013. ISSN 09218009. doi: 10.1016/j.ecolecon.2013.04.019. URL `http://linkinghub.elsevier.com/retrieve/pii/S0921800913001584`.

W. Köppen and R. Geiger. Klima der Erde (Climate of the earth) Wall Map. *Gotha: Klett-Perthes*, 1954.

V. Laparra, J. Malo, and G. Camps-Valls. Dimensionality Reduction via Regression in Hyperspectral Imagery. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1026–1036, 2015. ISSN 1932-4553. doi: 10.1109/jstsp.2015.2417833.

J. A. Lee, J. A. Lee, and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *NEUROCOMPUTING*, 72, 2009.

J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013. ISSN 0925-2312. doi: 10.1016/j.neucom.2012.12.036.

P. Legendre, D. Planas, and M.-J. Auclair. Succession des communautés de gastéropodes dans deux milieux différant par leur degré d'eutrophisation. *Canadian Journal of Zoology*, 62(11):2317–2327, November 1984. ISSN 0008-4301, 1480-3283. doi: 10.1139/z84-339. URL `http://www.nrcresearchpress.com/doi/10.1139/z84-339`.

P. Legendre and L. Legendre. Numerical ecology: Second English edition. *Developments in environmental modelling*, 20, 1998.

T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber. Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences*, 105(6):1786–1793, December 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0705414105.

E. N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, March 1963. doi: 10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2.

W. Lueks, B. Mokbel, M. Biehl, and B. Hammer. How to Evaluate Dimensionality Reduction? - Improving the Co-ranking Matrix. *arXiv:1110.3917 [cs]*, 2011. URL `http://arxiv.org/abs/1110.3917`. arXiv: 1110.3917.

M. D. Mahecha, M. Reichstein, H. Lange, N. Carvalhais, C. Bernhofer, T. Grünwald, D. Papale, and G. Seufert. Characterizing ecosystem-atmosphere interactions from short to interannual time scales. *Biogeosciences*, 4(5):743–758, September 2007a. ISSN 1726-4170. doi: https://doi.org/10.5194/bg-4-743-2007.

M. D. Mahecha, A. Martínez, G. Lischeid, and E. Beck. Nonlinear dimensionality reduction: Alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. *Ecological informatics*, 2(2):138–149, 2007b.

M. D. Mahecha, A. Martínez, G. Lischeid, and E. Beck. Nonlinear dimensionality reduction: Alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. *Ecological Informatics*, 2(2):138–149, June 2007c. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2007.05.002.

M. D. Mahecha, F. Gans, S. Sippel, J. F. Donges, T. Kaminski, S. Metzger, M. Migliavacca, D. Papale, A. Rammig, and J. Zscheischler. Detecting impacts of extreme events with ecological in situ monitoring networks. *Biogeosciences*, 14(18):4255–4277, September 2017. ISSN 1726-4170. doi: https://doi.org/10.5194/bg-14-4255-2017.

M. D. Mahecha, F. Gans, G. Brandt, R. Christiansen, S. E. Cornell, N. Fomferra, G. Kraemer, J. Peters, P. Bodesheim, G. Camps-Valls, J. F. Donges, W. Dorigo, L. Estupiñan-Suarez, V. H. Gutierrez-Velez, M. Gutwin, M. Jung, M. C. Londoño, D. G. Miralles, P. Papastefanou, and M. Reichstein. Earth system data cubes unravel global multivariate dynamics. *Earth System Dynamics Discussions*, pages 1–51, October 2019. ISSN 2190-4979. doi: https://doi.org/10.5194/esd-2019-62.

M. D. Mahecha, F. Gans, G. Brandt, R. Christiansen, S. E. Cornell, N. Fomferra, G. Kraemer, J. Peters, P. Bodesheim, G. Camps-Valls, J. F. Donges, W. Dorigo, L. M. Estupinan-Suarez, V. H. Gutierrez-Velez, M. Gutwin, M. Jung, M. C. Londoño, D. G. Miralles, P. Papastefanou, and M. Reichstein. Earth system data cubes unravel global multivariate dynamics.

Earth System Dynamics, 11(1):201–234, February 2020. ISSN 2190-4979. doi: https://doi.org/10.5194/esd-11-201-2020.

H. B. Mann. Nonparametric Tests Against Trend. *Econometrica*, 13(3): 245–259, 1945. ISSN 0012-9682. doi: 10.2307/1907187.

M. G. Marshall and G. Elzinga-Marshall. *Global Report 2017, Conflict, Governance, and State Fragility*. Center for Systemic Peace, 2017. URL http://www.systemicpeace.org/vlibrary/GlobalReport2017.pdf.

B. Martens, D. G. Miralles, H. Lievens, R. v. d. Schalie, R. A. M. d. Jeu, D. Fernández-Prieto, H. E. Beck, W. A. Dorigo, and N. E. C. Verhoest. GLEAM v3: satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, 10(5):1903–1925, May 2017. ISSN 1991-959X. doi: https://doi.org/10.5194/gmd-10-1903-2017. URL https://www.geosci-model-dev.net/10/1903/2017/.

S. Martin, W. M. Brown, and B. N. Wylie. Dr.l: Distributed Recursive (graph) Layout. Technical Report dRl; 002182MLTPL00, Sandia National Laboratories, 2007. URL http://www.osti.gov/scitech/biblio/1231060-dr-distributed-recursive-graph-layout.

R. Mathieu, L. Naidoo, M. A. Cho, B. Leblon, R. Main, K. Wessels, G. P. Asner, J. Buckley, J. Van Aardt, B. F. N. Erasmus, and I. P. J. Smit. Toward structural assessment of semi-arid African savannahs and woodlands: The potential of multitemporal polarimetric RADARSAT-2 fine beam images. *Remote Sensing of Environment*, 138:215–231, November 2013. ISSN 0034-4257. doi: 10.1016/j.rse.2013.07.011.

M. Mazziotta and A. Pareto. Use and Misuse of PCA for Measuring Well-Being. *Social Indicators Research*, 142(2):451–476, April 2019. ISSN 1573-0921. doi: 10.1007/s11205-018-1933-0.

M. McGillivray. The human development index: Yet another redundant composite development indicator? *World Development*, 19(10):1461–1468, October 1991. ISSN 0305-750X. doi: 10.1016/0305-750X(91)90088-Y.

L. McRae, R. Freeman, V. Marconi, and Canadian Electronic Library (Firm). *Living Planet Report 2016: Risk and Resilience in a New Era*. WWF, 2016. ISBN 978-2-940529-40-7. OCLC: 1001121301.

M. J. Metzger, R. G. H. Bunce, R. H. G. Jongman, R. Sayre, A. Trabucco, and R. Zomer. A high-resolution bioclimate map of the world: A unifying framework for global biodiversity research and monitoring. *Global*

Ecology and Biogeography, 22(5):630–638, May 2013. ISSN 1466-8238. doi: 10.1111/geb.12022.

S. Mika, B. Scholkopf, A. Smola, K. Muller, M. Scholz, and G. Ratsch. Kernel PCA and de-noising in feature spaces. In Kearns, MS and Solla, SA and Cohn, DA, editor, *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 11*, volume 11 of *Advances in Neural Information Processing Systems*, pages {536–542}, 1999. ISBN 0-262-11245-0. 12th Annual Conference on Neural Information Processing Systems (NIPS), DENVER, CO, NOV 30-DEC 05, 1998.

D. G. Miralles, A. J. Teuling, C. C. van Heerwaarden, and J. Vilà-Guerau de Arellano. Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nature Geoscience*, 7(5):345–349, May 2014. ISSN 1752-0908. doi: 10.1038/ngeo2141. URL `https://www.nature.com/articles/ngeo2141`.

S. Monni and A. Spaventa. Beyond GDP and HDI: Shifting the focus from paradigms to politics. *Development*, 56(2):227–231, June 2013. ISSN 1011-6370, 1461-7072. doi: 10.1057/dev.2013.30.

G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, February 2018. ISSN 1051-2004. doi: 10.1016/j.dsp.2017.10.011.

J.-P. Muller, P. Lewis, J. Fischer, P. North, and U. Framer. The ESA Glob-Albedo Project for mapping the Earth's land surface albedo for 15 years from European sensors. *Geophysical Research Abstracts*, 13:2, 2011.

K. Murthy and S. Bagchi. Spatial patterns of long-term vegetation greening and browning are consistent across multiple scales: Implications for monitoring land degradation. *Land Degradation & Development*, 29(8): 2485–2495, 2018. ISSN 1099-145X. doi: 10.1002/ldr.3019.

E. Najafi, I. Pal, and R. Khanbilvardi. Climate drives variability and joint variability of global crop yields. *Science of The Total Environment*, 662: 361–372, April 2019. ISSN 0048-9697. doi: 10.1016/j.scitotenv.2019.01.172.

K. N. Nasahara and S. Nagai. Review: Development of an in situ observation network for terrestrial ecological remote sensing: The Phenological Eyes Network (PEN). *Ecological Research*, 30(2):211–223, March 2015. ISSN 0912-3814. doi: 10.1007/s11284-014-1239-x.

N. Nicholls. The Changing Nature of Australian Droughts. *Climatic Change*, 63(3):323–336, April 2004. ISSN 1573-1480. doi: 10.1023/B:CLIM.0000018515.46344.6d.

S. E. Nicholson. A detailed look at the recent drought situation in the Greater Horn of Africa. *Journal of Arid Environments*, 103:71–79, April 2014. ISSN 0140-1963. doi: 10.1016/j.jaridenv.2013.12.003. URL http://www.sciencedirect.com/science/article/pii/S0140196313002322.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax2342.

OEDC. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD, Paris, 2008. ISBN 978-92-64-04345-9. OCLC: ocn244969711.

J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling. Climate Data Challenges in the 21st Century. *Science*, 331(6018):700–702, February 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1197869.

C. Papagiannopoulou, D. G. Miralles, M. Demuzere, N. E. C. Verhoest, and W. Waegeman. Global hydro-climatic biomes identified via multi-task learning. *Geoscientific Model Development Discussions*, pages 1–19, April 2018. ISSN 1991-962X. doi: 10.5194/gmd-2018-92. URL https://www.geosci-model-dev-discuss.net/gmd-2018-92/.

D. Papale, T. A. Black, N. Carvalhais, A. Cescatti, J. Chen, M. Jung, G. Kiely, G. Lasslop, M. D. Mahecha, H. Margolis, L. Merbold, L. Montagnani, E. Moors, J. E. Olesen, M. Reichstein, G. Tramontana, E. van Gorsel, G. Wohlfahrt, and B. Ráduly. Effect of spatial sampling from European flux towers for estimating carbon and water fluxes with artificial neural networks. *Journal of Geophysical Research: Biogeosciences*, 120(10):1941–1957, 2015. ISSN 2169-8961. doi: 10.1002/2015JG002997.

C. Parmesan. Ecological and Evolutionary Responses to Recent Climate Change. *Annual Review of Ecology, Evolution, and Systematics*, 37(1):637–669, November 2006. ISSN 1543-592X. doi: 10.1146/annurev.ecolsys.37.091305.110100.

T. M. Parris and R. W. Kates. Characterizing and Measuring Sustainable Development. *Annual Review of Environment and Resources*, 28(1):559–586, November 2003. ISSN 1543-5938, 1545-2050. doi: 10.1146/annurev.energy.28.050302.105551. URL http://www.annualreviews.org/doi/10.1146/annurev.energy.28.050302.105551.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics : A Primer*. Wiley, 2016. ISBN 978-1-119-18684-7.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

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.

A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls. Fair Kernel Learning. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 339–355. Springer International Publishing, 2017. ISBN 978-3-319-71249-9. doi: https://doi.org/10.1007/978-3-319-71249-9_21.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

S. Piao, X. Wang, T. Park, C. Chen, X. Lian, Y. He, J. W. Bjerke, A. Chen, P. Ciais, H. Tømmervik, R. R. Nemani, and R. B. Myneni. Characteristics, drivers and feedbacks of global greening. *Nature Reviews Earth & Environment*, 1(1):14–27, January 2020a. ISSN 2662-138X. doi: 10.1038/s43017-019-0001-x.

S. Piao, X. Wang, K. Wang, X. Li, A. Bastos, J. G. Canadell, P. Ciais, P. Friedlingstein, and S. Sitch. Interannual variation of terrestrial carbon cycle: Issues and perspectives. *Global Change Biology*, 26(1):300–318, 2020b. ISSN 1365-2486. doi: 10.1111/gcb.14884.

J. Porta, J. Verbeek, and B. Krose. Active appearance-based robot localization using stereo vision. *Autonomous Robots*, 18(1):59–80, 2005. doi: 10.1023/B:AURO.0000047287.00119.b6.

P. Pradhan, L. Costa, D. Rybski, W. Lucht, and J. P. Kropp. A Systematic Study of Sustainable Development Goal (SDG) Interactions: A SYSTEMATIC STUDY OF SDG INTERACTIONS. *Earth's Future*, 5(11):1169–1179, November 2017. ISSN 23284277. doi: 10.1002/2017EF000632.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

R. Ram. Composite indices of physical quality of life, basic needs fulfilment, and income: A 'principal component' representation. *Journal of Development Economics*, 11(2):227–247, 1982. URL `http://www.sciencedirect.com/science/article/pii/0304387882900050`.

M. Rao, Saw Htun, S. G. Platt, R. Tizard, C. Poole, Than Myint, and J. E. M. Watson. Biodiversity Conservation in a Changing Climate: A Review of Threats and Implications for Conservation Planning in Myanmar. *AMBIO*, 42(7):789–804, November 2013. ISSN 1654-7209. doi: 10.1007/s13280-013-0423-5.

M. Reichstein, M. Bahn, P. Ciais, D. Frank, M. D. Mahecha, S. I. Seneviratne, J. Zscheischler, C. Beer, N. Buchmann, D. C. Frank, D. Papale, A. Rammig, P. Smith, K. Thonicke, M. van der Velde, S. Vicca, A. Walz, and M. Wattenbach. Climate extremes and the carbon cycle. *Nature*, 500(7462):287–295, August 2013. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12350. URL `http://www.nature.com/articles/nature12350`.

M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, February 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0912-1.

M. Renner, C. Brenner, K. Mallick, H.-D. Wizemann, L. Conte, I. Trebs, J. Wei, V. Wulfmeyer, K. Schulz, and A. Kleidon. Using phase lags to evaluate model biases in simulating the diurnal cycle of evapotranspiration: A case study in Luxembourg. *Hydrology and Earth System Sciences*, 23(1):515–535, January 2019. ISSN 1027-5606. doi: https://doi.org/10.5194/hess-23-515-2019.

A. D. Richardson, B. H. Braswell, D. Y. Hollinger, P. Burman, E. A. Davidson, R. S. Evans, L. B. Flanagan, J. W. Munger, K. Savage, S. P. Urbanski, and S. C. Wofsy. Comparing simple respiration models for eddy flux and dynamic chamber data. *Agricultural and Forest Meteorology*, 141(2):219–234, December 2006. ISSN 0168-1923. doi: 10.1016/j.agrformet.2006.10.010.

M. B. Richman. Rotation of principal components. *Journal of Climatology*, 6 (3):293–335, 1986. ISSN 1097-0088. doi: 10.1002/joc.3370060305.

W. Rickels, J. Dovern, J. Hoffmann, M. F. Quaas, J. O. Schmidt, and M. Visbeck. Indicators for monitoring sustainable development goals: An application to oceanic development in the European Union. *Earth's Future*, 4(5):252–267, 2016. ISSN 2328-4277. doi: 10.1002/2016EF000353.

W. J. Ripple, C. Wolf, T. M. Newsome, M. Galetti, M. Alamgir, E. Crist, M. I. Mahmoud, and W. F. Laurance. World Scientists' Warning to Humanity: A Second Notice. *BioScience*, 67(12):1026–1028, December 2017. ISSN 0006-3568. doi: 10.1093/biosci/bix125.

D. Rosenfeld, Y. Zhu, M. Wang, Y. Zheng, T. Goren, and S. Yu. Aerosol-driven droplet concentrations dominate coverage and water of oceanic low-level clouds. *Science*, 363(6427), February 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aav0566.

W. Rouse, R. H. Haas, J. A. Well, and D. W. Deering. Monitoring vegetation systems in the great plains with ERTS. In *Third Earth Resources Technology Satellite-1 Symposium Technical Presentations Section a*, volume I, pages 309–317, Goddart Space Flight Center, Washington, D.C., 1973. National Aeronautics and Space Administration.

P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1987. ISBN 978-0-471-85233-9.

S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.290.5500.2323.

J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, 6(1):1–10, October 2015. ISSN 2041-1723. doi: 10.1038/ncomms9502.

Y. Ryu, J. A. Berry, and D. D. Baldocchi. What is global photosynthesis? History, uncertainties and opportunities. *Remote Sensing of Environment*, 223:95–114, March 2019. ISSN 00344257. doi: 10.1016/j.rse.2019.01.016.

S. Sarmah, G. Jia, and A. Zhang. Satellite view of seasonal greenness trends and controls in South Asia. *Environmental Research Letters*, 13(3):034026, March 2018. ISSN 1748-9326. doi: 10.1088/1748-9326/aaa866.

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8. URL `http://dl.acm.org/citation.cfm?id=645527.657464`.

D. Schimel and F. D. Schneider. Flux towers in the sky: Global ecology from space. *New Phytologist*, 224(2):570–584, 2019. ISSN 1469-8137. doi: 10.1111/nph.15934.

M. D. Schwartz. Monitoring global change with phenology: The case of the spring green wave. *International Journal of Biometeorology*, 38(1):18–22, March 1994. ISSN 0020-7128, 1432-1254. doi: 10.1007/BF01241799.

M. D. Schwartz. Green-wave phenology. *Nature*, 394(6696):839, August 1998. ISSN 1476-4687. doi: 10.1038/29670.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998. ISSN 08997667. doi: 10.1162/089976698300017467.

B. Schölkopf, R. Herbrich, and A. J. Smola. A Generalized Representer Theorem. In *Computational Learning Theory*, pages 416–426. Springer-Verlag, 2001. doi: 10.1007/3-540-44581-1_27.

L. See, S. Fritz, E. Dias, E. Hendriks, B. Mijling, F. Snik, P. Stammes, F. D. Vescovi, G. Zeug, P.-P. Mathieu, Y.-L. Desnos, and M. Rast. Supporting Earth-Observation Calibration and Validation: A new generation of tools for crowdsourcing and citizen science. *IEEE Geoscience and Remote Sensing Magazine*, 4(3):38–50, September 2016. ISSN 2168-6831, 2473-2397. doi: 10.1109/MGRS.2015.2498840.

P. K. Sen. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, 63(324):1379–1389, 1968. ISSN 0162-1459. doi: 10.2307/2285891.

S. Seth and M. McGillivray. Composite indices, alternative weights, and comparison robustness. *Social Choice and Welfare*, 51(4):657–679, 2018.

R. R. Shaker. A mega-index for the Americas and its underlying sustainable development correlations. *Ecological Indicators*, 89:466–479, June 2018. ISSN 1470160X. doi: 10.1016/j.ecolind.2018.01.050.

V. D. Silva and J. B. Tenenbaum. Global Versus Local Methods in Nonlinear Dimensionality Reduction. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 721–728. MIT Press, 2003. URL `http://papers.nips.cc/paper/2141-global-versus-local-methods-in-nonlinear-dimensionality-reduction.pdf`.

S. Sippel, M. Reichstein, X. Ma, M. D. Mahecha, H. Lange, M. Flach, and D. Frank. Drought, Heat, and the Carbon Cycle: A Review. *Current Climate Change Reports*, 4(3):266–286, September 2018. ISSN 2198-6061. doi: 10.1007/s40641-018-0103-4.

S. Sitch, P. Friedlingstein, N. Gruber, S. D. Jones, G. Murray-Tortarolo, A. Ahlström, S. C. Doney, H. Graven, C. Heinze, C. Huntingford, S. Levis, P. E. Levy, M. Lomas, B. Poulter, N. Viovy, S. Zaehle, N. Zeng, A. Arneth, G. Bonan, L. Bopp, J. G. Canadell, F. Chevallier, P. Ciais, R. Ellis, M. Gloor, P. Peylin, S. L. Piao, C. Le Quéré, B. Smith, Z. Zhu, and R. Myneni. Recent trends and drivers of regional sources and sinks of carbon dioxide. *Biogeosciences*, 12(3):653–679, February 2015. ISSN 1726-4189. doi: 10.5194/bg-12-653-2015.

B. Smith, I. C. Prentice, and M. T. Sykes. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: Comparing two contrasting approaches within European climate space. *Global Ecology and Biogeography*, 10(6):621–637, 2001.

J. Smits and I. Permanyer. The Subnational Human Development Database. *Scientific Data*, 6:190038, March 2019. ISSN 2052-4463. doi: 10.1038/sdata.2019.38.

J. Smits and R. Steendijk. The International Wealth Index (IWI). *Social Indicators Research*, 122(1):65–85, May 2015. ISSN 0303-8300, 1573-0921. doi: 10.1007/s11205-014-0683-x. URL `https://link.springer.com/article/10.1007/s11205-014-0683-x`.

R. R. Sokal and F. J. Rohlf. The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2):33–40, 1962. ISSN 0040-0262. doi: 10.2307/1217208.

X.-P. Song, M. C. Hansen, S. V. Stehman, P. V. Potapov, A. Tyukavina, E. F. Vermote, and J. R. Townshend. Global land change from 1982 to 2016. *Nature*, 560(7720):639, August 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0411-9.

S. Sonnenburg, G. Raetsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl, and V. Franc. The SHOGUN Machine Learning Toolbox. *Journal of Machine Learning Research*, 11:1799–1802, 2010. ISSN 1532-4435.

C. Spearman. "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292, 1904. ISSN 0002-9556. doi: 10.2307/1412107.

T. N. Srinivasan. Human Development: A New Paradigm or Reinvention of the Wheel? *The American Economic Review*, 84(2):238–243, 1994. ISSN 0002-8282.

W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig. pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, May 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm069.

A. Stanojević and J. Benčina. The Construction of an Integrated and Transparent Index of Wellbeing. *Social Indicators Research*, 143(3):995–1015, June 2019. ISSN 1573-0921. doi: 10.1007/s11205-018-2016-y.

E. Stanton. The Human Development Index: A History. Technical Report wp127, Political Economy Research Institute, University of Massachusetts at Amherst, 2007.

W. Steffen, K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. de Vries, C. A. de Wit, C. Folke, D. Gerten, J. Heinke, G. M. Mace, L. M. Persson, V. Ramanathan, B. Reyers, and S. Sörlin. Planetary boundaries: Guiding human development on a changing planet. *Science*, 347(6223), February 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1259855.

A. R. Stine, P. Huybers, and I. Y. Fung. Changes in the phase of the annual cycle of surface temperature. *Nature*, 457(7228):435–440, January 2009. ISSN 1476-4687. doi: 10.1038/nature07675.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6): 2769–2794, December 2007. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053607000000505.

J. Tang, D. D. Baldocchi, and L. Xu. Tree photosynthesis modulates soil respiration on a diurnal time scale. *Global Change Biology*, 11(8):1298–1304, 2005. ISSN 1365-2486. doi: 10.1111/j.1365-2486.2005.00978.x.

J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500): 2319–2323, December 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.290.5500.2319.

A. P. Tewkesbury, A. J. Comber, N. J. Tate, A. Lamb, and P. F. Fisher. A critical synthesis of remotely sensed optical image change detection

techniques. *Remote Sensing of Environment*, 160:1–14, April 2015. ISSN 0034-4257. doi: 10.1016/j.rse.2015.01.006.

The World Bank. World Development Indicators (WDI) | Data Catalog, 2018a. URL https://datacatalog.worldbank.org/dataset/world-development-indicators.

The World Bank. Sustainable Development Goals (SDG) | Data Catalog, 2018b. URL https://datacatalog.worldbank.org/dataset/sustainable-development-goals.

H. Theil. A Rank-Invariant Method of Linear and Polynomial Regression Analysis, I, II, III. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, 53:386–392, 521–525, 1397–1412, 1950.

W. S. Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, 1952. ISSN 0033-3123, 1860-0980. doi: 10.1007/BF02288916.

G. Tramontana, M. Jung, C. R. Schwalm, K. Ichii, G. Camps-Valls, B. Ráduly, M. Reichstein, M. A. Arain, A. Cescatti, G. Kiely, L. Merbold, P. Serrano-Ortiz, S. Sickert, S. Wolf, and D. Papale. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences*, 13(14):4291–4313, July 2016. ISSN 1726-4170. doi: https://doi.org/10.5194/bg-13-4291-2016. URL https://www.biogeosciences.net/13/4291/2016/.

UNDP. *Human Development Report 2016 Human Development for Everyone*. Human Development Reports. United Nations Development Programme, New York, NY, 2016. ISBN 978-92-1-126413-5.

UNDP. Human Development Reports | United Nations Development Programme, 2018. URL http://hdr.undp.org/.

UNDP. *Human Development Report 2019 Beyond Income, beyond Averages, beyond Today: Inequalities in Human Development in the 21st Century*. Human Development Reports. United Nations Development Programme, New York, NY, USA, September 2019. ISBN 978-92-1-126439-5.

United Nations General Assembly. Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development, June 2017a.

United Nations General Assembly. Work of the Statistical Commission pertaining to the 2030 Agendanda for Sustainable Development, July 2017b. URL `http://ggim.un.org/meetings/2017-4th_Mtg_IAEG-SDG-NY/documents/A_RES_71_313.pdf`.

L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9:2579–2605, 2008. ISSN 1532-4435. WOS:000262637600007.

L. Van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: A comparative review. *J Mach Learn Res*, 10:66–71, 2009.

L. van der Maaten, S. Schmidtlein, and M. D. Mahecha. Analyzing floristic inventories with multiple maps. *Ecological Informatics*, 9:1–10, May 2012. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2012.01.005.

A. I. J. M. van Dijk, H. E. Beck, R. S. Crosbie, R. A. M. de Jeu, Y. Y. Liu, G. M. Podger, B. Timbal, and N. R. Viney. The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resources Research*, 49(2):1040–1057, 2013. ISSN 1944-7973. doi: 10.1002/wrcr.20123.

J. H. van't Hoff. *Chemical Dynamics*. Number I in Lectures on Theoretical and Physical Chemistry. Edward Arnold, London, 1898.

J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization. *J. Mach. Learn. Res.*, 11:451–490, 2010. ISSN 1532-4435. WOS:000277186500001.

J. Verbesselt, R. Hyndman, G. Newnham, and D. Culvenor. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, 114(1):106–115, January 2010. ISSN 0034-4257. doi: 10.1016/j.rse.2009.08.014.

P. Verhulst. Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18:14–54, 1845.

P. Verhulst. Deuxième mémoire sur la loi d'accroissement de la population. *Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique*, 20:1–32, 1847.

U. von Luxburg. A Tutorial on Spectral Clustering. Technical Report TR-149, Max Planck Institute for Biological Cybernetics, Tuebingen, 2007.

M. Wackernagel, L. Onisto, P. Bello, A. Callejas Linares, I. Susana López Falfán, J. Méndez Garcıá, A. Isabel Suárez Guerrero, and M. Guadalupe Suárez Guerrero. National natural capital accounting with the ecological footprint concept. *Ecological Economics*, 29(3):375–390, June 1999. ISSN 0921-8009. doi: 10.1016/S0921-8009(98)90063-5.

R. H. Whittaker and P. L. Marks. Methods of Assessing Terrestrial Productivty. In H. Lieth and R. H. Whittaker, editors, *Primary Productivity of the Biosphere*, Ecological Studies, pages 55–118. Springer, Berlin, Heidelberg, 1975. ISBN 978-3-642-80913-2. doi: 10.1007/978-3-642-80913-2_4.

D. S. Wilks. Chapter 12 - Principal Component (EOF) Analysis. In D. S. Wilks, editor, *International Geophysics*, volume 100 of *Statistical Methods in the Atmospheric Sciences*, pages 519–562. Academic Press, January 2011. doi: 10.1016/B978-0-12-385022-5.00012-9.

L. Wingate, J. Ogée, E. Cremonese, G. Filippa, T. Mizunuma, M. Migliavacca, C. Moisy, M. Wilkinson, C. Moureaux, G. Wohlfahrt, A. Hammerle, L. Hörtnagl, C. Gimeno, A. Porcar-Castell, M. Galvagno, T. Nakaji, J. Morison, O. Kolle, A. Knohl, W. Kutsch, P. Kolari, E. Nikinmaa, A. Ibrom, B. Gielen, W. Eugster, M. Balzarolo, D. Papale, K. Klumpp, B. Köstner, T. Grünwald, R. Joffre, J.-M. Ourcival, M. Hellstrom, A. Lindroth, C. George, B. Longdoz, B. Genty, J. Levula, B. Heinesch, M. Sprintsin, D. Yakir, T. Manise, D. Guyon, H. Ahrends, A. Plaza-Aguilar, J. H. Guan, and J. Grace. Interpreting canopy development and physiology using a European phenology camera network at flux sites. *Biogeosciences*, 12(20):5995–6015, October 2015. ISSN 1726-4170. doi: https://doi.org/10.5194/bg-12-5995-2015.

K. Wolter and M. Timlin. *Monitoring ENSO in COADS with a Seasonally Adjusted Principal Component Index*. NOAA/NMC/CAC, NSSL, Oklahoma Clim. Survey, CIMMS and the School of Meteor., Univ. of Oklahoma, Norman, OK, January 1993.

K. Wolter and M. S. Timlin. El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). *International Journal of Climatology*, 31(7):1074–1087, June 2011a. ISSN 1097-0088. doi: 10.1002/joc.2336.

K. Wolter and M. S. Timlin. El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). *International Journal of Climatology*, 31(7):1074–1087, June 2011b. ISSN 1097-0088. doi: 10.1002/joc.2336.

T. Yan, H. Song, Z. Wang, M. Teramoto, J. Wang, N. Liang, C. Ma, Z. Sun, Y. Xi, L. Li, and S. Peng. Temperature sensitivity of soil respiration across multiple time scales in a temperate plantation forest. *Science of The Total Environment*, 688:479–485, October 2019. ISSN 0048-9697. doi: 10.1016/j.scitotenv.2019.06.318.

G. Yetman, S. Gaffin, and D. Balk. ISLSCP II global gridded gross domestic product (GDP), 1990. 2010. doi: 10.3334/ORNLDAAC/974.

A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. Strucchange: An R Package for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software*, 7(1):1–38, January 2002. ISSN 1548-7660. doi: 10.18637/jss.v007.i02.

N. Zeng, F. Zhao, G. J. Collatz, E. Kalnay, R. J. Salawitch, T. O. West, and L. Guanter. Agricultural Green Revolution as a driver of increasing atmospheric CO_2 seasonal amplitude. *Nature*, 515(7527):394–397, November 2014. ISSN 1476-4687. doi: 10.1038/nature13893.

Q. Zhang, R. P. Phillips, S. Manzoni, R. L. Scott, A. C. Oishi, A. Finzi, E. Daly, R. Vargas, and K. A. Novick. Changes in photosynthesis and soil moisture drive the seasonal soil respiration-temperature hysteresis relationship. *Agricultural and Forest Meteorology*, 259:184–195, September 2018. ISSN 0168-1923. doi: 10.1016/j.agrformet.2018.05.005.

B. Zhou, L. Gu, Y. Ding, L. Shao, Z. Wu, X. Yang, C. Li, Z. Li, X. Wang, Y. Cao, B. Zeng, M. Yu, M. Wang, S. Wang, H. Sun, A. Duan, Y. An, X. Wang, and W. Kong. The Great 2008 Chinese Ice Storm: Its Socioeconomic–Ecological Impact and Sustainability Lessons Learned. *Bulletin of the American Meteorological Society*, 92(1):47–60, January 2011. ISSN 00030007. doi: 10.1175/2010BAMS2857.1.

L. Zhou, Y. Tian, R. B. Myneni, P. Ciais, S. Saatchi, Y. Y. Liu, S. Piao, H. Chen, E. F. Vermote, C. Song, and T. Hwang. Widespread decline of Congo rainforest greenness in the past decade. *Nature*, 509(7498):86–90, May 2014. ISSN 1476-4687. doi: 10.1038/nature13265.

Z. Zhu, S. Piao, R. B. Myneni, M. Huang, Z. Zeng, J. G. Canadell, P. Ciais, S. Sitch, P. Friedlingstein, A. Arneth, C. Cao, L. Cheng, E. Kato, C. Koven,

Y. Li, X. Lian, Y. Liu, R. Liu, J. Mao, Y. Pan, S. Peng, J. Peñuelas, B. Poulter, T. A. M. Pugh, B. D. Stocker, N. Viovy, X. Wang, Y. Wang, Z. Xiao, H. Yang, S. Zaehle, and N. Zeng. Greening of the Earth and its drivers. *Nature Climate Change*, 6(8):791–795, August 2016. ISSN 1758-6798. doi: 10.1038/nclimate3004.

Z. Zhu. Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:370–384, August 2017. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2017.06.013.

J. Zscheischler, M. Reichstein, S. Harmeling, A. Rammig, E. Tomelleri, and M. D. Mahecha. Extreme events in gross primary production: A characterization across continents. *Biogeosciences*, 11(11):2909–2924, June 2014. ISSN 1726-4189. doi: 10.5194/bg-11-2909-2014.

Appendix A

Supporting Information Chapter 3

Description of Variables

Variables used describing the biosphere can be found in tab. 3.1. Here we provide a more complete description of all variables:

Black-sky albedo is the reflected fraction of total incoming radiation under direct hemispherical reflectance, i.e. direct illumination (Muller et al., 2011). This dataset is the broadband surface albedo including the visible, the near-infrared, and the shortwave-infrared spectrum (400–3000nm). It is derived from the SPOT4-VEGETATION, SPOT5-VEGETATION2, and the MERIS satellite sensors.

White-sky albedo is the reflected fraction of total incoming radiation under bihemispherical reflectance, i.e. diffuse illumination (Muller et al., 2011). Together with black-sky albedo it can be used to estimate the albedo under different illumination conditions. This dataset is the broadband surface albedo including the visible, the near, and the shortwave-infrared spectrum (400–3000nm). This dataset is derived from the SPOT4-VEGETATION, SPOT5-VEGETATION2, and the MERIS satellite sensors.

Evaporation $[mm/day]$ is the amount of water evaporated per day, depending on the amount of available water and energy. This dataset is based on the GLEAMv3 model (Martens et al., 2017), using satellite data from ESA CCI and SMOS to derive a number of variables.

Evaporative stress is modeled water stress for plants. 0 means that the vegetation has no water available for transpiration and 1 means that transpiration equals potential transpiration. This dataset is based on the GLEAMv3 model (Martens et al., 2017), using satellite data from ESA CCI and SMOS to derive a number of variables.

fAPAR is the fraction of absorbed photosynthetically active radiation, a proxy for plant productivity (Disney et al., 2016). This dataset is based on the GlobAlbedo dataset (`http://globalbedo.org`) and the MODIS fAPAR and leaf area index (LAI) products.

Gross primary productivity (GPP) $[gCm^{-2}day^{-1}]$ is the total amount

of carbon fixed by photosynthesis (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Terrestrial ecosystem respiration (TER) $[gCm^{-2}day^{-1}]$ the total amount of carbon respired by the ecosystem, including autotrophic and heterotrophic respiration (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Net ecosystem exchange (NEE) $[gCm^{-2}day^{-1}]$ is the total exchange of carbon of the ecosystem with the atmosphere NEE $=$ GPP $-$ TER (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Latent energy (LE) $[Wm^{-2}]$ is the amount of energy lost by the surface due to evaporation (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Sensible heat (H) $[Wm^{-2}]$ is the amount of energy lost by the surface due to radiation (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Root-zone soil moisture $[m^3m^{-3}]$ is the moisture content of the root zone. This dataset is based on the GLEAMv3 model (Martens et al., 2017), using satellite data from ESA CCI and SMOS to derive a number of variables.

Surface soil moisture $[mm^3mm^{-3}]$ the soil moisture content at the soil surface. This dataset is based on the GLEAMv3 model (Martens et al., 2017), using satellite data from ESA CCI and SMOS to derive a number of variables. Variables used describing the biosphere can be found in tab. 3.1, here we provide a more complete description of all variables:

Time–Space Patterns of Components 1–3

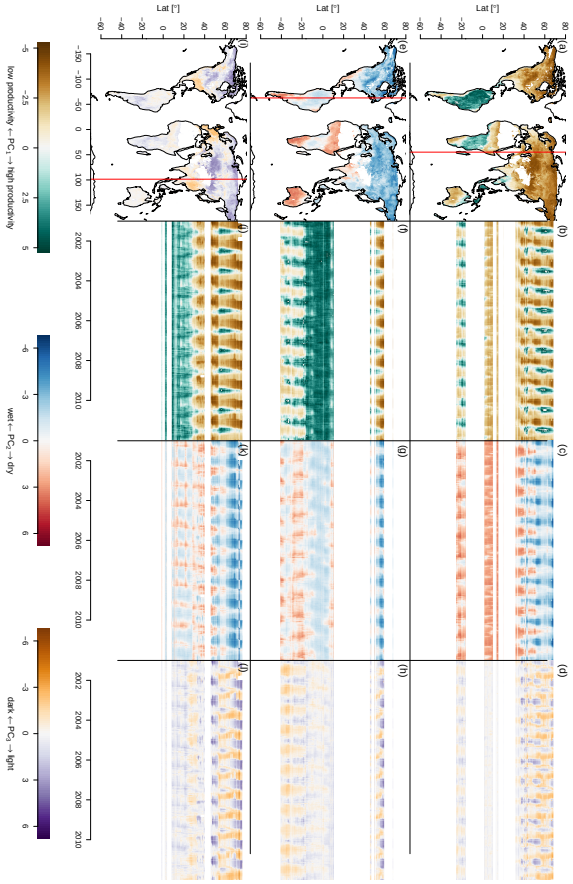

Figure A.1: Time and space patterns of PC_1–PC_3, where the cut points are the same as in fig. 3.8. The brown–green contrast shows the state of PC_1, from low to high productivity. The blue–red contrast shows the state of PC_2, from cold to dry. The brown–purple contrast shows the state of PC_3, from dark to light. Panels (a), (e), and (i) are maps showing the state of PC_1–PC_3, respectively, on the 1 January 2001. (b), (c), and (d) show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in (a). (f), (g), and (h) show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in (e). (j), (k), and (l) show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in (i).

