Remote Sensing of Environment 266 (2021) 112678

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Statistical inference for trends in spatiotemporal data

Anthony R. Ives [a,*], Likai Zhu [b], Fangfang Wang [c], Jun Zhu [d], Clay J. Morrow [a,e], Volker C. Radeloff [e]

[a] Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI 53706, United States of America
[b] Shandong Provincial Key Laboratory of Water and Soil Conservation and Environmental Protection, College of Resources and Environment, Linyi University, Linyi 276000, China
[c] Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609, United States of America
[d] Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, United States of America
[e] SILVIS Lab, Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, Madison, WI 53706, United States of America

ARTICLE INFO

Edited by: Marie Weiss

Keywords:
Statistical hypothesis testing
Spatiotemporal analysis
Global patterns in temporal trends
Spatial autocorrelation
Temporal autocorrelation
Large datasets

ABSTRACT

Global change analyses are facilitated by the growing number of remote-sensing datasets that have both broad spatial extent and repeated observations over decades. These datasets provide unprecedented power to detect patterns of time trends involving information from all pixels on a map. However, rigorously testing for time trends requires a solid statistical foundation to identify underlying patterns and test hypotheses. Appropriate statistical analyses are challenging because environmental data often have temporal and spatial autocorrelation, which can either obscure underlying patterns in the data or suggest false associations between patterns in the data and independent values used to explain them. Existing statistical methods that account for temporal and spatial autocorrelation are not practical for remote-sensing datasets that often contain millions of pixels. Here, we first analyze simulated data to show the need to account for both spatial and temporal autocorrelation in time-trend analyses. Second, we present a new statistical approach, PARTS (Partitioned Autoregressive Time Series), to identify underlying patterns and test hypotheses about time trends using all pixels in large remote-sensing datasets. PARTS is flexible and can include, for example, the effects of multiple independent variables, such as land-cover or latitude, on time trends. Third, we use PARTS to analyze global trends in NDVI, focusing on trends in pixels that have not experienced land-cover change. We found that despite the appearance of overall increases in NDVI in all continents, there is little statistical support for these trends except for Asia and Europe, and only in some land-cover classes. Furthermore, we found no overall latitudinal trend in greening for any continent, but some latitude by land-cover class interactions, implying that latitudinal patterns differed among land-cover classes. PARTS makes it possible to identify patterns and test hypotheses that involve the aggregate information from many pixels on a map, thereby increasing the value of existing remote-sensing datasets.

1. Introduction

The global environment is changing rapidly but not uniformly, and many trends differ spatially in their magnitude, intensity, and speed (Huang et al., 2017; Zhu et al., 2016). Remote-sensing time series make it possible to monitor trends thanks to long-term datasets of consistent observations of the Earth's surface. For example, the longest-running satellite program, Landsat, started in 1972 (Zhu et al., 2019), and the Advanced Very High Resolution Radiometer program started in 1978 (Tucker et al., 2005). These programs now contain nearly five decades of data, providing time series that are long enough to identify trends and to distinguish them from short-term fluctuations (de Beurs and Henebry, 2005). For example, many parts of the globe have exhibited a greening trend since the 1980s (Piao et al., 2019; Zhu et al., 2016) (Fig. 1). This greening trend appears to be most pronounced in northern high latitudes, as has been identified by both remote sensing (Myneni et al., 1997; Piao et al., 2011) and assessments of shrub cover (Ackerman et al., 2017; Fraser et al., 2014; Tape et al., 2006). However, greening trends in North America differ from Eurasia (Bi et al., 2013; Xu et al., 2013; Zhou et al., 2001), and there is considerable variation in shrubland expansion

* Corresponding author.
E-mail address: arives@wisc.edu (A.R. Ives).

https://doi.org/10.1016/j.rse.2021.112678
Received 1 February 2021; Received in revised form 1 August 2021; Accepted 23 August 2021
Available online 24 September 2021

and contraction at regional and local scales (Chen et al., 2021; Tape et al., 2012). Furthermore, strong greening trends in China and India are most likely due to afforestation and agricultural intensification (Chen et al., 2019), and in the arid Sahel region greening is most likely due to increasing precipitation (Dardel et al., 2014). In summary, there is a need to identify where the environment is changing, how much it is changing, and why.

When rigorously analyzing any time series of satellite data, it is useful to fit a statistical model. A statistical model makes it possible to identify and test explicit patterns in the data. A pattern of interest might be, for example, whether greening as measured by NDVI has been greater in the Arctic than at lower latitudes in the Northern Hemisphere, which appears to be the case most visibly in Alaska, parts of Canada, and parts of Siberia (Fig. 1a). While this pattern sounds simple, a statistical model reveals the true underlying complications. For example, can trends measured in adjacent pixels be treated as independent data points? The answer is clearly no, because adjacent points are likely to be affected by the same environmental fluctuations. But if adjacent points are not independent, is it impossible to include them together in a statistical test? If points have to be picked at distances far enough apart that they are independent, how far is "far enough," how many points might be left to be analyzed, and how much information is lost? While there are numerous statistical approaches for fitting statistical models for non-independent data (Cressie et al., 2015; Cressie and Kang, 2016; Cressie and Zammit-Mangion, 2016; Cressie, 1993; Finley et al., 2009; Harvey, 1993; Kang and Cressie, 2011, 2013; Tsay, 2014; Wikle et al., 2019), these methods are numerically intensive and struggle to scale up to the size of even coarse-resolution remote-sensing datasets.

Although a common approach when studying patterns in space and time is to analyze pixels individually (pixel-scale), researchers are generally interested in the underlying patterns that span many pixels on a map (map-scale). Analyzing map-scale patterns involves two statistical steps. The first step is estimation, in which a statistical model is built containing parameters that quantitatively describe the pattern of interest, and the data are used to estimate the values of these parameters. To make the estimates of the parameters informative about the patterns of real interest, the model must often include additional parameters to absorb those patterns that are not of immediate interest but might deceptively look like the patterns of real interest. As a specific example, suppose we are interested in whether increases in greenness differ

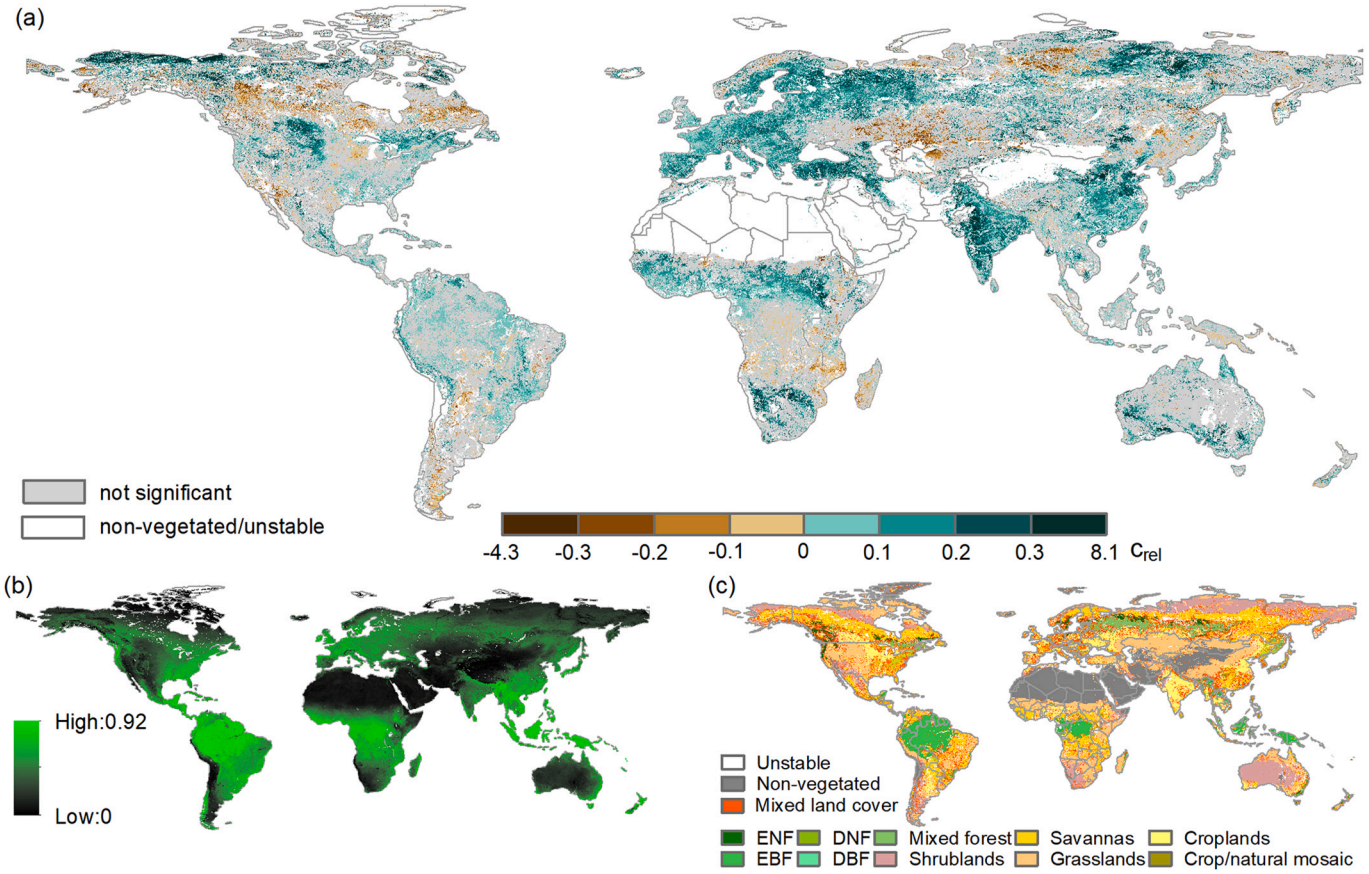

Fig. 1. Global patterns of trends in annual cumulative NDVI from 1982 to 2015 from NDVI3g data. (a) Estimates of time trends from regression with autocorrelated errors (AR, Eq. (2)) including only pixels for which there was no change in land-cover class (not categorized as "unstable" as depicted in panel c). The time trends are measured by fitting an autoregressive time-series model (*2.2.2 Regression with temporally autocorrelated errors*) to data from each pixel and dividing the trend coefficient by the pixel mean to give the relative change in NDVI. For clarity, pixels for which the null hypothesis of no trend was rejected (at the 0.05 significance level) in the pixel-level time series are shown in brown and green corresponding to decreasing and increasing trends, respectively. Pixels for which the null hypothesis was not rejected are shown in gray, and pixels that were unvegetated or had changes in land-cover class are shown in white (*2.4 Application to global NDVI data*). (b) Global pattern of the mean annual cumulative NDVI for the period 1982–2015. For the Northern Hemisphere, the annual cumulative NDVI was calculated from 8-day NDVIs from January to December of a given year. For the Southern Hemisphere, the cumulative NDVI was computed from July to December of the previous year and from January to June of the current year. Pixels with mean annual NDVI values less than 0 were set to zero. (c) Global pattern of land-cover classes from 2001 to 2015. This map is derived from the MODIS land-cover product (MCD12Q1 V006), which provides global land cover annually at 500-m spatial resolution (*2.4 Application to global NDVI data*). The stable land-cover class is defined as the land-cover class of a pixel that did not change from 2001 to 2015. We upscaled MODIS land-cover pixels with majority rule to match NDVI3g NDVI pixels. A mixed pixel has no land-cover representing more than 50% of the pixel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

among land-cover classes. Map-scale patterns in global NDVI (Fig. 1a) could reflect differences among land-cover classes (Fig. 1c), but also differences in regional temperature and precipitation. A statistical model could contain both parameters for land-cover classes and parameters that absorb regional variation in climate to statistically quantify the effects of land-cover class on changes in NDVI while factoring out the possible effects of regional climate variation.

