METHODS AND CHALLENGES IN TIMESERIES ANALYSIS OF VEGETATION IN THE GEOSPATIAL DOMAIN

A. Elia [1*], M. Pickering [2], M. Girardello [3], G. Oton [3], G. Ceccherini [3], S. Capobianco [2], M. Piccardo [2], G. Forzieri [4], M. Migliavacca [3], A. Cescatti [3]

[1] Arcadia SIT, Ispra, Italy - agata.elia@ext.ec.europa.eu, agata.elia1991@gmail.com

[2] JRC Consultant, Ispra, Italy - (mark.pickering1, samuele.capobianco, matteo.piccardo)@ext.ec.europa.eu

[3] Joint Research Centre, European Commission, Ispra, Italy – (marco.girardello, gonzalo.oton, guido.ceccherini, mirco.migliavacca, alessandro.cescatti)@ec.europa.eu

[4] Department of Civil and Environmental Engineering, University of Florence, Florence, Italy - giovanni.forzieri@unifi.it

KEY WORDS: Forests, Vegetation Indices, Climate, Timeseries Analysis, Random Forest, GEE

ABSTRACT:

The increasing availability of remotely sensed data have offered unprecedented possibilities for monitoring and analysis of environmental variables, including boosting recent studies in the field of ecosystem resilience relying on indicators derived from timeseries analysis, such as the temporal autocorrelation of vegetation indices. A forest ecosystem with decreased resilience will be more susceptible to external drivers and their change and could shift into an alternative system configuration by crossing a tipping point. Nevertheless, remote sensing data quantifying vegetation and forests properties inherently carry information related to the climate as well, which has to be accounted for before performing any modelling exercise. In this paper, we aim to present the general workflow and the challenges encountered in processing and analysing the historical, high-frequency and high-resolution timeseries of vegetation and climatic data. The final aim is training a machine learning model (Random Forest) in order to model and explore the performance and importance of a set of climatic and environmental metrics in predicting an indicator of the resilience of forests. In this case, the resilience of forests is quantified through the temporal autocorrelation (TAC) of the kernel NDVI (kNDVI). Climatic and environmental predictors include 2-meter air temperature, total precipitation, vapour pressure deficit, surface solar radiation, forest cover and soil organic carbon content. Results show a good performance of the Random Forest model and the ranking in the importance of the predicting variables captured in terms of background climate and climate variability. This application allows to separate and identify the main drivers of the temporal autocorrelation of kNDVI.

1. INTRODUCTION

The increasing availability and ease of access of global, historical, and high-frequency remote sensing data have offered unprecedented possibilities for monitoring and analysis of environmental variables. Recent studies in the field of ecosystem resilience relied on indicators derived from timeseries analysis, such as the temporal autocorrelation and the variance of a system signal (Dakos et al., 2015). The aforementioned availability of global, temporally and spatially dense timeseries of indicators of biomass and greenness of vegetation, such as the normalized difference vegetation index (NDVI), among others, has boosted ecosystem resilience scientific applications to forests as well. The ecological definition of resilience corresponds with the capacity of a system to absorb and recover from a disturbance (Forzieri et al., 2022). When dealing with ecosystems increasingly affected by natural and anthropogenic pressures such as forests, monitoring their health is particularly relevant.

Forest ecosystems play a crucial part in the global carbon cycle and in any climate change mitigation strategy, despite being increasingly affected by natural and anthropogenic pressures. While anthropogenic action on forests is mainly represented by stand replacement, natural perturbations include wind throws and fires, as well as extended insects and disease outbreaks, such as the recent outbreak affecting Central Europe (Bárta et al., 2021; Thonfeld et al., 2022). These natural disturbances are strictly interconnected with climate change. A forest ecosystem with decreased resilience will be more susceptible to external drivers and their change and could shift into an alternative system configuration by crossing a tipping point.

However, remote sensing data quantifying vegetation and forests properties inherently carry information related to the climate as well. If not accounted for, these confounding factors, such as short-term climate fluctuations, may hide the actual vegetation anomalies focus of a study and the importance of other drivers in vegetation itself. In addition, the comparison of the same vegetation property between different geographical areas naturally affected by different climates is hindered.