Mean Seasonal Cycle Extrema

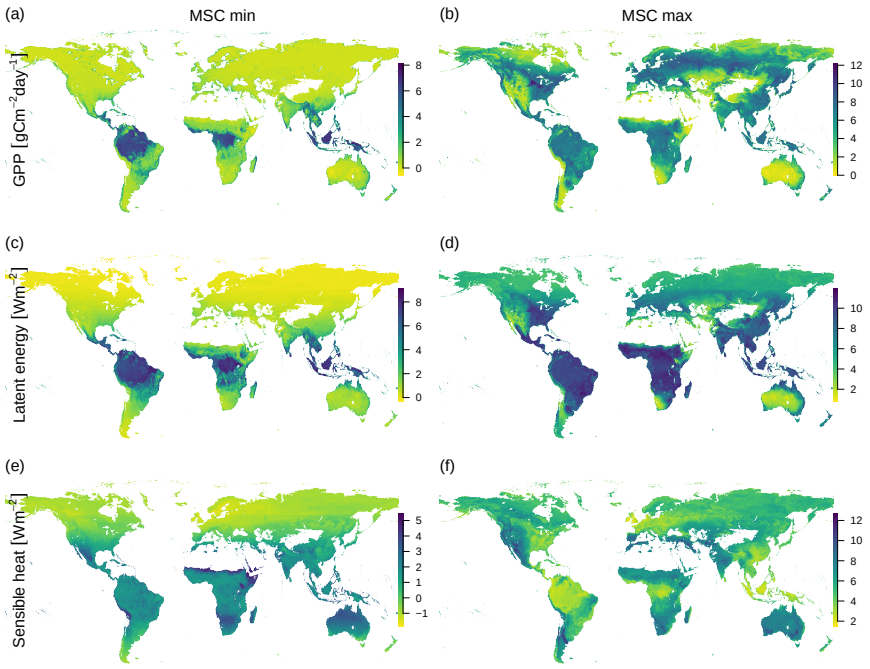

Figure A.2: The minimum (a, c, e) and maximum (b, d, f) mean seasonal cycles of GPP (a, b), latent heat (c, d), and sensible heat (e, f). This illustrates the similarity of possibly very different ecosystems in terms of productivity and limitations. During peak growing season, many midlatitude areas have a similar productivity and latent energy release as tropical rain forests (b, d). The highest maximum seasonal sensible heat loss can be found in dry areas around the world and is lowest in areas with a wet climate such as tropical rain forests and maritime climates (f).

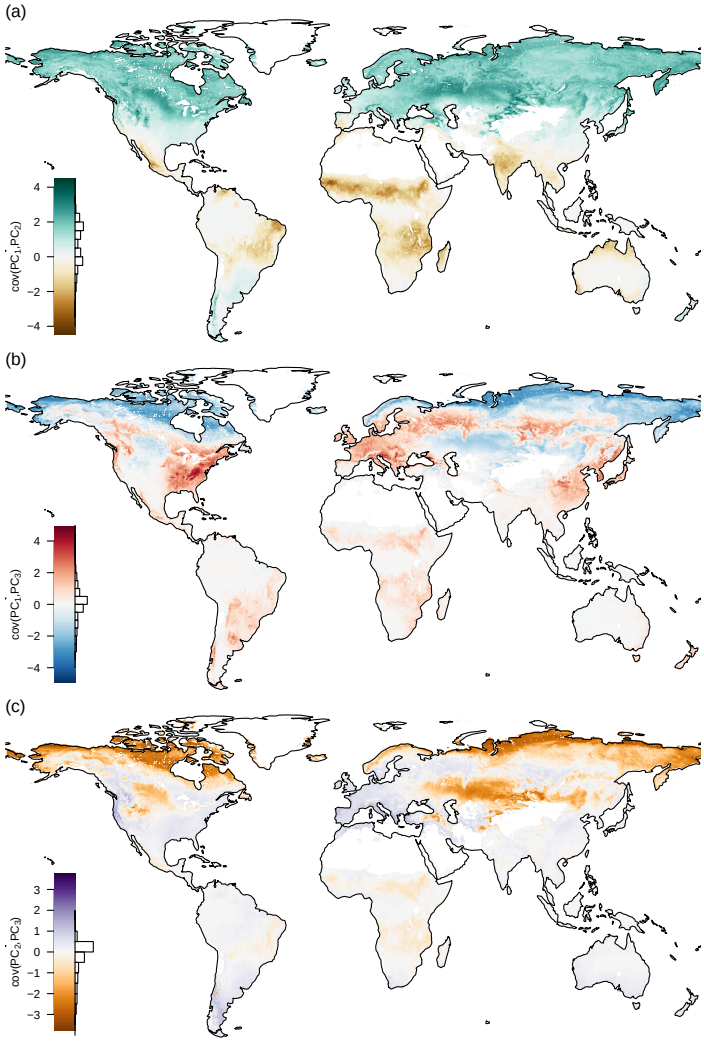

Figure A.3: Pairwise covariances of the first three principal components mean seasonal cycles by space. (a) $\mathrm{cov}(\mathrm{PC}_1, \mathrm{PC}_2)$, (b) $\mathrm{cov}(\mathrm{PC}_1, \mathrm{PC}_3)$, and (c) $\mathrm{cov}(\mathrm{PC}_2, \mathrm{PC}_3)$. The bar charts show the distribution of the covariances. It can be seen that although two principal components are globally uncorrelated by their way of construction, they covary locally.

Changes in the Seasonal Amplitude

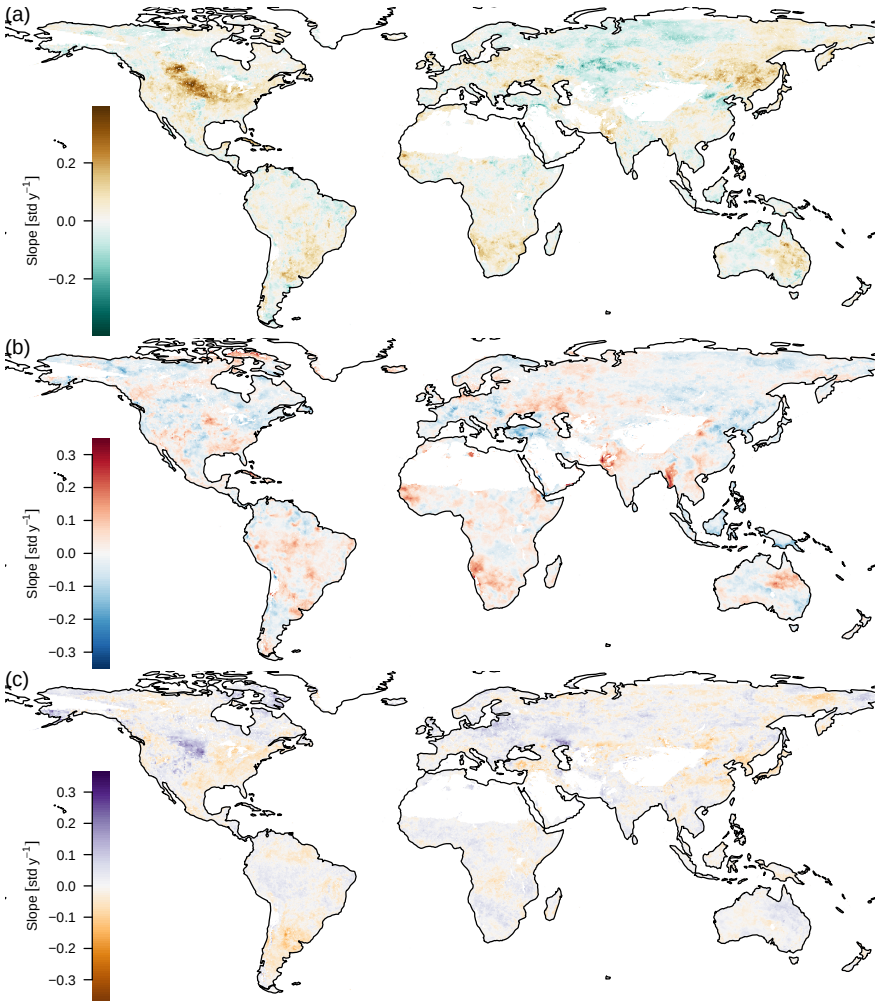

Figure A.4: Trends in the amplitude of the yearly cycle, 2001–2011. Only Theil–Sen estimators for significant slopes ($p < 0.05$, *unadjusted*) are shown. Because there is only a single amplitude per year and therefore only 11 data points per time series, the Benjamini–Hochberg adjusted p values are not significant.

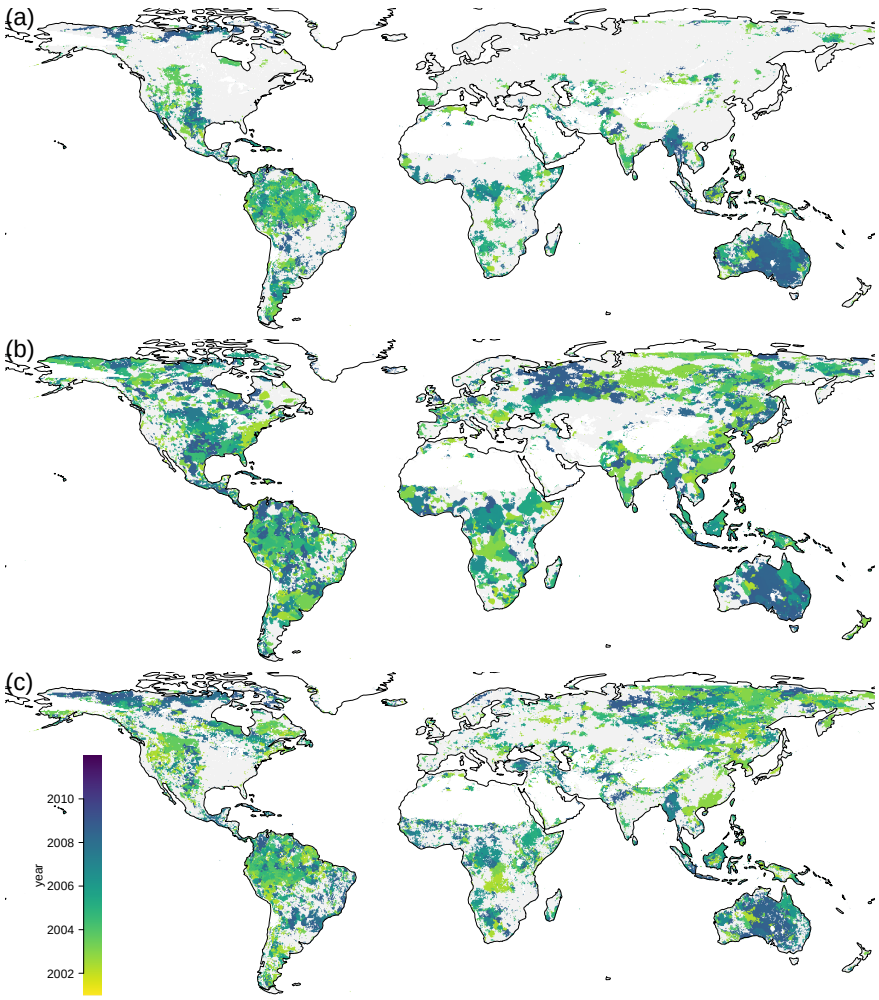

Figure A.5: Breakpoint detection, (a) on PC_1, (b) on PC_2, and (c) on PC_3. The color indicates the year of the biggest breakpoint if a significant breakpoint was found, with gray if there was no significant breakpoint found.

As the environmental conditions change, due to climate change and hu-

man intervention, the local ecosystems may change gradually or abruptly. Detecting these changes is very important for monitoring the impact of climate change and land use change on the ecosystems. We applied breakpoint detection to the trajectories (fig. A.5).

Breakpoints on the first component were found in the entire Amazon, and the largest breakpoint is dated to the year 2005 during the large drought event. The entire eastern part of Australia shows its largest breakpoint towards the end of the time series because of a La Niña event, which caused lower temperatures and higher rainfall than usual during the years 2010 and 2011.

Appendix B

Supporting Information Chapter 4

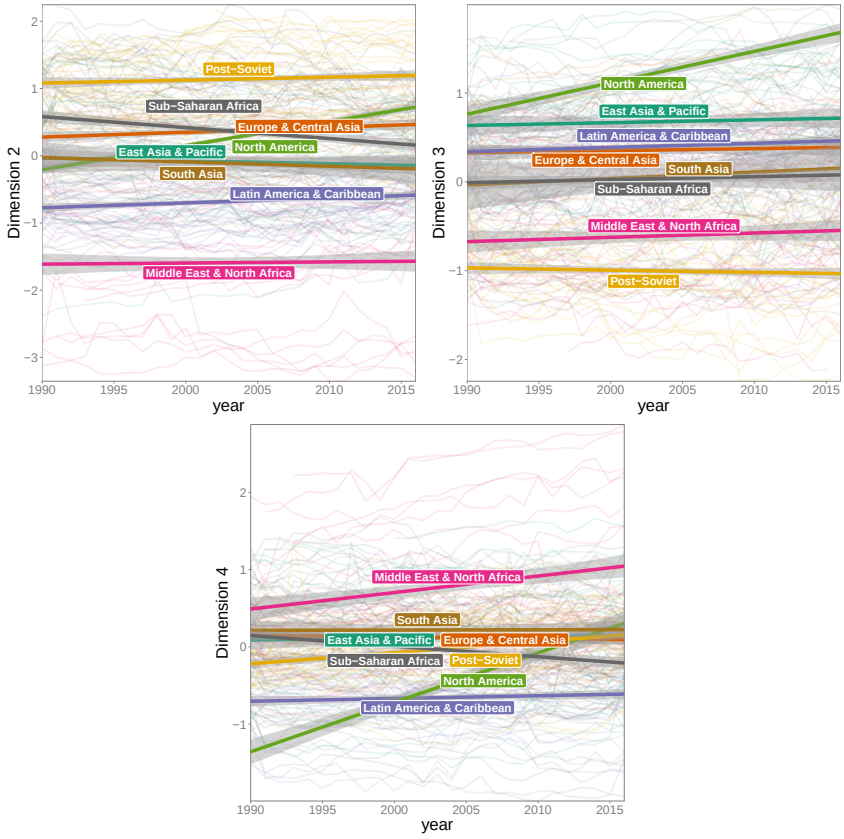

Figure B.1: Trends over time for Isomap dimensions 3–5. Compared to the first e-Isomap component, there are no strong trends observable.

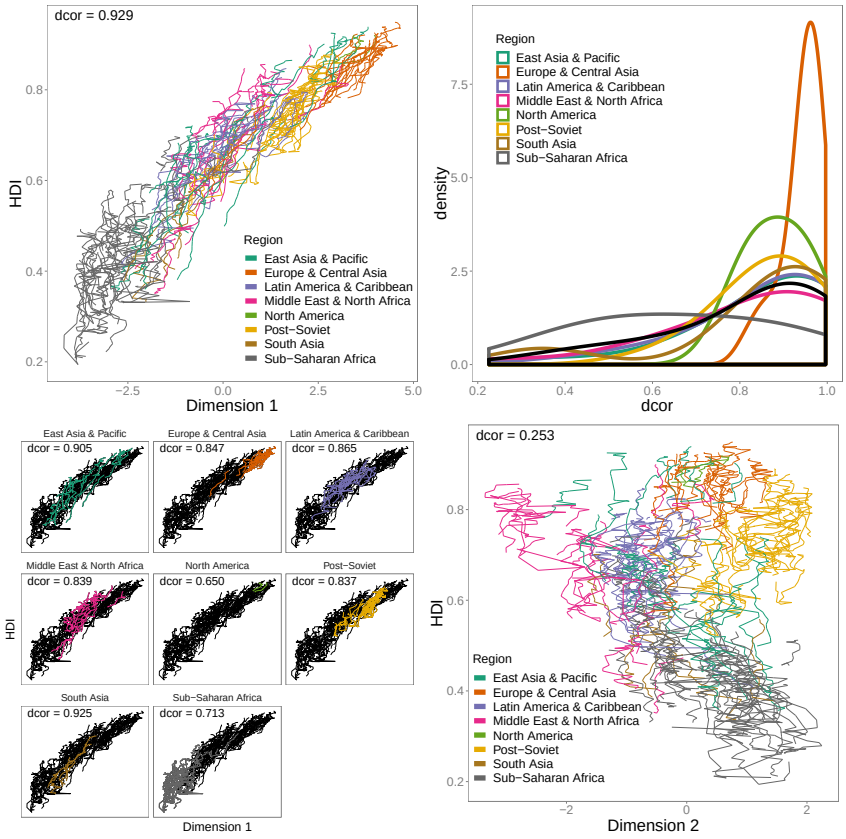

Figure B.2: Correlation between HDI and the data driven indicators. Top left: The HDI is strongly related to the first dimension, dcor over all show data points. Top right: Distribution of distance correlations HDI and dimension 1 of single trajectories, separated by regions, "Sub-Saharan Africa" has lower correlations, black is the distribution over all regions jointly. Bottom left: Relationships between all trajectories per region dimension 1 and HDI, the correlation is lower for "Sub-Saharan Africa", dcors over all colored data points of a region. Bottom right: Relationship between dimension 2 and the HDI, the correlation is much lower, dcor over all shown data points.

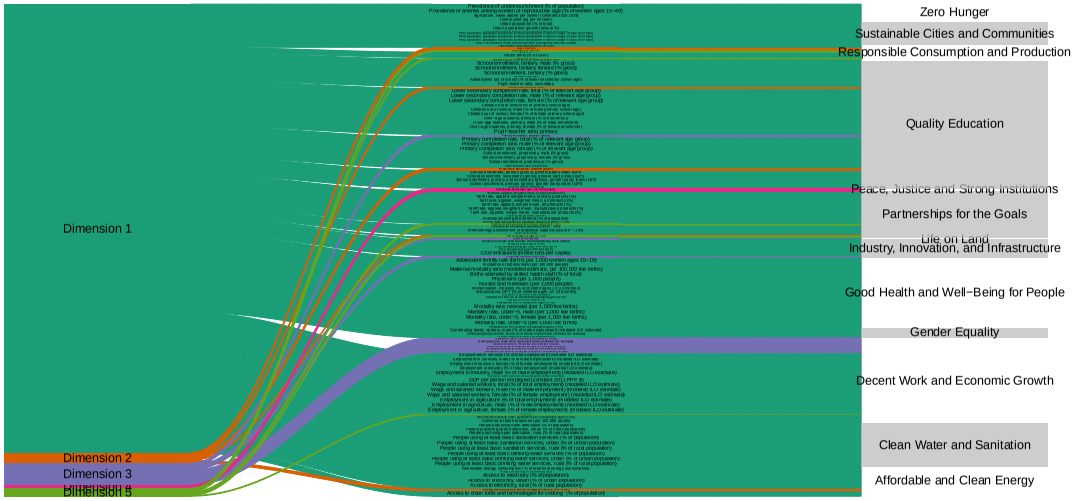

Figure B.3: Showing the importance of the dimensions for the SDGs, color code by dimension (left), WDIs (center) are connected to their corresponding SDG (right) and the dimension they have maximum distance correlation with (left). The thickness of the connection reflects the distance correlation strength with the dimension.

Table B.1: The main difference between classical indicators and data driven indicators is that the classical indicators highlight a single aspect of the system, whereas the data driven indicators try to represent as much of the information content of the data as possible. This makes classical indicators easier to interpret and to communicate, but limits their ability to faithfully represent the system in its entirety.

	Classical Indicators	**Data Driven Indicators**
Interpretability	Easy interpretation but requires ad hoc assumptions	Aspects are generated from data
Rankings	Simple, risk of oversimplifying, rankings may not be meaningful	Rankings are complicated if more than one axis is involved, lower risk of oversimplification
Aspects	Based on variables chosen by the creator	Aspects emerge from data, multiple aspects may emerge
Method	Hand crafted, infinite degrees of freedom, arbitrary	Choose the right method, parameter tuning
Political appeal	High, depending of the topic	Probably more difficult
Faithfulness in representing the data	Low overall representativeness/single aspects may be represented more faithfully	High, especially if more than one dimension is used

Appendix C

Article: dimRed and coRanking — Unifying Dimensionality Reduction in R

Kraemer, G., Reichstein, M., and Mahecha, M. D. (2018). dimRed and coRanking—Unifying Dimensionality Reduction in R. *The R Journal*, 10(1), 342–358. doi:10.32614/RJ-2018-039

This article was written to create a framework to easily compare different methods of dimensionality reduction and their ability to be used as indicators.

The article was published in the *R Journal* which has an impact factor of 1.3 and a 5 year impact factor of 2.5 and occupies a relative position of 41/123 in the category of "Statistics & Probability" and 76/105 in the category of "Computer Science, Interdisciplinary Applications" in the ISI Web of Knowledge database.

dimRed and coRanking—Unifying Dimensionality Reduction in R

by Guido Kraemer, Markus Reichstein, and Miguel D. Mahecha

Abstract "Dimensionality reduction" (DR) is a widely used approach to find low dimensional and interpretable representations of data that are natively embedded in high-dimensional spaces. DR can be realized by a plethora of methods with different properties, objectives, and, hence, (dis)advantages. The resulting low-dimensional data embeddings are often difficult to compare with objective criteria. Here, we introduce the **dimRed** and **coRanking** packages for the R language. These open source software packages enable users to easily access multiple classical and advanced DR methods using a common interface. The packages also provide quality indicators for the embeddings and easy visualization of high dimensional data. The **coRanking** package provides the functionality for assessing DR methods in the co-ranking matrix framework. In tandem, these packages allow for uncovering complex structures high dimensional data. Currently 15 DR methods are available in the package, some of which were not previously available to R users. Here, we outline the **dimRed** and **coRanking** packages and make the implemented methods understandable to the interested reader.

Introduction

Dimensionality Reduction (DR) essentially aims to find low dimensional representations of data while preserving their key properties. Many methods exist in literature, optimizing different criteria: maximizing the variance or the statistical independence of the projected data, minimizing the reconstruction error under different constraints, or optimizing for different error metrics, just to name a few. Choosing an inadequate method may imply that much of the underlying structure remains undiscovered. Often the structures of interest in a data set can be well represented by fewer dimensions than exist in the original data. Data compression of this kind has the additional benefit of making the encoded information better conceivable to our brains for further analysis tasks like classification or regression problems.

For example, the morphology of a plant's leaves, stems, and seeds reflect the environmental conditions the species usually grow in (e.g., plants with large soft leaves will never grow in a desert but might have an advantage in a humid and shadowy environment). Because the morphology of the entire plant depends on the environment, many morphological combinations will never occur in nature and the morphological space of all plant species is tightly constrained. Díaz et al. (2016) found that out of six observed morphological characteristics only two embedding dimensions were enough to represent three quarters of the totally observed variability.

DR is a widely used approach for the detection of structure in multivariate data, and has applications in a variety of fields. In climatology, DR is used to find the modes of some phenomenon, e.g., the first Empirical Orthogonal Function of monthly mean sea surface temperature of a given region over the Pacific is often linked to the El Niño Southern Oscillation or ENSO (e.g., Hsieh, 2004). In ecology the comparison of sites with different species abundances is a classical multivariate problem: each observed species adds an extra dimension, and because species are often bound to certain habitats, there is a lot of redundant information. Using DR is a popular technique to represent the sites in few dimensions, e.g., Aart (1972) matches wolfspider communities to habitat and Morrall (1974) match soil fungi data to soil types. (In ecology the general name for DR is ordination or indirect gradient analysis.) Today, hyperspectral satellite imagery collects so many bands that it is very difficult to analyze and interpret the data directly. Resuming the data into a set of few, yet independent, components is one way to reduce complexity (e.g., see Laparra et al., 2015). DR can also be used to visualize the interiors of deep neural networks (e.g., see Han et al., 2017), where the high dimensionality comes from the large number of weights used in a neural network and convergence can be visualized by means of DR. We could find many more example applications here but this is not the main focus of this publication.

The difficulty in applying DR is that each DR method is designed to maintain certain aspects of the original data and therefore may be appropriate for one task and inappropriate for another. Most methods also have parameters to tune and follow different assumptions. The quality of the outcome may strongly depend on their tuning, which adds additional complexity. DR methods can be modeled after physical models with attracting and repelling forces (Force Directed Methods), projections onto low dimensional planes (PCA, ICA), divergence of statistical distributions (SNE family), or the reconstruction of local spaces or points by their neighbors (LLE).

As an example for how changing internal parameters of a method can have a great impact, the breakthrough for Stochastic Neighborhood Embedding (SNE) methods came when a Student's *t*-

distribution was used instead of a normal distribution to model probabilities in low dimensional space to avoid the "crowding problem", that is, a sphere in high dimensional space has a much larger volume than in low dimensional space and may contain too many points to be represented accurately in few dimensions. The t-distribution, allows medium distances to be accurately represented in few dimensions by larger distances due to its heavier tails. The result is called in t-SNE and is especially good at preserving local structures in very few dimensions, this feature made t-SNE useful for a wide array of data visualization tasks and the method became much more popular than standard SNE (around six times more citations of van der Maaten and Hinton (2008) compared to Hinton and Roweis (2003) in Scopus (Elsevier, 2017)).

There are a number of software packages for other languages providing collections of methods: In Python there is scikit-learn (Pedregosa et al., 2011), which contains a module for DR. In Julia we currently find ManifoldLearning.jl for nonlinear and MultivariateStats.jl for linear DR methods. There are several toolboxes for DR implemented in Matlab (Van Der Maaten et al., 2009; Arenas-Garcia et al., 2013). The Shogun toolbox (Sonnenburg et al., 2017) implements a variety of methods for dimensionality reduction in C++ and offers bindings for a many common high level languages (including R, but the installation is anything but simple, as there is no CRAN package). However, there is no comprehensive package for R and none of the former mentioned software packages provides means to consistently compare the quality of different methods for DR.

For many applications it can be difficult to objectively find the right method or parameterization for the DR task. This paper presents the **dimRed** and **coRanking** packages for the popular programming language R. Together, they provide a standardized interface to various dimensionality reduction methods and quality metrics for embeddings. They are implemented using the S4 class system of R, making the packages both easy to use and to extend.

The design goal for these packages is to enable researchers, who may not necessarily be experts in DR, to apply the methods in their own work and to objectively identify the most suitable methods for their data. This paper provides an overview of the methods collected in the packages and contains examples as to how to use the packages.

The notation in this paper will be as follows: $X = [x_i]_{1\leq i\leq n}^T \in \mathbb{R}^{n\times p}$, and the observations $x_i \in \mathbb{R}^p$. These observations may be transformed prior to the dimensionality reduction step (e.g., centering and/or standardization) resulting in $X' = [x_i']_{1\leq i\leq n}^T \in \mathbb{R}^{n\times p}$. A DR method then embeds each vector in X' onto a vector in $Y = [y_i]_{1\leq i\leq n}^T \in \mathbb{R}^{n\times q}$ with $y_i \in \mathbb{R}^q$, ideally with $q \ll p$. Some methods provide an explicit mapping $f(x_i') = y_i$. Some even offer a inverse mapping $f^{-1}(y_i) = \hat{x}_i'$, such that one can reconstruct a (usually approximate) sample from the low-dimensional representation. For some methods, pairwise distances between points are needed, we set $d_{ij} = d(x_i, x_j)$ and $\hat{d}_{ij} = d(y_i, y_j)$, where d is some appropriate distance function.

When referring to `functions` in the **dimRed** package or base R simply the function name is mentioned, functions from other packages are referenced with their namespace, as with `package::function`.

Dimensionality Reduction Methods

In the following section we do not aim for an exhaustive explanation to every method in **dimRed** but rather to provide a general idea on how the methods work. An overview and classification of the most commonly used DR methods can be found in Figure 1.

In all methods, parameters have to be optimized or decisions have to be made, even if it is just about the preprocessing steps of data. The **dimRed** package tries to make the optimization process for parameters as easy as possible, but, if possible, the parameter space should be narrowed down using prior knowledge. Often decisions can be made based on theoretical knowledge. For example, sometimes an analysis requires data to be kept in their original scales and sometimes this is exactly what has to be avoided as when comparing different physical units. Sometimes decisions based on the experience of others can be made, e.g., the Gaussian kernel is probably the most universal kernel and therefore should be tested first if there is a choice.

All methods presented here have the embedding dimensionality, q, as a parameter (or `ndim` as a parameter for `embed`). For methods based on eigenvector decomposition, the result generally does not depend on the number of dimensions, i.e., the first dimension will be the same, no matter if we decide to calculate only two dimensions or more. If more dimensions are added, more information is maintained, the first dimension is the most important and higher dimensions are successively less important. This means, that a method based on eigenvalue decomposition only has to be run once if one wishes to compare the embedding in different dimensions. In optimization based methods this is generally not the case, the number of dimensions has to be chosen a priori, an embedding of 2 and 3 dimensions may vary significantly, and there is no ordered importance of dimensions. This means that comparing dimensions of optimization-based methods is computationally much more expensive.

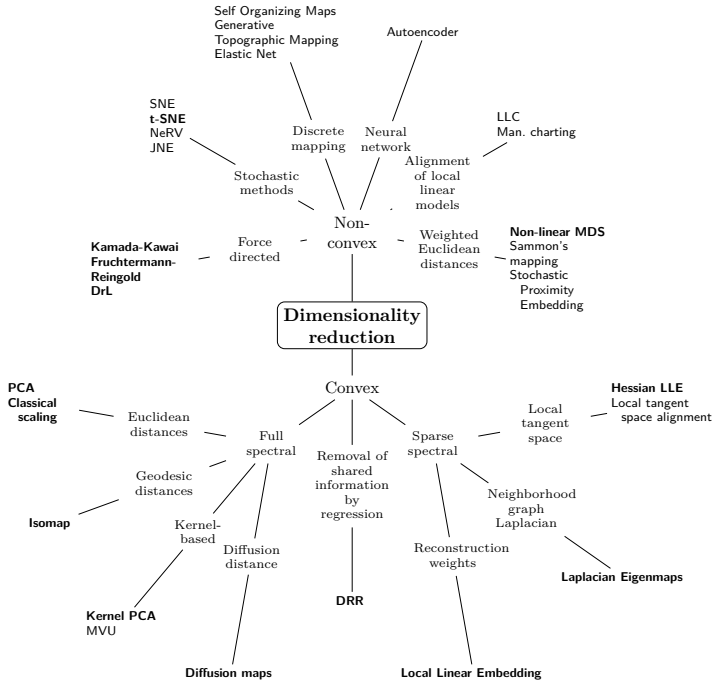

Figure 1: Classification of dimensionality reduction methods. Methods in bold face are implemented in **dimRed**. Modified from Van Der Maaten et al. (2009).

We try to give the computational complexity of the methods. Because of the actual implementation, computation times may differ largely. R is an interpreted language, so all parts of an algorithm that are implemented in R often will tend to be slow compared to methods that call efficient implementations in a compiled language. Methods where most of the computing time is spent for eigenvalue decomposition do have very efficient implementations as R uses optimized linear algebra libraries. Although, eigenvalue decomposition itself does not scale very well in naive implementations ($\mathcal{O}(n^3)$).

PCA

Principal Component Analysis (PCA) is the most basic technique for reducing dimensions. It dates back to Pearson (1901). PCA finds a linear projection (U) of the high dimensional space into a low dimensional space $Y = XU$, maintaining maximum variance of the data. It is based on solving the following eigenvalue problem:

$$(C_{XX} - \lambda_k I)u_k = 0 \tag{1}$$

where $C_{XX} = \frac{1}{n}X^T X$ is the covariance matrix, λ_k and u_k are the k-th eigenvalue and eigenvector, and I is the identity matrix. The equation has several solutions for different values of λ_k (leaving aside the trivial solution $u_k = 0$). PCA can be efficiently applied to large data sets, because it computationally scales as $\mathcal{O}(np^2 + p^3)$, that is, it scales linearly with the number of samples and R uses specialized linear algebra libraries for such kind of computations.

PCA is a rotation around the origin and there exist a forward and inverse mapping. PCA may suffer from a scale problem, i.e., when one variable dominates the variance simply because it is in a higher scale, to remedy this, the data can be scaled to zero mean and unit variance, depending on the use case, if this is necessary or desired.

Base R implements PCA in the functions `prcomp` and `princomp`; but several other implementations exist i.e., **pcaMethods** from Bioconductor which implements versions of PCA that can deal with missing data. The **dimRed** package wraps `prcomp`.

kPCA

Kernel Principal Component Analysis (kPCA) extends PCA to deal with nonlinear dependencies among variables. The idea behind kPCA is to map the data into a high dimensional space using a possibly non-linear function ϕ and then to perform a PCA in this high dimensional space. Some mathematical tricks are used for efficient computation.

If the columns of X are centered around 0, then the principal components can also be computed from the inner product matrix $K = X^T X$. Due to this way of calculating a PCA, we do not need to explicitly map all points into the high dimensional space and do the calculations there, it is enough to obtain the inner product matrix or kernel matrix $K \in \mathbb{R}^{n \times n}$ of the mapped points (Schölkopf et al., 1998).

Here is an example calculating the kernel matrix using a Gaussian kernel:

$$K = \phi(x_i)^T \phi(x_j) = \kappa(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \tag{2}$$

where σ is a length scale parameter accounting for the width of the kernel. The other trick used is known as the "representers theorem." The interested reader is referred to Schölkopf et al. (2001).

The kPCA method is very flexible and there exist many kernels for special purposes. The most common kernel function is the Gaussian kernel (Equation 2). The flexibility comes at the price that the method has to be finely tuned for the data set because some parameter combinations are simply unsuitable for certain data. The method is not suitable for very large data sets, because memory scales with $\mathcal{O}(n^2)$ and computation time with $\mathcal{O}(n^3)$.

Diffusion Maps, Isomap, Locally Linear Embedding, and some other techniques can be seen as special cases of kPCA. In which case, an out-of-sample extension using the Nyström formula can be applied (Bengio et al., 2004). This can also yield applications for bigger data, where an embedding is trained with a sub-sample of all data and then the data is embedded using the Nyström formula.

Kernel PCA in R is implemented in the **kernlab** package using the function `kernlab::kpca`, and supports a number of kernels and user defined functions. For details see the help page for `kernlab::kpca`.

The **dimRed** package wraps `kernlab::kpca` but additionally provides forward and inverse methods (Bakir et al., 2004) which can be used to fit out-of sample data or to visualize the transformation of

the data space.

Classical Scaling

What today is called Classical Scaling was first introduced by Torgerson (1952). It uses an eigenvalue decomposition of a transformed distance matrix to find an embedding that maintains the distances of the distance matrix. The method works because of the same reason that kPCA works, i.e., classical scaling can be seen as a kPCA with kernel $x^T y$. A matrix of Euclidean distances can be transformed into an inner product matrix by some simple transformations and therefore yields the same result as a PCA. Classical scaling is conceptually more general than PCA in that arbitrary distance matrices can be used, i.e., the method does not even need the original coordinates, just a distance matrix D. Then it tries to find and embedding Y so that $\hat{d}_{ij}$ is as similar to d_{ij} as possible.

The disadvantage is that is computationally much more demanding, i.e., an eigenvalue decomposition of a $n \times n$ matrix has to be computed. This step requires $\mathcal{O}(n^2)$ memory and $\mathcal{O}(n^3)$ computation time, while PCA requires only the eigenvalue decomposition of a $d \times d$ matrix and usually $n \gg d$. R implements classical scaling in the `cmdscale` function.

The **dimRed** package wraps `cmdscale` and allows the specification of arbitrary distance functions for calculating the distance matrix. Additionally a forward method is implemented.

Isomap

As Classical Scaling can deal with arbitrarily defined distances, Tenenbaum et al. (2000) suggested to approximate the structure of the manifold by using geodesic distances. In practice, a graph is created by either keeping only the connections between every point and its k nearest neighbors to produce a k-nearest neighbor graph (k-NNG), or simply by keeping all distances smaller than a value ε producing an ε-neighborhood graph (ε-NNG). Geodesic distances are obtained by recording the distance on the graph and classical scaling is used to find an embedding in fewer dimensions. This leads to an "unfolding" of possibly convoluted structures (see Figure 3).