The second statistical step is hypothesis testing, in which the estimate of a parameter is assessed against the probability of obtaining an estimate more extreme under a null hypothesis. A null hypothesis could be simple, such as whether a parameter equals zero. More-complex null hypotheses can be formulated to address other types of questions. For example, whether greening in NDVI differs among land-cover classes could be tested with the null hypothesis that the parameters summarizing land-class-specific greening did not differ. For another example, a null hypothesis could be that latitudinal increases in greening are the same for all land-cover classes. A hypothesis test produces a *P*-value for the map-scale pattern in question, rather than *P*-values for time trends within individual pixels. The map-scale *P*-value does not indicate the magnitude of a map-scale pattern; the magnitude is given by the point or interval estimates of the parameter. Nonetheless, *P*-values are related to the uncertainty in the parameter estimate, and they give a rigorous and general way to determine the strength of evidence that can be placed on conclusions about patterns in the data. We caution, however, against using *P*-values to determine whether a pattern is significant or not based on a threshold such as "$P < 0.05$" (Benjamin et al., 2018; Wasserstein et al., 2019); instead, *P*-values give the probability of observing something more extreme than the estimated value of a parameter in a statistical model under a specified null hypothesis, and this full context of what a *P*-value is should be considered.

Although the concepts of statistical estimation and hypothesis testing are familiar, there is no simple way of implementing estimation and hypothesis testing for large spatiotemporal datasets due to the statistical challenges of temporal and spatial autocorrelation. Temporal autocorrelation occurs when values of a variable between consecutive time steps are not statistically independent. In natural ecosystems, temporal autocorrelation is common, both within years and among years, because organisms need time to grow and sometimes to die. For example, a drought may cause years of browning followed by years of greening after the drought ends (de Beurs et al., 2015; Wessels et al., 2012). When testing the hypothesis that there are long-term trends in NDVI, however, the changes caused by a single drought event should be considered as stochastic, because the single event might not predict the long-term change in NDVI. Many of the approaches that are commonly used to test for statistical significance in satellite datasets do not account adequately for temporal autocorrelation, despite acknowledgments that temporal autocorrelation is a problem (Udelhoven, 2011; Zhou et al., 2001). Approaches that do not account fully for temporal autocorrelation, as we will show in our simulations, include least-squares regression (Dardel et al., 2014; Ju and Masek, 2016; Myneni et al., 1997; Piao et al., 2015) and the Mann-Kendall test with the Theil-Sen slope estimator (Fensholt et al., 2015; Fensholt and Proud, 2012; Zhu et al., 2016). Valid statistical tests that have been used in remote sensing include a modified seasonal Kendall test that accounts for temporal autocorrelation between seasons within a given year (but not for autocorrelation among years) (de Beurs et al., 2015; Hirsch and Slack, 1984) and the size-robust trend test (Bi et al., 2013; Fomby and Vogelsang, 2002; Vogelsang, 1998; Xu et al., 2013), which we will also discuss.

Spatial autocorrelation presents an even greater statistical challenge than temporal autocorrelation. Spatial autocorrelation occurs when pixels exhibit non-independent patterns, such as when time series in nearby pixels show similar (or predictably dissimilar) fluctuations. For example, multiple pixels in the same region may experience the same drought and show similar patterns of browning and subsequent greening. When this occurs, pixels are not independent. Given that satellite datasets typically contain millions of pixels, treating them as independent would lead almost any statistical test to be "significant" (Wikle et al., 2019), a concern that has been raised in the remote-sensing literature (de Beurs et al., 2015; Tomaszewska et al., 2020; Zhou et al., 2001). There are statistical methods for adjusting the significance level of pixel-scale statistical tests of time trends (Cortés et al., 2020; Cortés et al., 2021; Wilks, 2006, 2016); nonetheless, these are not frequently used, and although they give pixel-scale *P*-values corrected for spatial autocorrelation, they do not lead to map-scale statistical tests that aggregate the power from all pixels on a map.

Here, we develop and evaluate a new statistical approach, PARTS (Partitioned Autoregressive Time-Series) analysis, that makes it possible to test map-scale hypotheses about patterns of temporal trends from all pixels in large remote-sensing datasets. Our objectives are:

(i) to highlight the need to account for temporal and spatial autocorrelation using simulations for which the underlying processes are known. Using simulations makes it possible to assess the output of statistical models when we know the true (simulated) patterns.
(ii) to explain the statistical approach underlying PARTS. We validate the statistical approach using simulations presented here and in Ives et al. (2021).
(iii) to use PARTS to estimate parameter values and test hypotheses about global greening trends from 1982 to 2015 based on mean annual values of NDVI at 8-km resolution. These analyses focus on changes in greening that are not associated with land-cover changes by excluding pixels in which land cover has been unstable (Fig. 1c).

2. Methods

2.1. Simulation model

To understand how temporal and spatial autocorrelation can affect analyses of time trends, we designed simulations to generate realistic but simple datasets with varying strengths of both types of autocorrelation. The simulation model is

$$x_i(t) = a_i + c_i t + \varepsilon_i(t)$$
$$\varepsilon_i(t) = \beta_i \varepsilon_i(t-1) + \delta_i(t) \quad (1)$$

where $x_i(t)$ is the value of interest (e.g., NDVI) in pixel (location) i in year t, a_i is the intercept, and c_i is a coefficient that measures the effect of time t on $x_i(t)$, where $t = 1, 2, \ldots, T$. Environmental variation that affects changes in $x_i(t)$ from one year to the next is given by the random variable $\varepsilon_i(t)$. To account for possible temporal autocorrelation, $\varepsilon_i(t)$ is governed by a Gaussian (i.e., normal) autoregressive process generated from the normal random variable $\delta_i(t)$ that has mean zero and variance σ^2, with values independent through time so that $E[\delta_i(t)\delta_i(s)] = 0$ for $s \neq t$. The dependence of $\varepsilon_i(t)$ on $\beta_i\varepsilon_i(t-1)$ generates temporal autocorrelation. If $\beta_i = 0$, then $\varepsilon_i(t)$ and $\varepsilon_i(t-1)$ are not correlated, and the random fluctuations in $x_i(t)$ are independent from one year to the next. In many cases, however, we would expect random fluctuations to be positively correlated through time ($\beta_i > 0$), and negative autocorrelation ($\beta_i < 0$) is also possible.

To include spatial autocorrelation, we assume that the normal random variables $\delta_i(t)$ and $\delta_j(t)$ from pixels i and j are correlated such that $\text{cor}[\delta_i(t), \delta_j(t)] = \exp(-d_{ij}/r)$, where d_{ij} is the geographic distance between pixels and r is the "range" parameter. Values of $\delta_i(t)$ and $\delta_j(s)$ are independent when $s \neq t$. The larger r, the greater the distance at which pixels share environmental fluctuations and hence the greater the spatial autocorrelation. For simulations, we scale distances so that the maximum distance between pixels on a map equals 1. Therefore, a value of $r = 0.1$ implies that the range is 10% of the extent of the map. With $r = 0.1$, the correlation between $\delta_i(t)$ and $\delta_j(t)$ for sites i and j that are

separated by a distance of 0.1 is $\exp(-1) = 0.37$. Spatial autocorrelation in $\delta_i(t)$ imparts spatial autocorrelation to the random variable $\varepsilon_i(t)$ and hence to $x_i(t)$.

The time trend of interest is given by the parameter c_i. If $c_i > 0$, then there is a positive time trend in which the expectation of $x_i(t)$ increases linearly through time. The time trend is deterministic, rather than stochastic, because it represents a fixed change in the expected value of $x_i(t)$.

2.2. Time-series analyses

2.2.1. Least-squares linear regression, Mann–Kendall, and size-robust trend tests

In the remote-sensing literature, three methods are commonly used to fit pixel-level time series. The most common is least-squares linear regression (LS) in which $x_i(t)$ is regressed on t (e.g., Fensholt and Proud, 2012; Myneni et al., 1997; Piao et al., 2011). This method ignores the potential for temporal autocorrelation ($\beta_i \neq 0$). A commonly used alternative is a Mann-Kendall nonparametric significance test (MK) combined with the nonparametric Theil-Sen slope estimator (Fensholt et al., 2015; Zhu et al., 2016). The MK test can also be modified to account for seasonal effects for time series with multiple values per year: the seasonal effect is removed by comparing values from the same seasons among years (de Beurs and Henebry, 2005; de Beurs et al., 2015). While this method accounts for temporal autocorrelation between consecutive seasons, the data we analyze are annualized (Fig. 1a), and therefore we applied the standard MK test that does not account for temporal autocorrelation. Finally, we used the size-robust trend test (SR) (Fomby and Vogelsang, 2002; Vogelsang, 1998), which has been used with remote-sensing data (Bi et al., 2013; Xu et al., 2013). The SR test correctly accounts for temporal autocorrelation, although at the expense of statistical power. Also, while the SR test determines whether or not a trend is statistically significant at a given significance level (e.g., $P < 0.05$), it does not give an exact P-value (e.g., $P = 0.023$) like LS and MK tests.

2.2.2. Regression with temporally autocorrelated errors

To account for temporal autocorrelation, we fit pixel-level time series using the model

$$x_i(t) = a_i + c_i t + \varepsilon_i(t) \tag{2}$$

Here, $\varepsilon_i(t)$ is a univariate stationary first-order Gaussian autoregressive (AR) process with mean zero for each separate time series, and the vector of values of $\varepsilon_i(t)$ ($t = 1, 2, \ldots, T$) has a multivariate Gaussian distribution, $N(0, \sigma^2\mathbf{\Sigma}_i)$. The correlation matrix $\mathbf{\Sigma}_i$ contains the elements $\text{cor}[\varepsilon_i(t), \varepsilon_i(s)] = \beta_i^{|t-s|}$ for all t and s. This regression model with lag-1 autoregressive error terms for a single pixel is commonly fit using Maximum Likelihood (ML) (Box et al., 1994). However, ML gives biased estimates of c_i (results not shown), and therefore instead we use Restricted Maximum Likelihood (REML) (Ives et al., 2010). We refer to this method as AR (autoregression).

2.2.3. Simulations of pixel-level time series

To evaluate the performance of each method of time-series analysis, we simulated 5000 datasets for each combination of $c_i = 0, 0.5, 1, 2$ and $\beta_i = 0, 0.2, 0.4, 0.6$ for 30 years, with $a_i = 0$ (Eq. (1)). We applied all competing methods (LS, MK, SR, AR) to assess their type I error rates for the case when $c_i = 0$. The type I error rates are given by the proportion of simulations for which the null hypothesis is rejected given that it is true. Thus, for datasets simulated under the null hypothesis, a correct statistical test should reject the null hypothesis in 5% of the simulations when using a significance level of alpha = 0.05. If the null hypothesis is rejected in a fraction greater than 0.05, then the type I error rates are inflated, which corresponds to the P-values given by the hypothesis test being too low. We assessed the statistical power of the four methods when $c_i = 0.5$, 1, and 2 as their ability to reject the null hypothesis that $c_i = 0$. To test the methods when there are disturbance events (drought, fire, logging, etc.), we ran simulations in which $x_i(t)$ was decreased by 2 at either year 5 or year 20. These disturbance datasets were otherwise identical to the other simulations.