In order to explore the relationships of a set of environmental and climatic metrics with an indicator of the resilience of forests, a machine learning (ML) model is implemented. In this paper, we aim to present the general workflow and the challenges encountered in processing and analysing the timeseries of vegetation and climatic. The focus of this paper will be on a workflow implemented to analyse the aforementioned timeseries and on the methods and tools implemented to account for the background climate effect on vegetation. A key aspect to assess resilience of ecosystems is the treatment of the signal to remove long-term trends and seasonal cycle of the signal. Being aware of the variety and heterogeneity of methodologies existing in the field of timeseries analysis, in this contribution we will describe tools and methods for deseasonalization, detrending, growing season identification and accounting for short-term climate effects.

* Corresponding author

The final aim is to present one of the many workflows that can be implemented when dealing with timeseries of vegetation-related data in the geospatial domain, where climate plays a crucial role. The importance of the availability of open data and open-source tools and platforms in making this big data analysis possible is also strongly highlighted.

2. OBJECTIVE

The main objective of the presented paper is to illustrate a thorough methodology to pre-process and analyse timeseries of vegetation and climatic data in order to retrieve a vegetation signal and a derived metric of resilience, in this specific case in terms of lag-1 temporal autocorrelation, as close as possible to representing the actual response of the system accounting for short-term climate fluctuation.
In order to do so, it has to be considered firstly that the timeseries of vegetation and climatic variables are constituted of three main components: a long-term trend, the seasonality cycles and the remaining part representing the anomaly that deviates from the average conditions. In the following equation representing a decomposed timeseries, the Y represents either the vegetation or the climate variable timeseries (Sun et a;., 2022).

$$Y = \text{Trend} + \text{Seasonality} + \text{Anomaly}, \quad (1)$$

In order to obtain an accurate estimate of vegetation and its resilience, the vegetation timeseries used in any model needs to be stationary, hence without periodical seasonal cycles and long-term trends (Forzieri et al., 2022).

Secondly, a Random Forest (Breiman, 2001) model is used to explain the observed long-term lag-1 temporal autocorrelation (TAC) of vegetation, by accounting simultaneously for climate fluctuations and the temporal autocorrelation of the climate itself.

$$TAC = RF(X) + \varepsilon, \quad (2)$$

In the previous equation the X represents a series of climatic and environmental predictors used to estimate the vegetation long-term TAC and ε are the residuals of such estimation. This allows to explain TAC at different pixels accounting for the differences in climate and environmental conditions.

3. MATERIALS AND METHODS

All data and most of the tools leveraged for this study are open. In the following sections the main datasets and tools are described.

3.1 Tools

The data processing takes place mainly within Google Earth Engine (GEE) and the Joint Research Centre (JRC) Big Data Analytics Platform (BDAP).
Google Earth Engine is a cloud-based geospatial analysis platform providing a multi-petabyte catalogue of satellite imagery and geospatial datasets coupled with large analysis capabilities (Gorelick et al., 2017).
The JRC Big Data Analytics Platform is a petabyte-scale storage system coupled with a processing cluster. It includes open-source interactive data analysis tools, a remote data science desktop and distributed computing with specialized hardware for machine learning and deep learning tasks (Soille et al., 2018).
GEE is mainly used to pre-process MODIS data and secondary datasets. The ERA5 pre-processing and the core timeseries analysis are performed within the JEODPP, where main tools include R (R Core Team, 2022), Climate Data Operator (CDO) and netCDF Operators (NCO). The whole machine learning model is instead trained and run in R. The different platforms and tools implemented in the study highlight the heterogeneity of data as well involved, data availability and data formats, ranging from TIFF, netCDF and R objects.

3.2 Datasets

This section is going to illustrate main datasets used in the analysis divided by category and the main filtering and resampling steps applied to each dataset in order to create a coherent ensemble of data to be used in the subsequent model in terms of temporal and spatial resolution. All data have been retrieved over Europe.