Isomap's computational cost is dominated by the eigenvalue decomposition and therefore scales with $\mathcal{O}(n^3)$. Other related techniques can use more efficient algorithms because the distance matrix becomes sparse due to a different preprocessing.

In R, Isomap is implemented in the **vegan** package. `vegan::isomap` calculates an Isomap embedding and `vegan::isomapdist` calculates a geodesic distance matrix. The **dimRed** package uses its own implementation. This implementation is faster mainly due to using a KD-tree for the nearest neighbor search (from the **RANN** package) and to a faster implementation for the shortest path search in the k-NNG (from the **igraph** package). The implementation in **dimRed** also includes a forward method that can be used to train the embedding on a subset of data points and then use these points to approximate an embedding for the remaining points. This technique is generally referred to as landmark Isomap (De Silva and Tenenbaum, 2004).

Locally Linear Embedding

Points that lie on a manifold in a high dimensional space can be reconstructed through linear combinations of their neighborhoods if the manifold is well sampled and the neighbohoods lie on a locally linear patch. These reconstruction weights, W, are the same in the high dimensional space as the internal coordinates of the manifold. Locally Linear Embedding (LLE; Roweis and Saul, 2000) is a technique that constructs a weight matrix $W \in \mathbb{R}^{n \times n}$ with elements w_{ij} so that

$$\sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{n} w_{ij} x_j \right\|^2 \tag{3}$$

is minimized under the constraint that $w_{ij} = 0$ if x_j does not belong to the neighborhood and the constraint that $\sum_{j=1}^{n} w_{ij} = 1$. Finally the embedding is made in such a way that the following cost function is minimized for Y,

$$\sum_{i=1}^{n} \left\| y_i - \sum_{j=1}^{n} w_{ij} y_j \right\|^2. \tag{4}$$

This can be solved using an eigenvalue decomposition.

Conceptually the method is similar to Isomap but it is computationally much nicer because the weight matrix is sparse and there exist efficient solvers. In R, LLE is implemented by the package **lle**, the embedding can be calculated with `lle::lle`. Unfortunately the implementation does not make

use of the sparsity of the weight matrix W. The manifold must be well sampled and the neighborhood size must be chosen appropriately for LLE to give good results.

Laplacian Eigenmaps

Laplacian Eigenmaps were originally developed under the name spectral clustering to separate non-convex clusters. Later it was also used for graph embedding and DR (Belkin and Niyogi, 2003).

A number of variants have been proposed. First, a graph is constructed, usually from a distance matrix, the graph can be made sparse by keeping only the k nearest neighbors, or by specifying an ε neighborhood. Then, a similarity matrix W is calculated by using a Gaussian kernel (see Equation 2), if $c = 2\sigma^2 = \infty$, then all distances are treated equally, the smaller c the more emphasis is given to differences in distance. The degree of vertex i is $d_i = \sum_{j=1}^{n} w_{ij}$ and the degree matrix, D, is the diagonal matrix with entries d_i. Then we can form the graph Laplacian $L = D - W$ and, then, there are several ways how to proceed, an overview can be found in Luxburg (2007).

The **dimRed** package implements the algorithm from Belkin and Niyogi (2003). Analogously to LLE, Laplacian eigenmaps avoid computational complexity by creating a sparse matrix and not having to estimate the distances between all pairs of points. Then the eigenvectors corresponding to the lowest eigenvalues larger than 0 of either the matrix L or the normalized Laplacian $D^{-1/2}LD^{-1/2}$ are computed and form the embedding.

Diffusion Maps

Diffusion Maps (Coifman and Lafon, 2006) take a distance matrix as input and calculates the transition probability matrix P of a diffusion process between the points to approximate the manifold. Then the embedding is done by an eigenvalue decompositon of P to calculate the coordinates of the embedding. The algorithm for calculating Diffusion Maps shares some elements with the way Laplacian Eigenmaps are calculated. Both algorithms depart from the same weight matrix, Diffusion Map calculate the transition probability on the graph after t time steps and do the embedding on this probability matrix.

The idea is to simulate a diffusion process between the nodes of the graph, which is more robust to short-circuiting than the k-NNG from Isomap (see bottom right Figure 3). Diffusion maps in R are accessible via the `diffusionMap::diffuse()` function, which is available in the **diffusionMap** package. Additional points can be approximated into an existing embedding using the Nyström formula (Bengio et al., 2004). The implementation in **dimRed** is based on the `diffusionMap::diffuse` function.

non-Metric Dimensional Scaling

While Classical Scaling and derived methods (see section Classical Scaling) use eigenvector decomposition to embed the data in such a way that the given distances are maintained, non-Metric Dimensional Scaling (nMDS, Kruskal, 1964a,b) uses optimization methods to reach the same goal. Therefore a stress function,

$$S = \sqrt{\frac{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}, \tag{5}$$

is used, and the algorithm tries to embed y_i in such a way that the order of the d_{ij} is the same as the order of the $\hat{d}_{ij}$ Because optimization methods can fit a wide variety of problems, there are very loose limits set to the form of the error or stress function. For instance Mahecha et al. (2007) found that nMDS using geodesic distances can be almost as powerful as Isomap for embedding biodiversity patterns. Because of the flexibility of nMDS, there is a whole package in R devoted to Multidimensional Scaling, `smacof` (de Leeuw and Mair, 2009).

Several packages provide implementations for nMDS in R, for example **MASS** and **vegan** with the functions `MASS::isoMDS` and `vegan::monoMDS`. Related methods include Sammons Mapping which con be found as `MASS::sammon`. The **dimRed** package wraps `vegan::monoMDS`.

Force Directed Methods

The data X can be considered as a graph with weighted edges, where the weights are the distances between points. Force directed algorithms see the edges of the graphs as springs or the result of an electric charge of the nodes that result in an attractive or repulsive force between the nodes, the

algorithms then try to minimize the overall energy of the graph.

$$E = \sum_{i<j} k_{ij}(d_{ij} - \hat{d_{ij}})^2, \tag{6}$$

where k_{ij} is the spring constant for the spring connecting points i and j.

Graph embedding algorithms generally suffer from long running times (though compared to other methods presented here they do not scale as badly) and many local optima. This is why a number of methods have been developed that try to deal with some of the shortcomings, for example, the Kamada-Kawai (Kamada and Kawai, 1989), the Fruchtermann-Reingold (Fruchterman and Reingold, 1991), or the DrL (Martin et al., 2007) algorithms.

There are a number of graph embedding algorithms included in the **igraph** package, they can be accessed using the `igraph::layout_with_*` function family. The **dimRed** package only wraps the three algorithms mentioned above; there are many others which are not interesting for dimensionality reduction.

t-SNE

Stochastic Neighbor Embedding (SNE; Hinton and Roweis, 2003) is a technique that minimizes the Kullback-Leibler divergence of scaled similarities of the points i and j in a high dimensional space, p_{ij}, and a low dimensional space, q_{ij}:

$$KL(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{7}$$

SNE uses a Gaussian kernel (see Equation 2) to compute similarities in a high and a low dimensional space. The t-Distributed Stochastic Neighborhood Embedding (t-SNE; van der Maaten and Hinton, 2008) improves on SNE by using a t-Distribution as a kernel in low dimensional space. Because of the heavy-tailed t-distribution, t-SNE maintains local neighborhoods of the data better and penalizes wrong embeddings of dissimilar points. This property makes it especially suitable to represent clustered data and complex structures in few dimensions.

The t-SNE method has one parameter, perplexity, to tune. This determines the neighborhood size of the kernels used.

The general runtime of t-SNE is $\mathcal{O}(n^2)$, but an efficient implementation using tree search algorithms that scales as $\mathcal{O}(n \log n)$ exists and can be found in the **Rtsne** package in R. The t-SNE implementation in **dimRed** wraps the **Rtsne** package.

There exist a number of derived techniques for dimensionality reduction, e.g., NeRV (Venna et al., 2010) and JNE (Lee et al., 2013), that improve results but for which there do not yet exist packages on CRAN implementing them.

ICA

Independent Component Analysis (ICA) interprets the data X as a mixture of independent signals, e.g., a number of sound sources recorded by several microphones, and tries to "un-mix" them to find the original signals in the recorded signals. ICA is a linear rotation of the data, just as PCA, but instead of recovering the maximum variance, it recovers statistically independent components. A signal matrix S and a mixing matrix A are estimated so that $X = AS$.

There are a number of algorithms for ICA, the most widely used is fastICA (Hyvarinen, 1999) because it provides a fast and robust way to estimate A and S. FastICA maximizes a measure for non-Gaussianity called negentropy J (Comon, 1994). This is equivalent to minimizing mutual information between the resulting components. Negentropy J is defined as follows:

$$H(u) = -\int f(u) \log f(Y)\, \mathrm{d}u, \tag{8}$$

$$J(u) = H(u_{\text{gauss}}) - H(u), \tag{9}$$

where $u = (u_1, \ldots, u_n)^T$ is a random vector with density $f(\cdot)$ and u_{gauss} is a Gaussian random variable with the same covariance structure as u. FastICA uses a very efficient approximation to calculate negentropy. Because ICA can be translated into a simple linear projection, a forward and an inverse method can be supplied.

There are a number of packages in R that implement algorithms for ICA, the **dimRed** package

wraps the `fastICA::fastICA()` function from **fastICA**.

DRR

Dimensionality Reduction via Regression is a very recent technique extending PCA (Laparra et al., 2015). Starting from a rotated (PCA) solution $X' = XU$, it predicts redundant information from the remaining components using non-linear regression.

$$y_{\cdot i} = x'_{\cdot i} - f_i(x'_{\cdot 1}, x'_{\cdot 2}, \ldots, x'_{\cdot i-1}) \tag{10}$$

with $x_{\cdot i}$ and $y_{\cdot i}$ being the loading of observations on the i-th axis. In theory, any kind of regression can be used. the authors of the original paper choose Kernel Ridge Regression (KRR; Saunders et al., 1998) because it is a flexible nonlinear regression technique and computational optimizations for a fast calculation exist. DRR has another advantage over other techniques presented here, because it provides an exact forward and inverse function.

The use of KRR also has the advantage of making the method convex, here we list it under non-convex methods, because other types of regression may make it non-convex.

Mathematicaly, functions are limited to map one input to a single output point. Therefore, DRR reduces to PCA if manifolds are too complex; but it seems very useful for slightly curved manifolds. The initial rotation is important, because the result strongly depends on the order of dimensions in high dimensional space.

DRR is implemented in the package **DRR**. The package provides forward and inverse functions which can be used to train on a subset.

Quality criteria

The advantage of unsupervised learning is that one does not need to specify classes or a target variable for the data under scrutiny. Instead the chosen algorithm arranges the input data. For example, arranged into clusters or into a lower dimensional representation. In contrast to a supervised problem, there is no natural way to directly measure the quality of any output or to compare two methods by an objective measure like for instance modeling efficiency or classification error. The reason is that every method optimizes a different error function, and it would be unfair to compare t-SNE and PCA by means of either recovered variance or KL-Divergence. One fair measure would be the reconstruction error, i.e., reconstructing the original data from a limited number of dimensions, but as discussed above not many methods provide forward and inverse mappings.

However, there are a series of independent estimators on the quality of a low-dimensional embedding. The **dimRed** package provides a number of quality measures which have been proposed in the literature to measure performance of dimensionality reduction techniques.

Co-ranking matrix based measures

The co-ranking matrix (Lee and Verleysen, 2009) is a way to capture the changes in ordinal distance. As before, let $d_{ij} = d(x_i, x_j)$ be the distances between x_i and x_j, i.e., in high dimensional space and $\hat{d}_{ij} = d(y_i, y_j)$ the distances in low dimensional space, then we can define the rank of y_j with respect to y_i

$$\hat{r}_{ij} = |\{k : \hat{d}_{ik} < \hat{d}_{ij} \text{ or } (\hat{d}_{ik} = \hat{d}_{ij} \text{ and } 1 \leq k < j \leq n)\}|, \tag{11}$$

and, analogously, the rank in high-dimensional space as:

$$r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq n)\}|, \tag{12}$$

where the notation $|A|$ denotes the number of elements in a set A. This means that we simply replace the distances in a distance matrix column wise by their ranks. This means, that r_{ij} is an integer which indicates that x_i is the r_{ij}-th closest neighbor of x_j in the set X.

The co-ranking matrix Q then has elements

$$q_{kl} = |\{(i, j) : \hat{r}_{ij} = k \text{ and } r_{ij} = l\}|, \tag{13}$$

which is the 2d-histogram of the ranks. That is, q_{ij} is an integer which counts how many points of distance rank j became rank i. In a perfect DR, this matrix will only have non-zero entries in the diagonal, if most of the non-zero entries are in the lower triangle, then the DR collapsed far away

points onto each other; if most of the non-zero entries are in the upper triangle, then the DR teared close points apart. For a detailed description of the properties of the co-ranking matrix the reader is referred to Lueks et al. (2011).

The co-ranking matrix can be computed using function `coRanking::coranking()` and can be visualized using `coRanking::imageplot()`. A good embedding should scatter the values around the diagonal of the matrix. If the values are predominantly in the lower triangle, then the embedding collapses the original structure causing far away points to be much closer; if the values are predominantly in the upper triangle the points from the original structure are torn apart. Nevertheless this method requires visual inspection of the matrix. For an automated assessment of quality, a scalar value that assigns a quality to an embedding is needed.

A number of metrics can be computed from the co-ranking matrix. For example:

$$Q_{NX}(k) = \frac{1}{kn} \sum_{i=1}^{k} \sum_{j=1}^{k} q_{ij}, \tag{14}$$

which is the number of points that belong to the k-th nearest neighbors in both high- and low-dimensional space, normalized to give a maximum of 1 (Lee and Verleysen, 2009). This quantity can be adjusted for random embeddings, giving the Local Continuity Meta Criterion (Chen and Buja, 2009):

$$\mathrm{LCMC}(k) = Q_{NX}(k) - \frac{k}{n-1} \tag{15}$$

The above measures still depend on k, but LCMC has a well defined maximum at $k_{\max}$. Two measures without parameters are then defined:

$$Q_{\text{local}} = \frac{1}{k_{\max}} \sum_{k=1}^{k_{\max}} Q_{NX}(k) \text{ and} \tag{16}$$

$$Q_{\text{global}} = \frac{1}{n - k_{\max}} \sum_{k=k_{\max}}^{n-1} Q_{NX}(k). \tag{17}$$

These measure the preservation of local and global distances respectively. The original authors advised using Q_{local} over Q_{global}, but this depends on the application.

LCMC(k) can be normalized to a maximum of 1, yielding the following measure for a quality embedding (Lee et al., 2013):

$$R_{NX}(k) = \frac{(n-1)Q_{NX}(k) - k}{n - 1 - k}, \tag{18}$$

where a value of 0 corresponds to a random embedding and a value of 1 to a perfect embedding into the k-ary neighborhood. To transform $R_{NX}(k)$ into a parameterless measure, the area under the curve can be used:

$$\mathrm{AUC}_{\ln k}\left(R_{NX}(k)\right) = \left(\sum_{k=1}^{n-2} R_{NX}(k)\right) \Big/ \left(\sum_{k=1}^{n-2} 1/k\right). \tag{19}$$

This measure is normalized to one and takes k at a log-scale. Therefore it prefers methods that preserve local distances.

In R, the co-ranking matrix can be calculated using the the `coRanking::coranking` function. The **dimRed** package contains the functions `Q_local`, `Q_global`, `Q_NX`, `LCMC`, and `R_NX` to calculate the above quality measures in addition to `AUC_lnK_R_NX`.

Calculating the co-ranking matrix is a relatively expensive operation because it requires sorting every row of the distance matrix twice. It therefore scales with $\mathcal{O}(n^2 \log n)$. There is also a plotting function `plot_R_NX`, which plots the R_{NX} values with log-scaled K and adds the $AUC_{\ln K}$ to the legend (see Figure 2).

There are a number of other measures that can be computed from a co-ranking matrix, e.g., see Lueks et al. (2011); Lee and Verleysen (2009), or Babaee et al. (2013).

Cophenetic correlation

An old measure originally developed to compare clustering methods in the field of phylogenetics is cophenetic correlation (Sokal and Rohlf, 1962). This method consists simply of the correlation between the upper or lower triangles of the distance matrices (in dendrograms they are called cophenetic matrices, hence the name) in a high and low dimensional space. Additionally the distance measure and correlation method can be varied. In the **dimRed** package this is implemented in the

`cophenetic_correlation` function.

Some studies use a measure called "residual variance" (Tenenbaum et al., 2000; Mahecha et al., 2007), which is defined as

$$1 - r^2(D, \hat{D}),$$

where r is the Pearson correlation and D, $\hat{D}$ are the distances matrices consisting of elements d_{ij} and $\hat{d}_{ij}$ respectively.

Reconstruction error

The fairest and most common way to assess the quality of a dimensionality reduction when the method provides a inverse mapping is the reconstruction error. The **dimRed** package includes a function to calculate the root mean squared error which is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d(x_i', x_i)^2} \tag{20}$$

with $x_i' = f^{-1}(y_i)$, f^{-1} being the function that maps an embedded value back to feature space.

The **dimRed** package provides the `reconstruction_rmse` and `reconstruction_error` functions.

Test data sets

There are a number of test data sets that are often used to showcase a dimensionality reduction technique. Common ones being the 3d S-curve and the Swiss roll, among others. These data sets have in common that they usually have three dimensions, and well defined manifolds. Real world examples usually have more dimensions and often are much noisier, the manifolds may not be well sampled and exhibit holes and large pieces may be missing. Additionally, we cannot be sure if we can observe all the relevant variables.

The **dimRed** package implements a number of test datasets that are being used in literature to benchmark methods with the function `dimRed::loadDataSet()`. For artificial datasets the number of points and the noise level can be adjusted, the function also returns the internal coordinates.

The dimRed Package

The **dimRed** package collects DR methods readily implemented in R, implements missing methods and offers means to compare the quality of embeddings. The package is open source and available under the GPL3 license. Released versions of the package are available through CRAN (https://cran.r-project.org/package=dimRed) and development versions are hosted on GitHub (https://github.com/gdkrmr/dimRed). The **dimRed** package provides a common interface and convenience functions for a variety of different DR methods so that it is made easier to use and compare different methods. An overview of the packages main functions can be found in Table 1.

Function	Description
`embed`	Embed data using a DR method.
`quality`	Calculate a quality score from the result of `embed`.
`plot`	Plot a `"dimRedData"` or `"dimRedResult"` object, colors the points automatically, for exploring the data.
`plot_R_NX`	Compares the quality of various embeddings.
`dimRedMethodList`	Returns a character vector that contains all implemented DR methods.
`dimRedQualityList`	Returns a character vector that contains all implemented quality measures.

Table 1: The main interface functions of the **dimRed** package.

Internally, the package uses S4 classes but for normal usage the user does not need to have any knowledge on the inner workings of the S4 class system in R (cf. table 2). The package contains simple conversion functions from and to standard R-objects like a data.frame or a matrix. The `"dimRedData"` class provides an container for the data to be processed. The slot `data` contains a matrix

with dimensions in columns and observations in rows, the slot `meta` may contain a data frame with additional information, e.g., categories or other information of the data points.

Class Name	Function
`"dimRedData"`	Holds the data for a DR. Fed to `embed()`. An `as.dimRedData()` methods exists for `"data.frame"`, `"matrix"`, and `"formula"` exist.
`"dimRedMethod"`	Virtual class, ancestor of all DR methods.
`"dimRedResult"`	The result of `embed()`, the embedded data.

Table 2: The S4 classes used in the **dimRed** package.

Each embedding method is a class which inherits from `"dimRedMethod"` which means that it contains a function to generate `"dimRedResult"` objects and a list of standard parameters. The class `"dimRedResult"` contains the data in reduced dimensions, the original meta information along with the original data, and, if possible, functions for the forward and inverse mapping.

From a user-perspective the central function of the package is `embed` which is called in the form `embed(data,method,...)`, `data` can take standard R objects such as instances of `"data.frame"`, `"matrix"`, or `"formula"`, as input. The `method` is given as a character vector. All available methods can be listed by calling '`dimRedMethodList()`'. Method-specific parameters can be passed through `...`; when no method-specific parameters are given, defaults are chosen. The `embed` function returns an object of class `"dimRedResult"`.

For comparing different embeddings, **dimRed** contains the function `quality` which relies on the output of `embed` and a method name. This function returns a scalar quality score; a vector that contains the names of all quality functions is returned by calling '`dimRedQualityList()`'.

For easy visual examination, the package contains `plot` methods for `"dimRedData"` and `"dimRedResult"` objects in order to plot high dimensional data using parallel plots and pairwise scatter plots. Automatic coloring of data points is done using the available metadata.

Examples

The comparison of different DR methods, choosing the right parameters for a method, and the inspection of the results is simplified by **dimRed**. This section contains a number of examples to highlight the use of the package.

To compare methods of dimensionality reduction, first a test data set is loaded using `loadDataSet`, then the `embed` function is used for DR (`embed` can also handle standard R types like `matrix` and `data.frame`). This makes it very simple to apply different methods of DR to the same data e.g., by defining a character vector of method names and then iterating over these, say with `lapply`. For inspection, **dimRed** provides methods for the `plot` function to visualize the resulting embedding (Figure 2 b and d), internal coordinates of the manifold are represented by color gradients. To visualize how well embeddings represent different neighborhood sizes, the function `plot_R_NX` is used on a list of embedding results (Figure 2 c). The plots in figure 2 are produced by the following code:

```
## define which methods to apply
embed_methods <- c("Isomap", "PCA")
## load test data set
data_set <- loadDataSet("3D S Curve", n = 1000)
## apply dimensionality reduction
data_emb <- lapply(embed_methods, function(x) embed(data_set, x))
names(data_emb) <- embed_methods
## figure 2a, the data set
plot(data_set, type = "3vars")
## figures 2b (Isomap) and 2d (PCA)
lapply(data_emb, plot, type = "2vars")
## figure 2c, quality analysis
plot_R_NX(data_emb)
```

The function `plot_R_NX` produces a figure that plots the neighborhood size (k at a log-scale) against the quality measure $R_{NX}(k)$ (see Equation 18). This gives an overview of the general behavior of methods: if R_{NX} is high for low values of K, then local neighborhoods are maintained well; if R_{NX} is

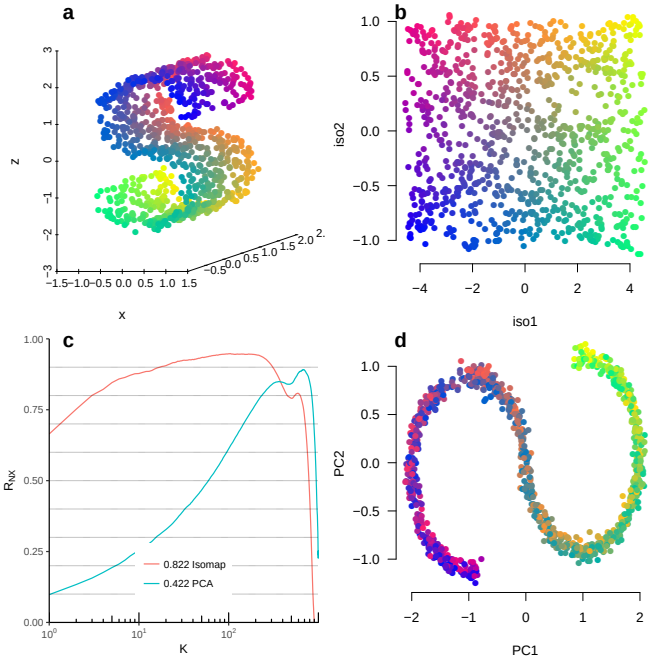

Figure 2: Comparing PCA and Isomap: (a) An S-shaped manifold, colors represent the internal coordinates of the manifold. (b) Isomap embedding, the S-shaped manifold is unfolded. (c) R_{NX} plotted agains neighborhood sizes, Isomap is much better at preserving local distances and PCA is better at preserving global Euclidean distances. The numbers on the legend are the $\text{AUC}_{1/K}$. (d) PCA projection of the data, the directions of maximum variance are preserved.

high for large values of K, then global gradients are maintained well. It also provides a way to directly compare methods by plotting more than one R_{NX} curve and an overall quality of the embedding by taking the area under the curve as an indicator for the overall quality of the embedding (see fig 19) which is shown as a number in the legend.

Therefore we can see from Figure 2c that t-SNE is very good a maintaining close and medium distances for the given data set, whereas PCA is only better at maintaining the very large distances. The large distances are dominated by the overall bent shape of the S in 3D space, while the close distances are not affected by this bending. This is reflected in the properties recovered by the different methods, the PCA embedding recovers the S-shape, while t-SNE ignores the S-shape and recovers the inner structure of the manifold.

Often the quality of an embedding strongly depends on the choice of parameters, the interface of **dimRed** can be used to facilitate searching the parameter space.

Isomap has one parameter k which determines the number of neighbors used to construct the k-NNG. If this number is too large, then Isomap will resemble an MDS (Figure 3 e), if the number is too small, the resulting embedding contains holes (Figure 3 c). The following code finds the optimal value, k_{max}, for k using the Q_{local} criterion, the results are visualized in Figure 3 a:

```
## Load data
ss <- loadDataSet("3D S Curve", n = 500)
## Parameter space
kk <- floor(seq(5, 100, length.out = 40))
## Embedding over parameter space
emb <- lapply(kk, function(x) embed(ss, "Isomap", knn = x))
```

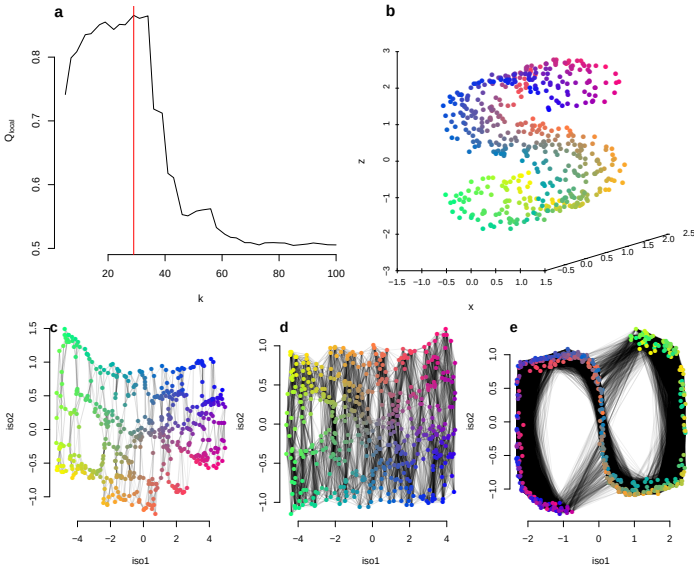

Figure 3: Using **dimRed** and the Q_{local} indicator to estimate a good value for the parameter k in Isomap. (a) Q_{local} for different values of k, the vertical red line indicates the maximum k_{max}. (b) The original data set, a 2 dimensional manifold bent in an S-shape in 3 dimensional space. Bottom row: Embeddings and k-NNG for different values of k. (c) When $k = 5$, the value for k is too small resulting in holes in the embedding, the manifold itself is still unfolded correctly. (d) Choose $k = k_{\text{max}}$, the best representation of the original manifold in two dimensions achievable with Isomap. (e) $k = 100$, too large, the k-NNG does not approximate the manifold any more.

```
## Quality over embeddings
qual <- sapply(emb, function(x) quality(x, "Q_local"))
## Find best value for K
ind_max <- which.max(qual)
k_max <- kk[ind_max]
```

Figure 3a shows how the Q_{local} criterion changes when varying the neighborhood size k for Isomap, the gray lines in Figure 3 represent the edges of the k-NN Graph. If the value for k is too low, the inner structure of the manifold will still be recovered, but it will be imperfect (Figure 3c, note that the holes appear in places that are not covered by the edges of the k-NN Graph), therefore the Q_{local} score is lower than optimal. If k is too large, the error of the embedding is much larger due to short circuiting and we observe a very steep drop in the Q_{local} score. The short circuiting can be observed in Figure 3e with the edges that cross the gap between the tips and the center of the S-shape.

It is also very easy to compare across methods and quality scores. The following code produces a matrix of quality scores and methods, where `dimRedMethodList` returns a character vector with all methods. A visualization of the matrix can be found in Figure 4.

```
embed_methods <- dimRedMethodList()
quality_methods <- c("Q_local", "Q_global", "AUC_lnK_R_NX",
                     "cophenetic_correlation")
scurve <- loadDataSet("3D S Curve", n = 2000)
quality_results <- matrix(
  NA, length(embed_methods), length(quality_methods),
  dimnames = list(embed_methods, quality_methods)
)
```

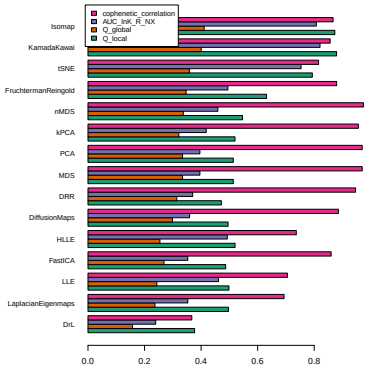

Figure 4: A visualization of the `quality_results` matrix. The methods are ordered by mean quality score. The reconstruction error was omitted, because a higher value means a worse embedding, while in the present methods a higher score means a better embedding. Parameters were not tuned for the example, therefore it should not be seen as a general quality assessment of the methods.

```
embedded_data <- list()
for (e in embed_methods) {
  embedded_data[[e]] <- embed(scurve, e)
  for (q in quality_methods)
    try(quality_results[e, q] <- quality(embedded_data[[e]], q))
}
```

This example showcases the simplicity with which different methods and quality criteria can be combined. Because of the strong dependencies on parameters it is not advised to apply this kind of analysis without tuning the parameters for each method separately. There is no automatized way to tune parameters in **dimRed**.

Conclusion

This paper presents the **dimRed** and **coRanking** packages and it provides a brief overview of the methods implemented therein. The **dimRed** package is written in the R language, one of the most popular languages for data analysis. The package is freely available from CRAN. The package is object oriented and completely open source and therefore easily available and extensible. Although most of the DR methods already had implementations in R, **dimRed** adds some new methods for dimensionality reduction, and **coRanking** adds methods for an independent quality control of DR methods to the R ecosystem. DR is a widely used technique. However, due to the lack of easily usable tools, choosing the right method for DR is complex and depends upon a variety of factors. The **dimRed** package aims to facilitate experimentation with different techniques, parameters, and quality measures so that choosing the right method becomes easier. The **dimRed** package wants to enable the user to objectively compare methods that rely on very different algorithmic approaches. It makes the life of the programmer easier, because all methods are aggregated in one place and there is a single interface and standardized classes to access the functionality.

Acknowledgments

We thank Dr. G. Camps-Valls and an anonymous reviewer for many useful comments. This study was supported by the European Space Agency (ESA) via the Earth System Data Lab project (http://earthsystemdatacube.org) and the EU via the H2020 project BACI, grant agreement No 640176.

Bibliography

P. J. M. V. D. Aart. Distribution Analysis of Wolfspiders (Araneae, Lycosidae) in a Dune Area By Means of Principal Component Analysis. *Netherlands Journal of Zoology*, 23(3):266–329, 1972. ISSN 1568-542X. URL https://doi.org/10.1163/002829673x00076. [p342]

J. Arenas-Garcia, K. B. Petersen, G. Camps-Valls, and L. K. Hansen. Kernel Multivariate Analysis Framework for Supervised Subspace Learning: A Tutorial on Linear and Kernel Multivariate Methods. *IEEE Signal Processing Magazine*, 30(4):16–29, 2013. ISSN 1053-5888. URL https://doi.org/10.1109/msp.2013.2250591. [p343]

M. Babaee, M. Datcu, and G. Rigoll. Assessment of dimensionality reduction based on communication channel model; application to immersive information visualization. In *Big Data 2013*, pages 1–6. IEEE Xplore, 2013. URL https://doi.org/10.1109/bigdata.2013.6691726. [p350]

G. H. Bakir, J. Weston, and P. B. Schölkopf. Learning to Find Pre-Images. In S. Thrun, L. K. Saul, and P. B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 449–456. MIT Press, 2004. URL https://doi.org/10.1007/978-3-540-28649-3_31. [p345]

M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373, 2003. ISSN 08997667. URL https://doi.org/10.1162/089976603321780317. [p347]

Y. Bengio, O. Delalleau, N. L. Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Learning Eigenfunctions Links Spectral Embedding and Kernel PCA. *Neural Computation*, 16(10):2197–2219, 2004. ISSN 0899-7667. URL https://doi.org/10.1162/0899766041732396. [p345, 347]

L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, 104(485):209–219, 2009. URL https://doi.org/10.1198/jasa.2009.0111. [p350]

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. ISSN 10635203. URL https://doi.org/10.1016/j.acha.2006.04.006. [p347]

P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, 1994. ISSN 01651684. URL https://doi.org/10.1016/0165-1684(94)90029-9. [p348]

J. de Leeuw and P. Mair. Multidimensional scaling using majorization: Smacof in r. *Journal of Statistical Software, Articles*, 31(3):1–30, 2009. ISSN 1548-7660. doi: 10.18637/jss.v031.i03. URL https://www.jstatsoft.org/v031/i03. [p347]

V. De Silva and J. B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, 2004. [p346]

S. Díaz, J. Kattge, J. H. C. Cornelissen, I. J. Wright, S. Lavorel, S. Dray, B. Reu, M. Kleyer, C. Wirth, I. Colin Prentice, E. Garnier, G. Bönisch, M. Westoby, H. Poorter, P. B. Reich, A. T. Moles, J. Dickie, A. N. Gillison, A. E. Zanne, J. Chave, S. Joseph Wright, S. N. Sheremet'ev, H. Jactel, C. Baraloto, B. Cerabolini, S. Pierce, B. Shipley, D. Kirkup, F. Casanoves, J. S. Joswig, A. Günther, V. Falczuk, N. Rüger, M. D. Mahecha, and L. D. Gorné. The global spectrum of plant form and function. *Nature*, 529(7585):167–171, 2016. ISSN 0028-0836. URL https://doi.org/10.1038/nature16489. [p342]

Elsevier. Scopus - Advanced search, 2017. URL https://www.scopus.com/. [p343]

T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Softw: Pract. Exper.*, 21(11):1129–1164, 1991. ISSN 1097-024X. URL https://doi.org/10.1002/spe.4380211102. [p348]

Y. Han, J. Kim, and K. Lee. Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. *IEEE-ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING*, 25(1):208–221, 2017. ISSN 2329-9290. [p342]

G. E. Hinton and S. T. Roweis. Stochastic Neighbor Embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 857–864. MIT Press, 2003. URL http://papers.nips.cc/paper/2276-stochastic-neighbor-embedding.pdf. [p343, 348]

W. W. Hsieh. Nonlinear multivariate and time series analysis by neural network methods. *Rev. Geophys.*, 42(1):RG1003, 2004. ISSN 1944-9208. URL https://doi.org/10.1029/2002rg000112. [p342]

A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999. ISSN 1045-9227. URL https://doi.org/10.1109/72.761722. [p348]

T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989. ISSN 0020-0190. URL https://doi.org/10.1016/0020-0190(89)90102-6. [p348]

J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964a. ISSN 0033-3123, 1860-0980. URL https://doi.org/10.1007/bf02289565. [p347]

J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964b. ISSN 0033-3123, 1860-0980. URL https://doi.org/10.1007/bf02289694. [p347]

V. Laparra, J. Malo, and G. Camps-Valls. Dimensionality Reduction via Regression in Hyperspectral Imagery. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1026–1036, 2015. ISSN 1932-4553. URL https://doi.org/10.1109/jstsp.2015.2417833. [p342, 349]

J. A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7–9):1431–1443, 2009. ISSN 0925-2312. URL https://doi.org/10.1016/j.neucom.2008.12.017. [p349, 350]

J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013. ISSN 0925-2312. URL https://doi.org/10.1016/j.neucom.2012.12.036. [p348, 350]

W. Lueks, B. Mokbel, M. Biehl, and B. Hammer. How to Evaluate Dimensionality Reduction? - Improving the Co-ranking Matrix. *arXiv:1110.3917 [cs]*, 2011. URL http://arxiv.org/abs/1110.3917. arXiv: 1110.3917. [p350]

U. v. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec. 2007. ISSN 0960-3174, 1573-1375. URL https://doi.org/10.1007/s11222-007-9033-z. [p347]

M. D. Mahecha, A. Martínez, G. Lischeid, and E. Beck. Nonlinear dimensionality reduction: Alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. *Ecological Informatics*, 2(2):138–149, 2007. ISSN 1574-9541. URL https://doi.org/10.1016/j.ecoinf.2007.05.002. [p347, 351]

S. Martin, W. M. Brown, and B. N. Wylie. Dr.l: Distributed Recursive (graph) Layout. Technical Report dRl; 002182MLTPL00, Sandia National Laboratories, 2007. URL http://www.osti.gov/scitech/biblio/1231060-dr-distributed-recursive-graph-layout. [p348]

R. A. A. Morrall. Soil microfungi associated with aspen in Saskatchewan: Synecology and quantitative analysis. *Can. J. Bot.*, 52(8):1803–1817, 1974. ISSN 0008-4026. URL https://doi.org/10.1139/b74-233. [p342]

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901. URL https://doi.org/10.1080/14786440109462720. [p345]

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. [p343]

S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000. ISSN 0036-8075, 1095-9203. URL https://doi.org/10.1126/science.290.5500.2323. [p346]

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8. URL http://dl.acm.org/citation.cfm?id=645527.657464. [p349]

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998. ISSN 08997667. URL https://doi.org/10.1162/089976698300017467. [p345]

B. Schölkopf, R. Herbrich, and A. J. Smola. A Generalized Representer Theorem. In *Computational Learning Theory*, pages 416–426. Springer-Verlag, 2001. URL https://doi.org/10.1007/3-540-44581-1_27. [p345]

R. R. Sokal and F. J. Rohlf. The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2):33–40, 1962. ISSN 0040-0262. URL https://doi.org/10.2307/1217208. [p350]

S. Sonnenburg, H. Strathmann, S. Lisitsyn, V. Gal, F. J. I. García, W. Lin, S. De, C. Zhang, frx, tklein23, E. Andreev, JonasBehr, sploving, P. Mazumdar, C. Widmer, P. D. . Zora, G. D. Toni, S. Mahindre, A. Kislay, K. Hughes, R. Votyakov, khalednasr, S. Sharma, A. Novik, A. Panda, E. Anagnostopoulos, L. Pang, A. Binder, serialhex, and B. Esser. Shogun-toolbox/shogun: Shogun 6.1.0, 2017. URL https://doi.org/10.5281/zenodo.1067840. [p343]

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000. ISSN 0036-8075, 1095-9203. URL https://doi.org/10.1126/science.290.5500.2319. [p346, 351]

W. S. Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, 1952. ISSN 0033-3123, 1860-0980. URL https://doi.org/10.1007/bf02288916. [p346]

L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9:2579–2605, 2008. ISSN 1532-4435. WOS:000262637600007. [p343, 348]

L. Van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: a comparative review. *J Mach Learn Res*, 10:66–71, 2009. [p343, 344]

J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization. *J. Mach. Learn. Res.*, 11:451–490, 2010. ISSN 1532-4435. WOS:000277186500001. [p348]

Guido Kraemer
Max Planck Institute for Biogeochemisty
Hans-Knöll-Str. 10, 07745 Jena
Jena
gkraemer@bgc-jena.mpg.de

Markus Reichstein
Max Planck Institute for Biogeochemisty
Hans-Knöll-Str. 10, 07745 Jena
Jena
mreichstein@bgc-jena.mpg.de

Miguel D. Mahecha
Max Planck Institute for Biogeochemisty
Hans-Knöll-Str. 10, 07745 Jena
Jena
mmahecha@bgc-jena.mpg.de

Appendix D

Article: Summarizing the State of the Terrestrial Biosphere in Few Dimensions

Kraemer, G., Camps-Valls, G., Reichstein, M., and Mahecha, M. D. (2020). Summarizing the state of the terrestrial biosphere in few dimensions. *Biogeosciences*, 17(9), 2397–2424. doi:10.5194/bg-2019-307.