2.3. PARTS

2.3.1. Statistical summary

Current statistical methods for spatiotemporal data generally evolved from spatial analyses, especially kriging (Krainski et al., 2019; Wikle et al., 2019). For example, Cressie and Kang (2010) and Kang and Cressie (2011) extend classical kriging to a Spatiotemporal Random Effects (STRE) model to interpolate and forecast temporal trends in data. Similar models have also been formulated for Bayesian analyses (Berrocal et al., 2010), and both frequentist and Bayesian spatiotemporal models emphasizing kriging have been applied to large spatial datasets (Cressie and Kang, 2010; Katzfuss and Cressie, 2011; Zammit-Mangion and Cressie, 2018). There are parallel efforts to fill gaps in weather station data (Finley et al., 2012) and to reduce noise and fill clouds with a spatiotemporal Savitsky-Golay filter (Cao et al., 2018).

PARTS takes a different approach by analyzing the time series of each pixel separately to give estimates of parameters quantifying time trends, and then using Generalized Least Squares (GLS) regression (Ives and Zhu, 2006; Judge et al., 1985) to analyze the spatial distribution of the parameter estimates. Thus, PARTS collapses the temporal dimension of the data into parameter estimates and then performs the spatial analysis on the parameter estimates. To computationally handle large maps, PARTS analyzes random partitions of the data separately and then combines the analyses from all partitions in a way that accounts for the data from different partitions being non-independent. There are at least four advantages of this overall approach. First, it makes the task of analyzing maps of millions of pixels feasible. While we do not know of any examples of analyses with maps containing more than one million pixels using existing statistical spatiotemporal methods (e.g., Wikle et al., 2019), PARTs can fit a model such as we present here (*3.2 Global trends in browning and greening*) for time trends across 30 years on a 1,000,000-pixel map using 8 cores at 2.79 GHz (CPU: AMD Ryzen 73,700× with a AMD X570 chipset; memory: dual channel DDR4 at 1800 MHz) in less than 30 min, and the computational burden scales linearly with increasing numbers of pixels for a given partition size. Second, the first step of analyzing individual time series using a variety of time-series models is already standard in the remote-sensing literature, and PARTS can be viewed as an extension of these methods. Third, because it is based on GLS, PARTS lends itself to standard hypothesis tests that are familiar in regression-style analyses, including analysis of variance (ANOVA) and analysis of covariance (ANCOVA). Fourth, PARTS requires fewer assumptions about the structure of the data than other spatiotemporal methods (e.g., Wikle et al., 2019) because it collapses the temporal dimension into a single variable. For example, a full spatiotemporal model would require specifying and estimating differences in the mean values (intercepts) of a response variable among pixels along with the time trends when fitting the entire dataset. In contrast, PARTS uses only the estimates of the pixel-scale time trends in the spatial analysis. Therefore, in PARTS there is no need to specify in the model how the intercept varies among pixels if the goal is to analyze time trends.

2.3.2. Statistical procedure

The PARTS procedure is:

1. Perform AR regression for the time series in each pixel to give pixel-specific estimates of the time trends, c_i.
2. Calculate the correlations between the residuals obtained from the fitted AR regression models for each pair of pixels, or each pair in a large subset of pixels. These correlations are then used to estimate

the spatial autocorrelation structure of the estimates of c_i (Ives et al., 2021).

3. Using the spatial autocorrelation structure from step 2, perform GLS of a spatial regression of the estimates of c_i against independent variables to give estimates of parameters and test hypotheses about map-scale patterns of time trends. It is also possible to estimate spatial autocorrelation in this step and skip step 2.
4. Steps 1–3 can be performed on maps of up to roughly 30,000 pixels. To analyze larger maps, analyses are performed by partitioning all pixels into subsets, analyzing each partition using step 3, and then combining the results for the partitions into an overall statistical test.

Steps 2–4 can be applied to any parameter estimated from time-series analyses of each pixel, not just c_i. Several points need to be explained in more detail.

First, for an overall model like that given by Eq. (1), the estimates of c_i are not independent among pixels due to spatial autocorrelation. Therefore, the "best" estimates of c_i (technically, the best linear unbiased estimate, BLUE, Judge et al., 1985) require estimating the values of c_i for all pixels simultaneously, and this would entail a large computational burden. The AR estimates of c_i given by $\widehat{c}_i$ are computed for each pixel separately and therefore they are not the BLUE estimates of c_i. Nonetheless, they are unbiased and in practice are numerically close to the BLUE estimates.

Second, we use GLS to regress values of the estimates $\widehat{c}_i$ against independent variables w_{i1}, w_{i2}, …, w_{ip} for p independent variables,

$$\widehat{c}_i = b_0 + b_1 w_{i1} + \ldots + b_p w_{ip} + \gamma_i. \quad (3)$$

GLS explicitly incorporates the spatial correlation matrix to account for the correlations between $\widehat{c}_i$ and $\widehat{c}_j$ by assuming the N spatial errors γ_i follow a multivariate Gaussian distribution with correlation matrix $\mathbf{V}$, N $(0, \sigma^2_\gamma \mathbf{V})$. We assume that the correlation between γ_i and γ_j, cor$[\gamma_i, \gamma_j]$, decays according to some function $v(d_{ij})$ of the distance d_{ij} between pixels i and j. For example, in the simulation model (Eq. (1)), $v(d_{ij}) = \exp(-d_{ij}/r)$. For fitting the NDVI data (Fig. 1a), we use the more-general exponential-power function $v(d_{ij}) = \exp(-(d_{ij}/r)^g)$. The second parameter g controls the shape of decline with distance; when $g = 2$, $v(d_{ij})$ is Gaussian; when $g = 1$, $v(d_{ij})$ is exponential; and values of $g < 1$ imply more leptokurtic functions.

Third, we also assume that there is a "nugget effect" that represents the proportion of local variation in γ_i that is not spatially autocorrelated. Specifically, we let $\mathbf{V} = (1 - nugget)\, v(\mathbf{D}) + nugget\, \mathbf{I}$, where $\mathbf{D}$ is the $N \times N$ distance matrix between all N pixels (i.e., $\mathbf{D}$ has elements d_{ij}), and $\mathbf{I}$ is the identity matrix. The nugget is estimated by maximum likelihood during step 3 of PARTS. We estimate the nugget effect because measurement error of the individual time series will appear as local variation, and because there could be real (biological) variation in time trends at the local scale. The parameters in $v(\mathbf{D})$ (e.g., r and g) can be estimated in step 2, before the GLS is performed in step 3. This approach uses the result that cor$[\gamma_i, \gamma_j]$ is roughly proportional to cor$[\varepsilon_i(t), \varepsilon_j(t)]$, with the proportionality exact when the strength of temporal autocorrelation is the same for all time series ($\beta_i = \beta_j$) (Ives et al., 2021). Independent estimation of parameters in $v(d_{ij})$ using cor$[\varepsilon_i(t), \varepsilon_j(t)]$ will speed computation time and can lead to better estimates of b_k ($k = 0, 1, \ldots, p$). The parameters in $v(d_{ij})$ giving the spatial autocorrelation (e.g., r and g) can alternatively be estimated simultaneously with the *nugget* in step 3 (Ives et al., 2021). We consider the effects of estimating r in step 3 when analyzing a real dataset (*3.2.1 Alaska*).

Fourth, because the GLS performs statistical tests using the estimates of the coefficients $\widehat{c}_i$ from the time-series analyses, it implicitly combines variation in trends among pixels from two sources: the variation in the estimates caused by temporal variation in the time series within pixels (spatiotemporal variation), and the variation in the underlying time trend among pixels (purely spatial variation). The relative magnitudes of these two sources of variation can be roughly obtained by comparing the standard errors of the estimates $\widehat{c}_i$ (step 1) with the standard deviation of the random error in Eq. (3), σ_γ (step 3). If the standard errors of $\widehat{c}_i$ are small compared to σ_γ, then much of the spatial variation in $\widehat{c}_i$ is caused by fixed spatial differences in trends among pixels. Conversely, if standard errors of $\widehat{c}_i$ are large compared to σ_γ, then variation in $\widehat{c}_i$ can be attributed to uncertainty in their estimates that is caused by temporal fluctuations over the course of the time series. Comparing the standard errors of the estimates $\widehat{c}_i$ with σ_γ provides a way to assess the source of variation in the GLS analysis.

Fifth, for maps with more than 30,000 pixels, hypothesis testing can be performed by partitioning the map into n_p subsets of pixels drawn randomly without replacement from throughout the map (step 4). The hypothesis is tested for each of the n_p partitions separately, and then the results from the n_p tests are combined to give a single test of the hypothesis for the entire map. The hypothesis testing can be done with standard approaches: *t*-tests, *F*-tests, and Likelihood Ratio Tests (LRT). We present two methods for combining tests. The first method combines the n_p tests by selecting the partition with the strongest result (lowest *P*-value) and then correcting this *P*-value for the n_p multiple comparisons using either a Hochberg (1988) (Bonferroni) or False Discovery Rate (FDR) adjustment (Benjamini and Hochberg, 1995). Wilks (2006, 2016) applies this approach for tests of single pixels, while here we apply it to tests of partitions of pixels. For correcting the single lowest *P*-value, both Hochberg and FDR corrections are valid for non-independent tests (Nichols and Hayasaka, 2003); this is important, because the pixels from different partitions are not independent and hence the n_p tests are not independent. Nonetheless, the consequence of non-independence is loss of statistical power to reject the null hypothesis across the entire dataset. The second method for combining the n_p tests is detailed in Ives et al. (2021). It involves obtaining the statistical distribution of the test statistic when measured on the correlated partitions; this distribution of the test statistic then gives an omnibus test for the hypothesis for the entire map. We use partition versions of *t*-tests for individual parameters and LRTs for hypotheses involving more than one parameter. We used both multiple comparison and partition methods, because the first is well-established in the literature, while the new, second method has greater statistical power (which we demonstrate with our simulations).

2.3.3. Spatiotemporal patterns

To illustrate the problem of spatial autocorrelation, we used Eq. (1) to simulate data on a 100 × 100 pixel map, with each pixel having a time series of length 30. We followed the common practice of showing only pixels that have statistically significant trends as determined by AR, using a significance level of alpha = 0.1 so that roughly 10% of pixels should be flagged as significant if the null hypothesis (no time trend) were true.