3.2.1 Vegetation: The long-term kNDVI timeseries was retrieved by processing the full timeseries of daily MODIS Terra and Aqua Surface Reflectance at 500m from 2003 to 2021 (Vermote et al., 2021). A simplified version of kNDVI is defined as:

$$KNDVI = \tanh(NDVI^3 / | NDVI |), \quad (3)$$

The kNDVI is a nonlinear generalization of the NDVI that shows stronger correlations than NDVI and NIRv with forest key parameters. kNDVI is also more resistant to saturation, bias, and complex phenological cycles, and it is more robust to noise and more stable across spatial and temporal scales (Camps-Valls et al., 2021).
MODIS daily data have been filtered in order to select only the highest quality data for the bands of interest (1 and 2) and MODIS pixels have been masked for clouds, clouds shadows, water, snow and ice by referring to the data product state QA (Quality Assessment) bit flags. After quality filtering, MODIS derived kNDVI daily data have been reduced as mean into a timeseries with an 8 days' time-step and resampled at 0.005° resolution to facilitate comparison with climatic data.
Each of the 0.005° kNDVI pixel in the timeseries was afterward masked with a binary forest mask including only pixel with at least 50% of forest cover. Afterwards, the kNDVI was averagely aggregated to 0.05° resolution.

The forest mask used was obtained from a forest cover percentage layer. The latter was obtained as the percentage of forest covered pixels in a 0.005° cell, accounting as forest covered pixels all the 30m pixels from the Hansen tree cover 2000 layer (Hansen et al., 2013), only where tree cover is higher or equal 30% and retaining only patches of at least 6 connected pixels, accounting for a surface higher than 0.5ha. In addition, any pixel undergoing a forest loss in the time period was removed in order to account for managed or disturbed forest patches, where disturbances may include forest fires or clear-cutting. Indeed, such events can artificially boost the lag-1 TAC of the vegetation signal.
This pre-processing resulted in a 0.05° resolution timeseries of 8-days averaged kNDVI, derived exclusively from forested pixels.

In order to account for the phenology, MODIS Land Cover Dynamics yearly data at 500m (Friedl et al., 2022) were used.

From these, the circular mean of greenup (as the date when EVI2 first crossed 15% of the segment EVI2 amplitude) and dormancy (as the date when EVI2 last crossed 15% of the segment EVI2 amplitude) was calculated, after yearly dates of greenup and dormancy have been translated in their respective angular measure of their respective day of the year. Average greenup and dormancy were obtained for the time period 2001-2021. Finally, the growing season products were aggregated at 0.05° resolution, accounting again only for pixels cover at least for 50% by forests.

3.2.2 Climate: Hourly ERA5-Land data at 10km resolution (Copernicus Climate Change Service (C3S), 2017) were used to retrieve the set of climatic and environmental predictors including 2-meter air temperature (t2m), total precipitation (tp), vapour pressure deficit (vpd) and surface solar radiation (ssr). These variables were computed as 8-days averages or sums according to the specific variable and as well resampled at 0.05° resolution in order to be coherent with the vegetation data.

3.2.3 Others: Additional datasets included in the final RF model are the forest cover (fc) layer previously described in the Vegetation section aggregated at 0.05° resolution and the soil organic carbon content layer at 30cm depth (socc30cm) from OpenLandMap Soil Organic Carbon Content at 250m (Hengl et al., 2018) and aggregated at 0.05° resolution.

3.3 Timeseries pre-processing

This section is going to illustrate the main pre-processing steps applied to each timeseries involved in the study, prior to the calculation of the lag-1 temporal and other statistics. Each step is important in timeseries analysis in order to maximise the representativeness of the resulting signal, being either of vegetation or climatic variables.

3.3.1 Deseasonalisation: The kNDVI and climatic variable were firstly deseasonalised by subtracting from each time-step the long-term average (2003-2021) of each 8 days timestep, resulting in removing the climatological average at each timestep. This removes the dominant seasonal component of the vegetation cycle.

3.3.2 Growing season selection: A climatological growing season was identified as the 20 years (2001-2021) circular mean of greenup and dormancy from the MODIS Land Cover Dynamics product. This climatological growing season was used in order to retain only 8 days timesteps that are within the greenup day of the year and the dormancy day of the year. This allows the background environment metrics to not be strongly affected by the dormient months of the vegetation.