This article was written to apply the system state indicator framework on biospheric data.

The article was published in *Biogeosciences* which has an impact factor of 3.951 and a 5 year impact factor of 4.699 and occupies a relative position of 33/164 in the category of "Ecology" and 29/196 in the category of "Geosciences, Multidisciplinary" in the ISI Web of Knowledge database.

Biogeosciences, 17, 2397–2424, 2020
https://doi.org/10.5194/bg-17-2397-2020

Biogeosciences Open Access EGU

Summarizing the state of the terrestrial biosphere in few dimensions

Guido Kraemer[1,2,3,4], **Gustau Camps-Valls**[2], **Markus Reichstein**[1,3], **and Miguel D. Mahecha**[1,3,4]

[1]Max Planck Institute for Biogeochemistry, Department for Biogeochemical Integration, 07745 Jena, Germany
[2]Image Processing Laboratory, Universitat de València, 46980 Paterna (València), Spain
[3]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany
[4]Remote Sensing Centre for Earth System Research, Leipzig University, 04103 Leipzig, Germany

Correspondence: Guido Kraemer (gkraemer@bgc-jena.mpg.de)

Received: 6 August 2019 – Discussion started: 21 August 2019
Revised: 21 February 2020 – Accepted: 25 March 2020 – Published: 5 May 2020

Abstract. In times of global change, we must closely monitor the state of the planet in order to understand the full complexity of these changes. In fact, each of the Earth's subsystems – i.e., the biosphere, atmosphere, hydrosphere, and cryosphere – can be analyzed from a multitude of data streams. However, since it is very hard to jointly interpret multiple monitoring data streams in parallel, one often aims for some summarizing indicator. Climate indices, for example, summarize the state of atmospheric circulation in a region. Although such approaches are also used in other fields of science, they are rarely used to describe land surface dynamics. Here, we propose a robust method to create global indicators for the terrestrial biosphere using principal component analysis based on a high-dimensional set of relevant global data streams. The concept was tested using 12 explanatory variables representing the biophysical state of ecosystems and land–atmosphere fluxes of water, energy, and carbon fluxes. We find that three indicators account for 82 % of the variance of the selected biosphere variables in space and time across the globe. While the first indicator summarizes productivity patterns, the second indicator summarizes variables representing water and energy availability. The third indicator represents mostly changes in surface albedo. Anomalies in the indicators clearly identify extreme events, such as the Amazon droughts (2005 and 2010) and the Russian heat wave (2010). The anomalies also allow us to interpret the impacts of these events. The indicators can also be used to detect and quantify changes in seasonal dynamics. Here we report, for instance, increasing seasonal amplitudes of productivity in agricultural areas and arctic regions. We assume that this generic approach has great potential for the analysis of land surface dynamics from observational or model data.

1 Introduction

Today, humanity faces negative global impacts of land use and land cover change (Song et al., 2018), global warming (IPCC, 2014), and associated losses of biodiversity (IPBES, 2019; Díaz et al., 2019), to only mention the most prominent transformations. Over the past decades, new satellite missions (e.g., Berger et al., 2012; Schimel and Schneider, 2019) along with the continuous collection of ground-based measurements (e.g., Wingate et al., 2015; Nasahara and Nagai, 2015; Baldocchi, 2020) and the integration of both (Papale et al., 2015; Babst et al., 2017; Jung et al., 2019) have increased our capacity to monitor the Earth's surface enormously. However, there are still large knowledge gaps limiting our capacity to monitor and understand the current transformations of the Earth system (Steffen et al., 2015; Rosenfeld et al., 2019; Yan et al., 2019; Piao et al., 2020).

Many recent changes due to increasing anthropogenic activity are manifested in long-term transformations. One prominent example is "global greening" that has been attributed to fertilization effects, temperature increases, and land use intensification (de Jong et al., 2011; Zhu et al., 2016; Piao et al., 2019). It is also known that phenological patterns change in the wake of climate change (Schwartz, 1998; Parmesan, 2006). However, these phenological patterns vary regionally. In "cold" ecosystems one may find decreased seasonal amplitudes on primary production due to warmer winters (Stine et al., 2009). Elsewhere, seasonal am-

Published by Copernicus Publications on behalf of the European Geosciences Union.

plitude may increase in agricultural areas, for example, due to the so-called "green revolution" (Zeng et al., 2014; Chen et al., 2019). Another change in terrestrial land surface dynamics is induced by increasing frequencies and magnitudes of extreme events (Barriopedro et al., 2011; Reichstein et al., 2013). The consequences for land ecosystems have yet to be fully understood (Flach et al., 2018; Sippel et al., 2018) and require novel detection and attribution methods tailored to the problem (Flach et al., 2017; Mahecha et al., 2007a). While extreme events are typically only temporary deviations from a normal trajectory, ecosystems may change their qualitative state permanently, for example shift from grassland to shrubland. Such shifts or tipping points can be induced by changing environmental conditions or direct human influence, and they pose yet another problem that needs to be considered (Lenton et al., 2008). The question we address here is how to uncover and summarize changes in land surface dynamics in a consistent framework. The idea is to simultaneously take advantage of a large array of global data streams, without addressing each observed phenomenon in a specific domain only. We seek to develop an integrated approach to uncover changes in the land surface dynamics based on a very generic approach.

The problem of identifying patterns of change in high-dimensional data streams is not new. Extracting the dominant features from high-dimensional observations is a well-known problem in many disciplines. One approach is to manually define indicators that are known to represent important properties such as the "Bowen ratio" (Bowen, 1926, find a more complete description of the concept in Sect. 3.3). Another one consists in using machine learning to extract unique, and ideally independent features from the data. In the climate sciences, for instance, it is common to summarize atmospheric states using empirical orthogonal functions (EOFs), also known as principal component analysis (PCA; Pearson, 1901). The rationale is that dimensionality reduction only retains the main data features, which makes them more easily accessible for analysis. One of the most prominent examples is the description of the El Niño–Southern Oscillation (ENSO) dynamics in the multivariate ENSO index (MEI; Wolter and Timlin, 2011), an indicator describing the state of the regional circulation patterns at a certain point in time. The MEI is a very successful index that can be easily interpreted and used in a variety of ways; most basically it provides a measure for the intensity and duration of the different quasi-cyclic ENSO events, but it can also be associated with its characteristic impacts, e.g., seasonal warming, changes in seasonal temperatures, and overall dryness in the Pacific Northwest of the United States (Abatzoglou et al., 2014); drought-related fires in the Brazilian Amazon (Aragão et al., 2018); and crop yield anomalies (Najafi et al., 2019).

In plant ecology, indicators based on dimensionality reduction methods are used to describe changes to species assemblages along unknown gradients (Legendre and Legendre, 1998; Mahecha et al., 2007a). The emerging gradients can be interpreted using additional environmental constraints, or based on internal plant community dynamics (van der Maaten et al., 2012). It is also common to compress satellite-based Earth observations via dimensionality reduction to get a notion of the underlying dynamics of terrestrial ecosystems. For instance, Ivits et al. (2014) showed that one can understand the impacts of droughts and heat waves based on a compressed view of the relevant vegetation indices. In general, dimensionality reduction is the method of choice to compress high-dimensional observations in a few (ideally) independent components with little loss of information (Van Der Maaten et al., 2009; Kraemer et al., 2018).

Understanding changes in land–atmosphere interactions is a complex problem, as all aforementioned patterns of change may occur and interact: land cover change may alter biophysical properties of the land surface such as (surface) albedo with consequences for the energy balance (Song et al., 2018). Long-term trends in temperature, water availability, or fertilization may impact productivity patterns and biogeochemical processes (Zhu et al., 2016; Sitch et al., 2015). In fact, these land surface dynamics have implications for multiple dimensions and require monitoring of biophysical state variables such as leaf area index, albedo, etc., as well as associated land–atmosphere fluxes of carbon, water, and energy.

Here, we aim to summarize these high-dimensional surface dynamics and make them accessible for subsequent interpretations and analyses such as mean seasonal cycles (MSCs), anomalies, trend analyses, breakpoint analyses, and the characterization of ecosystems. Specifically, we seek a set of uncorrelated, yet comprehensive, state indicators. We want to have a set of very few indicators that represent the most dominant features of the above-described temporal ecosystem dynamics. These indicators should also be uncorrelated, so that one can study the system state by looking and interpreting each indicator independently. The approach should also give an idea of the general complexity contained in the available data streams. If more than a single indicator is required to describe land surface dynamics accurately, then these indicators shall describe very different aspects. While one indicator may describe global patterns of change, others could be only relevant in certain regions, for certain types of ecosystems, or for specific types of impacts. The indicators shall have a number of desirable properties: (1) represent the overall state of observations comprising the system in space and time, (2) carry sufficient information to allow for reconstructing the original observations faithfully from these indicators, (3) be of much lower dimensionality than the number of observed variables, and (4) allow intuitive interpretations.

In this work, we first introduce a method to create such indicators, and then we apply the method to a global set of variables describing the biosphere. Finally, to prove the effectiveness of the method, we interpret the resulting set of indicators and explore the information contained in the indicators by analyzing them in different ways and relating them to well-known phenomena.

Table 1. Variables used describing the biosphere. For a description of the variables, see Appendix A.

Variable	Details	Source
Black-sky albedo	Directional reflectance	Muller et al. (2011)
Evaporation	($\mathrm{mm\,d^{-1}}$)	Martens et al. (2017)
Evaporative stress	Modeled water stress	Martens et al. (2017)
fAPAR	fraction of absorbed photosynthetically active radiation	Disney et al. (2016)
Gross primary productivity (GPP)	($\mathrm{gC\,m^{-2}\,d^{-1}}$)	Tramontana et al. (2016), Jung et al. (2019)
Latent energy (LE)	($\mathrm{W\,m^{-2}}$)	Tramontana et al. (2016), Jung et al. (2019)
Net ecosystem exchange (NEE)	($\mathrm{gC\,m^{-2}\,d^{-1}}$)	Tramontana et al. (2016), Jung et al. (2019)
Root-zone soil moisture	($\mathrm{m^3\,m^{-3}}$)	Martens et al. (2017)
Sensible heat (H)	($\mathrm{W\,m^{-2}}$)	Tramontana et al. (2016), Jung et al. (2019)
Surface soil moisture	($\mathrm{mm^3\,mm^{-3}}$)	Martens et al. (2017)
Terrestrial ecosystem respiration (TER)	($\mathrm{gC\,m^{-2}\,d^{-1}}$)	Tramontana et al. (2016), Jung et al. (2019)
White-sky albedo	Diffuse reflectance	Muller et al. (2011)

2 Methods

2.1 Data

Table 1 gives an overview of the data streams used in this analysis (for a more detailed description see Appendix A). For an effective joint analysis of more than a single variable, the variables have to be harmonized and brought to a single grid in space and time. The Earth System Data Lab (ESDL; https://www.earthsystemdatalab.net, last access: 23 April 2020; Mahecha et al., 2020) curates a comprehensive set of data streams to describe multiple facets of the terrestrial biosphere and associated climate system. The data streams are harmonized as analysis-ready data on a common spatiotemporal grid (equirectangular grid 0.25° in space and 8 d in time, 2001–2011), forming a 4D hypercube, which we call a "data cube". The ESDL not only curates Earth system data, but also comes with a toolbox to analyze these data efficiently. For this study, we chose all available variables in the ESDL v1.0 (the most recent version available at the time of analysis), divided the available variables into meteorological and biospheric variables and discarded the atmospheric variables. We also discarded variables with distributions that are badly suited for a linear PCA (e.g., burned area contains mostly zeros) and variables with too many missing values. The only dataset that was added post hoc was fAPAR, which represents an important aspect of vegetation which was not available in the data cube at the time of analysis (it is part of the most recent version of the data cube).

The datasets taken from Tramontana et al. (2016) and Jung et al. (2019) are derived from flux tower measurements (Baldocchi, 2020). The flux towers are not equally distributed in climate space; i.e., there are many flux towers in temperate areas but much fewer in tropic and arctic regions, which may lead to less accurate data in these regions. These datasets also exclude large arid areas such as the Sahara and Gobi deserts and parts of the Arabian Peninsula which may affect the resulting loadings of the PCA slightly.

In this study, each variable was normalized globally to zero mean and unit variance to account for the different units of the variables, i.e., transform the variables to have standard deviations from the mean as the common unit. Because the area of the pixel changes with latitude in the equirectangular coordinate system used by the ESDL, the pixels were weighted according to the represented surface area. Only spatiotemporal pixels without any missing values were considered in the calculation of the covariance matrix.

2.2 Dimensionality reduction with PCA

As a method for dimensionality reduction, we used a modified principal component analysis to summarize the information contained in the observed variables. PCA transforms the set of d centered and, in this case, standardized variables into a subset of p, $1 \leq p \leq d$, principal components (PCs). Each component is uncorrelated with the other components, while the first PCs explain the largest fraction of variance in the data.

The data streams consist of $d = 12$ observed variables at the same time and location. Each observation is defined in a d-dimensional space, $\boldsymbol{x}_i \in \mathbb{R}^d$, and we define the dataset by collecting all samples in the matrix $\mathbf{X} = [\boldsymbol{x}_1|\cdots|\boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$. The observations are repeated in space and time and lie on a grid of lat × long × time. In our case, we have $n = |\mathrm{lat}| \times |\mathrm{long}| \times |\mathrm{time}| = 720 \times 1440 \times 506 = 524,620,800$ observations, where $|\cdot|$ denotes the cardinality of the dimension. Note that the actual number of observations was lower, $n = 106,360,156$, because we considered land points only and removed missing values.

The fundamental idea of PCA is to project the data to a space of lower dimensionality that preserves the covariance structure of the data. Hence, the fundament of a PCA is the computation of a covariance matrix, $\mathbf{Q}$. When all variables

are centered to global zero mean and normalized to unit variance, the covariance matrix can in principle be estimated as

$$\mathbf{Q} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T = \frac{1}{n-1}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T. \quad (1)$$

However, in our case the data cube lies on a regular 0.25° grid and estimating $\mathbf{Q}$ as above would lead to overestimating the influence of dynamics in relatively small pixels of high latitudes compared to lower latitudes where each data point represents a larger area. Hence, one needs a weighted approach to calculate the covariance matrix,

$$\mathbf{Q} = \frac{1}{w}\sum_{i=1}^{n} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T, \quad (2)$$

where $w_i = \cos(\mathrm{lat}_i)$ and lat_i is the latitude of observation i, $w = \sum_{i=1}^{n} w_i$ is the total weight, and n is the total number of observations. Equation (2) has the additional property that it can be computed sequentially on very big datasets, such as our Earth System Data Cube, by a consecutively adding observations to an initial estimate.

Note that the actual calculation of the covariance matrix is even more complicated, because summing up many floating-point numbers one by one can lead to large inaccuracies due to precision issues of floating-point numbers and instabilities of the naive algorithm (Higham, 1993; the same holds for the implementations of the `sum` function in most software used for numerical computing). Here, we used the Julia package `WeightedOnlineStats.jl` (https://doi.org/10.5281/zenodo.3360311, repository: https://github.com/gdkrmr/WeightedOnlineStats.jl/, last access: 23 April 2020) (implemented by the first author of this paper), which uses numerically stable algorithms for summation, higher-precision numbers, and a map-reduce scheme that further minimizes floating-point errors.

Based on this weighted and numerically stable covariance matrix, the PCA can be computed using an eigendecomposition of the covariance matrix,

$$\mathbf{Q} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T \in \mathbb{R}^{d\times d}. \quad (3)$$

In this case, the covariance matrix $\mathbf{Q}$ is equal to the correlation matrix because we standardized the variables to unit variance. $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues, $\lambda_1, \ldots, \lambda_d$, in the diagonal in decreasing order and $\mathbf{V} \in \mathbb{R}^{d\times d}$, the matrix with the corresponding eigenvectors in columns. $\mathbf{V}$ can project the new incoming input data $\boldsymbol{x}_i$ (centered and standardized) onto the retained PCs,

$$\mathbf{y}_i = \mathbf{V}^T \boldsymbol{x}_i \in \mathbb{R}^d, \quad (4)$$

where $\mathbf{y}_i$ is the projection of the observation $\boldsymbol{x}_i$ onto the d PCs.

The canonical measure of the quality of a PCA is the fraction of explained variance by each component, σ_i^2, calculated as

$$\sigma_i^2 = \frac{\lambda_i}{\sum_{i=1}^{d}\lambda_i}. \quad (5)$$

To get a more complete measure of the accuracy of the PCA, we used the "reconstruction error" in addition to the fraction of explained variance. PCA allows a simple projection of an observation onto the first p PCs and a consecutive reconstruction of the observations from this p-dimensional projection. This is achieved by

$$\mathbf{Y}_p = \mathbf{V}_p^T \mathbf{X} \in \mathbb{R}^{p\times n} \text{ and } \mathbf{X}_p = \mathbf{V}_p \mathbf{Y}_p \in \mathbb{R}^{d\times n}, \quad (6)$$

where $\mathbf{Y}_p$ is the projection onto the first p PCs, $\mathbf{V}_p$ the matrix with columns consisting of the eigenvectors belonging to the p largest eigenvalues, and $\mathbf{X}_p$ the observations reconstructed from the first p PCs.

The reconstruction error, $\boldsymbol{e}_i$, was calculated for every point, $\boldsymbol{x}_i$, in the space–time domain based on the reconstructions from the first p principal components:

$$\boldsymbol{e}_i = \mathbf{V}_p \mathbf{V}_p^T \boldsymbol{x}_i - \boldsymbol{x}_i \in \mathbb{R}^d. \quad (7)$$

As this error is explicit in space, time, and variable, it allows for disentangling the contribution of each of these domains to the total error. This can be achieved by estimating the (weighed) mean square error,

$$\mathrm{MSE} = \frac{1}{w}\sum_i w_i \boldsymbol{e}_i^2. \quad (8)$$

This approach can give a better insight into the compositions of the error than a single global error estimate based on the eigenvalues.

2.3 Pixel-wise analyses of time series

The principal components estimated as described above are ideally low-dimensional representations of the land surface dynamics that require further interpretation. These components have temporal dynamics that need to be understood in detail. One crucial question is how the dynamics of a system of interest deviate from its expected behavior at some point in time. A classical approach is inspecting the "anomalies" of a time series, i.e., the deviation from the mean seasonal cycle at a certain day of year.

Another key description of such system dynamics are trends. We estimated trends of the indicators as well as of their seasonal amplitude using the Theil–Sen estimator. The advantage of the Theil–Sen estimator is its robustness to up to 29.3 % of outliers (Theil, 1950a, b, c; Sen, 1968), while ordinary least-squares regression is highly sensitive to such values. The calculation of the estimator consists simply in computing the median of the slopes spanned by all possible pairs of points,

$$\mathrm{slope}_{ij} = \frac{z_i - z_j}{t_i - t_j}, \quad (9)$$

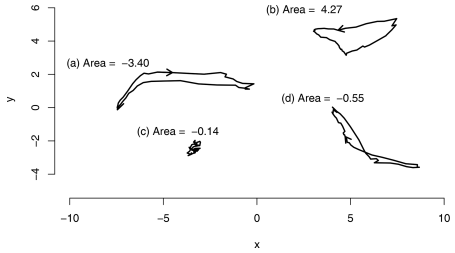

Figure 1. Example polygons and their areas, Eq. (10); the arrows indicate the directionality. **(a)** Clockwise polygon with a negative area. **(b)** Counterclockwise polygon with a positive area. **(c)** Chaotic polygon with a very low area. **(d)** Polygon with a single intersection and both a clockwise and counterclockwise portion. The clockwise portion is slightly larger than the counterclockwise portion; therefore the area is slightly negative.

where z_i is the value of the response variable at time step i and t_i the time at time step i. In our experiments, we computed the slopes separately per pixel and principal component with time as the predictor and the value of the principal component as the response variable.

To test the slopes for significance, we used the Mann–Kendall statistics (Mann, 1945; Kendall, 1970) and adjusted the resulting p values with the Benjamini–Hochberg method to control for the false discovery rate (Benjamini and Hochberg, 1995). Slopes with an adjusted $p < 0.05$ were deemed significant.

To identify disruptions in trajectories, breakpoint detection provides a good framework for analysis. For the estimation of breakpoints, the generalized fluctuation test framework (Kuan and Hornik, 1995) was used to test for the presence of breakpoints. The framework uses recursive residuals (Brown et al., 1975) such that a breakpoint is identified when the mean of the recursive residuals deviates from zero. We used the implementation in Zeileis et al. (2002). For practical reasons, here we only focus on the largest breakpoint.

The analysis of a different type of dynamic considers bivariate relations. In the context of oscillating signals it is particularly instructive to quantify their degree of phase shift and direction – even if both signals are not linearily related. A "hysteresis" would be such a pattern describing how the pathways $A \rightarrow B$ and $B \rightarrow A$ between states A and B differ (Beisner et al., 2003). We estimated hysteresis by calculating the area inside the polygon formed by the mean seasonal cycle of the combinations of two components.

$$\text{Area} = \frac{1}{2}\sum_{i=1}^{n} x_i(y_{i+1} - y_{i-1}), \tag{10}$$

where $n = 46$, the number of time steps in a year, and x_i and y_i are the mean seasonal cycle of two PCs at time step i. The polygon is circular; i.e., the indices wrap around the edges of the polygon so that $x_0 = x_n$ and $x_{n+1} = x_1$. This formula gives the actual area inside the polygon only if it is non-self-intersecting and the vertices run counterclockwise. If the vertices run clockwise, the area is negative. If the polygon is shaped like an 8, the clockwise and counterclockwise parts will cancel each other (partially) out. Trajectories that have larger amplitudes will also tend to have larger areas as illustrated in Fig. 1.

3 Results and discussion

In the following, we first briefly present and discuss the quality of the global dimensionality reduction (Sect. 3.1) and interpret the individual components from an ecological point of view (Sect. 3.2). We summarize the global dynamics that we uncovered in the low-dimensional space (Sect. 3.3). We characterize the contained seasonal dynamics (Sect. 3.4), including spatial patterns of hysteresis (Sect. 3.5). We then describe global anomalies of the identified trajectories (Sect. 3.6) and discuss the identified anomalies in depth based on local phenomena (Sect. 3.7). Finally, we present global trends and their breakpoints (Sect. 3.7).

3.1 Quality of the PCA

Figure 2a shows the explained fraction of variance (Eq. 5) for the global PCA based on the entire data cube. The two leading components explain 73 % of the variance from the 12 variables; additional components contribute relatively little additional variance (PC_3 contributes 9 % and all subsequent PCs less than 7 %) each. This results in a "knee" at component 3, which suggests that two indicators are sufficient to capture the major global dynamics of the terrestrial land surface, but we will also consider the third components in the following analyses (Cattell, 1966).

We estimated the reconstruction error sequentially up to the first three principal components (Fig. 3). Regions that do not fit the model well show a higher reconstruction error. Considering one component only, the highest reconstruction errors appear in high latitudes but decrease strongly with each additional component and nearly vanish if the third component is included.

3.2 Interpretation of the PCA

The first PC summarizes variables that are closely related to primary productivity (GPP, LE, NEE, fAPAR) and therefore are highly interrelated (see Fig. 2b). The energy for photosynthesis comes from solar radiation, and fAPAR is an indicator for the fraction of light used for photosynthesis. The available photosynthetic radiation is used by photosynthesis to fix CO_2 and to produce sugars that maintain the metabolism of the plant. The total uptake of CO_2 is reflected in GPP, which is also closely related to water con-

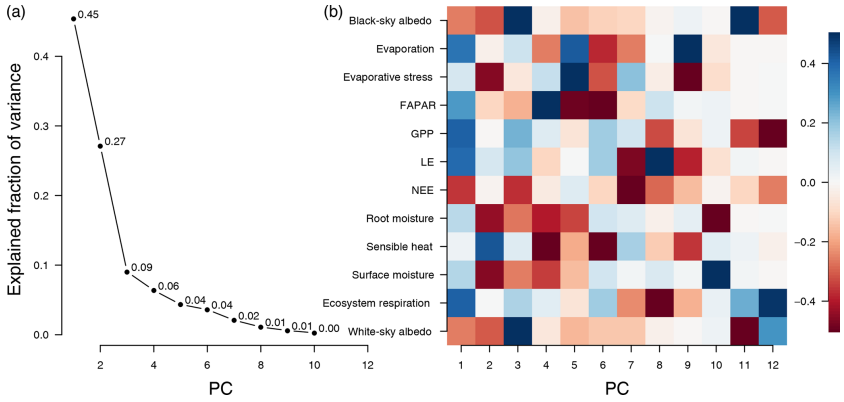

Figure 2. **(a)** Fraction of explained variance of the PCA by component. The knee at component three suggests that components four and higher do not contribute much to total variance. **(b)** Rotation matrix of the global PCA model (also called *loadings*, Eq. 4). The columns of the rotation matrix describe the linear combinations of the (centered and standardized) original variables that make up the principal components. PC_1 is dominated by primary-productivity-related variables, PC_2 by variables describing water availability, and PC_3 by variables describing albedo. Values of the rotation matrix are clamped to the range $[-0.5, 0.5]$; the actual range of the values is $[-0.73, 0.74]$ and $[-0.46, 0.54]$ for the first three components.

sumption. The flow of water within the plant is not only essential to enable photosynthesis but also drives the transport of nutrients from the roots. The uplift of water in the plant is ultimately driven by transpiration – together with evaporation from soil surfaces one can observe the integrated latent energy needed for the phase transition (LE). However, ecosystems also respire; CO_2 is produced by plants in energy-consuming processes as well as by the decomposition of dead organic materials via soil microbes and other heterotrophic organisms. This total respiration can be observed as terrestrial ecosystem respiration (TER). The difference between GPP and TER is the net ecosystem exchange (NEE) rate of CO_2 between ecosystems and the atmosphere (Chapin et al., 2006). GPP and TER are also well represented in the first dimension.

The second component represents variables related to the surface hydrology of ecosystems (see Fig. 2b). Surface moisture, evaporative stress, root-zone soil moisture, and sensible heat are all essential indicators for the state of plant-available water. While surface moisture is a rather direct measure, evaporative stress is a modeled quantity summarizing the level of plant stress: a value of zero means that there is no water available for transpiration, while a value of 1 means that transpiration equals the potential transpiration (Martens et al., 2017). Root-zone soil moisture is the moisture content of the root zone in the soil, the moisture directly available for root uptake. If this quantity is below the wilting point, there is no water available for uptake by the plants. Sensible heat is the exchange of energy by a change in temperature; if there is enough water available, then most of the surface heat will be lost due to evaporation (latent heat), and with decreasing water availability more of the surface heat will be lost due to sensible heat, making this an indicator of dryness as well.

We observe that the third component is most strongly related to albedo (Fig. 2b). Albedo describes the overall reflectiveness of a surface. Here we refer to broadband (400–3000 nm) surface albedo; for an exact definition see Appendix A. Light surfaces, such as snow and sand, reflect most of the incoming radiation, while surfaces that have a high liquid water content or active vegetation absorb most of the incoming radiation. Local changes to albedo can be due to many causes, e.g., snowfall, vegetation greening and browning, or land use change.

The relation of PC_3 to productivity and hydrology is opposite to what we would expect from an albedo axis. Because vegetation uses radiation as an energy source, albedo is negatively correlated with the productivity of vegetation, hence the negative correlation of albedo with PC_1. Given that water also absorbs radiation, we can observe a negative correlation of albedo with PC_2 (see Fig. 2b). We observe that PC_1 and PC_2 are positively correlated with PC_3 on the positive portion of their axes (see Fig. 4d and f), which means counterintuitively that the index representing albedo is positively correlated with primary productivity and moisture content. Finally we can observe that PC_1 and PC_2 have a much higher reconstruction error in snow-covered regions, which

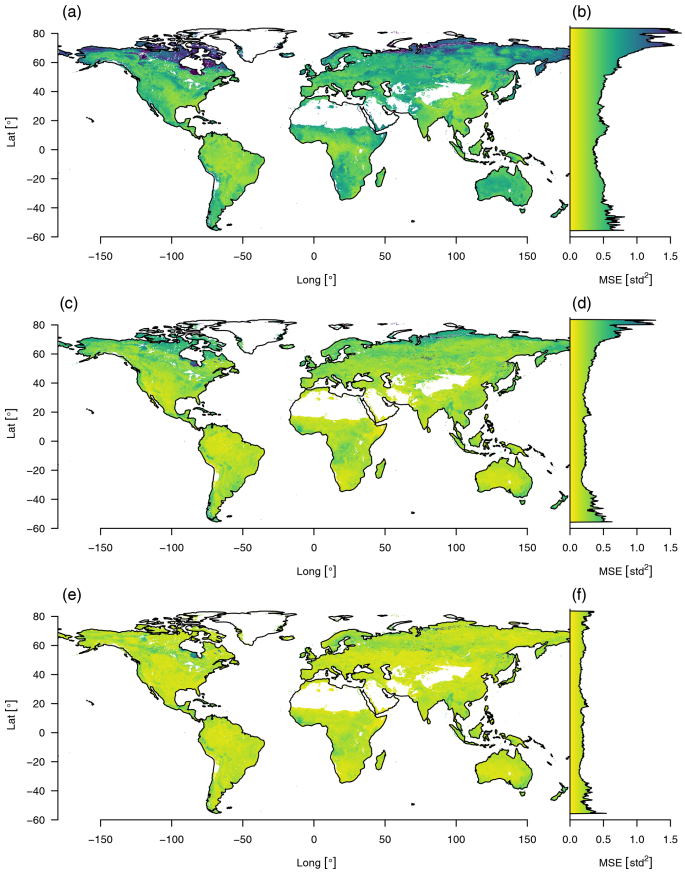

Figure 3. Reconstruction error of the data cube using varying numbers of principal components aggregated by the mean squared error. Reconstruction errors aggregated over all time steps and variables are shown in the left column: **(a)** using only the first component, **(c)** using the first two, **(e)** and using the first three. Corresponding right plots **(b, d, f)** show the mean reconstruction error aggregated by latitude.

is strongly improved by adding PC_3 (see Fig. 3f). Therefore the third component should be regarded mostly as a binary variable that introduces snow cover, as the other information that is usually associated with albedo is already contained in the first two components.

3.3 Distribution of points in PCA space

The bivariate distribution of the first two principal components forms a "triangle" (gray background in Fig. 4a). At the high end of PC_1 we find one point of the triangle in which ecosystems have a high primary productivity (high values of GPP, fAPAR, LE, TER, and evaporation), mostly limited by radiation. On the lower end of the first principal component we find the other two points of the triangle describing two alternative states of low productivity. These can happen either when the second principal component coincides with temperature limitation (the negative extreme of the second principal component) as seen in the lower left corner of the distribution

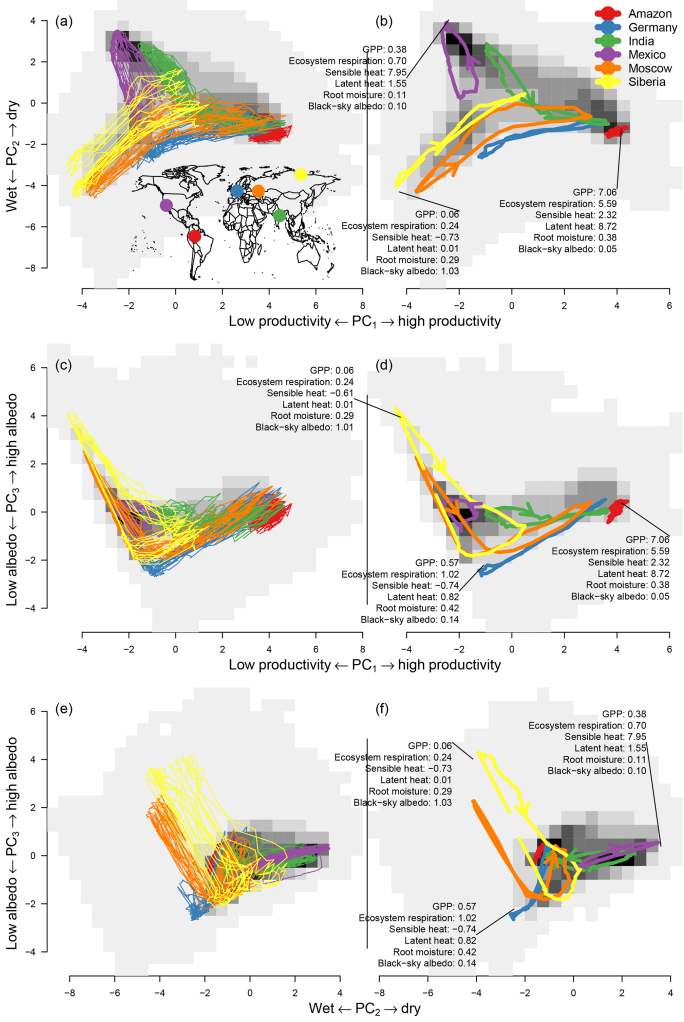

Figure 4. Trajectories of some points (colored lines) and the area-weighted density over principal components one and two (the gray background shading shows the density) for **(a, c, e)** the raw trajectories and **(b, d, f)** the mean seasonal cycle. The trajectories are shown in the space of PC_1–PC_2 (first row), PC_1–PC_3 (second row), and PC_2–PC_3 (third row). The trajectories were chosen to cover a large area in the space of the first two principal components. Some of the trajectories have an arrow indicating the direction. The numbers illustrate the value of some variables; for units see Table 1. Description of the points is as follows. Red: tropical rain forest, 2.625° S, 67.625° W; blue: maritime climate, 52.375° N, 7.375° E; green: monsoon climate, 22.375° N, 82.375° E; purple: subtropical, 34.875° N, 117.625° W; orange: continental climate, 52.375° N, 44.875° E; yellow: arctic climate, 72.375° N, 119.875° E.

in Fig. 4a and b or due to water limitation (positive extreme of the second principal component, the upper left corner in Fig. 4a). This pattern reflects the two essential global limitations of GPP in terrestrial ecosystems (Anav et al., 2015).