2.4. Application to global NDVI data

To illustrate PARTS, we analyzed trends in mean annual NDVI data globally. Mean annual cumulative NDVI is strongly correlated with vegetation productivity (Chen et al., 2014; Pettorelli et al., 2005) and a strong predictor of species richness (Hobi et al., 2017; Radeloff et al., 2019). Trends in mean annual NDVI reflect the response of ecosystems to climate change and human activities such as grassland degradation, crop status, and forest management (Chen et al., 2019; Fensholt et al., 2012; Wessels et al., 2012). We use the GIMMS3g dataset because it is one of the most widely used remote-sensing datasets for time-series and trend analyses. The dataset is processed to suppress multiple sources of "noise", and using mean annual data in effect increases smoothing in time and space. This smoothing makes the temporal and spatial autocorrelation more visible in the data. Although we illustrate our methods with the mean annual GIMMS3g dataset, the issues of spatial and temporal autocorrelation also arise in other more "noisy" data. The "noise" in other datasets might decrease the visibility of temporal and spatial

autocorrelation, but it will also decrease the strength of evidence for possible patterns that a researcher might want to estimate and test: the noise relative to the signal will increase, but the signal will still contain temporal and spatial autocorrelation. Therefore, there is no less reason to use methods that account for temporal and spatial autocorrelation in "noisy" data.

We generated mean annual NDVI values across the globe from 1982 to 2015 using the NDVI3g dataset produced from Advanced Very High Resolution Radiometer (AVHRR) instruments. The NDVI3g data have been corrected to minimize the adverse effects of various factors such as sensor calibration loss, orbit drift, and volcanic eruption compared to previous AVHRR NDVI data (Pinzon and Tucker, 2014). The NDVI3g data are highly consistent with an earlier version of the comparable MODIS product (Collection 5) but span more years and have a longer history (Fensholt and Proud, 2012), which is why we analyzed NDVI3g instead of MODIS Collection 6 vegetation data (Heck et al., 2019; Zhang et al., 2017). NDVI3g has been widely used to investigate vegetation productivity and phenological change, and to map land cover (e.g., Chen et al., 2014; Fensholt and Proud, 2012; He et al., 2017). We acquired the raw NDVI3g dataset (version 1.0) from the NASA Ames Ecological Forecasting Lab (https://ecocast.arc.nasa.gov/data/pub/gimms/) which has a spatial resolution of 0.083 degree (~8 km) and a temporal interval of 15 days (two NDVI values per month).

To calculate mean annual NDVI, we preprocessed the raw data to obtain both NDVI values and their quality flags. We then applied the quality flags to their corresponding NDVI values and used linear interpolation to predict the pixels with possible snow/cloud cover (flag = 3). We calculated the annual NDVI as the sum of positive NDVI values over 12 months (Hobi et al., 2017; Pettorelli et al., 2005; Radeloff et al., 2019). For pixels in the Northern Hemisphere, the 12 months ran from January to December, while for pixels in the Southern Hemisphere, they ran from July to June of the next year (Fig. 1b).

For our trend analyses, we excluded non-vegetated areas and areas with unstable land-cover classes, thereby excluding changes in NDVI caused by changes in land-cover class. Land-cover class was determined using the MODIS land-cover product (MCD12Q1 V006), which gives global maps of land cover at annual time steps and 500-m resolution for 2001–2019 (Friedl et al., 2002; Sulla-Menashe et al., 2019). We first determined whether land-cover change was stable at the 500-m pixel scale, classifying pixels as unstable if its land-cover class changed between 2001 and 2015. Then, to match the spatial resolution of NDVI3g data, we used the majority method to determine the land-cover class of the coarser NDVI3g data pixels. Specifically, a NDVI3g pixel was assigned to the dominant land-cover class of all finer MODIS pixels within it. If the dominant land-cover type was unstable, we removed the pixel from further analyses; for Africa, Asia, Australia, Europe, North America and South America, 9.6%, 8.8%, 9.0%, 6.0%, 7.0% and 8.5% of the NDVI3g pixels were so removed. If the percentage of the dominant land-cover class was less than 50%, the NDVI3g pixel was assigned to the mixed land-cover class (Fig. 1c); for Africa, Asia, Australia, Europe, North America and South America, 14.7%, 22.6%, 11.5%, 36.7%, and 25.5% of the NDVI3g pixels were categorized as stable with mixed land-cover class. All analyses of land-cover dynamics (e.g., unstable or stable) were conducted on Google Earth Engine.

We performed AR time-series analyses for each pixel, 1982–2015. To give estimates of c_i relative to the mean cumulative NDVI over the time series, we divided $\widehat{c}_i$ by this mean and denoted the result c_{rel} (dropping the subscript i for clarity). An outlier test was used to identify pixels with very high or low values of c_{rel}: values were considered outliers if their probability was less than $0.1/N$ assuming a Gaussian distribution, where N is the number of pixels on the map. These typically occurred at boundaries, either between land and water, or between land and barren areas. At these boundaries, the mean annual NDVI was small, so c_{rel} was large.

We first performed a detailed analysis of Alaska, USA, west of −141° longitude. We analyzed all stable land-cover classes that occurred in at least 0.5% of pixels, resulting in three land-cover classes. To analyze patterns in time trends among land-cover classes, we further limited analyses to pixels that contained at least 50% of a single land-cover class; this reduced the number of pixels from 29,089 to 20,694. For analyses that did not include land-cover classes, we used the same dataset so that results would be compatible with analyses including land-cover classes. We analyzed Alaska because its area is small enough that patterns are easily seen, yet large enough to present a statistical challenge.

We then analyzed the global data separately for each continent. For analyses that did not include land-cover classes, we used pixels that included mixed land-cover classes. For analyses that included land-cover classes, we included only pixels that contained at least 50% of a single land-cover class.

The software we used to analyze the data and simulations is available as the R package (R Core Team, 2021) called remotePARTS at https://github.com/morrowcj/remotePARTS.

3. Results

3.1. Simulations

Our simulations highlighted the effects of both temporal and spatial autocorrelation on trend analyses, and provided validation of PARTS for analyzing remote-sensing datasets.

3.1.1. Temporal analyses

Using simulated time series, we compared least-squares regression (LS), the Mann-Kendall/Theil-Sen method (MK), size-robust trend analysis (SR), and regression with autoregressive errors (AR). When simulations contained no temporal autocorrelation ($\beta_i = 0$, Eq. (1)), all four methods gave acceptable type I error rates when there was no time trend ($c_i = 0$); at a significance level of alpha = 0.05, roughly 5% of the 5000 simulated datasets were identified as having significant trends (Table 1). However, when there was strong temporal autocorrelation ($\beta_i = 0.8$), LS and MK rejected the null hypothesis that $c_i = 0$ for roughly 50% of the simulations. In contrast, SR showed no inflated type I error

Table 1
Proportion of simulated time series with 30 time points for which a trend was deemed statistically significant, depending on differing values of simulated trend c_i and temporal autocorrelation β_i.

c_i	β_i	LS	MK	SR	AR
0	0.0	0.05	0.05	0.04	0.06
0	0.2	0.10	0.10	0.03	0.07
0	0.4	0.18	0.17	0.03	0.07
0	0.6	0.30	0.28	0.03	0.09
0	0.8	0.51	0.48	0.03	0.14
0.5	0.0	0.12	0.11	0.08	0.13
0.5	0.2	0.17	0.16	0.06	0.11
0.5	0.4	0.24	0.22	0.04	0.11
0.5	0.6	0.34	0.32	0.03	0.11
0.5	0.8	0.51	0.48	0.03	0.15
1.0	0.0	0.33	0.32	0.20	0.34
1.0	0.2	0.36	0.33	0.12	0.25
1.0	0.4	0.37	0.34	0.07	0.19
1.0	0.6	0.44	0.41	0.05	0.17
1.0	0.8	0.52	0.49	0.03	0.15
2.0	0.0	0.87	0.84	0.60	0.84
2.0	0.2	0.81	0.78	0.38	0.67
2.0	0.4	0.75	0.72	0.21	0.48
2.0	0.6	0.66	0.63	0.10	0.33
2.0	0.8	0.62	0.58	0.05	0.22

When $c_i = 0$, 5% of the time series is expected to be rejected under the significance level of alpha = 0.05. We applied least-squares regression (LS), a Mann-Kendall/Theil-Sen test (MK), the size-robust trend test (SR), and regression with autocorrelated errors (AR). For each combination of c_i and β_i, 5000 simulations were performed. When $c_i = 0$, estimates of c_i were equally likely to be positive or negative, including those associated with values of $P < 0.05$.

rates, while the inflation for AR was much less than LS and MK. When the null hypothesis was false (c_i = 0.5, 1, and 2) and there was no temporal autocorrelation ($\beta_i = 0$), SR had reduced power to reject the null hypothesis compared to LS, MK, and AR. When there was temporal autocorrelation ($\beta_i > 0$), AR had greater power than SR, although increasing temporal autocorrelation reduced the power for both methods (Table 1). Because the type I error rates of LS and MK were so inflated when there was temporal autocorrelation, it is inappropriate to assess their statistical power; they had high rejection rates even when the null hypothesis was true.

The reason for the inflated type I error rates for LS and MK can be seen in graphs of time series from the simulations (Fig. 2). For an example with a temporal trend but no autocorrelation ($c_i = 2$, $\beta_i = 0$, Fig. 2a), LS, MK, and AR correctly identified the trend, while SR did not. In this case, the predictions from LS, MK, and AR for the future 20 years would be reliable. However, when there was no trend but strong autocorrelation ($c_i = 0$, $\beta_i = 0.9$, Fig. 2b), LS and MK identified the autocorrelation as a trend, even though the increase caused by the autocorrelation was transient and did not predict the following 20 years.

From the same simulations, we also computed the mean and standard deviation of the estimates of the time trend $\hat{c}_i$ (Table 2). While all methods gave approximately unbiased estimates, the standard deviation of the estimates increased with temporal autocorrelation (β_i). AR had the most accurate estimates (i.e., almost no bias and low standard deviations), while SR had the least accurate estimates.

Temporal shocks to the time series caused similar challenges for all methods. When negative shocks occurred early in the time series, all methods were likely to estimate a positive time trend, and shocks towards the end of the time series led to frequent negative estimates (Table 3; Fig. 2c,d). Overall, however, type I error rates were no more inflated than for the case of only autocorrelation in the absence of shocks (compare Tables 1 and 3 for the same values of β_i).