3.3.3 Detrending: As a final step in the pre-processing of the timeseries, a linear regression model has been fitted to each pixel's timeseries and the resulting fitted values have been subtracted from the kNDVI and climatic variables values, in order to remove any linear trends from the timeseries, because these trends result for long-term climate trends (i.e., global warming).

3.3.4 TAC and other statistics calculation: The previously explained pre-processing steps leaves anomalies from the seasonal cycle in the vegetation and climatic timeseries, as shown in Figure 1. Figure 1 shows the evolution of the timeseries of kNDVI for a representative pixel undergoing each step of the presented pre-processing.

The vegetation and climatic anomalies are finally used to compute the long-term (2003-2021) lag-1 temporal autocorrelation (TAC) of both kNDVI anomalies and each climatic variable anomalies involved in the study.

In addition, the average and coefficient of variation (CoV) of kNDVI and climatic data are computed for the climatic growing season.

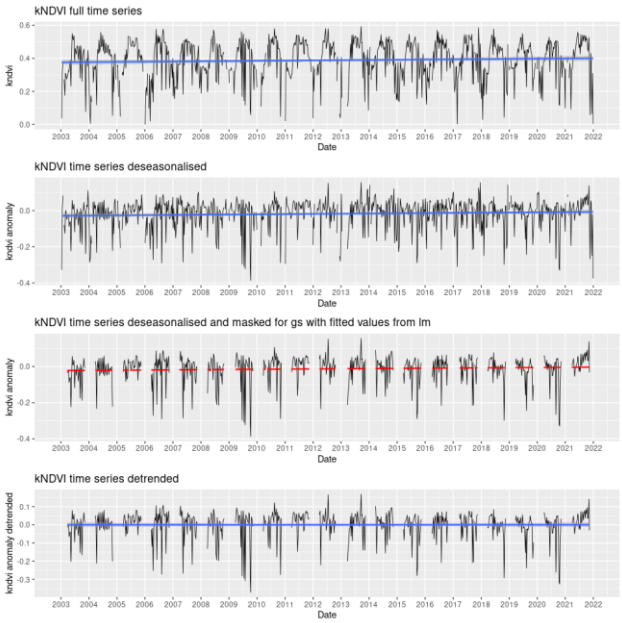

Figure 1. kNDVI timeseries for a sample pixel at each pre-processing step.

3.4 Random Forest (RF) model

In order to predict the long-term kNDVI TAC accounting for the impact of climate and other environmental factors, a RF regression model was implemented using as predicted variable the long-term kNDVI anomaly TAC and as predictors: the long-term TAC of each climatic anomaly, the average and coefficient of variation (CoV) of each climatic variable computed on the climatic growing season, as well as the average and coefficient of variation (CoV) of kNDVI computed on the climatic growing season. The CoV represents of the variability of climate, whilst the average represents the background climate. In addition to these vegetation and climatic variables, additional datasets including the forest cover and soil organic carbon content were included in the RF model. Predicting the long-term lag-1 kNDVI TAC with a series of climatic and environmental predictors allows to separate and identify the drivers of forest resilience in terms of long-term TAC.

The RF model was trained with 500 trees and with a proportion of training equal to 70% of the whole dataset and 30% for the testing.

4. RESULTS AND DISCUSSION

In this section, the results of the pre-processing applied to the timeseries will be illustrated specifically for the kNDVI timeseries. In addition, the performance and outcome in terms of predictors ranked by variable importance of the trained RF model will be presented.

Figure 2 illustrates the long-term lag-1 TAC of kNDVI without any pre-processing applied to the timeseries and the same lag-1 TAC of kNDVI following the timeseries pre-processing, as illustrated in the previous section.
As it can be seen from the two maps, the long-term lag-1 TAC of kNDVI strongly changes whether calculated over a non-pre-processed timeseries and after pre-processing. In the first case, the TAC is strongly dominated by seasonality, while in the second case the TAC shows patterns that are instead more coherent with general climatic conditions.