Both components form a subspace in which most of the variability of ecosystems takes place. Component one describes productivity and component two the limiting factors to productivity. Therefore, we can see that most ecosystems with high values on component one (a high productivity) are at the approximate center of component two. When ecosystems are found outside the center of component two, they have lower values on component one (lower productivity) because they are limited by water or temperature (see Fig. 4b).

To further interpret the triangle we analyze how the Bowen ratio embeds in the space of the first two dimensions. Energy fluxes from the surface into the atmosphere can represent either a radiative transfer (sensible heat) or evaporation (latent heat). Their ratio is the "Bowen ratio", $B = \frac{H}{LE}$, (Bowen, 1926; see also Fig. 5). When water is available most of the available energy will be dissipated by evaporation, $B < 1$, resulting in a high latent heat flux. Otherwise, the transfer by latent heat will be low and most of the incoming energy has to be dissipated via sensible heat, $B > 1$. In higher latitudes, there is relatively limited incoming radiation and temperatures are low; therefore there is not much energy to be dissipated and both heat fluxes are low. A high sensible heat flux is an indicator of water limitation.

3.4 Seasonal dynamics

The leading principal components represent most of the variability of the space spanned by the observed variables, summarizing the state of a spatiotemporal pixel efficiently. This means that the PCs track the state of a local ecosystem over time (Fig. 4a) or, in the case of the mean seasonal cycle, time of the year (Fig. 4b). For a representation of the state of the first three components in time and space, see Appendix Fig. B1.

A first inspection reveals a substantial overlap of seasonal cycles of very different regions of the world. We also see that very different ecosystems may reach very similar states in the course of the season, even though their seasonal dynamics are very different. For instance, a midlatitude pixel (blue trajectory in Fig. 4) shows very similar characteristics to tropical forests during peak growing season. This indicates that an ecosystem of the midlatitudes can reach similar levels of productivity and water availability as a tropical rain forest (see also Appendix Fig. C1). Likewise, for the first two components, many high-latitude areas show similar characteristics to midlatitude areas during winter (low latent and sensible energy release as well as low GPP), and many dry areas such as deserts show similar characteristics to areas with a pronounced dry season, e.g. the Mediterranean.

Depending on their position on Earth, ecosystem states can shift from limitation to growth during the year (Fig. 4b, e.g. Forkel et al., 2015). For example, the orange trajectory in Fig. 4, an area close to Moscow, shifts from a temperature-limited state in winter to a state of very high productivity during summer. Other ecosystems remain in a single limitation state with only slight shifts, such as the red trajectory in Fig. 4. In the corner of maximum productivity of the distribution, we find tropical forests characterized by a very low seasonality. We also observe that very different ecosystems can have very similar characteristics during their peak growing season; e.g. green (located in northeast India), blue (northwest Germany), and orange (located close to Moscow) trajectories have very similar characteristics during peak growing season compared to the red trajectory.

The third component shows a different picture. Due to a consistent winter snow cover in higher latitudes, the albedo is much higher and the amplitude of the mean seasonal cycle is much larger than in other ecosystems. Other areas show comparatively little variance on the third component and their relation to productivity and moisture content is even positively correlated to the third component, which is the opposite of what is expected from an albedo axis.

The global pattern of the first principal component follows the productivity cycles during summer and winter (Fig. 6, left column) of the Northern Hemisphere, with positive values (high productivity, green) during summer and negative values (low productivity, brown) during winter. The tropics show high productivity all year. The global pattern shows the well-known green wave (Schwartz, 1994, 1998) because the first dimension integrates over all variables that correlate with plant productivity.

The second principal component (Fig. 6, middle column) tracks water deficiency: red and light red areas indicate water deficiency, light blue areas excess water, and dark blue areas water growth limitation due to cold. Areas which are temperature limited during winter but have a growing season during summer, such as boreal forests, change from dark blue in winter to light blue during the growing season. Areas which have low productivity during a dry season change their coloring from red to light red during the growing season, e.g the northwest of Mexico and southwest of the United States.

The third principal component (Fig. 6, right column) tracks surface reflectance. Therefore we can see the highest values in the arctic region during winter, and other areas vary much less in their reflectance throughout the year. Again, the third component shows a counterintuitive behavior in the midlatitudes, as it is positively correlated with productivity and therefore shows the opposite behavior of what would be expected from an indicator tracking albedo.

Although the principal components are globally uncorrelated, they covary locally (see Fig. D1). Ecosystems with a dry season have a negative covariance between PC_1 and PC_2, while ecosystems that cease productivity in winter have a positive covariance. Cold arid steppes and boreal climates show a negative covariance between PC_1 and PC_3. While other ecosystems that have a strong seasonal cycle show a

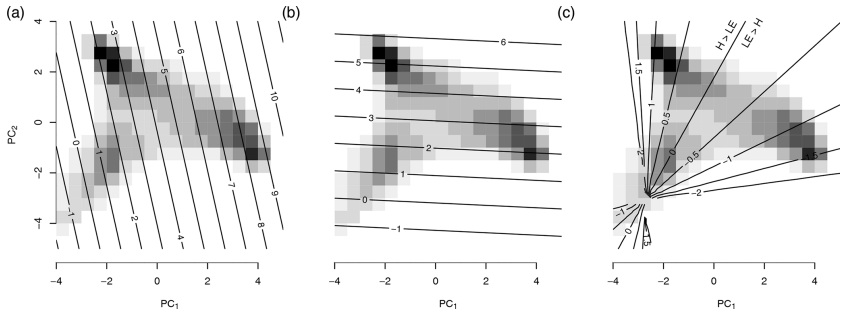

Figure 5. The background shading shows the distribution of the mean seasonal cycle of the spatial points (see Fig. 4). The contour lines represent the reconstruction of the variables from the first two principal components. The reconstructed variables are **(a)** latent heat (LE), **(b)** sensible heat (H), and **(c)** $\log_{10}\left(\frac{\text{Sensible Heat}}{\text{Latent Heat}}\right)$, the $\log_{10}$ of the Bowen ratio. Note that the LE and H have been considered in the construction of the PCs and hence are a linear function of the PCs. The Bowen ratio, instead, was not considered here and clearly responds in a nonlinear form.

positive correlation, many tropical ecosystems do not show a large covariance. A very similar picture is painted between the covariance of PC_2 and PC_3: boreal and steppe ecosystems show a negative covariance, while most other ecosystems show a more or less pronounced positive covariance, again depending on the strength of the seasonality.

Observing the mean seasonal cycle of the principal components gives us a tool to characterize ecosystems and may also serve as a basis for further analysis, such as a global comparison of ecosystems (Metzger et al., 2013; Mahecha et al., 2017).

3.5 Hysteresis

The alternative return path between ecosystem states forming the hysteresis loops arises from the ecosystem tracking seasonal changes in the environmental condition, e.g. summer–winter or dry–rainy seasons (Fig. 4b). Hysteresis is a common occurrence in ecological systems (Folke et al., 2004; Blonder et al., 2017; Renner et al., 2019). For instance, a hysteresis loop can be found when plotting soil respiration against soil temperature (Tang et al., 2005). The sensitivity of soil respiration to soil temperature changes seasonally due to changing soil moisture and photosynthesis (by supplying carbon to the rhizosphere), producing a seasonally changing hysteresis effect (Gaumont-Guay et al., 2006; Richardson et al., 2006; Zhang et al., 2018). Biological variables also show a hysteresis effect in their relations with atmospheric variables; e.g. Mahecha et al. (2007b) found a hysteresis effect between seasonal NEE, temperature, and a number of other ecosystem and climate-related variables. Here we look at the mean seasonal cycles of pairs of indicators and the area they enclose.

The orange trajectory (area close to Moscow) in Fig. 4b shows that the paths between maximum and minimum productivity can be very different, in contrast to the blue trajectory located in the northwest of Germany which also has a very pronounced yearly cycle but shows no such effect. Figure 4 also indicates that the area inside the mean seasonal cycles of PC_1–PC_2 and PC_1–PC_3 shows important characteristics while hysteresis in PC_2–PC_3 is a much less pronounced feature; i.e., we can only see a pronounced area inside the yellow curve in Fig. 4f.

The trajectories that show a more pronounced counterclockwise hysteresis effect in PC_1–PC_2 (Fig. 7a) are areas with a warm and temperate climate and partially those that have a snow climate with warm summers, i.e., areas that have pronounced growing, dry, and wet seasons and therefore shift their limitations more strongly during the year. That means the moisture reserves are depleted during growing season, and therefore the return path has higher values on the second principal component (the climatic zones are taken from the Köppen–Geiger classification; Kottek et al., 2006). We can also see that areas with dry winters tend to have a clockwise hysteresis effect, e.g. many areas in East Asia. Due to the humid summers there is no increasing water limitation during the summer months which causes a decrease for PC_2 instead of an increase. Other areas with clockwise hysteresis can be found in winter dry areas in the Andes and the winter dry areas north and south of the African rain forests. Tropical rain forests do not show any hysteresis effect due to their low seasonality. In general we can say that the area inside the mean seasonal cycle trajectory of PC_1–PC_2 depends mostly on water availability in the growing and non-growing seasons, i.e., the contrast of wet summer and dry winter vs. dry summer and wet winter.

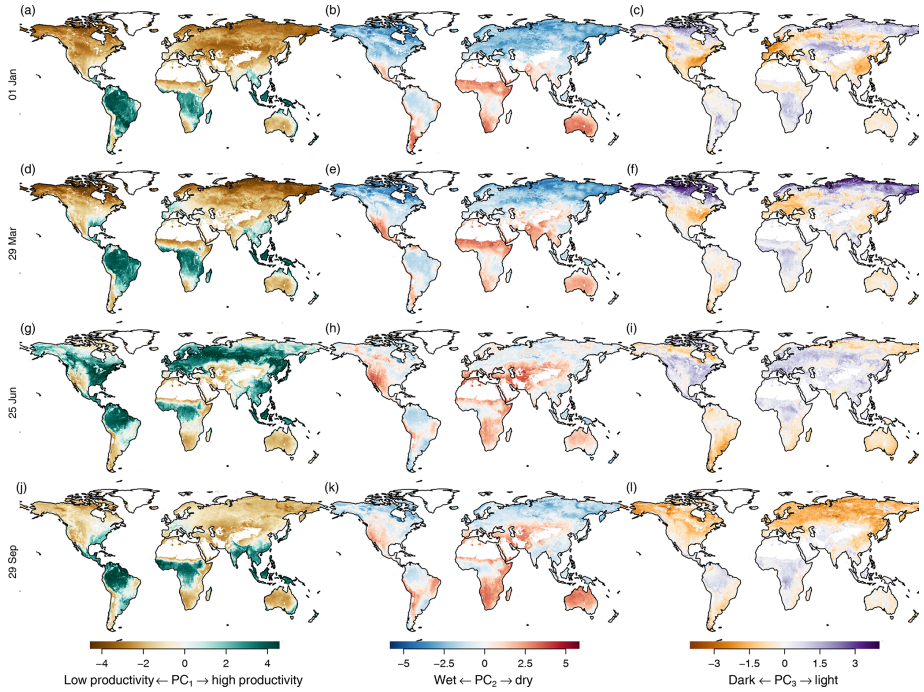

Figure 6. Mean seasonal cycle of the first three principal components (in columns) during the seasons (in rows). Left column: first principal component. Middle column: second principal component. Right column: third principal component. Rows from top to bottom: equally spaced intervals during the year. Values have been clamped to 0.7 times their range to increase contrast.

The hysteresis effect on PC_1–PC_3 (Fig. 7b) shows a pronounced counterclockwise MSC trajectory mostly in warm temperate climates with dry summers, while it shows a clockwise MSC trajectory in most other areas; again tropical rain forests are an exception due to their low seasonality. The most pronounced clockwise MSC trajectories can be found in tundra climates in arctic latitudes, where we have a consistent winter snow cover and a very short growing period. A counterclockwise rotation can be found in summer dry areas, such as the Mediterranean and California, but also some more humid areas, such as the southeast United States and the southeast coast of Australia. In these areas we can find a decrease for PC_3 during the non-growing phase which probably corresponds to a drying out of the vegetation and soils.

The hysteresis effect on PC_2–PC_3 (Fig. 7c) mostly depends on latitude. There is a large counterclockwise effect in the very northern parts, due to the large amplitude of PC_3. The amplitude gets smaller further south until the rotation reverses in winter dry areas at the northern and southern extremes of the tropics and disappears at the equatorial humid rain forests.

We can see that the hysteresis of pairs of indicators represents large-scale properties of climatic zones. The enclosed area and the direction of the rotation provide interesting information. Hysteresis can provide information on the seasonal availability of water, seasonal dry periods, or snowfall. With the method presented here, we can not observe intersecting trajectories, which would probably provide even more interesting insights (e.g. the green trajectory in Fig. 4b).

3.6 Anomalies of the trajectories

The deviation of the trajectories from their mean seasonal cycle should reveal anomalies and extreme events. These anomalies have a directional component which makes them interpretable the same way the original PCs are. Therefore

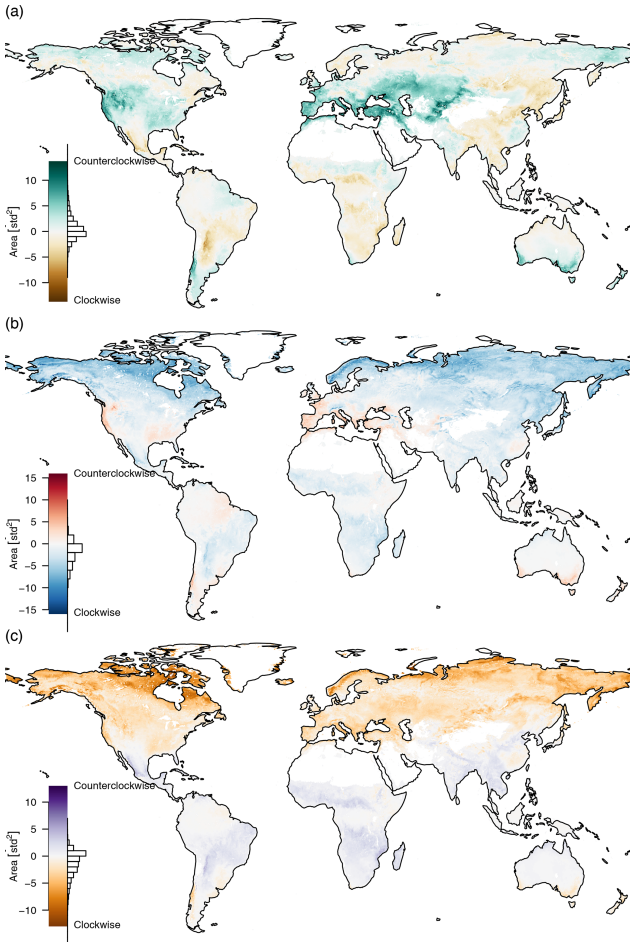

Figure 7. The area inside the mean seasonal cycles of **(a)** PC_1–PC_2, **(b)** PC_1–PC_3, and **(c)** PC_2–PC_3. The area is positive if the direction is counterclockwise and negative if the direction is clockwise. Most of the trajectories need a strong seasonal cycle to show a pronounced hysteresis effect. If the mean seasonal cycle intersects, the areas cancel each other out, e.g. the green trajectory of Fig. 4b.

one can infer the state of the ecosystem during an anomaly. For instance the well-known Russian heat wave in summer 2010 (Flach et al., 2018) appears in Fig. 8 as a dark brown spot in the southern part of the affected area, indicating lower productivity, and as a thin green line in the northern parts, indicating increased productivity. This confirms earlier reports in which only the southern agricultural ecosystems were negatively affected by the heat wave, while the northern predominantly forest ecosystems rather benefited from the heat wave in terms of primary productivity (Flach et al., 2018).

Another example of an extreme event that we find in the PCs is the very wet November rainy season of 2006 in the

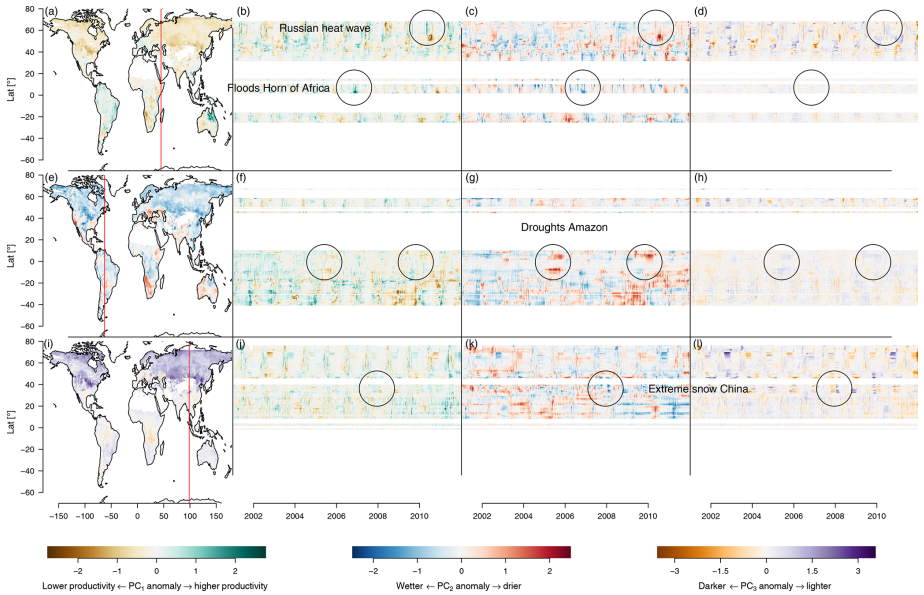

Figure 8. Anomalies of the first three principal components. The brown–green contrast shows the anomalies on PC_1, a relative low productivity or greening, respectively. The blue–red contrast shows the anomalies on PC_2, a relative wetness or dryness, respectively. The brown–purple contrast shows the anomaly on PC_3, a relative deviation in albedo. Panels **(a)**, **(e)**, and **(i)** are maps showing the anomalies of PC_1–PC_3, respectively, on 1 January 2001. Panels **(b)**, **(c)**, and **(d)** show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in **(a)**. The effects of the floods on the Horn of Africa (2006) and the Russian heat wave (2010) are highlighted by circles. Panels **(f)**, **(g)**, and **(h)** show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in **(e)**. Strong droughts in the Amazon during 2005 and 2010 can be observed as large red spots on the fringes of the Amazon basin (highlighted by circles). Panels **(j)**, **(k)**, and **(l)** show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in **(i)**. A strong snowfall event affecting central and southern China is marked as circles.

Horn of Africa after a very dry rainy season in the previous year. This event was reported to bring heavy rainfall and flooding events which caused an emergency for the local population but also increased ecosystem productivity (Nicholson, 2014). The rainfall event appears as green and blue spots in Fig. 8b and c, preceded by the drought events which appear as red and brown spots.

Figure 8f and g also show the strong drought events in the Amazon, particularly the droughts of 2005 and 2010 (Doughty et al., 2015; Feldpausch et al., 2016) appear strongly north and south of the Amazon basin. The central Amazon basin does not show these strong events, because the observable response of the ecosystem was buffered due to the large water storage capacity in the central Amazon basin.

Another extreme event that can be seen is the extreme snow and cold event affecting central and south China in January 2008, causing the temporary displacement of 1.7 million people and economic losses of approximately USD 21 billion (Hao et al., 2011). This event shows up clearly on PC_2 and PC_3 as cold and light anomalies, respectively (see Fig. 8k and f).

3.7 Single trajectories

Observing single trajectories can give insight into past events that happened at a certain place, such as extreme events or permanent changes in ecosystems. The creation of trajectories is an old method used by ecologists, mostly on species assembly data of local communities, to observe how the composition changes over time (e.g. Legendre et al., 1984; Ardisson et al., 1990). In this context, we observe how the states of the ecosystems inside the grid cell shift over time, which comprises a much larger area than a local community but is probably also less sensitive to very localized impacts than a community-level analysis. One of the main differences of the method applied here from the classical ecological indicators

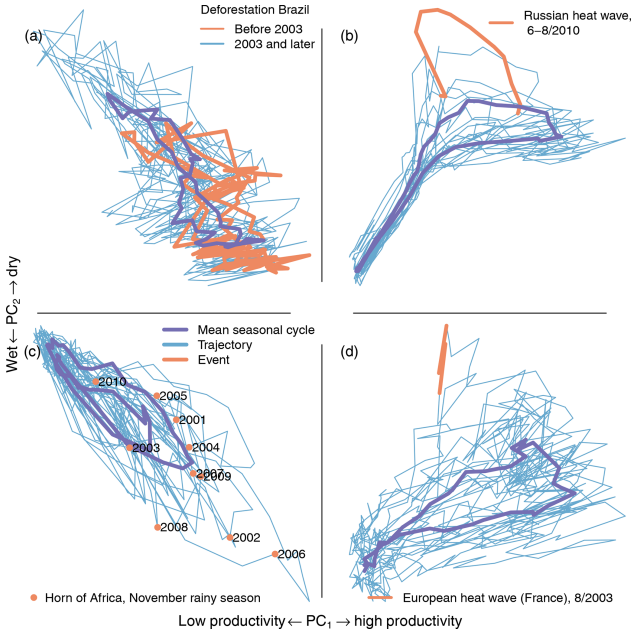

Figure 9. Trajectories of the first two principal components for single pixels. **(a)** Deforestation increases the seasonal amplitude of the first two PCs (Brazilian rain forest, 9.5° S, 63.5° W). The red line shows the trajectory before 2003 and the blue line the trajectory 2003 and later. A strong increase in seasonal amplitude can be observed after 2003. **(b)** The heat wave is clearly visible in the trajectory (red, Russian heat wave, summer 2010, 56° N, 45.5° E). **(c)** Rainfall in the short rainy season (November–December) influences agricultural yield and can cause flooding (extreme flooding after drought, November 2006, 3° N, 45.5° E). **(d)** The European heat wave in summer 2003 was one of the strongest on record (France, 47.2° N, 3.8° E). The mean seasonal cycle of the trajectories is shown in purple.

is that the trajectories observed here are embedded into the space spanned by a single global PCA, and therefore we can compare a much broader range of ecosystems directly.

The seasonal amplitude of the trajectory in the Brazilian Amazon increases due to deforestation and crop growth cycles. Figure 9a shows an area in the Brazilian Amazon in Rondônia (9.5° S, 63.5° W) which was affected by large-scale land use change and deforestation. It can be seen that the seasonal amplitude increases strongly after the beginning of 2003. This increased amplitude could be due to any of the following reasons or a combination of them: deforestation decreases water storage capability and dries out soils, causing larger variability in ecosystem productivity. Therefore, during periods of no rain, large-scale deforestation can cause a shift in local-scale circulation patterns, causing lower local precipitation (Khanna et al., 2017). Crop growth and harvest cause an increased amplitude in the cycle of productivity. An analysis of the trajectory can point to the nature of the change; however finding the exact causes for the change requires a deeper analysis.

The 2010 Russian heat wave has a very clear signal in the trajectories. Figure 9b shows the deviation of the trajectory during the Russian heat wave (red line) in an area east of Moscow (56° N, 45.5° E). In the southern grass- and croplands, the heat wave caused the productivity to drop significantly during summer due to a depletion of soil moisture. In the northern forested parts affected, the heat wave caused an increase in ecosystem productivity during spring due to higher temperatures combined with sufficient water availability. This shows the compound nature of this extreme event (see Fig. 8a and Flach et al., 2018). The analysis of the trajectory points directly towards the different types of extremes and responses that happened in the biosphere during the heat wave.

Variability of rainfall during the November rainy season in the Horn of Africa (3° N, 45.5° E, Fig. 9c) shows the tra-

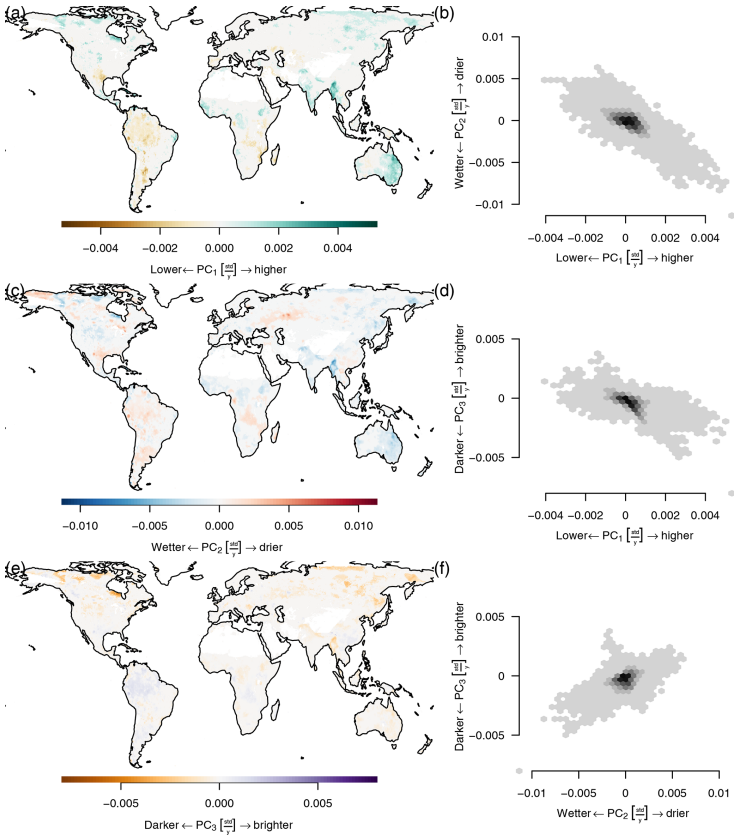

Figure 10. **(a, c, e)** Trends in PC_1–PC_3, respectively (2001–2011). **(b, d, f)** Bivariate distribution of trends. Trends were calculated using the Theil–Sen estimator. Panels **(a)**, **(c)**, and **(e)** show significant trends only ($p < 0.05$, Benjamini–Hochberg adjusted).

jectory and points in November of the observed time. The November rain has implications for food security because the second crop season depends on it. In 2006, the rainfall events were unusually strong and caused widespread flooding and disaster but also higher ecosystem productivity (see also Fig. 8). This was especially devastating because it followed a long drought that caused crop failures. Note also the two rainy seasons in the mean seasonal cycle (purple line in Fig. 9c).

The 2003 European heat wave is reflected in the trajectories just like the 2010 Russian heat wave. Figure 9d shows the trajectory during the August 2003 heat wave in Europe (France, 47.2° N, 3.8° E). The heat wave was unprecedented and caused large-scale environmental, health, and economic losses (Ciais et al., 2005; García-Herrera et al., 2010; Miralles et al., 2014). The 2010 heat wave was stronger than the 2003 heat wave but the strongest parts of the 2010 heat wave were in eastern Europe (see Fig. 8), while the center of the 2003 heat wave was located in France.

As we have seen here, observing single trajectories in reduced space can give us important insights into ecosystem states and changes that occur. While the trajectories can point us towards abnormal events, they can only be the starting

points for deeper analysis to understand the details of such state changes.

3.8 Trends in trajectories

The accumulation of CO_2 in the atmosphere should cause an increase in global productivity of plants due to CO_2 fertilization, while larger and more frequent droughts and other extremes may counteract this trend. Satellite observations and models have shown that during the last decades the world's ecosystems have greened up during growing seasons. This is explained by CO_2 fertilization, nitrogen deposition, climate change, and land cover change (Zhu et al., 2016; Huang et al., 2018; Anav et al., 2015). Tropical forests especially showed strong greening trends during the growing season.

General patterns of trends that can be observed are a positive trend (higher productivity) on the first principal component in many arctic regions. Many of these regions also show a wetness trend, with the notable exception of the western parts of Alaska, which have become drier. This is important, because wildfires play a major role in these ecosystems (Jolly et al., 2015; Foster et al., 2019). These changes are also accompanied by a decrease for PC_3 due to a loss in snow cover. A large-scale dryness trend can also be observed across large parts of western Russia. Increasing productivity can also be observed for large parts of the Indian subcontinent and eastern Australia. Negative trends in the first component can also be observed: they are generally smaller and appear in regions around the Amazon and the Congo Basin, but also in parts of western Australia. The main difference from previous analyses on the observations presented here is that Zhu et al. (2016), for example, looked only at trends during the growing season, while this analysis uses the entire time series to calculate the slope.

In the Amazon basin, we find a dryness trend accompanied by a decrease in productivity and a slight increase in PC_3. In the Congo Basin, we find a wetness trend and an increasing productivity in the northern parts, while the southern part and woodland south of the Congo Basin show a strong dryness trend with decreased productivity. This is different to the findings of Zhou et al. (2014), who found a widespread browning of vegetation in the entire Congo Basin for the April–May–June seasons during the period 2000–2012. The findings of Zhou et al. (2014) are not reflected in our data, especially compared to the areas surrounding the Congo Basin. We can find only minor browning effects inside the basin, and our findings are more in line with the global greening (Zhu et al., 2016), which shows a browning mostly outside the Congo Basin.

In eastern Australia we find a strong wetness and greenness trend which is due to Australia having a "millennium drought" since the mid-1990s with a peak in 2002 (Nicholls, 2004; Horridge et al., 2005) and extreme floods in 2010–2011 (Hendon et al., 2014).

Large parts of the Indian subcontinent show a trend towards higher productivity and an overall wetter climate. The greening trend in India happens mostly over irrigated cropland. However browning trends over natural vegetation have been observed but do not emerge in our analysis (Sarmah et al., 2018). A very notable greening and wetness trend can be observed in Myanmar due to an increase in intense rainfall events and storms, although the central part experienced some strong droughts at the same time (Rao et al., 2013). In Myanmar we also find one of the strongest trends in PC_3 outside of the Arctic.

In large parts of the Arctic, a trend towards higher productivity can be observed. Vegetation models attribute this general increase in productivity to CO_2 fertilization and climate change. The changes also cause changes to the characteristics of the seasonal cycles (Forkel et al., 2016). Stine et al. (2009) found a decreased seasonal amplitude of surface temperature over northern latitudes due to winter warming.

The seasonal amplitude of atmospheric CO_2 concentrations has been increasing due to climate change, causing longer growing seasons and changing vegetation cover in northern ecosystems (Forkel et al., 2016; Graven et al., 2013; Keeling et al., 1996). Therefore we checked for trends in the seasonal amplitude, but because each time series only consists of 11 values (one amplitude per year), after adjusting the p values for false discovery rate, we could not find a significant slope. However, there were many significant slopes with the unadjusted p values; see the appendix, Fig. E1.

Another way to detect changes to the biosphere consists in the detection of breakpoints, which has been applied successfully to detect changes in global normalized difference vegetation index (NDVI) time series (de Jong et al., 2011; Forkel et al., 2013) or generally to detect changes in time series (Verbesselt et al., 2010). A proof-of-concept analysis can be found in Fig. F1. We hope that applying this method to indicators instead of variables can detect a wider range of breakpoints analyzing a single time series.

3.9 Relations to other PCA-type analyses

One of the most popular applications of PCA in meteorology are EOFs, which typically apply PCA to a single variable, i.e., on a dataset with the dimensions lat × long × time, although EOFs can be calculated from multiple variables. EOFs can be calculated in S mode and R mode. If we matricize our data cube so that we have time in rows and lat × long × variables in columns, then S mode PCA works on the correlation matrix of the combined variable and space dimension. In T mode, the PCA works on the correlation matrix formed by the time dimension (Wilks, 2011). The PCA presented here works slightly differently. (1) We performed a different matricization (lat × long × time in rows and variables in columns) and then (2) the PCA works on the correlation matrix formed by the variables. Therefore in this framework we could call this a V mode PCA.

Ecological analyses usually use PCA with matrices of the shape object × descriptors. When calculating the PCA on the correlation matrix formed by the objects, then it is called a *Q* mode analysis. When the PCA is applied to the correlation matrix formed by the variables, then it is called an *R* mode analysis (Legendre and Legendre, 1998). The PCA carried out in this study is closest to an *R* mode analysis. In the present case the descriptors are the various data streams and the objects are the spatiotemporal pixels.

Using PCA as a method for dimensionality reduction means that we are assuming linear relations among features. A nonlinear method could possibly be more efficient in reducing the number of variables but would also have significant disadvantages. In particular, nonlinear methods typically require tuning specific parameters, objective criteria are often lacking, a proper weighting of observations is difficult, the methods are often not reversible, and it is harder to interpret the resulting indicators due to their nonlinear nature (Kraemer et al., 2018). The salient feature of PCA is that an inverse projection is well defined and allows for a deeper inspection of the errors, which is not the case for nonlinear methods which learn a highly flexible transformation that is hard to invert. Therefore interpretability of the transform in meaningful physical units in the input space is often not possible. In the machine-learning community, this problem is known as the "pre-imaging problem" (Mika et al., 1999; Arenas-Garcia et al., 2013) and is a matter of current research.

4 Conclusions

To monitor the complexity of the changes occurring in times of an increasing human impact on the environment, we used PCA to construct indicators from a large number of data streams that track ecosystem state in space and time on a global scale. We showed that a large part of the variability of the terrestrial biosphere can be summarized using three indicators. The first emerging indicator represents carbon exchange, the second indicator shows the availability of water in the ecosystem, while the third indicator mostly represents a binary variable that indicates the presence of snow cover. The distribution in the space of the first two principal components reflects the general limitations of ecosystem productivity. Ecosystem production can be limited by either water or energy.

The first three indicators can detect many well-known phenomena without analyzing variables separately due to their compound nature. We showed that the indicators are capable of detecting seasonal hysteresis effects in ecosystems, as well as breakpoints, e.g. large-scale deforestation. The indicators can also track other changes to the seasonal cycle such as patterns of changes to the seasonal amplitudes and trends in ecosystems. Deviations from the mean seasonal cycle of the trajectories indicate extreme events such as the large-scale droughts in the Amazon during 2005 and 2010 and the Russian heat wave of 2010. The events are detected in a similar fashion as with classical multivariate anomaly detection methods while directly providing information on the underlying variables.

Using multivariate indicators, we gain a high level overview of phenomena in ecosystems, and the method therefore provides an interesting tool for analyses where it is required to capture a wide range of phenomena which are not necessarily known a priori. Future research should consider nonlinearities, adding data streams describing other important biosphere variables (e.g. related to biodiversity and habitat quality), and including different subsystems, such as the atmosphere or the anthroposphere.

Appendix A: Description of variables

Variables used describing the biosphere can be found in Table 1. Here we provide a more complete description of all variables.

Black-sky albedo is the reflected fraction of total incoming radiation under direct hemispherical reflectance, i.e., direct illumination (Muller et al., 2011). This dataset is the broadband surface albedo including the visible, the near-infrared, and the shortwave-infrared spectrum (400–3000 nm). It is derived from the SPOT4-VEGETATION, SPOT5-VEGETATION2, and MERIS satellite sensors.

White-sky albedo is the reflected fraction of total incoming radiation under bihemispherical reflectance, i.e., diffuse illumination (Muller et al., 2011). Together with black-sky albedo it can be used to estimate the albedo under different illumination conditions. This dataset is the broadband surface albedo including the visible, the near-infrared, and the shortwave-infrared spectrum (400–3000 nm). This dataset is derived from the SPOT4-VEGETATION, SPOT5-VEGETATION2, and MERIS satellite sensors.

Evaporation ($\mathrm{mm\,d^{-1}}$) is the amount of water evaporated per day, depending on the amount of available water and energy. This dataset is based on the GLEAMv3 model (Martens et al., 2017), using satellite data from ESA CCI and SMOS to derive a number of variables.

Evaporative stress is modeled water stress for plants. Zero means that the vegetation has no water available for transpiration and 1 means that transpiration equals potential transpiration. This dataset is based on the GLEAMv3 model (Martens et al., 2017), using satellite data from ESA CCI and SMOS to derive a number of variables.

fAPAR is the fraction of absorbed photosynthetically active radiation, a proxy for plant productivity (Disney et al., 2016). This dataset is based on the GlobAlbedo dataset (http://globalbedo.org, last access: 23 April 2020) and the MODIS fAPAR and leaf area index (LAI) products.

Gross primary productivity (GPP) is ($\mathrm{gC\,m^{-2}\,d^{-1}}$) the total amount of carbon fixed by photosynthesis (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Terrestrial ecosystem respiration (TER) is ($\mathrm{gC\,m^{-2}\,d^{-1}}$) the total amount of carbon respired by the ecosystem, including autotrophic and heterotrophic respiration (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Net ecosystem exchange (NEE) is ($\mathrm{gC\,m^{-2}\,d^{-1}}$) the total exchange of carbon of the ecosystem with the atmosphere NEE = GPP − TER (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Latent energy (LE) is ($\mathrm{W\,m^{-2}}$) the amount of energy lost by the surface due to evaporation (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Sensible heat (H) is ($\mathrm{W\,m^{-2}}$) the amount of energy lost by the surface due to radiation (Tramontana et al., 2016). This dataset is derived from upscaling eddy covariance tower observations to a global scale using machine-learning methods.

Root-zone soil moisture is ($\mathrm{m^3\,m^{-3}}$) the moisture content of the root zone. This dataset is based on the GLEAMv3 model (Martens et al., 2017), using satellite data from ESA CCI and SMOS to derive a number of variables.