To summarize, our simulations of time series show that temporal autocorrelation can "fool" statistical tests and cause inflated type I error rates. Inflated type I error rates are especially a problem for methods like LS and MK that do not account for temporal autocorrelation, but even AR had moderately inflated type I errors when temporal autocorrelation was high. Nonetheless, AR gave the most accurate estimates of the time trend, having both low bias and low standard errors (Table 2). Furthermore, it performed better than other methods when confronted with data containing "shocks" (Table 3). For the PARTS method, accuracy of the estimates is most important, because PARTS does not use the *P*-values calculated from pixel-level time series. For our simulations, the LS estimates are similar to the AR estimates in accuracy, although for shorter time series LS can perform noticeably worse. Furthermore,

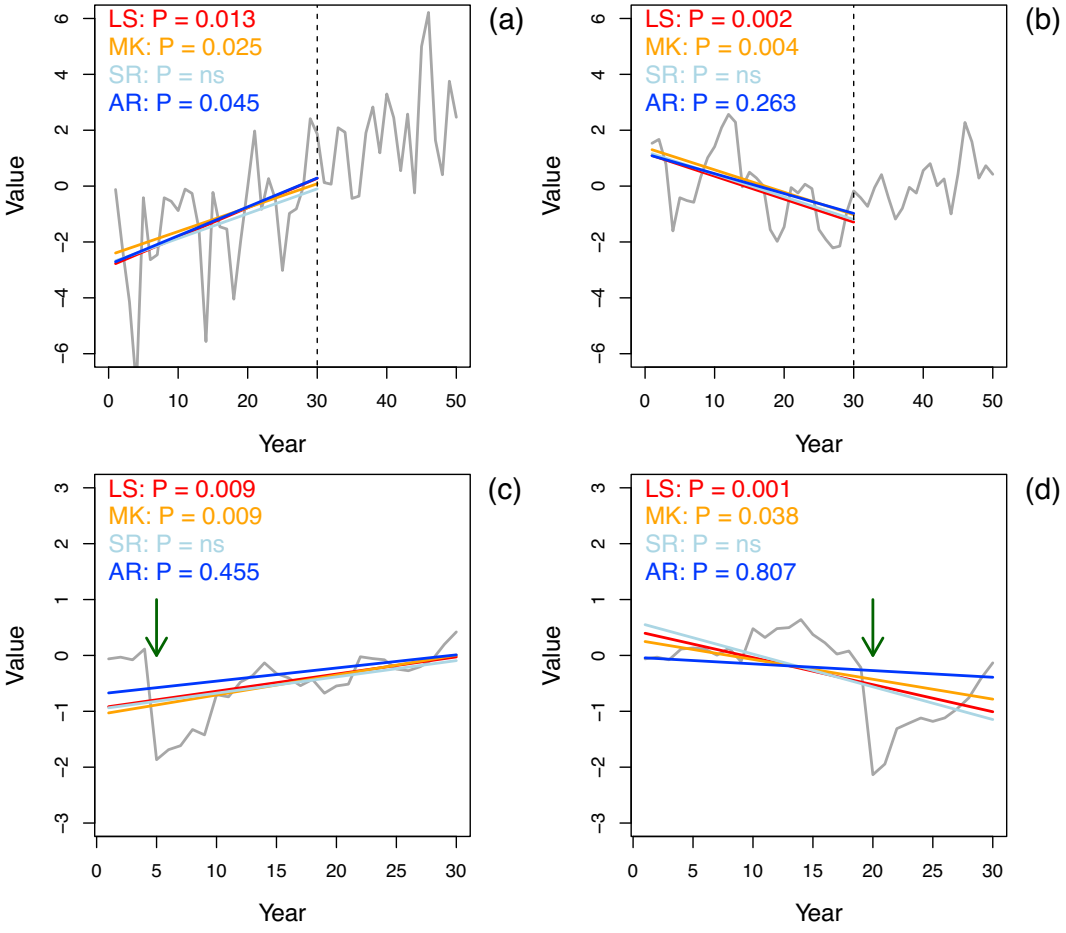

Fig. 2. Four example time series fit with least-squares regression (LS), a Mann-Kendall/Theil-analysis (MK), the size-robust trend test (SR), and regression with autocorrelated errors (AR). (a) When there was a time trend ($c_i = 2$) and no temporal autocorrelation ($\beta_i = 0$), all methods fit to the first 30 time points except SR correctly predicted the remaining 20 time points of the time series. (b) When there was no time trend ($c_i = 0$) and temporal autocorrelation ($\beta_i = 0.9$), LS and MK both incorrectly rejected the null hypothesis with high probability (low *P*-values) when fitted to the first 30 points and therefore incorrectly predicted the last 20 points, whereas SR and AR did not reject the null hypothesis at a significance level of alpha = 0.05. We also considered the case when there is no time trend and strong temporal autocorrelation ($c_i = 0$, $\beta_i = 0.9$), and an event at time 5 (c) and time 20 (d) decreased the value by 2. LS and MK identified strong positive (c) and negative (d) trends, whereas RS and AR identified no trends. Note that SR only reports whether *P*-values are below a threshold of 0.05. Parameters for the simulations were selected to give examples that visually illustrate the overall simulation results in Table 1.

Table 2
From 5000 simulated time series with 30 time points, the mean (standard deviation) of the time trend estimate, $\hat{c}_i$, for different simulated values of trend c_i and temporal autocorrelation β_i.

c_i	β_i	LS	MK	SR	AR
0	0.0	0.00 (0.63)	0.00 (0.66)	0.00 (0.72)	0.00 (0.63)
0	0.2	0.01 (0.77)	0.02 (0.79)	0.02 (0.89)	0.01 (0.77)
0	0.4	0.01 (0.99)	0.01 (1.01)	0.01 (1.14)	0.01 (1.00)
0	0.6	0.03 (1.4)	0.03 (1.42)	0.04 (1.64)	0.02 (1.41)
0	0.8	0.05 (2.45)	0.04 (2.49)	0.06 (2.83)	0.05 (2.32)
0.5	0.0	0.50 (0.63)	0.50 (0.65)	0.52 (0.72)	0.50 (0.63)
0.5	0.2	0.49 (0.78)	0.49 (0.8)	0.51 (0.89)	0.50 (0.78)
0.5	0.4	0.48 (1.02)	0.48 (1.04)	0.50 (1.18)	0.48 (1.03)
0.5	0.6	0.48 (1.41)	0.48 (1.44)	0.51 (1.63)	0.47 (1.42)
0.5	0.8	0.49 (2.43)	0.50 (2.47)	0.50 (2.81)	0.49 (2.34)
1.0	0.0	1.00 (0.63)	1.00 (0.66)	1.04 (0.72)	1.00 (0.64)
1.0	0.2	0.98 (0.78)	0.99 (0.8)	1.01 (0.89)	0.99 (0.79)
1.0	0.4	0.99 (1.00)	0.99 (1.02)	1.03 (1.15)	0.99 (1.00)
1.0	0.6	1.03 (1.47)	1.02 (1.49)	1.07 (1.7)	1.03 (1.46)
1.0	0.8	0.95 (2.37)	0.96 (2.41)	0.97 (2.77)	0.96 (2.25)
2.0	0.0	2.02 (0.63)	2.01 (0.66)	2.09 (0.72)	2.02 (0.64)
2.0	0.2	1.99 (0.77)	1.99 (0.79)	2.06 (0.88)	1.99 (0.78)
2.0	0.4	1.99 (1.00)	2.00 (1.02)	2.07 (1.16)	1.99 (1.01)
2.0	0.6	2.00 (1.42)	1.99 (1.44)	2.07 (1.66)	1.99 (1.42)
2.0	0.8	1.93 (2.43)	1.93 (2.47)	1.98 (2.84)	1.96 (2.28)

We applied least-squares regression (LS), a Mann-Kendall/Theil-Sen test (MK), the size-robust trend test (SR), and regression with autocorrelated errors (AR).

Table 3
Proportion of 5000 simulated time series with 30 points for which a trend was identified as statistically significant, depending on differing values of temporal autocorrelation β_i and the year in which a shock event occurred that reduced the value of the variable by 2.

Year of event	β_i	LS	MK	SR	AR
5	0	0.06 (88%)	0.05 (82%)	0.05 (67%)	0.07 (85%)
5	0.2	0.12 (87%)	0.11 (85%)	0.04 (68%)	0.08 (90%)
5	0.4	0.21 (82%)	0.19 (80%)	0.04 (68%)	0.08 (83%)
5	0.6	0.32 (80%)	0.30 (80%)	0.03 (69%)	0.09 (80%)
5	0.8	0.50 (72%)	0.46 (71%)	0.03 (65%)	0.12 (68%)
20	0	0.04 (32%)	0.05 (36%)	0.04 (35%)	0.05 (31%)
20	0.2	0.09 (31%)	0.09 (34%)	0.03 (35%)	0.06 (33%)
20	0.4	0.17 (29%)	0.16 (31%)	0.03 (35%)	0.06 (31%)
20	0.6	0.31 (31%)	0.29 (32%)	0.03 (34%)	0.08 (28%)
20	0.8	0.52 (27%)	0.49 (28%)	0.03 (32%)	0.13 (27%)

The data were simulated without a time trend, and therefore a correct statistical test will reject the null hypothesis in 5% of the simulated datasets at the significance level of alpha = 0.05. We applied least-squares regression (LS), a Mann-Kendall/Theil-Sen test (MK), the size-robust trend test (SR), and regression with autocorrelated errors (AR). Values in parentheses are the percentage of simulated datasets with $P < 0.05$ for which the estimate of c_i was positive.

because it is common practice to use *P*-values to visualize "strong" versus "weak" time trends on maps (as we have done in Fig. 1a), and because the computational burden of AR is not much greater than LS, we recommend AR.

3.1.2. *Spatiotemporal patterns*

Spatial autocorrelation generates visual patterns even when the patterns are random; in all three simulations in Fig. 3, there are no time trends. When there is no spatial autocorrelation ($r = 0$, Fig. 3a), the 10% of pixels with *P*-values <0.10 pepper the map. In contrast, when there is spatial autocorrelation ($r = 0.1$, Fig. 3b,c), the 10% of pixels with *P*-values <0.10 form clusters. The locations of these clusters are random, however, since different random simulations show different clustering patterns (Fig. 3b vs. Fig. 3c). The clusters are simply non-independent type I errors. The degree of spatial autocorrelation in this simulation ($r = 0.1$) is not unrealistic, being roughly the same as estimated for the Alaska dataset we analyze in detail (Section 3.2.1 *Alaska*).

3.1.3. *Simulation study of PARTS*

We used PARTS to test whether simulated land-cover classes had different time trends (Fig. 4). In the first simulation (Fig. 4a,b), there were no time trends in any land-cover class, while in the second simulation (Fig. 4c,d) the time trends increased from land-cover class 1 to 4. The same seed for the random number generator was used for both simulations so that the spatial patterns are visually similar (Fig. 4a,c), making it easier to see the differences in time trends between simulations.

A naive approach to analyzing these data is to perform an ANOVA on the estimated time trends $\hat{c}_i$ to detect differences among land-cover classes. ANOVA falsely rejected the null hypothesis of no differences among land-cover classes in the first simulation ($F_{3,9996} = 159.9$, $P < 10^{-15}$; Fig. 4b). ANOVA also gave the incorrect order of land-cover classes according to their time trends in the second simulation (Fig. 4d). Thus, the "standard" approach of analyzing pixels as if they were independent resulted in strong statistical support for patterns that do not exist in the simulated data (Fig. 4a) and missed the patterns that do exist (Fig. 4c).