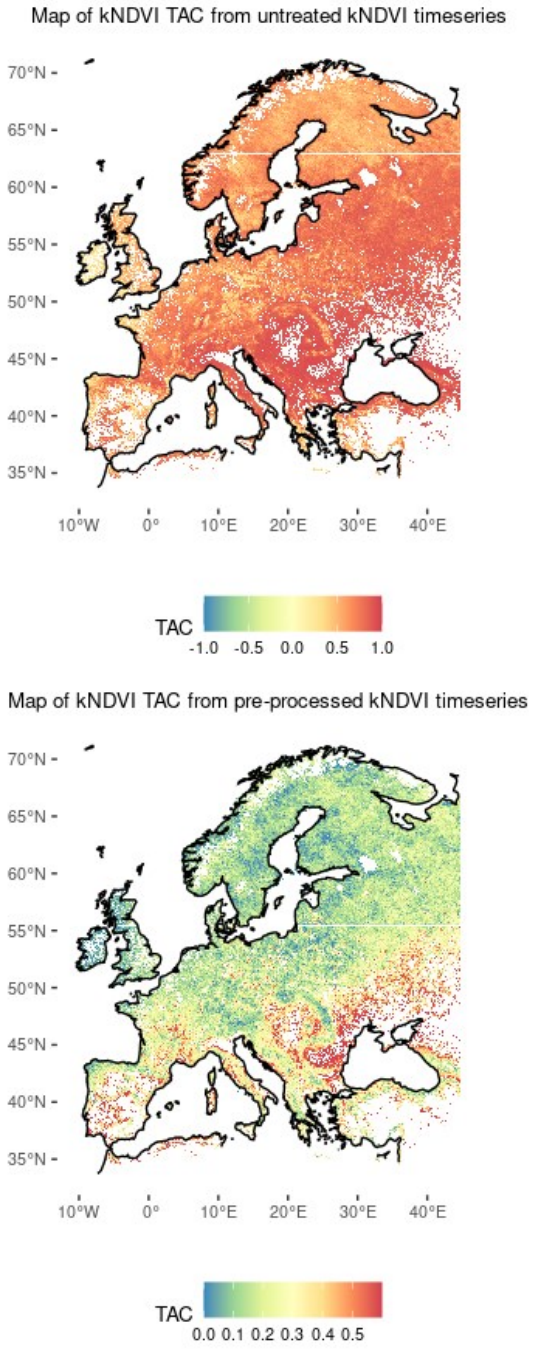

Figure 2. Long-term lag-1 TAC of kNDVI computed over an untreated timeseries (above) and over a pre-processed timeseries (below).

Figure 3 presents the performance of the trained RF model, in terms of predicted versus observed values of the kNDVI long-term lag-1 TAC on the testing data and in performance indicators of the RF model itself. The RF model achieves a R^2 of 0.898, with a mean squared error (MSE) of 0.0045 and an average overestimation (PBIAS) of -0.782.

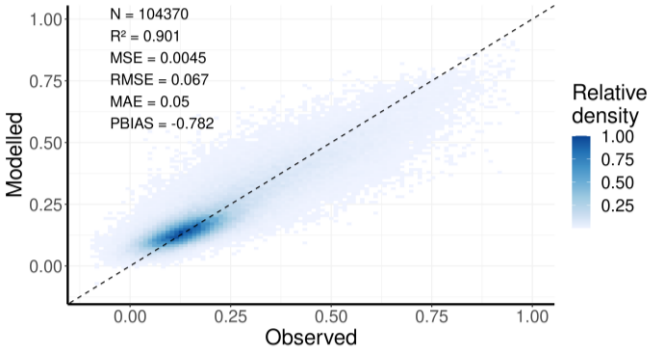

Figure 3. Observed versus modelled long-term lag-1 TAC. Number of binned records (N), coefficient of determination (R^2), mean squared error (MSE) and percent bias (PBIAS) of the RF model are shown in the labels.

In order to quantify and rank the influence of individual environmental and climatic factors on TAC, variable importance metrics have been extracted. These metrics allow to separate and identify main drivers of the kNDVI TAC. Figure 4 shows the environmental and climatic predictors ranked by variable importance in the Random Forest model. The percentage increase in MSE represents the mean decrease of accuracy in predictions when a variable is permuted, meaning the mean increase in MSE contribution by variable divided by its variability.