Surface soil moisture is ($\mathrm{mm^3\,mm^{-3}}$) the soil moisture content at the soil surface. This dataset is based on the GLEAMv3 model (Martens et al., 2017), using satellite data from ESA CCI and SMOS to derive a number of variables.

Appendix B: Time–space patterns of Components 1–3

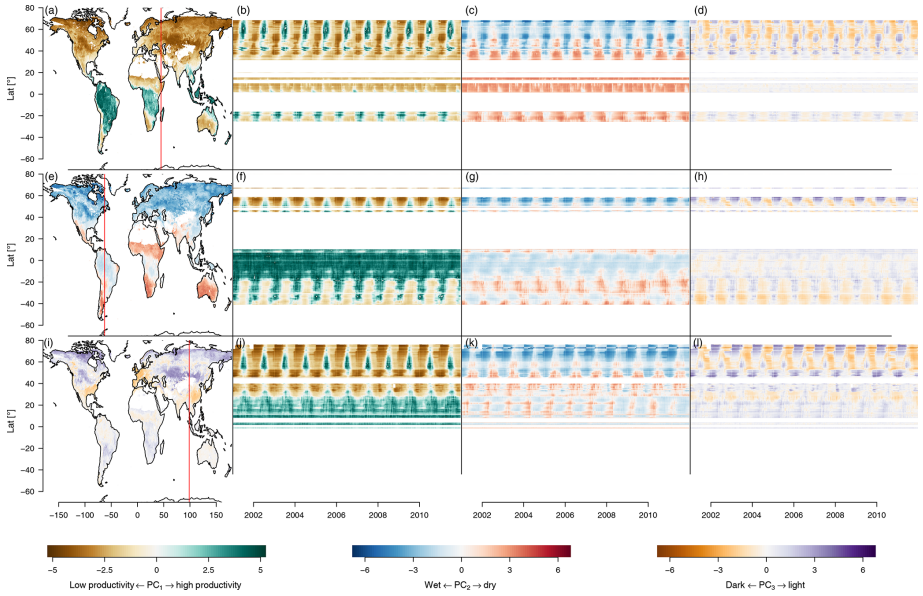

Figure B1. Time and space patterns of PC_1–PC_3, where the cut points are the same as in Fig. 8. The brown–green contrast shows the state of PC_1, from low to high productivity. The blue–red contrast shows the state of PC_2, from cold to dry. The brown–purple contrast shows the state of PC_3, from dark to light. Panels **(a)**, **(e)**, and **(i)** are maps showing the state of PC_1–PC_3, respectively, on 1 January 2001. Panels **(b)**, **(c)**, and **(d)** show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in **(a)**. Panels **(f)**, **(g)**, and **(h)** show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in **(e)**. Panels **(j)**, **(k)**, and **(l)** show longitudinal cuts of PC_1–PC_3, respectively, at the red vertical line in **(i)**.

Appendix C: Mean seasonal cycle extrema

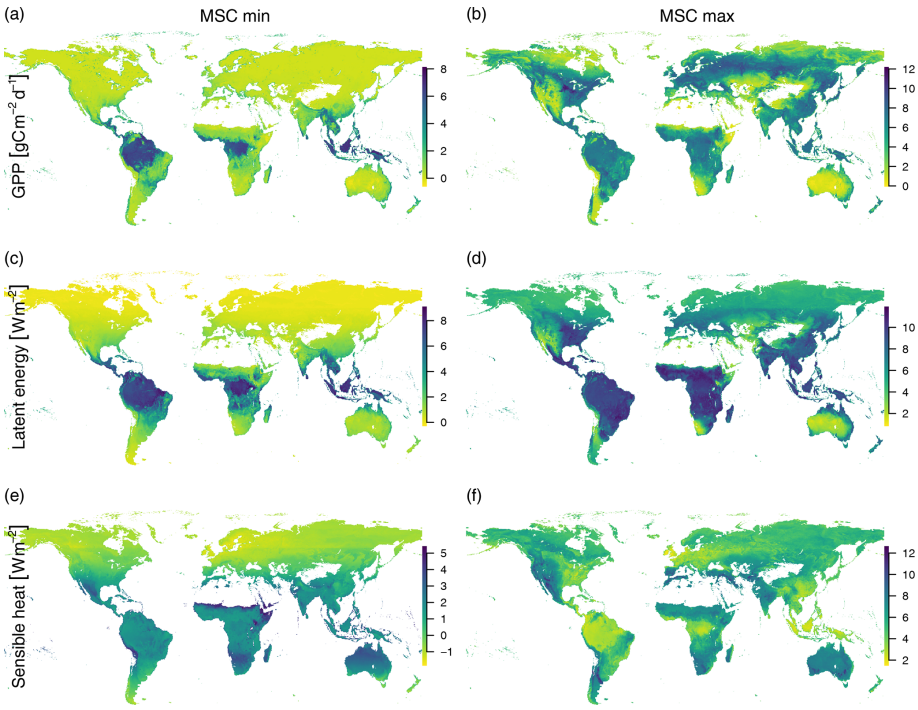

Figure C1. The minimum **(a, c, e)** and maximum **(b, d, f)** mean seasonal cycles of GPP **(a, b)**, latent heat **(c, d)**, and sensible heat **(e, f)**. This illustrates the similarity of possibly very different ecosystems in terms of productivity and limitations. During peak growing season, many midlatitude areas have a similar productivity and latent energy release as tropical rain forests **(b, d)**. The highest maximum seasonal sensible heat loss can be found in dry areas around the world and is lowest in areas with a wet climate such as tropical rain forests and maritime climates **(f)**.

Appendix D: Spatial covariances of the components

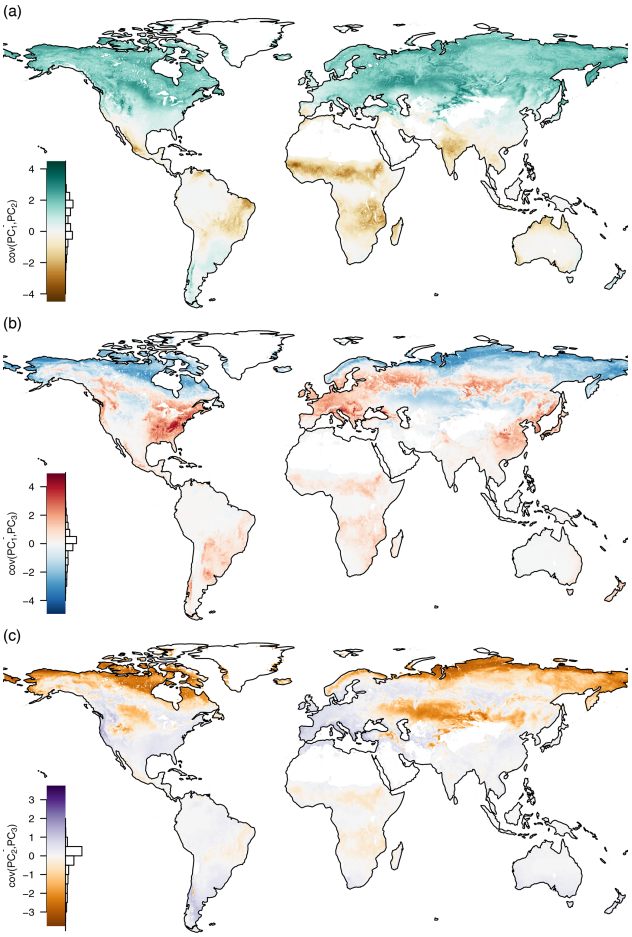

Figure D1. Pairwise covariances of the first three principal components mean seasonal cycles by space. **(a)** $\mathrm{cov}(\mathrm{PC}_1, \mathrm{PC}_2)$, **(b)** $\mathrm{cov}(\mathrm{PC}_1, \mathrm{PC}_3)$, and **(c)** $\mathrm{cov}(\mathrm{PC}_2, \mathrm{PC}_3)$. The bar charts show the distribution of the covariances. It can be seen that although two principal components are globally uncorrelated by their way of construction, they covary locally.

Appendix E: Changes in the seasonal amplitude

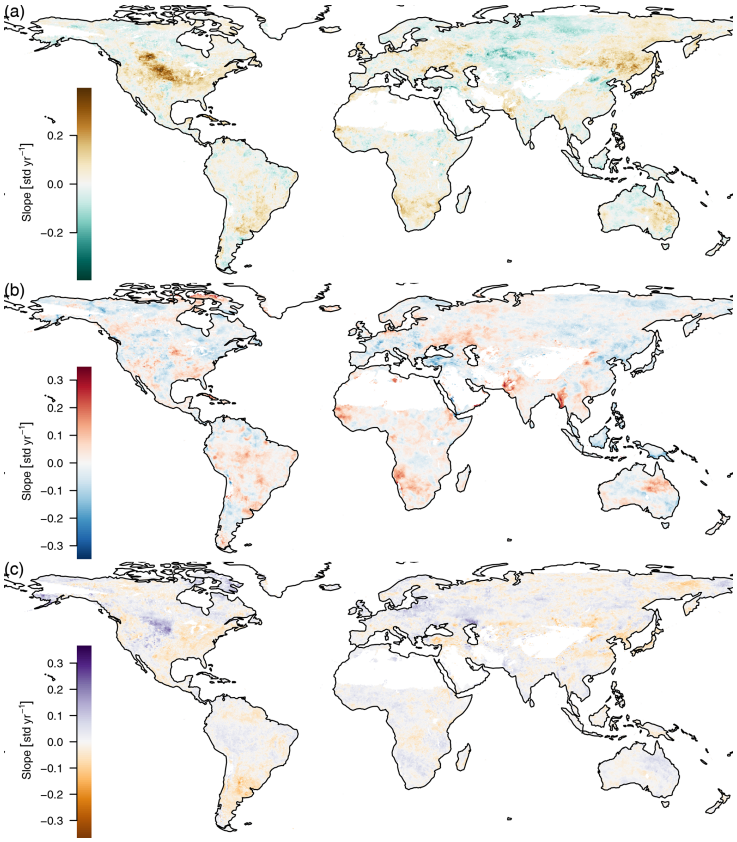

Figure E1. Trends in the amplitude of the yearly cycle, 2001–2011. Only Theil–Sen estimators for significant slopes ($p < 0.05$, *unadjusted*) are shown. Because there is only a single amplitude per year and therefore only 11 data points per time series, the Benjamini–Hochberg adjusted p values are not significant.

Appendix F: Breakpoints in trajectories

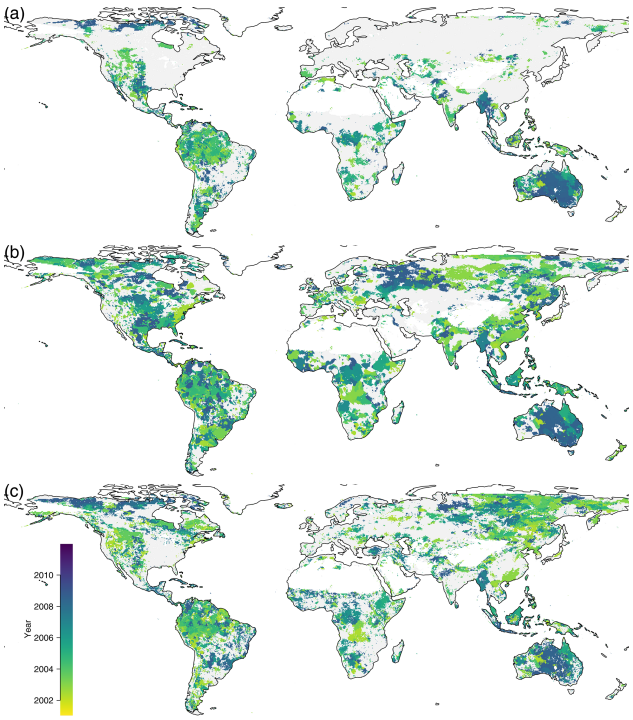

Figure F1. Breakpoint detection, **(a)** on PC_1, **(b)** on PC_2, and **(c)** on PC_3. The color indicates the year of the biggest breakpoint if a significant breakpoint was found, with gray if there was no significant breakpoint found.

As the environmental conditions change, due to climate change and human intervention, the local ecosystems may change gradually or abruptly. Detecting these changes is very important for monitoring the impact of climate change and land use change on the ecosystems. We applied breakpoint detection to the trajectories (Fig. F1).

Breakpoints on the first component were found in the entire Amazon, and the largest breakpoint is dated to the year 2005 during the large drought event. The entire eastern part of Australia shows its largest breakpoint towards the end of the time series because of a La Niña event, which caused lower temperatures and higher rainfall than usual during the years 2010 and 2011.

Code and data availability. The data are available and can be processed at https://www.earthsystemdatalab.net/index.php/interact/data-lab/, last access: 30 March 2020. The exact dataset and a docker container to reproduce the analysis can be found under https://doi.org/10.5281/zenodo.3733766 (Kraemer et al., 2020). The code to reproduce this analysis is available under https://doi.org/10.5281/zenodo.3733783 (Kraemer, 2020) and https://github.com/gdkrmr/summarizing_the_state_of_the_biosphere, last access: 23 April 2020.

Author contributions. GK and MDM designed the study in collaboration with MR and GCV. GK conducted the analysis and wrote the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank Fabian Gans and German Poveda for useful discussions. We thank Jake Nelson for proofreading a previous version of the manuscript. We thank Gregory Duveiller and the three anonymous reviewers for very helpful suggestions and Kirsten Thonicke for editorial advice that improved the manuscript greatly.

Financial support. This study is funded by the Earth System Data Lab – a project by the European Space Agency. Miguel D. Mahecha and Markus Reichstein have been supported by the Horizon 2020 EU project BACI under grant agreement no. 640176. Gustau Camps-Valls' work has been supported by the EU under the ERC consolidator grant SEDAL-647423.

The article processing charges for this open-access publication were covered by the Max Planck Society.

Review statement. This paper was edited by Kirsten Thonicke and reviewed by Gregory Duveiller and three anonymous referees.

References

Abatzoglou, J. T., Rupp, D. E., and Mote, P. W.: Seasonal Climate Variability and Change in the Pacific Northwest of the United States, J. Clim., 27, 2125–2142, https://doi.org/10.1175/JCLI-D-13-00218.1, 2014.

Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., Murray-Tortarolo, G., Papale, D., Parazoo, N. C., Peylin, P., Piao, S., Sitch, S., Viovy, N., Wiltshire, A., and Zhao, M.: Spatiotemporal patterns of terrestrial gross primary production: A review: GPP Spatiotemporal Patterns, Rev. Geophys., 53, 785–818, https://doi.org/10.1002/2015RG000483, 2015.

Aragão, L. E. O. C., Anderson, L. O., Fonseca, M. G., Rosan, T. M., Vedovato, L. B., Wagner, F. H., Silva, C. V. J., Silva Junior, C. H. L., Arai, E., Aguiar, A. P., Barlow, J., Berenguer, E., Deeter, M. N., Domingues, L. G., Gatti, L., Gloor, M., Malhi, Y., Marengo, J. A., Miller, J. B., Phillips, O. L., and Saatchi, S.: 21st Century Drought-Related Fires Counteract the Decline of Amazon Deforestation Carbon Emissions, Nat. Commun., 9, 146–149, https://doi.org/10.1038/s41467-017-02771-y, 2018.

Ardisson, P.-L., Bourget, E., and Legendre, P.: Multivariate Approach to Study Species Assemblages at Large Spatiotemporal Scales: The Community Structure of the Epibenthic Fauna of the Estuary and Gulf of St. Lawrence, Can. J. Fish. Aquat. Sci., 47, 1364–1377, https://doi.org/10.1139/f90-156, 1990.

Arenas-Garcia, J., Petersen, K. B., Camps-Valls, G., and Hansen, L. K.: Kernel Multivariate Analysis Framework for Supervised Subspace Learning: A Tutorial on Linear and Kernel Multivariate Methods, IEEE Signal Processing Magazine, 30, 16–29, https://doi.org/10.1109/MSP.2013.2250591, 2013.

Babst, F., Poulter, B., Bodesheim, P., Mahecha, M. D., and Frank, D. C.: Improved tree-ring archives will support earth-system science, Nat. Ecol. Evol., 1, 1–2, 2017.

Baldocchi, D. D.: How Eddy Covariance Flux Measurements Have Contributed to Our Understanding of Global Change Biology, Glob. Change Biol., 26, 242–260, https://doi.org/10.1111/gcb.14807, 2020.

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., and García-Herrera, R.: The Hot Summer of 2010: Redrawing the Temperature Record Map of Europe, Science, 332, 220–224, https://doi.org/10.1126/science.1201224, 2011.

Beisner, B., Haydon, D., and Cuddington, K.: Alternative Stable States in Ecology, Front. Ecol. Environ., 1, 376–382, https://doi.org/10.1890/1540-9295(2003)001[0376:ASSIE]2.0.CO;2, 2003.

Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, J. Roy. Stat. Soc. B, 57, 289–300, 1995.

Berger, M., Moreno, J., Johannessen, J. A., Levelt, P. F., and Hanssen, R. F.: ESA's Sentinel Missions in Support of Earth System Science, Remote Sens. Environ., 120, 84–90, https://doi.org/10.1016/j.rse.2011.07.023, 2012.

Blonder, B., Moulton, D. E., Blois, J., Enquist, B. J., Graae, B. J., Macias-Fauria, M., McGill, B., Nogué, S., Ordonez, A., Sandel, B., and Svenning, J.-C.: Predictability in Community Dynamics, Ecol. Lett., 20, 293–306, https://doi.org/10.1111/ele.12736, 2017.

Bowen, I. S.: The Ratio of Heat Losses by Conduction and by Evaporation from Any Water Surface, Phys. Rev., 27, 779–787, https://doi.org/10.1103/PhysRev.27.779, 1926.

Brown, R. L., Durbin, J., and Evans, J. M.: Techniques for Testing the.ournal of the Roy. Stat. Soc. B, 37, 149–192, 1975.

Cattell, R. B.: The Scree Test For The Number Of Factors, Multivar. Behav. Res., 1, 245–276, https://doi.org/10.1207/s15327906mbr0102_10, 1966.

Chapin, F. S., Woodwell, G. M., Randerson, J. T., Rastetter, E. B., Lovett, G. M., Baldocchi, D. D., Clark, D. A., Harmon, M. E., Schimel, D. S., Valentini, R., Wirth, C., Aber, J. D., Cole, J. J., Goulden, M. L., Harden, J. W., Heimann, M., Howarth, R. W., Matson, P. A., McGuire, A. D., Melillo, J. M., Mooney, H. A., Neff, J. C., Houghton, R. A., Pace, M. L., Ryan, M. G., Running, S. W., Sala, O. E., Schlesinger, W. H., and Schulze, E.-D.: Reconciling Carbon-Cycle Concepts, Terminol. Method. Ecosys., 9, 1041–1050, https://doi.org/10.1007/s10021-005-0105-7, 2006.

Chen, C., Park, T., Wang, X., Piao, S., Xu, B., Chaturvedi, R. K., Fuchs, R., Brovkin, V., Ciais, P., Fensholt, R., Tømmervik, H., Bala, G., Zhu, Z., Nemani, R. R., and Myneni, R. B.: China and India Lead in Greening of the World through Land-Use Management, Nature Sustainability, 2, 122–129, https://doi.org/10.1038/s41893-019-0220-7, 2019.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogée, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A., Chevallier, F., Noblet, N. D., Friend, A. D., Friedlingstein, P., Grünwald, T., Heinesch, B., Keronen, P., Knohl, A., Krinner, G., Loustau, D., Manca, G., Matteucci, G., Miglietta, F., Ourcival, J. M., Papale, D., Pilegaard, K., Rambal, S., Seufert, G., Soussana, J. F., Sanz, M. J., Schulze, E. D., Vesala, T., and Valentini, R.: Europe-Wide Reduction in Primary Productivity Caused by the Heat and Drought in 2003, Nature, 437, 529–533, https://doi.org/10.1038/nature03972, 2005.

de Jong, R., de Bruin, S., de Wit, A., Schaepman, M. E., and Dent, D. L.: Analysis of Monotonic Greening and Browning Trends from Global NDVI Time-Series, Remote Sens. Environ., 115, 692–702, https://doi.org/10.1016/j.rse.2010.10.011, 2011.

Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., Arneth, A., Balvanera, P., Brauman, K. A., Butchart, S. H. M., Chan, K. M. A., Garibaldi, L. A., Ichii, K., Liu, J., Subramanian, S. M., Midgley, G. F., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., Polasky, S., Purvis, A., Razzaque, J., Reyers, B., Chowdhury, R. R., Shin, Y.-J., Visseren-Hamakers, I., Willis, K. J., and Zayas, C. N.: Pervasive Human-Driven Decline of Life on Earth Points to the Need for Transformative Change, Science, 366, 6471, https://doi.org/10.1126/science.aax3100, 2019.

Disney, M., Muller, J.-P., Kharbouche, S., Kaminski, T., Voßbeck, M., Lewis, P., and Pinty, B.: A New Global fAPAR and LAI Dataset Derived from Optimal Albedo Estimates: Comparison with MODIS Products, Remote Sens., 8, 1–29, https://doi.org/10.3390/rs8040275, 2016.

Doughty, C. E., Metcalfe, D. B., Girardin, C. a. J., Amézquita, F. F., Cabrera, D. G., Huasco, W. H., Silva-Espejo, J. E., Araujo-Murakami, A., da Costa, M. C., Rocha, W., Feldpausch, T. R., Mendoza, A. L. M., da Costa, A. C. L., Meir, P., Phillips, O. L., and Malhi, Y.: Drought impact on forest carbon dynamics and fluxes in Amazonia, Nature, 519, 78–82, https://doi.org/10.1038/nature14213, 2015.

Feldpausch, T. R., Phillips, O. L., Brienen, R. J. W., Gloor, E., Lloyd, J., Lopez-Gonzalez, G., Monteagudo-Mendoza, A., Malhi, Y., Alarcón, A., Dávila, E. A., Alvarez-Loayza, P., Andrade, A., Aragao, L. E. O. C., Arroyo, L., C, G. A. A., Baker, T. R., Baraloto, C., Barroso, J., Bonal, D., Castro, W., Chama, V., Chave, J., Domingues, T. F., Fauset, S., Groot, N., Coronado, E. H., Laurance, S., Laurance, W. F., Lewis, S. L., Licona, J. C., Marimon, B. S., Marimon-Junior, B. H., Bautista, C. M., Neill, D. A., Oliveira, E. A., dos Santos, C. O., Camacho, N. C. P., Pardo-Molina, G., Prieto, A., Quesada, C. A., Ramírez, F., Ramírez-Angulo, H., Réjou-Méchain, M., Rudas, A., Saiz, G., Salomão, R. P., Silva-Espejo, J. E., Silveira, M., ter Steege, H., Stropp, J., Terborgh, J., Thomas-Caesar, R., van der Heijden, G. M. F., Martinez, R. V., Vilanova, E., and Vos, V. A.: Amazon Forest Response to Repeated Droughts, Global Biogeochem. Cy., 30, 964–982, https://doi.org/10.1002/2015GB005133, 2016.

Flach, M., Gans, F., Brenning, A., Denzler, J., Reichstein, M., Rodner, E., Bathiany, S., Bodesheim, P., Guanche, Y., Sippel, S., and Mahecha, M. D.: Multivariate anomaly detection for Earth observations: a comparison of algorithms and feature extraction techniques, Earth Syst. Dynam., 8, 677–696, https://doi.org/10.5194/esd-8-677-2017, 2017.

Flach, M., Sippel, S., Gans, F., Bastos, A., Brenning, A., Reichstein, M., and Mahecha, M. D.: Contrasting biosphere responses to hydrometeorological extremes: revisiting the 2010 western Russian heatwave, Biogeosciences, 15, 6067–6085, https://doi.org/10.5194/bg-15-6067-2018, 2018.

Folke, C., Carpenter, S., Walker, B., Scheffer, M., Elmqvist, T., Gunderson, L., and Holling, C.: Regime Shifts, Resilience, and Biodiversity in Ecosystem Management, Annu. Rev. Ecol. Evol. S., 35, 557–581, https://doi.org/10.1146/annurev.ecolsys.35.021103.105711, 2004.

Forkel, M., Carvalhais, N., Verbesselt, J., Mahecha, M., Neigh, C., Reichstein, M., Forkel, M., Carvalhais, N., Verbesselt, J., Mahecha, M. D., Neigh, C. S. R., and Reichstein, M.: Trend Change Detection in NDVI Time Series: Effects of Inter-Annual Variability and Methodology, Remote Sens., 5, 2113–2144, https://doi.org/10.3390/rs5052113, 2013.

Forkel, M., Migliavacca, M., Thonicke, K., Reichstein, M., Schaphoff, S., Weber, U., and Carvalhais, N.: Codominant Water Control on Global Interannual Variability and Trends in Land Surface Phenology and Greenness, Glob. Change Biol., 21, 3414–3435, https://doi.org/10.1111/gcb.12950, 2015.

Forkel, M., Carvalhais, N., Rodenbeck, C., Keeling, R., Heimann, M., Thonicke, K., Zaehle, S., and Reichstein, M.: Enhanced Seasonal CO_2 Exchange Caused by Amplified Plant Productivity in Northern Ecosystems, Science, 351, 696–699, https://doi.org/10.1126/science.aac4971, 2016.

Foster, A. C., Armstrong, A. H., Shuman, J. K., Shugart, H. H., Rogers, B. M., Mack, M. C., Goetz, S. J., and Ranson, K. J.: Importance of Tree- and Species-Level Interactions with Wildfire, Climate, and Soils in Interior Alaska: Implications for Forest Change under a Warming Climate, Ecol. Modell., 409, 108765, https://doi.org/10.1016/j.ecolmodel.2019.108765, 2019.

García-Herrera, R., Díaz, J., Trigo, R. M., Luterbacher, J., and Fischer, E. M.: A Review of the European Summer Heat Wave of 2003, Crit. Rev. Env. Sci. Tec., 40, 267–306, https://doi.org/10.1080/10643380802238137, 2010.

Gaumont-Guay, D., Black, T. A., Griffis, T. J., Barr, A. G., Jassal, R. S., and Nesic, Z.: Interpreting the Dependence of Soil Respiration on Soil Temperature and Water Content in a Boreal Aspen Stand, Agr. Forest Meteorol., 140, 220–235, https://doi.org/10.1016/j.agrformet.2006.08.003, 2006.

Graven, H. D., Keeling, R. F., Piper, S. C., Patra, P. K., Stephens, B. B., Wofsy, S. C., Welp, L. R., Sweeney, C., Tans, P. P., Kelley, J. J., Daube, B. C., Kort, E. A., Santoni, G. W., and Bent, J. D.: Enhanced Seasonal Exchange of CO2 by Northern Ecosystems Since 1960, Science, 341, 1085–1089, https://doi.org/10.1126/science.1239207, 2013.

Hao, Z., Zheng, J., Ge, Q., and Wang, W.: Historical Analogues of the 2008 Extreme Snow Event over Central and Southern China, Clim. Res., 50, 161–170, https://doi.org/10.3354/cr01052, 2011.

Hendon, H. H., Lim, E.-P., Arblaster, J. M., and Anderson, D. L. T.: Causes and Predictability of the Record Wet East Australian Spring 2010, Clim. Dynam., 42, 1155–1174, https://doi.org/10.1007/s00382-013-1700-5, 2014.

Higham, N. J.: The Accuracy of Floating Point Summation, SIAM J. Sci. Comput., 14, 783–799, https://doi.org/10.1137/0914050, 1993.

Horridge, M., Madden, J., and Wittwer, G.: The Impact of the 2002–2003 Drought on Australia, J. Policy Model., 27, 285–308, https://doi.org/10.1016/j.jpolmod.2005.01.008, 2005.

Huang, K., Xia, J., Wang, Y., Ahlström, A., Chen, J., Cook, R. B., Cui, E., Fang, Y., Fisher, J. B., Huntzinger, D. N., Li, Z., Michalak, A. M., Qiao, Y., Schaefer, K., Schwalm, C., Wang, J., Wei, Y., Xu, X., Yan, L., Bian, C., and Luo, Y.: Enhanced Peak Growth of Global Vegetation and Its Key Mechanisms, Nat. Ecol. Evol., 2, 1897–1905, https://doi.org/10.1038/s41559-018-0714-0, 2018.

IPBES: Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, Summary for policymakers, IPBES, 39 pp., 2019.

IPCC: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Tech. rep., IPCC, Geneva, Swizerland, 2014.

Ivits, E., Horion, S., Fensholt, R., and Cherlet, M.: Drought Footprint on European Ecosystems between 1999 and 2010 Assessed by Remotely Sensed Vegetation Phenology and Productivity, Glob. Change Biol., 20, 581–593, https://doi.org/10.1111/gcb.12393, 2014.

Jolly, W. M., Cochrane, M. A., Freeborn, P. H., Holden, Z. A., Brown, T. J., Williamson, G. J., and Bowman, D. M. J. S.: Climate-Induced Variations in Global Wildfire Danger from 1979 to 2013, Nat. Commun., 6, 7537, https://doi.org/10.1038/ncomms8537, 2015.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M.: The FLUXCOM Ensemble of Global Land-Atmosphere Energy Fluxes, Sci. Data, 6, 1–14, https://doi.org/10.1038/s41597-019-0076-8, 2019.

Keeling, C. D., Chin, J. F. S., and Whorf, T. P.: Increased Activity of Northern Vegetation Inferred from AtmosphericCO_2 Measurements, Nature, 382, 146–149, https://doi.org/10.1038/382146a0, 1996.

Kendall, M. G.: Rank Correlation Methods, Griffin, London, 202 pp., 1970.

Khanna, J., Medvigy, D., Fueglistaler, S., and Walko, R.: Regional dry-season climate changes due to three decades of Amazonian deforestation, Nat. Clim. Change, 7, 200–204, https://doi.org/10.1038/nclimate3226, 2017.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F.: World Map of the Köppen-Geiger climate classification updated, Meteorol. Z., 15, 259–263, https://doi.org/10.1127/0941-2948/2006/0130,

Kraemer, G.: gdkrmr/summarizing_the_state_of_the_biosphere v1.1.1, Zenodo, https://doi.org/10.5281/zenodo.3733783, 2020.

Kraemer, G., Reichstein, M., and Mahecha, M. D.: dimRed and coRanking – Unifying Dimensionality Reduction in R, R J., 10, 342–358, https://doi.org/10.32614/RJ-2018-039, 2018.

Kraemer, G., Camps-Valls, G., Reichstein, M., and Mahecha, M. D.: Summarizing the state of the terrestrial biosphere in few dimensions, Zenodo, https://doi.org/10.5281/zenodo.3733766, 2020.

Kuan, C.-M. and Hornik, K.: The Generalized Fluctuation Test: A Unifying View, Economet. Rev., 14, 135–161, https://doi.org/10.1080/07474939508800311, 1995.

Legendre, P. and Legendre, L.: Numerical Ecology: Second English Edition, Dev. Environ. Model., 20, 852 pp., 1998.

Legendre, P., Planas, D., and Auclair, M.-J.: Succession des communautés de gastéropodes dans deux milieux différant par leur degré d'eutrophisation, Can. J. Zool., 62, 2317–2327, https://doi.org/10.1139/z84-339, 1984.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J.: Tipping Elements in the Earth's Climate System, P. Natl. Acad. Sci. USA, 105, 1786–1793, https://doi.org/10.1073/pnas.0705414105, 2008.

Mahecha, M. D., Martínez, A., Lischeid, G., and Beck, E.: Nonlinear Dimensionality Reduction: Alternative Ordination Approaches for Extracting and Visualizing Biodiversity Patterns in Tropical Montane Forest Vegetation Data, Ecol. Inform., 2, 138–149, https://doi.org/10.1016/j.ecoinf.2007.05.002, 2007a.

Mahecha, M. D., Reichstein, M., Lange, H., Carvalhais, N., Bernhofer, C., Grünwald, T., Papale, D., and Seufert, G.: Characterizing Ecosystem-Atmosphere Interactions from Short to Interannual Time Scales, Biogeosciences, 4, 743–758, https://doi.org/10.5194/bg-4-743-2007, 2007b.

Mahecha, M. D., Gans, F., Sippel, S., Donges, J. F., Kaminski, T., Metzger, S., Migliavacca, M., Papale, D., Rammig, A., and Zscheischler, J.: Detecting Impacts of Extreme Events with Ecological in Situ Monitoring Networks, Biogeosciences, 14, 4255–4277, https://doi.org/10.5194/bg-14-4255-2017, 2017.

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J. F., Dorigo, W., Estupinan-Suarez, L. M., Gutierrez-Velez, V. H., Gutwin, M., Jung, M., Londoño, M. C., Miralles, D. G., Papastefanou, P., and Reichstein, M.: Earth system data cubes unravel global multivariate dynamics, Earth Syst. Dynam., 11, 201–234, https://doi.org/10.5194/esd-11-201-2020, 2020.

Mann, H. B.: Nonparametric Tests Against Trend, Econometrica, 13, 245–259, https://doi.org/10.2307/1907187, 1945.

Martens, B., Miralles, D. G., Lievens, H., Schalie, R. v. d., Jeu, R. A. M. d., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geosci. Model Dev., 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

Metzger, M. J., Bunce, R. G. H., Jongman, R. H. G., Sayre, R., Trabucco, A., and Zomer, R.: A High-resolution Bioclimate Map of the World: A Unifying Framework for Global Biodiversity Research and Monitoring, Glob. Ecol. Biogeogr., 22, 630–638, https://doi.org/10.1111/geb.12022, 2013.

Mika, S., Scholkopf, B., Smola, A., Muller, K., Scholz, M., and Ratsch, G.: Kernel PCA and De-Noising in Feature Spaces, in: Advances In Neural Information Processing Systems, edited by: Kearns, M. S., Solla, S. A., and Cohn, D. A., Vol. 11 of *Advances in Neural Information Processing Systems*, 12th Annual Conference on Neural Information Processing Systems (NIPS), Denver, CO, 30 November–5 December 1998, 536–542, 1999.

Miralles, D. G., Teuling, A. J., van Heerwaarden, C. C., and Vilà-Guerau de Arellano, J.: Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation,

Nat. Geosci., 7, 345–349, https://doi.org/10.1038/ngeo2141, 2014.

Muller, J.-P., Lewis, P., Fischer, J., North, P., and Framer, U.: The ESA GlobAlbedo Project for mapping the Earth's land surface albedo for 15 years from European sensors, Geophys. Res. Abstr., 13, EGU2011-10969, 2011.

Najafi, E., Pal, I., and Khanbilvardi, R.: Climate Drives Variability and Joint Variability of Global Crop Yields, Sci. Total Environ.t, 662, 361–372, https://doi.org/10.1016/j.scitotenv.2019.01.172, 2019.

Nasahara, K. N. and Nagai, S.: Review: Development of an in Situ Observation Network for Terrestrial Ecological Remote Sensing: The Phenological Eyes Network (PEN), Ecol. Res., 30, 211–223, https://doi.org/10.1007/s11284-014-1239-x, 2015.

Nicholls, N.: The Changing Nature of Australian Droughts, Climatic Change, 63, 323–336, https://doi.org/10.1023/B:CLIM.0000018515.46344.6d, 2004.

Nicholson, S. E.: A detailed look at the recent drought situation in the Greater Horn of Africa, J. Arid Environ., 103, 71–79, https://doi.org/10.1016/j.jaridenv.2013.12.003, 2014.

Papale, D., Black, T. A., Carvalhais, N., Cescatti, A., Chen, J., Jung, M., Kiely, G., Lasslop, G., Mahecha, M. D., Margolis, H., Merbold, L., Montagnani, L., Moors, E., Olesen, J. E., Reichstein, M., Tramontana, G., van Gorsel, E., Wohlfahrt, G., and Ráduly, B.: Effect of Spatial Sampling from European Flux Towers for Estimating Carbon and Water Fluxes with Artificial Neural Networks, J. Geophys. Res.-Biogeo., 120, 1941–1957, https://doi.org/10.1002/2015JG002997, 2015.

Parmesan, C.: Ecological and Evolutionary Responses to Recent Climate Change, Ann. Rev. Ecol. Evol. S., 37, 637–669, https://doi.org/10.1146/annurev.ecolsys.37.091305.110100, 2006.

Pearson, K.: On Lines and Planes of Closest Fit to Systems of Points in Space, Philos. Mag., 2, 559–572, 1901.

Piao, S., Wang, X., Park, T., Chen, C., Lian, X., He, Y., Bjerke, J. W., Chen, A., Ciais, P., Tømmervik, H., Nemani, R. R., and Myneni, R. B.: Characteristics, drivers and feedbacks of global greening, Nat. Rev. Earth Environ., 1, 14–27, https://doi.org/10.1038/s43017-019-0001-x, 2019.

Piao, S., Wang, X., Wang, K., Li, X., Bastos, A., Canadell, J. G., Ciais, P., Friedlingstein, P., and Sitch, S.: Interannual Variation of Terrestrial Carbon Cycle: Issues and Perspectives, Glob. Change Biol., 26, 300–318, https://doi.org/10.1111/gcb.14884, 2020.