We also fit the simulated data with PARTS (steps 1–3). For the first simulation (Fig. 4a, b), PARTS correctly reported neither an overall time trend ($t_{9999} = -0.12$, $P = 0.90$) nor differences among land-cover classes ($F_{3,9996} = 0.96$, $P = 0.41$). The appropriate standard errors for coefficients estimated from data with spatial autocorrelation depend on the hypothesis being tested. The appropriate standard error for the hypothesis that a coefficient is different from zero is the commonly used standard error (Fig. 4b, black error bars). However, when testing the hypothesis that coefficients are different from each other (i.e., land-cover classes have different trends), the appropriate standard errors

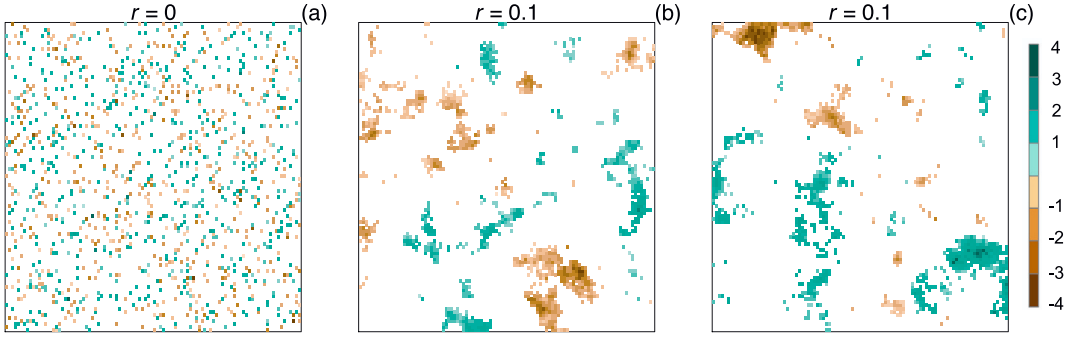

Fig. 3. Random time series on a 100 × 100 pixel map fit using AR (a) when there was no spatial autocorrelation and (b, c) when there was spatial autocorrelation. (b) and (c) only differ in the random numbers drawn in the simulation. In all cases, AR identified 10% of the pixels as having statistically significant trends at the significance level of alpha = 0.1; these are shown as brown (negative trends) and green (positive trends) pixels. In (b) and (c), the range of the spatial autocorrelation of the error variation was 0.1 (roughly 14 pixels). For each pixel, there was mild temporal autocorrelation ($\beta_i = 0.2$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

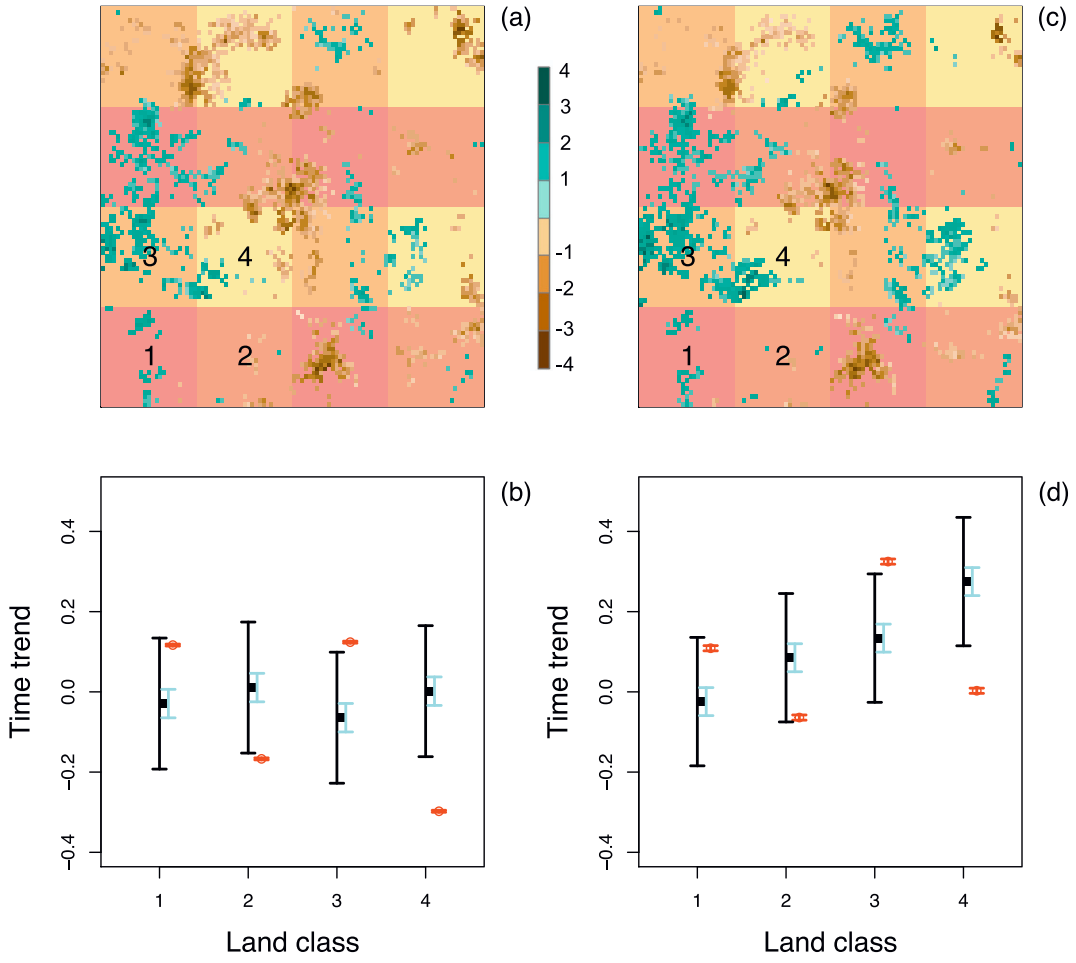

Fig. 4. AR estimates of time trends, $\hat{c}_i$, fit to a simulation of time series of length 30 when there is temporal autocorrelation ($\beta_i = 0.2$) and spatial autocorrelation ($r = 0.1$). Land-cover classes are shown as shaded blocks, with land-cover class 4 being the lightest shade (labelled in the lower-left corner of each map). Only those values of $\hat{c}_i$ that were significant at the alpha = 0.1 level are shown (brown to green scale). In (a) and (b) there were no time trends ($c_i = 0$), and in (c) and (d) the time trends in the simulation model were $c_i = 0$, 0.1, 0.2, and 0.3 for land-cover class 1–4. In (b) and (d), the estimates and standard errors of slopes for each land-cover class given by GLS are shown in black, the GLS conditional standard errors are shown in blue, and the mean and standard error of $\hat{c}_i$ (ignoring spatial autocorrelation) given by an ANOVA are shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

depend on the values of the other coefficients and in this case are narrower (Fig. 4b, blue error bars). The standard errors are consistent with the statistical conclusions of no overall time trends (the coefficients do not differ from zero; black error bars) and no differences among land-cover classes (the coefficients do not differ from each other; blue error bars). Finally, the estimates of the range $\hat{r} = 0.103$ and non-spatial variance *nugget* = 0.060 were close to the values used to simulate the data (0.10 and 0.04, respectively).

For the second simulation (Fig. 4c, d), PARTS correctly rejected the null hypothesis that c_i was the same in all land-cover classes ($F_{3,9996} = 12.68$, $P < 10^{-6}$). In contrast, the hypothesis that there was no overall time trend was not rejected ($t_{9999} = 0.74$, $P = 0.46$). It might seem paradoxical that for this simulation, it is statistically easier to detect differences in time trends among land-cover classes than to determine whether the time trend on average over all pixels (regardless of land-cover class) differs from zero. This seeming paradox is caused by information that is available from pixels that are in close proximity. If nearby pixels in different land-cover classes show differences in their trends, then the statistical test will pick up these differences. In contrast, for the null hypothesis that on average for all pixels there is no trend, the statistical analysis does not have this type of contrasting information between nearby pixels.

Fig. 4 gives only a single simulation example, but Ives et al. (2021) include a detailed study of similar simulations. The simulation study shows the PARTS method reports *P*-values that do not give inflated type I errors and has good statistical power to reject the null hypothesis when it is false.

3.2. Global trends in greening and browning

3.2.1. Alaska

Alaska shows a visual pattern of positive trends in NDVI (c_{rel}) being more common at higher latitudes (Fig. 5a,b), consistent with "Arctic greening" (Jia et al., 2003; Ju and Masek, 2016). Compared to LS, AR identified fewer pixels with time trends ($P < 0.05$), which is consistent with the simulation findings that AR gave less inflated type I error rates; the differences in maps produced by AR and LS are depicted in fig. S1. On average, there was moderate temporal autocorrelation ($\beta_i = 0.40 \pm 0.23$ (SD); Fig. 5c). Furthermore, areas of grassland and shrubland appear to have more positive c_{rel} than areas of savanna (Fig. 5a,d). We tested four statistical null hypotheses: (i) the mean time trend in NDVI is zero when the mean is taken across the entire map of Alaska; (ii) the mean time trends in NDVI for each of the land-cover classes are equal, implying that land-cover classes do not differ; (iii) in a regression of time trends (regardless of land-cover class) on latitude, the coefficient for latitude is zero; and (iv) in a regression of time trends on latitude and land-cover class, the interaction between latitude and land-cover class is zero, implying that the effects of latitude are the same among land-cover

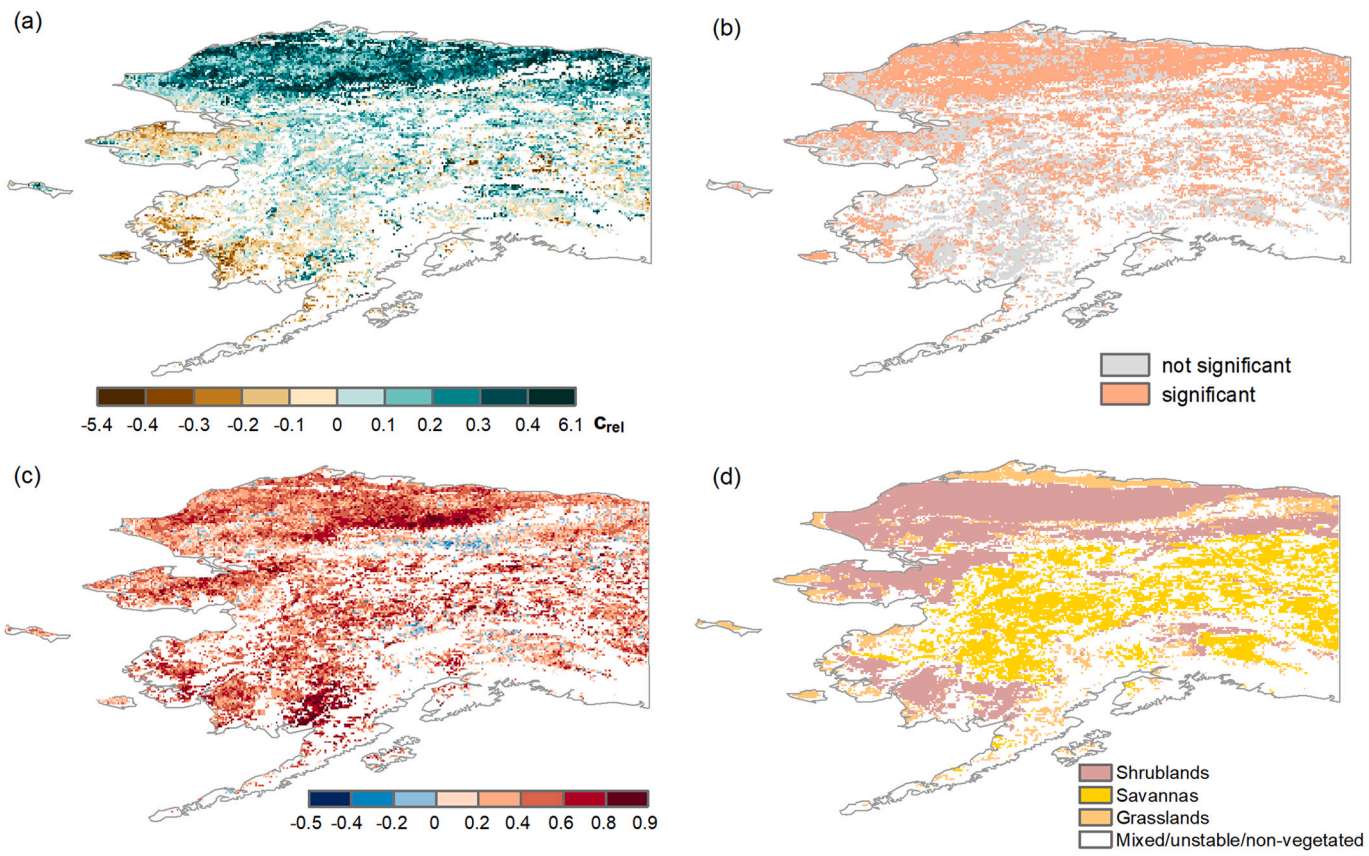

Fig. 5. For Alaska west of −141° longitude, patterns of time trends in NDVI, 1982–2015. (a) Time trends in NDVI measured by c_{rel}, the estimate of c_i from regression with autocorrelated errors (AR; Eq. (2)) divided by the mean NDVI in each pixel. White corresponds to pixels either that have NDVI values too low to analyze or that contain mixed or unstable land-cover classes. (b) Pixels for which the null hypothesis of $c_{rel} = 0$ was rejected by AR ($P < 0.05$). (c) The strength of temporal autocorrelation (parameter β_i; Eq. (2)), and (d) land-cover classes for pixels containing at least 50% of a single class. There are a total of 20,694 pixels with estimates of c_{rel}: grassland (2924), savanna (7361) and shrubland (10,409). Deciduous forest (1 pixel), evergreen needle forest (83 pixels) and mixed forest (36 pixels) were removed, as were 33 pixels that were outliers.