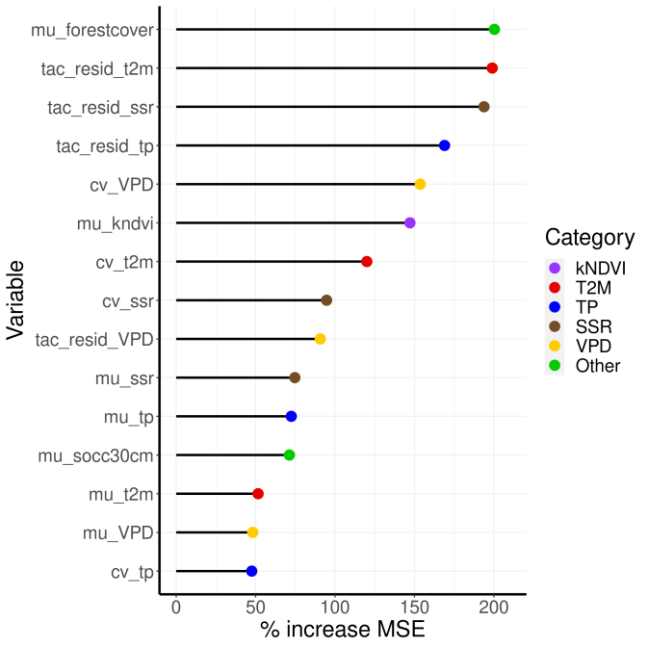

Figure 4. Predictors and corresponding variable importances of the RF model of TAC. Statistics indicated by mean (mu), coefficient of variation (cv) and temporal autocorrelation (tac). Acronyms of the environmental and climatic variables indicated in the Material and Methods section.

From Figure 4, it is clear how the most important predictor is forest cover, followed by the temporal autocorrelation of the climatic variables (2-meter air temperature, surface solar radiation and total precipitation).

Finally, in Figure 5 a map of the predicted values of long-term lag-1 TAC of kNDVI extended over the whole dataset and a map of the residuals from the Random Forest model are presented.

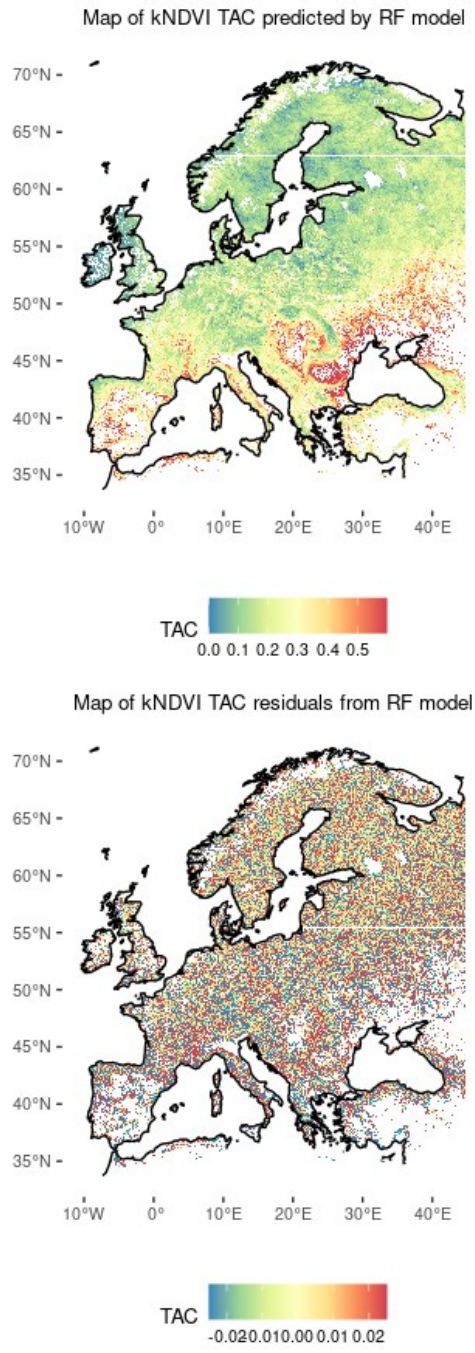

Figure 5. Predicted long-term lag-1 TAC of kNDVI and residuals of the RF model.