Rao, M., Saw Htun, Platt, S. G., Tizard, R., Poole, C., Than Myint, and Watson, J. E. M.: Biodiversity Conservation in a Changing Climate: A Review of Threats and Implications for Conservation Planning in Myanmar, AMBIO, 42, 789–804, https://doi.org/10.1007/s13280-013-0423-5, 2013.

Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., Zscheischler, J., Beer, C., Buchmann, N., Frank, D. C., Papale, D., Rammig, A., Smith, P., Thonicke, K., van der Velde, M., Vicca, S., Walz, A., and Wattenbach, M.: Climate extremes and the carbon cycle, Nature, 500, 287–295, https://doi.org/10.1038/nature12350, 2013.

Renner, M., Brenner, C., Mallick, K., Wizemann, H.-D., Conte, L., Trebs, I., Wei, J., Wulfmeyer, V., Schulz, K., and Kleidon, A.: Using Phase Lags to Evaluate Model Biases in Simulating the Diurnal Cycle of Evapotranspiration: A Case Study in Luxembourg, Hydrol. Earth Syst. Sci., 23, 515–535, https://doi.org/10.5194/hess-23-515-2019, 2019.

Richardson, A. D., Braswell, B. H., Hollinger, D. Y., Burman, P., Davidson, E. A., Evans, R. S., Flanagan, L. B., Munger, J. W., Savage, K., Urbanski, S. P., and Wofsy, S. C.: Comparing Simple Respiration Models for Eddy Flux and Dynamic Chamber Data, Agr. Forest Meteorol., 141, 219–234, https://doi.org/10.1016/j.agrformet.2006.10.010, 2006.

Rosenfeld, D., Zhu, Y., Wang, M., Zheng, Y., Goren, T., and Yu, S.: Aerosol-Driven Droplet Concentrations Dominate Coverage and Water of Oceanic Low-Level Clouds, Science, 363, eaav0566, https://doi.org/10.1126/science.aav0566, 2019.

Sarmah, S., Jia, G., and Zhang, A.: Satellite View of Seasonal Greenness Trends and Controls in South Asia, Environ. Res. Lett., 13, 034026, https://doi.org/10.1088/1748-9326/aaa866, 2018.

Schimel, D. and Schneider, F. D.: Flux Towers in the Sky: Global Ecology from Space, New Phytol., 224, 570–584, https://doi.org/10.1111/nph.15934, 2019.

Schwartz, M. D.: Monitoring Global Change with Phenology: The Case of the Spring Green Wave, Int. J. Biometeorol., 38, 18–22, https://doi.org/10.1007/BF01241799, 1994.

Schwartz, M. D.: Green-Wave Phenology, Nature, 394, 839–840, https://doi.org/10.1038/29670, 1998.

Sen, P. K.: Estimates of the Regression Coefficient Based on Kendall's Tau, J. Am. Stat. Assoc., 63, 1379–1389, https://doi.org/10.2307/2285891, 1968.

Sippel, S., Reichstein, M., Ma, X., Mahecha, M. D., Lange, H., Flach, M., and Frank, D.: Drought, Heat, and the Carbon Cycle: A Review, Current Climate Change Reports, 4, 266–286, https://doi.org/10.1007/s40641-018-0103-4, 2018.

Sitch, S., Friedlingstein, P., Gruber, N., Jones, S. D., Murray-Tortarolo, G., Ahlström, A., Doney, S. C., Graven, H., Heinze, C., Huntingford, C., Levis, S., Levy, P. E., Lomas, M., Poulter, B., Viovy, N., Zaehle, S., Zeng, N., Arneth, A., Bonan, G., Bopp, L., Canadell, J. G., Chevallier, F., Ciais, P., Ellis, R., Gloor, M., Peylin, P., Piao, S. L., Le Quéré, C., Smith, B., Zhu, Z., and Myneni, R.: Recent Trends and Drivers of Regional Sources and Sinks of Carbon Dioxide, Biogeosciences, 12, 653–679, https://doi.org/10.5194/bg-12-653-2015, 2015.

Song, X.-P., Hansen, M. C., Stehman, S. V., Potapov, P. V., Tyukavina, A., Vermote, E. F., and Townshend, J. R.: Global Land Change from 1982 to 2016, Nature, 560, 639–643, https://doi.org/10.1038/s41586-018-0411-9, 2018.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C. A., Folke, C., Gerten, D., Heinke, J., Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B., and Sörlin, S.: Planetary Boundaries: Guiding Human Development on a Changing Planet, Science, 347, 1259855-1–1259855-10, https://doi.org/10.1126/science.1259855, 2015.

Stine, A. R., Huybers, P., and Fung, I. Y.: Changes in the Phase of the Annual Cycle of Surface Temperature, Nature, 457, 435–440, https://doi.org/10.1038/nature07675, 2009.

Tang, J., Baldocchi, D. D., and Xu, L.: Tree Photosynthesis Modulates Soil Respiration on a Diurnal Time Scale, Glob. Change Biol., 11, 1298–1304, https://doi.org/10.1111/j.1365-2486.2005.00978.x, 2005.

Theil, H.: A Rank-Invariant Method of Linear and Polynomial Regression Analysis, I, II, III, Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, 53, 386–392, 1950a.

Theil, H.: A Rank-Invariant Method of Linear and Polynomial Regression Analysis, I, II, III, Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, 53, 521–525, 1950b.

Theil, H.: A Rank-Invariant Method of Linear and Polynomial Regression Analysis, I, II, III, Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, 53, 1397–1412, 1950c.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

Van Der Maaten, L., Postma, E., and Van den Herik, J.: Dimensionality Reduction: A Comparative Review, J. Mach. Learn. Res., 10, 66–71, 2009.

van der Maaten, L., Schmidtlein, S., and Mahecha, M. D.: Analyzing Floristic Inventories with Multiple Maps, Ecol. Inform., 9, 1–10, https://doi.org/10.1016/j.ecoinf.2012.01.005, 2012.

Verbesselt, J., Hyndman, R., Newnham, G., and Culvenor, D.: Detecting Trend and Seasonal Changes in Satellite Image Time Series, Remote Sens. Environ., 114, 106–115, https://doi.org/10.1016/j.rse.2009.08.014, 2010.

Wilks, D. S.: Chapter 12 – Principal Component (EOF) Analysis, in: International Geophysics, edited by Wilks, D. S., vol. 100 of *Statistical Methods in the Atmospheric Sciences*, Academic Press, 519–562, https://doi.org/10.1016/B978-0-12-385022-5.00012-9, 2011.

Wingate, L., Ogée, J., Cremonese, E., Filippa, G., Mizunuma, T., Migliavacca, M., Moisy, C., Wilkinson, M., Moureaux, C., Wohlfahrt, G., Hammerle, A., Hörtnagl, L., Gimeno, C., Porcar-Castell, A., Galvagno, M., Nakaji, T., Morison, J., Kolle, O., Knohl, A., Kutsch, W., Kolari, P., Nikinmaa, E., Ibrom, A., Gielen, B., Eugster, W., Balzarolo, M., Papale, D., Klumpp, K., Köstner, B., Grünwald, T., Joffre, R., Ourcival, J.-M., Hellstrom, M., Lindroth, A., George, C., Longdoz, B., Genty, B., Levula, J., Heinesch, B., Sprintsin, M., Yakir, D., Manise, T., Guyon, D., Ahrends, H., Plaza-Aguilar, A., Guan, J. H., and Grace, J.: Interpreting Canopy Development and Physiology Using a European Phenology Camera Network at Flux Sites, Biogeosciences, 12, 5995–6015, https://doi.org/10.5194/bg-12-5995-2015, 2015.

Wolter, K. and Timlin, M. S.: El Niño/Southern Oscillation Behaviour since 1871 as Diagnosed in an Extended Multivariate ENSO Index (MEI.Ext), Int. J. Climatol., 31, 1074–1087, https://doi.org/10.1002/joc.2336, 2011.

Yan, T., Song, H., Wang, Z., Teramoto, M., Wang, J., Liang, N., Ma, C., Sun, Z., Xi, Y., Li, L., and Peng, S.: Temperature Sensitivity of Soil Respiration across Multiple Time Scales in a Temperate Plantation Forest, Sci. Total Environ., 688, 479–485, https://doi.org/10.1016/j.scitotenv.2019.06.318, 2019.

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C.: Strucchange: An R Package for Testing for Structural Change in Linear Regression Models, J. Stat. Softw., 7, 1–38, https://doi.org/10.18637/jss.v007.i02, 2002.

Zeng, N., Zhao, F., Collatz, G. J., Kalnay, E., Salawitch, R. J., West, T. O., and Guanter, L.: Agricultural Green Revolution as a Driver of Increasing Atmospheric CO_2 Seasonal Amplitude, Nature, 515, 394–397, https://doi.org/10.1038/nature13893, 2014.

Zhang, Q., Phillips, R. P., Manzoni, S., Scott, R. L., Oishi, A. C., Finzi, A., Daly, E., Vargas, R., and Novick, K. A.: Changes in Photosynthesis and Soil Moisture Drive the Seasonal Soil Respiration-Temperature Hysteresis Relationship, Agr. Forest Meteorol., 259, 184–195, https://doi.org/10.1016/j.agrformet.2018.05.005, 2018.

Zhou, L., Tian, Y., Myneni, R. B., Ciais, P., Saatchi, S., Liu, Y. Y., Piao, S., Chen, H., Vermote, E. F., Song, C., and Hwang, T.: Widespread Decline of Congo Rainforest Greenness in the Past Decade, Nature, 509, 86–90, https://doi.org/10.1038/nature13265, 2014.

Zhu, Z., Piao, S., Myneni, R. B., Huang, M., Zeng, Z., Canadell, J. G., Ciais, P., Sitch, S., Friedlingstein, P., Arneth, A., Cao, C., Cheng, L., Kato, E., Koven, C., Li, Y., Lian, X., Liu, Y., Liu, R., Mao, J., Pan, Y., Peng, S., Peñuelas, J., Poulter, B., Pugh, T. A. M., Stocker, B. D., Viovy, N., Wang, X., Wang, Y., Xiao, Z., Yang, H., Zaehle, S., and Zeng, N.: Greening of the Earth and Its Drivers, Nat. Clim. Change, 6, 791–795, https://doi.org/10.1038/nclimate3004, 2016.

Appendix E

Article: The Low Dimensionality of Development

Kraemer, G., Reichstein, M., Camps-Valls, G., Smits, J., and Mahecha, M. D. (2020). The Low Dimensionality of Development. *Social Indicators Research*, . doi:10.1007/s11205-020-02349-0

This article was written to apply the system state indicator framework on social data. In order to deal with missing values an extension of Isomap had to be developed.

The article was published in *Social Indicators Research* which has an impact factor of 1.703 and a 5 year impact factor of 2.353 and occupies a relative position of 35/104 in the category of "Social Sciences, Interdisciplinary" and 50/148 in the category of "Sociology" in the ISI Web of Knowledge datbase.

Social Indicators Research
https://doi.org/10.1007/s11205-020-02349-0

ORIGINAL RESEARCH

The Low Dimensionality of Development

Guido Kraemer[1,2,3,4] · **Markus Reichstein**[1] · **Gustau Camps-Valls**[3] · **Jeroen Smits**[5] · **Miguel D. Mahecha**[1,2,4]

Accepted: 20 April 2020

Abstract
The World Bank routinely publishes over 1500 "World Development Indicators" to track the socioeconomic development at the country level. A range of indices has been proposed to interpret this information. For instance, the "Human Development Index" was designed to specifically capture development in terms of life expectancy, education, and standard of living. However, the general question which independent dimensions are essential to capture all aspects of development still remains open. Using a nonlinear dimensionality reduction approach we aim to extract the core dimensions of development in a highly efficient way. We find that more than 90% of variance in the WDIs can be represented by solely five uncorrelated dimensions. The first dimension, explaining 74% of variance, represents the state of education, health, income, infrastructure, trade, population, and pollution. Although this dimension resembles the HDI, it explains much more variance. The second dimension (explaining 10% of variance) differentiates countries by gender ratios, labor market, and energy production patterns. Here, we differentiate societal structures when comparing e.g. countries from the Middle-East to the Post-Soviet area. Our analysis confirms that most countries show rather consistent temporal trends towards wealthier and aging societies. We can also find deviations from the long-term trajectories during warfare, environmental disasters, or fundamental political changes. The data-driven nature of the extracted dimensions complements classical indicator approaches, allowing a broader exploration of global development space. The extracted independent dimensions represent different aspects of development that need to be considered when proposing new metric indices.

Keywords Sustainable development · Sustainability indicators · Dimensionality reduction · PCA · Isomap

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11205-020-02349-0) contains supplementary material, which is available to authorized users.

✉ Guido Kraemer
guido.kraemer@uni-leipzig.de

Extended author information available on the last page of the article

Published online: 20 May 2020

1 Introduction

During the last decades, humanity has achieved on average longer life spans, decreased child mortality, better access to health care and economic growth (UNDP 2019). In emerging countries like China and India many people have escaped extreme poverty (less than 1.90 US$ per person per day) in the wake of persistent economic growth (UNDP 2016). To measure development, a wide range of variables are routinely made available by the World Bank, describing multiple facets of societal conditions. These "World Development Indicators" (WDIs, revision of 3/5/2018; The World Bank 2018b) have become a key data resource that today contains more than 1500 variables with annual values for most countries of the world.

A widely accepted method for assessing development consists in the construction of indicators (hereafter called "classical" indicators) based on expert knowledge that allow ranking countries by their development status and tracking them over time. A multitude of such classical indicators have been developed over the past few decades (Parris and Kates 2003; Shaker 2018; Ghislandi et al. 2018), focusing on different aspects of development. For instance, the United Nations Development Programme (UNDP) uses a Multidimensional Poverty Index, a Gender Development Index, a Gender Inequality Index, amongst others for reporting on human development (UNDP 2019). The UNDP's most prominent indicator is the Human Development Index (HDI), which is the geometric mean of indicators describing life expectancy, education, and income (UNDP 2019). However, there are many other efforts to produce relevant indicators, such as the Genuine Progress Index (Kubiszewski et al. 2013), the Global Footprint and Biocapacity indicators (McRae et al. 2016), and the POLITY scores (Marshall and Elzinga-Marshall 2017), to name just a few. These classical approaches are well suited for describing and communicating selected aspects of development, e.g. the HDI has been specifically developed "to shift the focus of development economics from national income accounting to people-centered policies" (UNDP 2018).

An alternative to approach to constructing indices consists in using purely data-driven methods, such as "Principal Component Analysis" (PCA; Pearson 1901) or "Factor Analysis" (FA; Spearman 1904). PCA linearly compresses a set of variables of interest. The resulting principal component or components represent the main dimensions of variability and can then be interpreted as an emerging indicator (OEDC 2008). This approach has been used to create indicators of well-being from sets of co-varying variables (Mazziotta and Pareto 2019). While PCA refers to a well defined method which tries to summarize the variance of an entire dataset, FA refers to a family of methods which assumes a multivariate linear model to explain the influences of a number of latent factors on observed variables. PCA and FA have been used extensively in the social sciences, e.g. to create indicators of well-being (Stanojević and Benčina 2019) or to construct wealth indices (Filmer and Scott 2012; Smits and Steendijk 2015). An advantage of such data-driven methods is that they follow well defined mathematical behaviors and are not subjective, while there is no well established method for the creation of classical indicators (Shaker 2018). A disadvantage of these methods is that they do not consider the polarity of the variables nor allow for expert based weighting (Mazziotta and Pareto 2019). A detailed comparison between classical indicators and data driven indicators can be found in SI Table 1.

The rationale for dimensionality reduction methods like PCA is that often the intrinsic dimension of a dataset is much lower than the number of variables describing it. In climate science, for example, a set of co-varying variables observed over a region in the equatorial

pacific can be compressed into the Multivariate ENSO Index (MEI, Wolter and Timlin 1993, 2011), to describe the state of the El Niño Southern Oscillation (ENSO)—the principal climate mode that determines e.g. food security in many regions of the world. In image vision, the number of main features from a set of images is much less than the number of pixels per image. For example Tenenbaum et al. (2000) shows that pictures taken from the same object at different angles have the viewing angle as the main features of the set of images. These main features are called "intrinsic dimensions" because they are sufficient to describe the essential nature of the entire dataset, the number of such intrinsic dimensions is called the "intrinsic dimensionality" of the dataset (Bennett 1969).

Development is a complex concept though, which is reflected in the large number of variables included in the WDI database. The large number of indicators let us expect substantial redundant information (Shaker 2018; Rickels et al. 2016). This issue has also been discussed in the context of the Sustainable Development Goals (SDGs; The World Bank 2018a). Since their introduction by the United Nations in 2015, the SDGs have become a widely accepted framework to guide policymakers. Today 17 SDGs address the issues of poverty, hunger, health, education, climate change, gender inequality, water, sanitation, energy, urbanization, environment and social justice. To monitor the SDGs, 169 specific targets have been developed which are measured using 232 different indicators included in the WDIs (The World Bank 2018a; United Nations General Assembly 2017a), leading to substantial interactions across and within the targets that need to be analyzed (Costanza et al. 2016). Hence, the question emerges how to extract the key information jointly contained in the WDIs that leads to a succinct, objective, and tangible picture of development.

In this paper, we aim to elucidate the most important dimensions of development contained in the WDI dataset, using a data-driven approach. Specifically, we aim to answer the question, how many independent indicators are necessary to summarize development space and what is their interpretation. We exploit the potential of nonlinear dimensionality reduction to identify dimensions that represent these (typically mutually dependent) variables, while preserving relevant properties of the underlying data. The rationale is that we expect strong interactions between the different WDIs which may not be linear.

Understanding what intrinsic dimensionality our current indicators of development have, could have important implications for policy makers. If the intrinsic dimensionality of development proves to be high, one would indeed need to track many indicators synchronously to understand the interplay of different aspects of development. On the contrary, in the case of a low-dimensional development space, it would be sufficient to track either the emerging dimensions, or the closely related variables to monitor development across countries and time. In fact there is already substantial evidence that supports our hypothesis of a low-dimensional development space. For instance Pradhan et al. (2017) found strong correlations between all SDGs, suggesting that the intrinsic dimensionality of the SDGs is relatively low, but this has not been quantified yet.

This article is divided into five sections. Section 2 presents a data-driven approach to extract nonlinear components from the WDI database, Sect. 3 presents the resulting dimensions, their interpretations, global distributions, trends and trajectories. Section 4 discusses the relation of the indicators produced by the method presented here with previous indicator approaches, and finally Sect. 5 gives some concluding remarks.

2 Data and Methods

2.1 Data

To understand the structure and dimensionality of development we rely on the WDI dataset, which is the primary World Bank collection of development indicators, compiled from officially-recognized international sources. The WDIs comprise a total of 1549 variables with yearly data between 1960 and 2016 for 217 countries. As such, it represents the most current and accurate global development database available (The World Bank 2018b).

Even though the WDI dataset is the most comprehensive set of development indicators available, it contains many missing values. Only for the most developed countries the dataset is (nearly) complete. For many other countries—particularly low and middle income countries—many indicators are partly or completely missing. This is problematic, as for most dimension reduction methods a dataset without missing observations is required. To make our analyses possible, we therefore had to select a subset of indicators, countries and years with few missing observations and to fill in the remaining missing observations using gapfilling techniques (see next section). To avoid arbitrariness of the subset selection, a scoring approach was used (see Sect. 2.2) and the 1000 subsets with the highest scores were selected. These 1000 subsets contained a total of 621 variables, 182 countries and the years ranging from 1990 to 2016. The subsets cover almost all categories of variables. The categories with their respective number of variables in the entire WDI dataset and the subsets are "Economic Policy & Debt" (120 out of 518), "Education" (73 out of 151), "Environment" (74 out of 138), "Financial Sector" (29 out of 54), "Gender" (1 out of 21), "Health" (123 out of 226), "Infrastructure" (19 out of 41), "Poverty" (0 out of 24), "Private Sector & Trade" (103 out of 168), "Public Sector" (31 out of 83), and "Social Protection & Labor" (48 out of 161). Jointly these subsets are representative for the original dataset while avoiding large gaps.

2.2 Gapfilling

The dimensionality reduction approach we have chosen (see Sect. 2.3) relies on a full matrix of distances between the different country–year data points. However, given the large amount of data gaps this global distance matrix cannot be computed directly. In the following, we develop an approach to find subsets of the WDI database which we can gapfill and use for estimating distances among data points.

In order to choose subsets of the WDI database covering a wide range of WDIs, countries, and years, but also having as few missing values as possible, the following method was applied: A series of subset was created from the full WDI dataset using a combination of thresholds for the maximum fraction of missing values for the WDIs, f_v, and countries, f_c, as well as a starting year, y_{start}, and an ending year, y_{end}. We assigned a score to each of the resulting subsets by using a grid search over the parameters, $f_v, f_c \in (0.05, 0.15, \ldots, 0.65)$ and $y_{\text{start}}, y_{\text{end}} \in (1960, 1961, \ldots, 2017), y_{\text{start}} < y_{\text{end}}$. The size of this parameter space is 80997, each with a different combination of missing value thresholds and starting and ending year combinations. The 1000 subsets with the highest scores were finally chosen to build the global distance matrix. For an overview of the entire method, see Fig. 1.

Each subset was created from the full WDI dataset by choosing consecutive years with starting year, y_{start}, and ending year, y_{end}; WDIs with a higher missing value fraction, p_v, than the corresponding threshold were dropped ($p_v > f_v$). Then, countries with

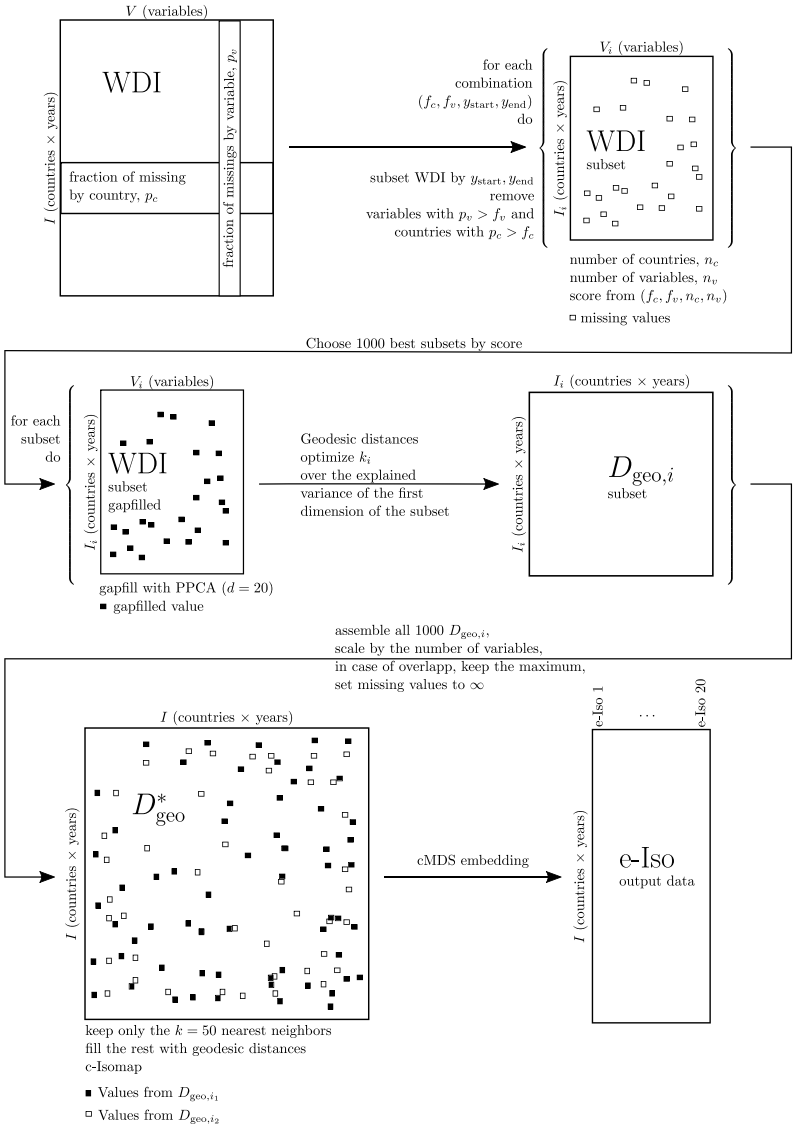

Fig. 1 Schematic presentation of the ensemble Isomap (e-Iso) algorithm, for details see text

higher missing value fractions, p_c, than the corresponding threshold were dropped as well ($p_c > f_c$). The number of remaining countries, n_c, and WDIs, n_v, was recorded and the resulting subsets were filtered to retain more observations (the number of countries times

the number of years) than variables, leaving a total of 77,610 subsets of the WDI for score calculation.

To account for different scales of the parameters, the values had to be rescaled, i.e. we calculated n'_v from n_v by scaling the values from subsets linearly to a minimum of 0 and a maximum of 1, analogously for n'_c, f'_c, and f'_v. The final score was then calculated as

$$\text{score} = \sqrt{n'_v n'_c} - \sqrt{f'_c f'_v}.$$

This score calculates the geometric means of the variables of interest. The geometric mean has the advantage over the arithmetic mean that it is very sensitive to single bad values. As we want to maximize the number of countries and WDIs chosen and have as few missing values possible, the final score is the difference between the geometric means. For further processing, the subsetted WDI data matrices with the 1000 highest scores were selected.

Finally the subsetted WDI data matrices with the 1000 highest scores were selected and a gapfilling procedure using Probabilistic PCA (Stacklies et al. 2007) was performed on the centered and standardized (z-transformed) variables using the leading 20 dimensions.

2.3 Dimensionality Reduction

Dimensionality reduction describes a family of multivariate methods that find alternative representations of data by constructing linear or, in our case, nonlinear combinations of the original variables so that important properties are maintained in as few dimensions as possible. A plethora of algorithms is currently available for dimensionality reduction, both linear and nonlinear (Arenas-Garcia et al. 2013; Van Der Maaten et al. 2009; Kraemer et al. 2018), but PCA is dominating in applied sciences because of ease of use and interpretation.

One method to find an embedding from a known distance matrix is "classical Multidimensional Scaling" (CMDS; Torgerson 1952), this method is equivalent to PCA if the distance matrix is computed from the observations using Euclidean distance. CMDS finds coordinates in a reduced Euclidean space of dimension i minimizing

$$\|\tau(D) - \tau(D_i)\|_2,$$

where D is the matrix of Euclidean distances of observations and D_i the matrix of Euclidean distances of the embedded points. $\tau(D) = -\frac{1}{2}HSH$, is the "double centering operator", with $S = [D_{ij}^2]$, $H = [\delta_{ij} - \frac{1}{n}]$, and $\|X\|_2 = \sqrt{\sum_{ij} X_{ij}^2}$ the L_2-norm. CMDS and therefore PCA tend to maintain the large scale gradients of the data and cannot cope with nonlinear relations between the covariates.

"Isometric Feature Mapping" (Isomap; Tenenbaum et al. 2000) extends CMDS, but instead of Euclidean distances, it respects geodesic distances, i.e. the distances measured along a manifold of possibly lower dimensionality,

$$\|\tau(D_{\text{geo}}) - \tau(D_i)\|_2.$$

Specifically, Isomap uses geodesic distances, $D_{\text{geo}} = [d_{\text{geo}}(x_i, x_j)]$, which are the distances between two points following a k-nearest neighbor graph of points sampled from the manifold.

Isomap is guaranteed to recover the structure of nonlinear manifolds whose intrinsic geometry is that of a convex region of Euclidean space (Tenenbaum et al. 2000). Isomap

unfolds curved manifold which makes the method more efficient than PCA in reducing the number of necessary dimensions in the presence of nonlinearities.

To construct the geodesic distances, a graph is created by connecting each point to its k nearest neighbors and distances are measured along this graph. If the data samples the manifold well enough, then the distances along the graph will approximate the geodesic distances along the manifold. The value of k will determine the quality of the embedding and has to be tuned.

We applied Isomap on the 1000 previously generated subsets of the WDI database. To find the optimum value k for each subset, k_i, Isomap was calculated first with $k_i = 5$ and the residual variance for the embedding of the first component was calculated (see below). This process was repeated increasing the values of k_i by 5 in each step until there was no decrease in the residual variance for the first component any more (Mahecha et al. 2007). In order to get an intuition of Isomap, we recommend the original publication of the Isomap method (Tenenbaum et al. 2000) which contains an excellent didactic explanation of the method.

2.4 Ensemble PCA and Ensemble Isometric Feature Mapping

An observation consists of a country name and year. To calculate a linear embedding (ensemble PCA) over the union of all countries, years and variables chosen before, we used a Probabilistic PCA ($d = 80$, where d is the number of dimensions used in the probabilistic PCA) to gapfill all the observations and variables occurring in the subsets of the WDI dataset and applied a normal PCA to the gapfilled dataset.

We developed "Ensemble Isometric Feature Mapping" (e-Isomap) to produce the final nonlinear embedding based on the different gapfilled subsets of data. E-Isomap combines $m = 1000$ geodesic distance matrices created from the subsets of the previous step and constructs an global ensemble geodesic distance matrix, D^*_{geo}, from the geodesic distance matrices of the m Isomaps.

Let the total set of observations be $I = \{1, \ldots, n\}$ (a country–year combination) and the observed variables $V = \{1, \ldots, p\}$ (the WDIs). We first perform one Isomap $i \in \{1, \ldots, m\}$ per subset of I and V, I_i and V_i respectively, where $|V_i|$ is the number of variables for Isomap i. The geodesic distance matrix for Isomap i is $D_{\text{geo},i} = (d_{\text{geo},i}(x_j, x_k))_{j,k}$ with $j, k \in I_i$. If a pair of observations (x_j, x_k) does not occur in Isomap i, it is treated as a missing value. First the geodesic distance matrices are scaled element-wise to account for the different number of variables used,

$$d'_{\text{geo},i}(x_j, x_k) = d_{\text{geo},i}(x_j, x_k)\sqrt{\frac{|V|}{|V_i|}},$$

which are then combined into a single geodesic distance matrix D^*_{geo} by using the maximum distance value,

$$d^*_{\text{geo}}(x_j, x_k) = \max_i d'_{\text{geo},i}(x_j, x_k).$$

Missing values are ignored if all values are missing for a pair (x_j, x_k) and they are treated as infinite distances. Taking the maximum avoids short-circuiting distances and as long as there are few missing values. This provides an accurate approximation of the internal distances.

Finally the k nearest neighbor graph G is constructed from the distance matrix, and each edge $\{x_i, x_j\}$ is weighted by $\frac{|x_i - x_j|}{\sqrt{M(i)M(j)}}$, where $M(i)$ is the mean distance of x_i to its k nearest neighbors. This last step is called c-Isomap (Silva and Tenenbaum 2003) and it contracts sparsely sampled regions of the manifold and expands densely sampled regions, the c-Isomap step proved to give a more evenly distributed embedding. Finally the geodesic distances are calculated on G and classical scaling is performed to find the final embeddings.

2.5 Quality Measurement of an Embedding and Influence of Variables

The quality for the embedding is estimated by calculating the residual variance (Tenenbaum et al. 2000) computed as

$$\text{residual variance}_i = 1 - r^2(\hat{D}, D_i) = 1 - \text{explained variance}_i,$$

where D_i is the matrix of Euclidean distances of the first i embedded components and $\hat{D}$ is the matrix of Euclidean distances for PCA and the matrix of geodesic distances for Isomap in original space. Note that because D_i and $\hat{D}$ are symmetric, we only use one triangle for the calculation of the residual variance. This notion of explained variance is different from the one usually used for PCA, which is derived from the eigenvalue spectrum, but the measure used here has the advantage that it gives comparable results for arbitrary data such as the HDI and Isomap.

To assess the influence of single variables on the final e-Isomap dimensions, we calculated the distance correlation (dcor, Szkely et al. 2007), which is a measure of dependence between variables that takes nonlinearities into account. Due to the strong nonlinearities in the dataset and the embedding method, a simple linear correlation would not have provided sufficient information about the relationships between variables and the embedding dimensions.

3 Results

3.1 Required Number of Dimensions

Our results suggest that the "development space" described by the WDI data is of low intrinsic dimensionality. Using e-Isomap we needed five dimensions only to explain 90% of the variance of global development (see Fig. 2). The first dimension alone explains 74% of the variance in the WDI data; Dimension 2 explains 9.9% of the variance and dimensions 3–5 explain less than 3% of the variance each. Although the explained variance of dimensions 2–5 seems small compared to that of the first dimension, each of these dimensions still represents a distinct, well defined and highly significant aspect of development, as we will show later. Therefore the raw variances should not be used as the sole measure to discard dimensions.

The finding that such a high compression can be achieved with e-Isomap indicates that the WDIs are highly interdependent and that the underlying processes are highly nonlinear (see Fig. 2). This is also confirmed by an analogous analysis using linear PCA which cannot compress the data with the same efficiency: the first PCA dimension only explains 10% of the variance, and 12 dimensions are required to express more than 90% of the variance. The cumulative explained variances for the first five e-Isomap dimensions are 74%, 84%,

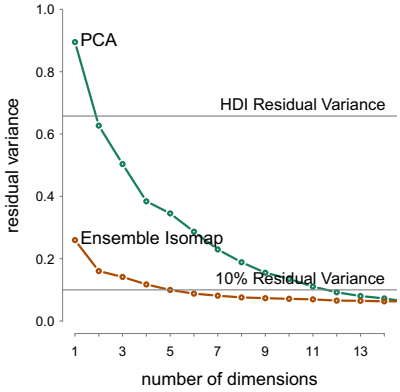

Fig. 2 The residual variance for the first 14 components. The circled lines represent the residual variance of the Ensemble Isomap and the PCA. Isomap is much more efficient in compressing dimensionality of the data requiring only 5 components to describe more than 90% of the variance, while PCA requires 12 components to describe 90% of variance. The upper grey horizontal line represents the residual variance for the HDI (66%) and the lower one the 10% residual variance boundary

86%, 88%, and 90%, which is much more than the respective PCA dimensions (10%, 37%, 50%, 61%, and 65%).

To understand if the HDI can compress the data in the same way, we compute the variance of the HDI in the same way. We find that the HDI captures 34% of the variance (see Fig. 2), which is less than half of the variance captured by the first dimension extracted via nonlinear dimensionality reduction but more than three times the variance explained by the first PCA dimension. If the target is reducing the WDI data to a single dimension, the best performing method is e-Isomap, followed by the HDI, while PCA does not perform this task very well. In other words, the first e-Isomap dimension seems to be a more powerful summary of the WDI data than the HDI.

3.2 Intrinsic Dimensions of Development

Our results suggest that the dimensions resulting from the e-Isomap can be indeed interpreted analogously to traditional indicators of development. The main difference from classical indicators is that these dimensions emerge directly from the data. Hence, the interpretation of these indicators has to be achieved *a posteriori*. We also find that the relationship between the WDIs and the dimensions is highly nonlinear (see Fig. 3) requiring the use of nonlinear measurements of correlation. Here we relate the extracted dimensions to the original data using distance correlation. See Fig. 4, for a complete and interactive table in the supporting information.[1]

We find that dimension 1 essentially represents progress in education, life expectancy, health, and relates to the population pyramid (see Fig. 4). Additionally, dimension 1 is associated with infrastructure and income-related indicators. Other indicators that strongly correlate with this dimension are related to pollution and primary production and include tariffs and imports as well as trade, the climate impact of GDP (gross domestic product), and development aid received. Because dimension 1 embraces education, health, and life expectancy, it is conceptually similar to the HDI. In fact, dimension 1 has a strong nonlinear correlation with the HDI (dcor = 0.93), and can be

[1] http://bgc-jena.mpg.de/~gkraemer/consolidated_cor_table.

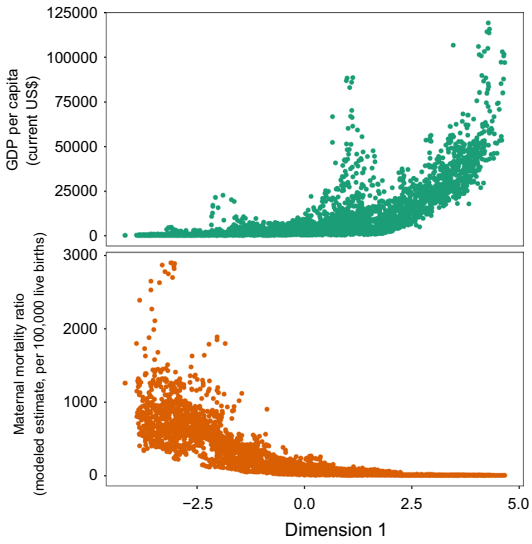

Fig. 3 Illustrating the nonlinear relation between dimension 1 and GDP per capita and maternal mortality rates. Top: There is a positive correlation between GDP per capita and dimension 1. On the positive end of dimension 1 the per capita income increases strongly, while it increases very slowly on the negative side of dimension 1. Bottom: There is a negative correlation between the maternal mortality rate and dimension 1. The maternal mortality rate decreases strongly on the negative end but does not decrease any more on the positive end

interpreted as a measure of development *sensu* HDI, even though it includes much more than the aspects measured by the HDI. We also find that the correlation is much lower for most sub-Saharan countries (Fig. S2).