classes. These hypotheses are stated as if they were regression analyses or analyses of covariance (ANCOVA), highlighting that PARTS is in essence a method for performing regression analyses at the scale of an entire map.

We applied PARTS both without (steps 1–3) and with data partitions (steps 1–4), using 5, 10, 15, and 20 partitions (Table 4). The overall conclusion was that hypotheses (i) and (iv) were not rejected, implying that there was no overall trend in NDVI, and there was no land-cover class by latitude interaction. However, null hypotheses (ii) and (iii) were rejected, implying differences in trends among land-cover classes and with latitude. These conclusions are based on the analyses that partition the dataset and then combine the statistical results (P_{part}). Applying a single GLS analysis to all of the data (P_{GLS} for $n_p = 1$) failed to reject any null hypothesis. This result is the opposite from what we found in our simulations in which GLS applied to the entire map gave tests that were at least as powerful as those obtained with partitions (see Table 3 and Fig. 3 in Ives et al., 2021). This contrast is the result of adjacent pixels being very highly autocorrelated, as shown by the much smaller nugget effect estimated from GLS (Table 4). These high correlations between adjacent pixels are only detectable when analyzing all pixels, because partitioning the data will remove a large fraction of the pairwise correlations between pixels that are adjacent to each other. To verify this explanation, we removed every other row and column from the data, which caused the nugget effect to increase to the same level as found in the case of 10 partitions (results not shown); this confirmed that high correlations between adjacent pixels were the cause of the small nuggets. Because a small nugget will result in a higher estimate of the spatial autocorrelation, it will reduce statistical power. High spatial autocorrelation between adjacent pixels could be an artifact of image processing, such as those caused by atmospheric scattering and consequent adjacency effects (Semenov et al., 2011), and sensor effects such as linear-shift-invariant blurring, sampling effects, and shift-invariant, signal-independent additive white noise in the AVHRR data (Reichenbach et al., 1995). Therefore, because partitioning removes these effects, we consider the partition results more reliable than those from a single GLS. Because adjacency issues are likely to be common in many remote-sensing datasets, we recommend partitioning even in small datasets that do not require partitioning for computational reasons.

The PARTS approach of combining statistical results among partitions (P_{part}) had better performance than multiple-comparison corrections (P_{hoch} and P_{fdr}). We repeated the model for land-cover classes (hypothesis ii) five times (Table 4(ii)). Because the partitions were selected randomly, the results differed somewhat among repetitions. Nonetheless, the range of values of P_{part} were generally less than P_{hoch} and P_{fdr}, thereby giving more repeatable results. Also, as found in simulations (Ives et al., 2021), the statistical power given by P_{part} was generally higher than P_{hoch} and P_{fdr}.

A closer look at the analyses gives more details about the character of autocorrelation underlying the statistical results. Focusing on the model for land-cover classes (hypothesis ii) with 10 partitions, the average standard error of the estimates of c_{rel} was 0.071, whereas the standard deviation of the unexplained variation in the partitioned GLS model was estimated as $\widehat{\sigma_\gamma} = 0.187$. This result implies that there were fixed (non-temporally varying) differences in trends among pixels, because the spatial variation in c_{rel} (0.187) exceeded that caused solely by pixel-level uncertainty in c_{rel} (0.071). In other words, the variation among values of

Table 4
For Alaska west of −141° longitude, statistical tests of differences among time trends, c_{rel}, in mean annual NDVI, 1982–2015.

	n_p	*nugget*	P_{GLS}/P_{part}	P_{hoch}	P_{fdr}
(i) intercept	1	0.050	1.000	–	–
	5	0.162	0.571	0.731	0.731
	10	0.237	0.450	0.729	0.729
	15	0.260	0.427	0.744	0.698
	20	0.274	0.400	0.748	0.621
(ii) land-cover class	1	0.050	0.600	–	–
	5	0.181	0.202	0.310	0.296
			(0.075-0.237)	(0.134–0.611)	(0.134–0.418)
	10	0.212	0.009	0.011	0.011
			(0.004–0.020)	(0.006–0.030)	(0.006–0.030)
	15	0.247	0.012	0.034	0.024
			(0.004–0.017)	(0.002–0.100)	(0.002–0.062)
	20	0.265	0.006	0.053	0.048
			(0.005–0.027)	(0.0003–0.168)	(0.0003–0.138)
(iii) latitude	1	0.050	0.052	–	–
	5	0.164	0.006	0.011	0.011
	10	0.209	0.002	0.002	0.001
	15	0.260	0.0003	0.001	0.0004
	20	0.247	0.0001	0.0002	0.0001
(iv) land × latitude	1	0.050	0.610	–	–
	5	0.183	0.230	0.284	0.284
	10	0.252	0.152	0.243	0.242
	15	0.298	0.132	0.414	0.324
	20	0.324	0.094	0.032	0.032

Four null hypotheses were tested: (i) the mean slope does not differ from zero (intercept $b_0 = 0$ in Eq. (3)), (ii) there is no difference among land-cover classes, (iii) there is no difference with latitude, and (iv) there is no land-cover class by latitude interaction. For the last test, the null hypothesis was the model including land-cover class and latitude, but no interaction between them. Hypotheses were tested for the entire dataset ($n_p = 1$) and for partitions into $n_p = 5$, 10, 15, and 20 subsets. Tests were performed with an *F*-test ($n_p = 1$, P_{GLS}) or LRT ($n_p > 1$, P_{part}), and by selecting the partition with the lowest *P*-value and correcting for multiple comparisons (P_{hoch} and P_{fdr}). For land-cover class (ii), the partition analyses were repeated five times, and the *P*-values are median (range).

c_{rel} across the map was due, in part, to purely spatial differences among time trends that cannot be attributed to spatiotemporal variation. Because the spatial autocorrelation matrix was estimated from the correlations among temporal residuals across pixels (PARTS step 2) and hence did not use information about possible spatial variation in the fixed time trends among pixels, we performed a study by varying the spatial range parameter *r* in PARTS step 3 while keeping other parameters of the model fixed (Table 5). The maximum likelihood estimate of *r* from PARTS step 3 was $\hat{r} = 142$ km, slightly less than the originally estimated value of $\hat{r} = 185$ km from PARTS step 2. Nonetheless, the difference in log likelihood was small (1.26), and the *P*-value for the test of differences among land-cover classes only changed from 0.009 to 0.0108. Thus, the two methods for estimating *r* (from PARTS step 2 and step 3) give the same conclusions. The R package remotePARTS can estimate *r* and *g* using either method. When computationally feasible, we recommend using both methods as a check on the conclusions and as a way to extract information about the character of autocorrelation (spatiotemporal versus purely spatial) in a dataset.

Table 5
For Alaska west of −141° longitude, the effect of the value of the range parameter, *r*, on the GLS model likelihood and statistical test of differences among land-cover classes.

r (km)	ΔlogLik	*P*	Grassland	Savanna	Shrubland
123	−0.67	0.0116	0.053	0.024	0.044
132	−0.21	0.0112	0.053	0.024	0.044
142	0.00	0.0108	0.053	0.024	0.044
154	−0.07	0.0103	0.053	0.024	0.043
168	−0.47	0.0097	0.052	0.023	0.043
185*	−1.26	0.0090	0.052	0.023	0.043
205	−2.53	0.0083	0.052	0.023	0.043
231	−4.42	0.0074	0.052	0.022	0.042
264	−7.17	0.0065	0.052	0.022	0.042
308	−11.16	0.0055	0.052	0.022	0.042
370	−17.07	0.0045	0.052	0.022	0.042

The PARTS analysis was performed with 10 partitions using the same partitions that gave the median *P*-value for the test of differences among land-cover classes (Table 4(ii)). The *nugget* effect and parameter *g* for the spatial autocorrelation matrix were fixed while the range *r* was varied. For each value of *r*, the model was refitted. ΔlogLik gives the difference in log likelihoods between each model and the best model ($r = 142$ km). The value of $r = 185$ km (marked with *) was estimated from the correlations among residuals from the time-series analysis (PARTS step 2). Coefficients for the time trends c_{rel} are given for the three land-cover classes.

We also compared PARTS (with $n_p = 10$) to a "standard" method: we estimated time trends for each pixel using LS and then performed ANOVA on the resulting estimates of the time trends (Fig. 6). Roughly 70% of the pixels were identified by LS as being significant compared to roughly 55% for AR (see also Fig. S1). Furthermore, the ANOVA showed higher estimates of trends for grassland and shrubland with very narrow standard error bars, giving a highly significant overall positive trend combining land-cover classes ($t_{20674} = 77.76$, $P < 10^{-15}$) and differences among land-cover classes ($F_{3,\ 20{,}672} = 2716$, $P < 10^{-15}$). The very small standard errors of the trend estimates for each land-cover class (Fig. 6a, red bars) obscures the true variability in c_{rel} within land-cover classes. We also tested the null hypothesis that there was no effect of latitude (Fig. 6c), which was similarly highly significant in a standard regression analysis ($t_{20674} = 106.6$, $P < 10^{-15}$). These "standard" results that do not account for spatial autocorrelation give falsely low *P*-values, and results of this type of analysis should not be trusted.

The extent of spatial autocorrelation identified by PARTS is given by the parameters fit to the exponential-power function, $\exp(-(d_{ij}/r)^g)$, and the nugget. The estimates of $\hat{r} = 185$ km and $\hat{g} = 0.571$, and nugget = 0.212 from the land-cover class model with $n_p = 10$ (Table 4) imply that the spatial autocorrelation between pixels drops to 0.10 only above distances of 600 km. The maximum extent of the Alaska map is 2000 km, implying that if we limited the analyses to pixels with autocorrelation less than 0.10, we could analyze at most 16 pixels.