5. CONCLUSION

The study presents an overview of the pre-processing of historical, high-frequency and high-resolution timeseries of vegetation and climatic data with the aim of training a Random Forest model to explain a general resilience indicator, the long-term temporal autocorrelation of kNDVI, with climatic and environmental variables. The overall pre-processing highlighted the importance of retrieving stationary timeseries before the data are input in the machine learning model. The application of the RF model highlighted the strong influence of climatic variables as drivers of vegetation temporal autocorrelation.

REFERENCES

Bárta V., Lukeš P., Homolová L., 2021: Early detection of bark beetle infestation in Norway spruce forests of Central Europe using Sentinel-2. *International Journal of Applied Earth Observation and Geoinformation*, Volume 100, 2021, 102335, ISSN 1569-8432, doi.org/10.1016/j.jag.2021.102335.

Breiman, L., 2001: Random forests. Machine learning, 45, 5-32.

Camps-Valls G., Campos-Taberner M., Moreno-Martínez Á., Walther S., Duveiller G., Cescatti A., Mahecha M.D., Muñoz-Marí J., García-Haro F.J., Guanter L., Jung M., Gamon J.A., Reichstein M., Running S.W., 2021: A unified vegetation index for quantifying the terrestrial biosphere. *Science Advances*, 7 (9) (2021), Article eabc7447.

Copernicus Climate Change Service (C3S), 2017: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), cds.climate.copernicus.eu/cdsapp#!/home.

Dakos V., Carpenter S.R., Van Nes E.H., Scheffer M., 2015: Resilience indicators: prospects and limitations for early warnings of regime shifts. *Phil. Trans.* R. Soc. B 370: 20130263.

Forzieri G., Dakos V., McDowell N.G. et al., 2022: Emerging signals of declining forest resilience under climate change. *Nature* 608, 534–539 (2022). doi.org/10.1038/s41586-022-04959-9.

Friedl M., Gray J., Sulla-Menashe D., 2022: MODIS/Terra+Aqua Land Cover Dynamics Yearly L3 Global 500m SIN Grid V061 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2023-05-05 from doi.org/10.5067/MODIS/MCD12Q2.061.

Gorelick N., Hancher M., Dixon M., Ilyushchenko S., Thau D., Moore R., 2017: Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.*, 202, 18–27.

Hansen M. C., Potapov P. V., Moore R., Hancher M., Turubanova S. A., Tyukavina A., Thau D., Stehman S. V., Goetz S. J., Loveland T. R., Kommareddy A., Egorov A., Chini L., Justice C. O., Townshend J. R. G., 2013: High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* 342 (15 November): 850-53. 10.1126/science.1244693 Data available on-line at: https://glad.earthengine.app/view/global-forest-change.

Hengl T., Wheeler I., 2018: Soil organic carbon content in x 5 g / kg at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (Version v02) [Data set]. Zenodo. 10.5281/zenodo.1475457.

R Core Team, 2022: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Soille P., Burger A., De Marchi D., Kempeneers P., Rodriguez D., Syrris V., Vasilev V., 2018: A versatile data-intensive computing platform for information retrieval from big geospatial data. *Futur. Gener. Comput. Syst.*, 81, pp. 30-40.

Sun N., Liu N., Zhao X., Zhao J., Wang H., Wu D., 2022: Evaluation of Spatiotemporal Resilience and Resistance of Global Vegetation Responses to Climate Change. *Remote Sens.* 2022, 14, 4332. 10.3390/rs14174332.

Thonfeld F., Gessner U., Holzwarth S., Kriese J., da Ponte E., Huth J., Kuenzer C., 2022: A First Assessment of Canopy Cover Loss in Germany's Forests after the 2018–2020 Drought Years. *Remote Sensing*. 2022; 14(3):562. doi.org/10.3390/rs14030562.

Vermote E., Wolfe R., 2021: MODIS/Terra Surface Reflectance Daily L2G Global 1km and 500m SIN Grid V061 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2023-05-05 from doi.org/10.5067/MODIS/MOD09GA.061.