Dimension 2 (9.9% of the variance) is strongly related to gender ratios in the general population and the labor market, as well as primary energy production and consumption and the fraction of 25–29 year old people. This dimension spans a gradient between the extremes of dimension 1 and former Soviet allied countries on one end, and rich mostly oil exporting nations on the other end (see Fig. 5). On the positive extreme on this axis are countries that have a very high participation of women in the labor market (e.g Mozambique has the highest participation of women in the labor force with around 55%, similar to countries like Lithuania with a rate of approx 50%) on the negative extreme we can find countries with a very low participation of women in the labor market: Rich countries like the United Arabian Emirates have a female labor force of around 12%, just as poorer countries like Yemen that has a participation rate of women of around 8%, and low death rates. Crude death rates also correlate well with this dimension and do not separate regions, e.g. Latvia in 1994 had a crude death rate of 16.6/1000 people, Denmark in 1993 a crude death rate of 12.1 per 1000 people, while similar crude death rates can be found in undeveloped countries (Democratic Republic of the Congo, 1996, 16.655 death per 1000 people; or Liberia, 2005, 12.128 deaths per 1000 people), on the low extreme we find mostly rich oil exporting nations (e.g. Qatar and the United Arabian Emirates with values around 1.5 deaths per 1000 people).

The third to fifth dimensions explain much less variance but are still important in that they account for variables not found in the first two dimensions: Dimension 3 (1.9% of the variance) is a labor market gradient representing descriptors like ratios of labor force, employment, and unemployment. Dimension 4 (2.4% of the variance) summarizes homicide rates, methane emissions and food exports. Dimension 5 (1.8% of the

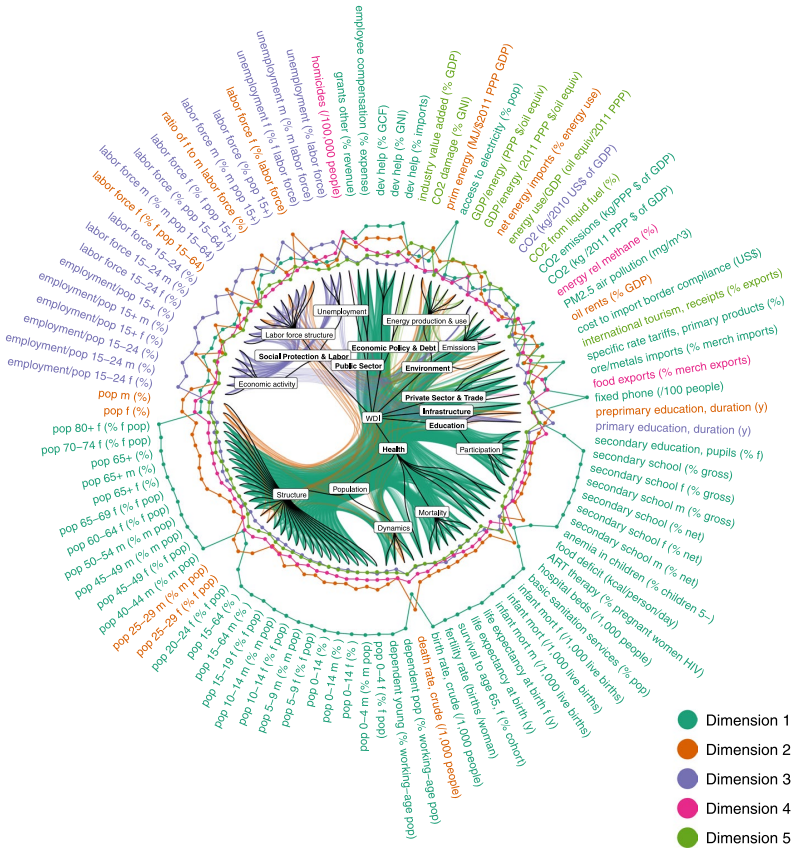

Fig. 4 The importance of single WDIs to the dimensions. Outer circle: indicator names, colored by the maximum distance correlation dimension. Middle circle: distance correlation values with dimensions 1–5. Black center dendrogram: thematic ordering of the indicators by the World Bank. Colored lines in the center circle: connect all indicators that have maximum distance correlation with the same dimension, bundled along the center dendrogram for easier visualization. The figure only shows the WDIs that have the highest distance correlations with dimensions 1–5

variance) represents the CO_2 impact of GDP, tourism and value added to products by industry.

3.3 Global Trends

Development is dynamic. Over time each country moves along a characteristic trajectory in development space. Along the first dimension, clear trends can be observed. Most countries have a positive slope (see Fig. 6). Given that dimension 1 essentially spans a gradient between wealthy and poor countries, this reveals the overall global

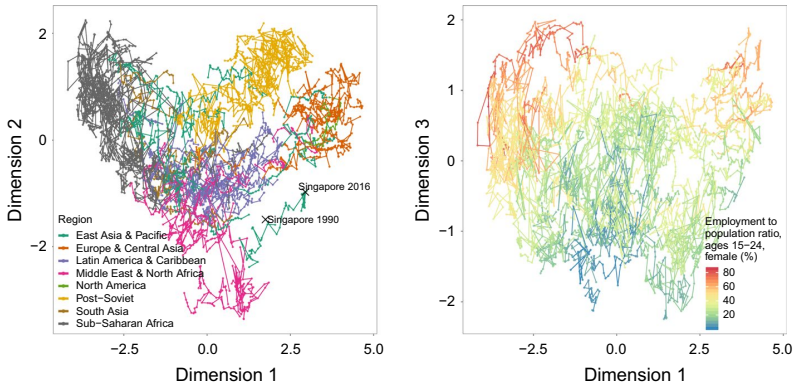

Fig. 5 E-Isomap dimensions 1–3. Left: Dimension 1 and 2, colored by World Bank regions and former East Block and allies in Eastern Europe. Right: Dimensions 1 and 3 colored by Employment to population ratio, ages 15–24, female (%). Dimension 1 (the horizontal axis on both panes) is a general wealth gradient, on the far left side are poor countries, mostly classified as "Sub-Saharan Africa" while on the right side are the developed countries with most Western European countries on the far right. Dimension 2 (vertical axis on the left pane) spans mostly the percentage of female population and labor force participation of women. Dimension 3 (vertical axis on the right pane) spans employment ratios, employment ratios for women and labor force participation of young working age women. There is an interactive online version available (http://bgc-jena.mpg.de/~gkraemer/consolidated_dimred/)

trend towards a wealthier world (Gapminder Foundation 2018). Only a few countries have negative slopes. Comparing the slopes of "Sub-Saharan Africa" with the rest of the world reveals a widening gap in the development gradient *sensu* HDI. Dimensions 2–4 do not show such pronounced overall trends.

Dimension 2 shows positive trends in most of the "Western World" and North Africa and negative trends in most parts of Asia and Sub-Saharan Africa. The positive trends in the "Western World" countries are due to an increased participation of women in the labor market, declining death rates in countries with young populations, and climbing death rates in countries with aging societies. Many developing countries in Sub-Saharan Africa and Asia show negative trends, which seems to be a common interaction between dimensions 1 and 2 on the far negative end of dimension 1.

Dimension 3 shows mostly employment/unemployment ratios, but there are no really strong general trends observable. We note that eastern and western Europe show fundamentally different trends, most of eastern Europe has predominantly negative trends, while in the rest of Europe there are few significant slopes reflecting the increase in unemployment in Eastern Europe. Other notable countries include Peru, Ethiopia, and Azerbaijan, where unemployment rates have strongly decreased; these countries show strong positive trends.

Dimension 4 shows energy-related methane emissions, which have increased in most parts of the northern hemisphere and decreased in most other parts of the world, as well as homicide rates, which have decreased in large parts of the world, but increased in parts of Latin America. The data on homicide rates in large parts of Africa are very sparse.

Dimension 5 shows tourism and the ecological impact of GDP. In general, more GDP is produced per unit of energy. This trend seems to be stronger in the Western World.

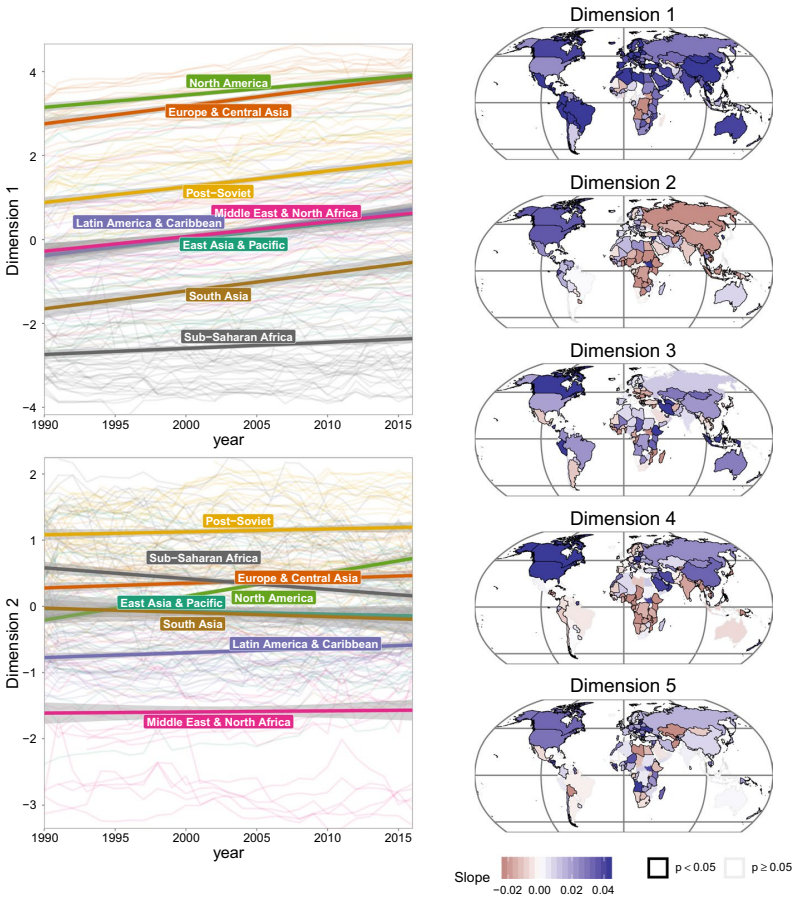

Fig. 6 Trends in the first e-Isomap axis over time. Left: transparent lines are the trajectories of all countries over time (dimensions 1 and 2, other dimension see SI Fig. 1), colored by geographic regions, the straight lines are linear regressions over all data points of a region and the 95% confidence intervals over their coefficients. Right: World maps of slopes over time of dimension 1–5, the color represents the value of Sen's slope, countries with significant slopes have black borders

3.4 Trajectories

Changes in the direction of trajectories are very likely to be a major disruption of a given development path. Some examples can be found in Fig. 7. For example, the earthquake in Haiti in 2010 coincides with a major disruption in the trajectory. The financial crisis and the onset of austerity measures can be noted from a dent in 2008 in the trajectory of Greece. A few years after massive privatizations in Argentina the trajectory of Argentina changes drastically. Major disruptions in the trajectory of the United States happen during

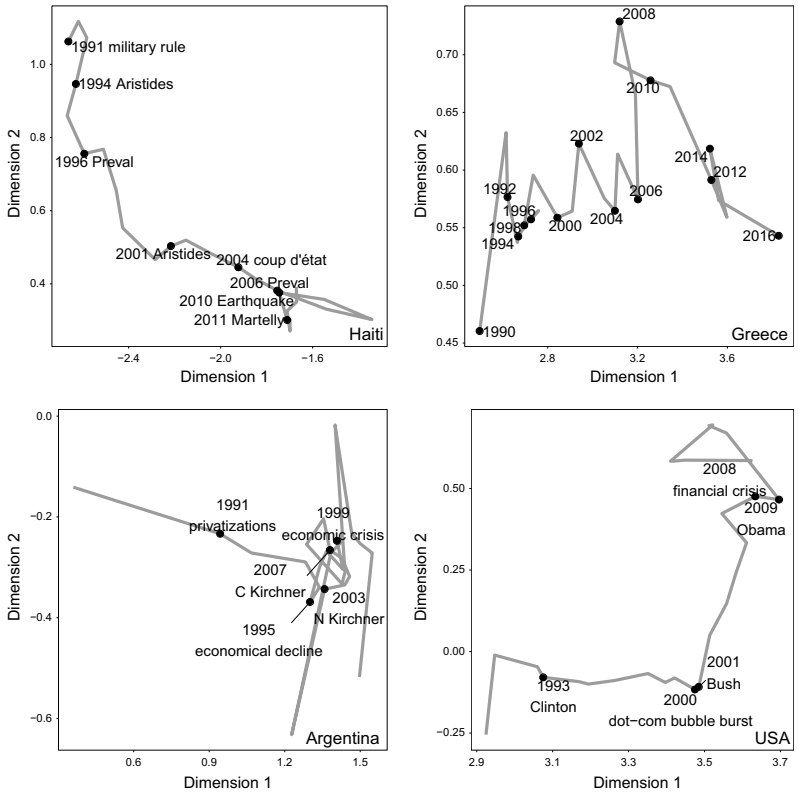

Fig. 7 Example trajectories. Haiti: Large jumps in the trajectory and kinks are preceded by changes in government and natural disasters. Greece: complete change of direction of the trajectory seem to appear during the financial crisis 2007–08. Argentina: The trajectory reflects important changes in economic policies. USA: The trajectory reflects the economic crises and changing presidencies

the burst of the dot-com bubble in 2000–2001 and the financial crisis in 2008. Attribution of changes to the trajectories to only these events can be challenging, and would require a formal causal framework (Pearl et al. 2016; Peters et al. 2017). For instance, in the case of the US, the changes in the trajectory could equally be attributed to changes in the presidency or to politics after 9/11/01. In the case of Argentina, it is not clear if the changes were caused by changes in politics during the Kirchner presidencies, problems that set in later after the privatizations, or a mixture of both, and remain of purely speculative nature.

In the overall view, some countries appear to change their centers of attraction recovered space of human development, e.g. Singapore in 1990 appears to be similar to the rich oil exporting Arab countries, but its trajectory suggests that it is currently gravitating towards most of the wealthy European countries, see Fig. 5. Countries that share similar history also seem to be close in the final dimensions, e.g. former Soviet countries, rich oil exporting nations, western European nations.

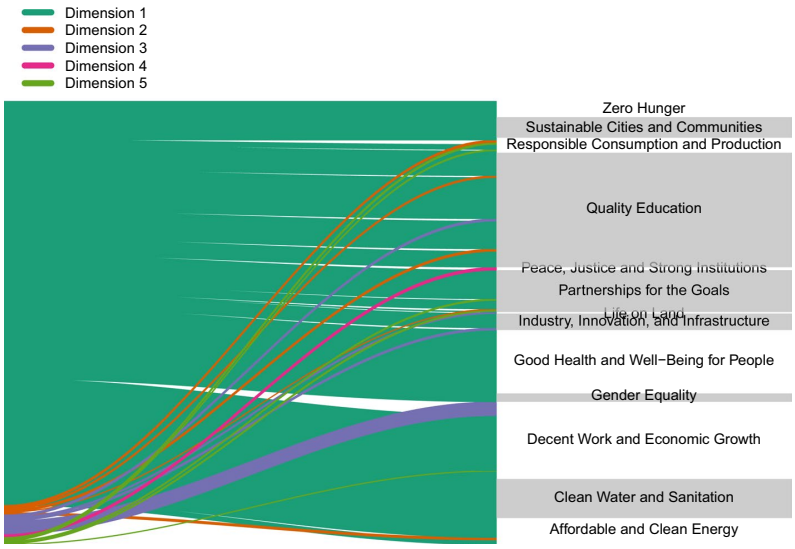

Fig. 8 Showing the importance of the dimensions for the SDGs, color code by dimension (left, unlabelled) are connected to the SDGs (right) through the corresponding WDIs (not shown, see text for details). The thickness of the connection reflects the distance correlation between the WDIs and the dimensions. See SI Fig. 3 for a more detailed version of the figure

3.5 Sustainable Development

To understand the relevance of the emerging dimensions for the different SDGs, we again use distance correlation and the WDIs that the World Bank uses to track the SDGs (United Nations General Assembly 2017b). We consider only the dimension with the maximum distance correlation to each WDI which is used to track an SDG. The results are shown in Fig. 8.

As most goals are poverty related, they load most strongly on the first dimension. The goals "Decent Work and Economic Growth" and "Industry, Innovation, and Infrastructure" also load on dimension 3, as this dimension describes the labor market. Dimension 2 describes educational and energy aspects and is related to "Affordable and Clean Energy" and "Quality Education". We found a relationship between dimension 4 and the SDG "Peace, Justice and Strong Institutions" due to the homicide rate indicator. Dimension 5 was important to the "Partnership SDG and Responsible Consumption and Production", due to relatedness of non-renewable energy sources and statistical reporting indicators.

Surprisingly, dimension two does not have any influence on the SDG "Achieve gender equality and empower all women and girls" despite describing aspects of gender equality. The reason for this may be that the SDG "Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all" is described by many of the variables loading on dimension two.

We can also see wich SDGs are well represented by the data (the height of the SDG in Fig. 8) and which ones are not. For example, best represented are SDGs that represent

traditional ideas of development, such as “Quality Education”, “Decent Work and Economic Growth”, “Good Health and Well-Being for People”, while environmental SDGs such as “Life on Land”, “Life Below Water”, or “Climate Action” are not well or not at all represented.

4 Discussion

The assessment of development on the basis of a few key indicators has often proven very useful, but has also been controversial. As early as in the 1960s, GDP was recognized to be a very incomplete measure of development (Ram 1982; McGillivray 1991; Göpel 2016). Later, a large number of indicator approaches emerged, each constructed to describe specific aspects of development (Parris and Kates 2003; Shaker 2018). The large number of measured variables and derived indicators that are used today to describe development could suggest that global development is a high dimensional process requiring many indicators to describe it accurately. This perception contrasts with our finding that three quarters of the variability of the development space can be explained by only one dimension, and five dimensions recover 90% of variance. This indicates that the dimensionality of development is much lower than one would expect. Or, to put in other terms, the fact that many properties of development are highly correlated (Ghislandi et al. 2019) also means that one can summarize them efficiently in very few dimensions.

The notion that development is of low-dimensionality, however, does by no means imply that it is a “simple” process. In general it is well-known that low-dimensional spaces can still contain and depict very complex and unpredictable dynamics: Prominent examples are the logistic map (Verhulst 1845, 1847), describing population dynamics in a space of a single dimension, or the Lorenz (1963) attractor in physics, describing hydrodynamic flow in a three dimensional space.

The question whether data-driven indicators as presented here can be an alternative to classical indicators has been widely discussed (Ram 1982; OEDC 2008; Gapminder Foundation 2018). One argument in favour of such an approach is to overcome the lack of objectivity, which is a common criticism of classical indicators (Monni and Spaventa 2013; Göpel 2016). Consequently, PCA is increasingly used for the creation of wealth indicators (Filmer and Pritchett 2001; Smits and Steendijk 2015; Shaker 2018), as well as other approaches to identify suitable variable weights (Seth and McGillivray 2018). In our study we show that the PCA approach is less effective due to the strong nonlinear relations among the covariates present in the dataset.

Nonlinear data dimensionality reduction, however, makes the assessment of the identified dimensions difficult and hard to trace back to the underlying processes. Dimension 1, for example, includes both basic health and wealth variables. Figure 3 illustrates the reason for this. On the negative end of dimension 1, the maternal mortality rate is high and per capita income is low. When moving upwards along this dimension, first maternal mortality rates drop steeply, while the per capita income hardly changes. When moving towards the positive end of dimension 1, maternal mortality cannot decrease much further, as it is already close to zero, but the per capita income starts to increase strongly (Fig. 3). Combining both effects, dimension 1 manages to incorporate wealth as well as mortality related variables into a single (nonlinear) indicator. Each indicator can have a strong influence on a subset of a dimension (e.g. maternal mortality rate on the negative side of dimension 1) and a very low impact on other subsets (e.g. maternal mortality rate on the positive end of

dimension 1). Still, the fact that these factors co-vary in a way that we can represent them in a single dimension can guide the development of novel metric indices.

While dimension 1 allows for a relatively straightforward interpretation, we see in dimension 2 that there are more complex patterns to discuss. We find that Post-Soviet countries, Western European countries and Sub-Saharan African countries all lay on similar high coordinate values in dimension 2 (Fig. 5). Looking at variables that correlate strongly with dimension 2, we find that the participation of women in the labor market can be similar for very different states of dimension 1. We probably also uncover certain socio-cultural divides: most countries classified as "Middle East & North Africa" show a very low participation of women in the labor marked while in other parts of the world participation of women in the labor marked is much higher and does not depend on the geopolitical region of a country or its development status (see Fig. 5). For example in many European countries 45–50% of the working population is female, the same or even less than in most Sub-Saharan countries. Another variable that is orthogonal to development are crude death rates, where a rich country like Germany can have very similar rates to many countries in central Africa. Death rates in the WDI database are not resolved by age groups, given the aging societies in the developed world and the very young societies in many African countries, the death rates affect mostly older age groups in countries with high values on dimension 1, while it affects many younger age groups in the African countries.

In general, data-driven approaches to index construction can be criticized for not taking the polarity, i.e. the "direction", into account (Mazziotta and Pareto 2019). This means that it remains subject to a subsequent interpretation whether a high value of a principal component (or non-linearly derived component) is a sign of a positive state in a certain domain or the opposite. The reason is that the underlying eigenvectors can be of arbitrary sign. However, we have shown (in Fig. 4) that an interpretation is possible, and the analysis of trends and trajectories can remedy this issue. Collapsing many aspects of development into a single dimension, which in turn forms the main gradient along which countries move over time, essentially expresses (nonlinear) covariations that should not be studied in isolation. For example, higher employment rates and an increased per capita income often go hand in hand. Here we showed that these connections between the 621 measured variables are so strong that a single dimension suffices to represent 74% of the variance. In this sense, we also see our approach as an opportunity to generate novel hypotheses on development that can guide policy making e.g. towards achieving the SDGs.

A general criticism of machine learning approaches is that underlying data biases are propagated and exacerbated. For instance, if the training data contain biases against minority groups, e.g. gender or race, these groups will systematically be put in a disadvantage by the algorithm (Barocas and Selbst 2016). Latest research tries to detect such biases (Obermeyer et al. 2019) and to avoid them during the training phase (Pérez-Suay et al. 2017). Therefore the implications of every machine learning based analysis have to be seen in the light of the dataset used for training. Here we summarize the WDI database, which represents the efforts of the World Bank to collect information on development at the global scale. The high variance explained by variables representing basic infrastructure, per capita income, and the population pyramid therefore reflects the (historic) emphasis that has been given to these kinds of basic indicators. For instance financial accounting has been ubiquitous, there are large scale efforts to monitor infrastructure and poverty, and census data is globally available.

Our analysis does not reveal an "environmental axis", a component that is essential to sustainable development (Steffen et al. 2015). We can therefore also read our analysis as a gap analysis and conclude that future versions of the WDI data base should put more

emphasis on environmental data that are now widely available (Mahecha et al. 2020). Another essential component are inequalities (UNDP 2019). While some aspects are recovered by our analysis, such as between country inequalities on dimension one and some aspects of gender inequality on dimension 2, others do not emerge, e.g. income inequalities inside a country.

The best represented SDGs are those related to traditional ideas of development, while "Life on Land", "Life Below Water" or "Climate Action" are not well or not at all represented. This shows a clear bias towards classical development data, and a lack of environmental data in the WDI data base. The reasons for this lie in the topics that have been emphasized for development historically (Griggs et al. 2013).

An analysis like the present one can be informative for policy making in various ways. It reveals general constraints of the development manifold, i.e. which combinations of WDIs are possible, which trajectories in the development space have been observed and which ones not. In particular, the trajectories can inform policy makers regarding the general present and past position of a country in this space beyond a single metric like the HDI (or our dimension 1). This means that also the less obvious changes, e.g. the changes of post-Soviet countries along dimension 2, can be taken into consideration.

Focusing on these dimensions is not trivial. It allows to target a few orthogonal aspects of developments only, instead of screening hundreds of individual WDIs. Another way how this analysis can guide policies is by seeing the results in the context of the dataset and pointing out weaknesses and underrepresented dimensions in the dataset, such as the environment and within-country inequalities. The key difference between our approach and the classical approaches is that we try to describe development space in its entirety, and hence the extracted components are neutral and agnostic to any societal or political agenda.

In particular the trajectory of single countries can yield essential information on important events for a country. The trajectories analyzed in this paper all showed changes that are obvious to the human eye, such as temporary deviations or changes in speed and directions. We could find connections for all of the observed changes in the trajectories of Fig. 7 with important socioeconomic or environmental events, although we were not able to automatically detect changes in all trajectories due to the different characteristics of each change. Future research is needed, to better understand the anomalies in the extracted trajectories.

In our opinion, a main advantage of data-driven approaches compared to classical indicator approaches is that the number of necessary indicators emerges naturally and the resulting indicators represent orthogonal features. The main disadvantage is the loss of indicators that represent very specific aspects of the data. Obviously, dimensionality reduction can only summarize the available data which also means that data incompleteness, data errors, and reporting biases are inherited—as it is also the case for classical indicators. Still, the proposed approach can help in the planning of adding measures of development and testing their redundancy with respect to the existing indicators, simplifying e.g. reporting of complementary dimensions of development.

A general limitation of the data under scrutiny is their aggregation at the country level. This means that our analyses cannot account for the often large socioeconomic differences and developments *within* a country. Also localized disasters may not influence the trajectory of a large economy as a whole, e.g. a large hurricane causing damage in Florida will only have a very marginal influence on the trajectory of the United States. Today there are efforts to collect data on sub-national levels which would alleviate this problem, see e.g. Smits and Permanyer (2019). However these efforts are relatively recent and there are still not many variables available.

5 Conclusions

In this study we investigated the "World Development Indicators" from 1990 to 2016 using a method of nonlinear dimensionality reduction. Our study led to three key insights. Firstly, the WDI database is of very low intrinsic dimensionality: We found that the WDIs are strongly interconnected, but we also showed that these connections are highly nonlinear. This is the reason why linear indices based on PCA cannot compress the information on human development that efficiently, while our approach only needs five dimensions to represent 90% of the data variance. The first dimension partly resembles the HDI, but also reveals much more differentiated patters in low-income countries. The subsequent dimensions show orthogonal aspects such as the participation of women in the labor market and complex demographic dynamics. Quantifying such interactions uncovered by this approach can lead to new approaches to quantify different aspects of development. Exploring the meaning of the emerging dimensions allows us to understand which aspects of development are underrepresented in current databases. The second insight is that development as described by the dimensional space, remains to be a highly complex process that involves strong nonlinear interactions. We have elaborated some of these aspects, but a more profound exploration of the five-dimensional development space is still needed. Clearly, our approach can only account for the information in the data and ignore any additional aspects such as environmental issues that are clearly critical for sustainable development. The third insight is that single countries' trajectories in the low dimensional space show abrupt changes that coincide with major environmental hazards or socioeconomic anomalies. As these changes in the trajectories can be of different nature, automatized detection is non-trivial and may require further causal explorations. Overall, our analysis gives new insights into the general structure of development which is of low dimensionality, but highly nonlinear and interconnected. Future work is need to understand the observed trajectories in development space in much more detail, as well as to exploit them for achieving the Sustainable Development Goals.

Acknowledgements Open access funding provided by Projekt DEAL. G.K. acknowledges the support of the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - FZT 118. G.K., M.R., and M.D.M. thank the European Union for funding the H2020 Project BACI under Grant Agreement No. 640176 and the ESA for support via the "Earth System Data Lab" project. G.C.V. work has been supported by EU under the ERC consolidator Grant SEDAL-647423.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Arenas-Garcia, J., Petersen, K. B., Camps-Valls, G., & Hansen, L. K. (2013). Kernel multivariate analysis framework for supervised subspace learning: a tutorial on linear and kernel multivariate methods. *IEEE Signal Processing Magazine*, *30*(4), 16–29. https://doi.org/10.1109/MSP.2013.2250591.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal*,. https://doi.org/10.2139/ssrn.2477899.

Bennett, R. (1969). The intrinsic dimensionality of signal collections. *IEEE Transactions on Information Theory*, *15*(5), 517–525. https://doi.org/10.1109/TIT.1969.1054365.

Costanza, R., Fioramonti, L., & Kubiszewski, I. (2016). The UN sustainable development goals and the dynamics of well-being. *Frontiers in Ecology and the Environment*, *14*(2), 59–59. https://doi.org/10.1002/fee.1231.

Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data-or tears: an application to educational enrollments in states of India. *Demography*, *38*(1), 115–132. https://doi.org/10.1353/dem.2001.0003.

Filmer, D., & Scott, K. (2012). Assessing asset indices. *Demography*, *49*(1), 359–392. https://doi.org/10.1007/s13524-011-0077-5.

Gapminder Foundation (2018) *Gapminder: Unveiling the beauty of statistics for a fact based world view*. Retrieved May 17, 2020, from https://www.gapminder.org/.

Ghislandi, S., Sanderson, W. C., & Scherbov, S. (2018). A simple measure of human development: The human life indicator. *Population and Development Review*,. https://doi.org/10.1111/padr.12205.

Ghislandi, S., Sanderson, W. C., & Scherbov, S. (2019). A simple measure of human development: The human life indicator. *Population and Development Review*, *45*(1), 219.

Göpel, M. (2016). *The Great Mindshift, The Anthropocene: Politik–Economics–Society–Science* (Vol. 2). Cham: Springer. https://doi.org/10.1007/978-3-319-43766-8.

Griggs, D., Stafford-Smith, M., Gaffney, O., Rockström, J., Öhman, M. C., Shyamsundar, P., et al. (2013). Sustainable development goals for people and planet. *Nature*, *495*(7441), 305–307. https://doi.org/10.1038/495305a.

Kraemer, G., Reichstein, M., & Mahecha, M. D. (2018). dimRed and coRanking—Unifying dimensionality reduction in R. *The R Journal*, *10*(1), 342–358.

Kubiszewski, I., Costanza, R., Franco, C., Lawn, P., Talberth, J., Jackson, T., et al. (2013). Beyond GDP: Measuring and achieving global genuine progress. *Ecological Economics*, *93*, 57–68. https://doi.org/10.1016/j.ecolecon.2013.04.019.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2.

Mahecha, M. D., Martínez, A., Lischeid, G., & Beck, E. (2007). Nonlinear dimensionality reduction: alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. *Ecological Informatics*, *2*(2), 138–149.

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., et al. (2020). Earth system data cubes unravel global multivariate dynamics. *Earth System Dynamics*, *11*(1), 201–234. https://doi.org/10.5194/esd-11-201-2020.

Marshall, M. G., & Elzinga-Marshall, G. (2017). *Global Report 2017, conflict, governance, and state fragility. Center for Systemic Peace*. Retrieved May 17, 2020, from http://www.systemicpeace.org/vlibrary/GlobalReport2017.pdf.

Mazziotta, M., & Pareto, A. (2019). Use and misuse of PCA for measuring well-being. *Social Indicators Research*, *142*(2), 451–476. https://doi.org/10.1007/s11205-018-1933-0.

McGillivray, M. (1991). The human development index: Yet another redundant composite development indicator? *World Development*, *19*(10), 1461–1468. https://doi.org/10.1016/0305-750X(91)90088-Y.

McRae, L., Freeman, R., Marconi, V., & Canadian Electronic Library (Firm) (2016) Living planet report 2016: Risk and resilience in a new era. WWF, oCLC: 1001121301

Monni, S., & Spaventa, A. (2013). Beyond GDP and HDI: Shifting the focus from paradigms to politics. *Development*, *56*(2), 227–231. https://doi.org/10.1057/dev.2013.30.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.

OEDC. (2008). Handbook on constructing composite indicators: Methodology and user guide. OECD, Paris, oCLC: ocn244969711

Parris, T. M., & Kates, R. W. (2003). Characterizing and measuring sustainable development. *Annual Review of Environment and Resources*, *28*(1), 559–586. https://doi.org/10.1146/annurev.energy.28.050302.105551.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. Hoboken: Wiley.
Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(6), 559–572.
Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., & Camps-Valls, G. (2017). Fair kernel learning. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, & S. Džeroski (Eds.), *Lecture notes in computer science, machine learning and knowledge discovery in databases* (pp. 339–355). New York: Springer. https://doi.org/10.1007/978-3-319-71249-9_21.
Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. New York: MIT press.
Pradhan, P., Costa, L., Rybski, D., Lucht, W., & Kropp, J. P. (2017). A systematic study of sustainable development goal (SDG) interactions: A systematic study of SDG interactions. *Earth's Future*, *5*(11), 1169–1179. https://doi.org/10.1002/2017EF000632.
Ram, R. (1982). Composite indices of physical quality of life, basic needs fulfilment, and income: A 'principal component' representation. *Journal of Development Economics*, *11*(2), 227–247.
Rickels, W., Dovern, J., Hoffmann, J., Quaas, M. F., Schmidt, J. O., & Visbeck, M. (2016). Indicators for monitoring sustainable development goals: An application to oceanic development in the European Union. *Earth's Future*, *4*(5), 252–267. https://doi.org/10.1002/2016EF000353.
Seth, S., & McGillivray, M. (2018). Composite indices, alternative weights, and comparison robustness. *Social Choice and Welfare*, *51*(4), 657–679.
Shaker, R. R. (2018). A mega-index for the Americas and its underlying sustainable development correlations. *Ecological Indicators*, *89*, 466–479. https://doi.org/10.1016/j.ecolind.2018.01.050.
Silva, V. D., & Tenenbaum, J. B. (2003). Global versus local methods in nonlinear dimensionality reduction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (pp. 721–728). New York: MIT Press.
Smits, J., & Permanyer, I. (2019). The subnational human development database. *Scientific Data*, *6*, 190038. https://doi.org/10.1038/sdata.2019.38.
Smits, J., & Steendijk, R. (2015). The international wealth index (IWI). *Social Indicators Research*, *122*(1), 65–85. https://doi.org/10.1007/s11205-014-0683-x.
Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107.
Stacklies, W., Redestig, H., Scholz, M., Walther, D., & Selbig, J. (2007). pcaMethods—A bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, *23*(9), 1164–1167. https://doi.org/10.1093/bioinformatics/btm069.
Stanojević, A., & Benčina, J. (2019). The construction of an integrated and transparent index of wellbeing. *Social Indicators Research*, *143*(3), 995–1015. https://doi.org/10.1007/s11205-018-2016-y.
Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., et al. (2015). Planetary boundaries: Guiding human development on a changing planet. *Science*,. https://doi.org/10.1126/science.1259855.
Szkely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505.
Tenenbaum, J. B., Silva, Vd, & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319.
The World Bank.(2018a). *Sustainable development goals (SDG)|data catalog*. Retrieved May 3, 2018, from https://datacatalog.worldbank.org/dataset/sustainable-development-goals.
The World Bank. (2018b). *World development indicators (WDI)|data catalog*. Retrieved May 3, 2018, from https://datacatalog.worldbank.org/dataset/world-development-indicators.
Torgerson, W. S. (1952). Multidimensional scaling: I theory and method. *Psychometrika*, *17*(4), 401–419. https://doi.org/10.1007/BF02288916.
UNDP. (2016). *Human development report. Human development for everyone, United Nations Development Programme*. New York, NY: Human Development Reports.
UNDP. (2018). Human Development Reports|United Nations Development Programme. http://hdr.undp.org/
UNDP. (2019). Human Development Report 2019 Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century. United Nations Development Programme, New York, NY, USA: Human Development Reports.
United Nations General Assembly. (2017a). *Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development*. https://doi.org/10.1891/9780826190123.0013

United Nations General Assembly. (2017b). *Work of the Statistical Commission pertaining to the 2030 Agendanda for Sustainable Development*. http://ggim.un.org/meetings/2017-4th_Mtg_IAEG-SDG-NY/documents/A_RES_71_313.pdf

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, *10*, 66–71.

Verhulst, P. (1845). Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, *18*, 14–54.

Verhulst, P. (1847). Deuxième mémoire sur la loi d'accroissement de la population. *Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique*, *20*, 1–32.

Wolter, K., & Timlin, M. (1993). Monitoring ENSO in COADS with a Seasonally Adjusted Principal Component Index. NOAA/NMC/CAC, NSSL, Oklahoma Clim. Survey, CIMMS and the School of Meteor., University of Oklahoma, Norman, OK

Wolter, K., & Timlin, M. S. (2011). El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). *International Journal of Climatology*, *31*(7), 1074–1087. https://doi.org/10.1002/joc.2336.

Affiliations

Guido Kraemer[1,2,3,4] · Markus Reichstein[1] · Gustau Camps-Valls[3] · Jeroen Smits[5] · Miguel D. Mahecha[1,2,4]

Gustau Camps-Valls
gustau.camps@uv.es
http://isp.uv.es

Miguel D. Mahecha
miguel.mahecha@uni-leipzig.de

[1] Max Planck Institute for Biogeochemistry, 07745 Jena, Germany

[2] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany

[3] Image Processing Lab, Universitat de Valéncia, 46980 Paterna, Valéncia, Spain

[4] Remote Sensing Centre for Earth System Research, Leipzig University, 04103 Leipzig, Germany

[5] Global Data Lab, Radboud University, 6500 HK Nijmegen, The Netherlands