3.2.2. *Continents*

For the global data, we analyzed the full datasets for the six continents separately; pixels were 0.083 degree (~8 km), resulting in 91,459 (Australia) to 570,250 (Asia) pixels per continent. We used PARTS with partitions of 2000 pixels, leading to varying numbers of partitions (n_p)

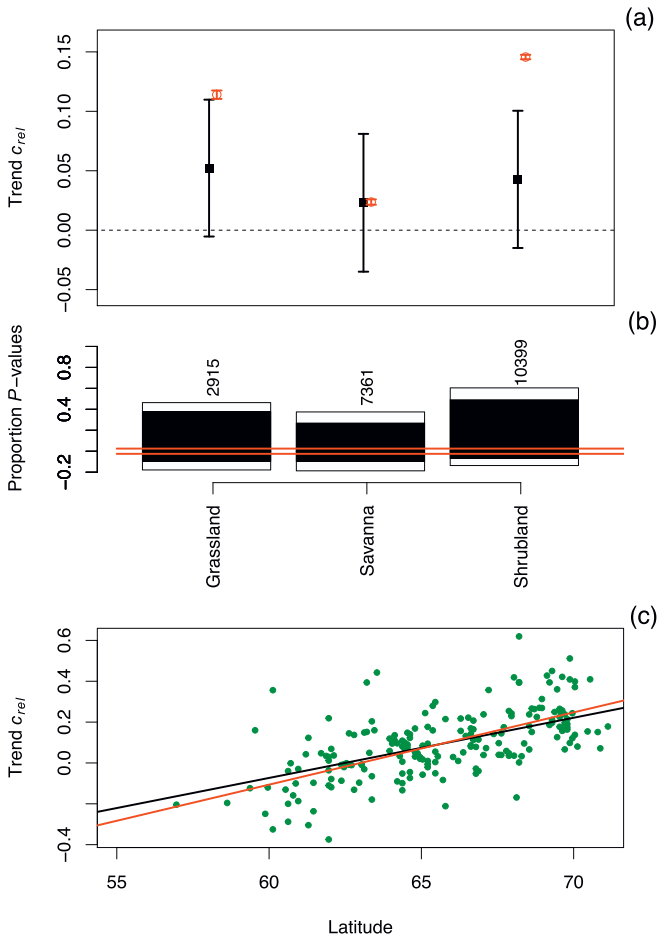

Fig. 6. Comparison of statistical analyses of time trends, c_{rel}, in NDVI for Alaska west of −141° longitude (Fig. 5). (a) Effects of land-cover class on c_{rel} given by PARTS analysis (black) dividing the map into 10 partitions each containing 2094 pixels. Bars give the unconditional standard errors of the estimates. An ANOVA of the LS estimates of c_{rel} gives greater estimates of the overall time trends in grassland and shrubland, with very small standard errors (red points and bars). (b) Proportions of pixels containing significant time trends ($P < 0.05$, either positive or negative) in each land-cover class, where P-values were calculated from either LS (white bars) or AR (black bars). Numbers of analyzed pixels in each land-cover class are given above the bars, and the red horizontal lines at −0.025 and 0.025 give the thresholds beyond which pixels should be identified as significant under the null hypothesis of no time trend. (c) Regression of the AR estimates of c_{rel} against latitude, showing only 200 points for clarity. The red line gives the relationship for standard regression, which is highly significant ($t_{20692} = 52.41$, $P < 10^{-16}$), and the black line is from PARTS with 10 partitions ($P = 0.0013$; Table 4). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

per continent. With this partition size, analyzing a model for the largest map (Asia) with an 8-core, 2.79 GHz processor (CPU: AMD Ryzen 73,700× with an AMD X570 chipset; memory: dual channel DDR4 at 1800 MHz) took less than 15 min (PARTS steps 1–4). Fig. 1a shows the pixel-level AR estimates of c_{rel} for all continents. Although we used the AR method, the patterns of browning and greening are similar to those produced by LS, but LS identified many more pixels than AR as "significant" (Fig. S2). We tested the same four null hypotheses we illustrated for Alaska, repeating PARTS three times for each analysis to assess variability in results (Table 6). As for Alaska (Table 5), we also fit the range parameter r using the GLS (PARTS step 3), in addition to the correlations among time-series residuals (PARTS step 2), to confirm that this did not affect the conclusions. For the case of differences among land-cover classes (hypothesis ii), the largest proportional change in r was for Australia (from 540 km in PARTS step 2 to 100 km in PARTS step 3), yet the P-value for the hypothesis test changed little ($P = 0.0084$ to $P = 0.0045$). There were similarly negligible changes in the P-values for the other continents, so these are not reported.

In the test for whether there is a trend in NDVI averaging over all pixels, PARTS failed to reject the null hypothesis in every continent except Asia, where there was weak support for an overall trend ($P_{part} = 0.01$–0.02; Table 6). This failure to identify a significant overall trend in five of six continents occurred even though the estimates of the trends were positive in most land-cover classes for every continent (Fig. 7). The differences in trends among land-cover classes were significant for all continents (black points in Fig. 7, Table 6). Therefore, different types of land cover are greening, or in a few cases browning, at different rates. However, there were only five and two land-cover classes in Asia and Europe, respectively, that have trends statistically greater than zero (solid points rather than open points in Fig. 7). The patterns among land-cover classes were distinctly different among continents. For example, in Europe the land-cover class showing the greatest positive trend in NDVI was evergreen needle forest, whereas evergreen needle forest had among the lowest trends in North America. These patterns are found for pixels that have stable land-cover in which a single land-cover type makes up >50% of the pixel. Therefore, these patterns reflect changes in mean annual NDVI that is not attributable to large changes in land-cover classes over time.

To compare with PARTS, we also performed a "standard" analysis by computing the pixel-level time trends using LS and then performing an ANOVA to test for differences in LS slopes among land-cover classes (Fig. 7, red). In the ANOVA, the differences among land-cover classes were highly significant for all continents ($P < 10^{-16}$). Furthermore, the estimates of land-cover classes were sometimes substantially different from the PARTS estimates. For example, the ANOVA estimate of the trend in cropland in Asia was twice the value of the PARTS estimate, while the ANOVA estimate for deciduous broadleaf forest was much lower. Over all continents, the PARTS estimates were less variable among land-cover classes than the ANOVA estimates.

PARTS also found no significant effect of latitude on greening in any continent (Table 6). However, for all continents except Australia there was a strong latitude × land-cover class interaction (based on P_{part}). Thus, while there were significant effects of latitude at the continental scale, these occurred only within land-cover classes.

4. Discussion

Remote-sensing data contain tremendous amounts of information, making it possible to answer questions about changes in the world with remarkable spatial detail. The amount of information, however, creates two challenges: how to pose hypotheses at the right scale of interest, and how to test these hypotheses with appropriate statistics that are computationally feasible. PARTS meets these challenges and makes it possible to fit regression-style models to test hypotheses about how time trends depend on spatially varying independent variables such as land-cover class and latitude. Patterns through time and in space may be caused by factors that are not part of the hypothesis being tested, and that is why it is necessary to account for "unexplained" temporal and spatial autocorrelation – patterns of variation in errors that are not explained by independent variables in the model. PARTS accounts for both temporal and spatial autocorrelation, making it able to reveal patterns underlying the data and to guard against falsely identifying patterns that are not in the data.

PARTS breaks down the problem of analyzing large spatiotemporal datasets into two stages: fitting the time series within each pixel to estimate pixel-level time trends, and then analyzing the pattern of time trends among pixels. The first stage is standard in the remote-sensing literature; therefore, the scientific advance given by PARTS is the ability to analyze time trends at the scale of entire maps. Nonetheless,

Table 6
Statistical tests of differences among relative time trends, c_{rel}, in mean annual NDVIs, 1982–2015, for six continents.

Model	Continent	n	n_p	*nugget*	P_{part}	P_{hoch}	P_{fdr}
(i) intercept	Africa	214,121	106	0.03	0.16, 0.17, 0.18	0.33, 0.36, 0.37	0.27, 0.29, 0.29
	Asia	570,250	283	0.35	0.01, 0.01, 0.02	0.06, 0.07, 0.09	0.03, 0.03, 0.04
	Australia	91,459	45	0.18	0.24, 0.24, 0.26	0.36, 0.36, 0.37	0.34, 0.34, 0.36
	Europe	174,498	86	0.23	0.06, 0.07, 0.07	0.07, 0.39, 0.46	0.07, 0.14, 0.15
	North America	345,264	172	0.34	0.44, 0.44, 0.49	0.91, 0.94, 0.97	0.77, 0.77, 0.79
	South America	190,832	94	0.11	0.39, 0.40, 0.41	0.75, 0.96, 1.00	0.62, 0.63, 0.64
(ii) land-cover class	Africa	182,587	90	0.03	0.0010, 0.0014, 0.0014	0.0001, 0.0471, 0.0499	0.0001, 0.0246, 0.0499
	Asia	441,312	219	0.35	0.0002, 0.0002, 0.0002	0.0028, 0.0096, 0.0197	0.0020, 0.0096, 0.0098
	Australia	80,958	40	0.18	0.0066, 0.0082, 0.0096	0.0214, 0.0477, 0.1220	0.0214, 0.0469, 0.0625
	Europe	110,385	54	0.23	<0.0001, <0.0001, <0.0001	<0.0001, <0.0001, 0.0004	<0.0001, <0.0001, 0.0004
	North America	257,344	128	0.36	<0.0001, <0.0001, <0.0001	0.0006, 0.0034, 0.0062	0.0003, 0.0021, 0.0045
	South America	152,549	75	0.09	<0.0001, <0.0001, <0.0001	<0.0001, <0.0001, <0.0001	<0.0001, <0.0001, <0.0001
(iii) latitude	Africa	214,121	106	0.03	0.42, 0.44, 0.44	0.90, 0.93, 0.98	0.74, 0.78, 0.79
	Asia	570,250	283	0.35	1.00, 1.00, 1.00	1.00, 1.00, 1.00	1.00, 1.00, 1.00
	Australia	91,459	45	0.18	0.60, 0.61, 0.62	0.83, 0.89, 0.91	0.81, 0.83, 0.84
	Europe	174,498	86	0.23	0.08, 0.08, 0.09	0.02, 0.03, 0.06	0.02, 0.03, 0.05
	North America	345,264	172	0.34	0.73, 0.75, 0.76	1.00, 1.00, 1.00	1.00, 1.00, 1.00
	South America	190,832	94	0.11	0.15, 0.15, 0.16	0.58, 0.71, 0.89	0.30, 0.31, 0.31
(iv) land × latitude	Africa	182,587	90	0.03	0.0047, 0.0074, 0.0112	0.0002, 0.0012, 0.0472	0.0002, 0.0007, 0.0472
	Asia	441,312	219	0.36	0.0003, 0.0004, 0.0004	<0.0001, 0.0004, 0.0010	<0.0001, 0.0004, 0.0010
	Australia	80,958	40	0.18	0.0163, 0.0555, 0.0785	0.0227, 0.1526, 0.1562	0.0227, 0.1526, 0.1530
	Europe	110,385	54	0.25	0.0003, 0.0006, 0.0007	0.0001, 0.0058, 0.0461	0.0001, 0.0033, 0.0277
	North America	257,344	128	0.37	0.0006, 0.0007, 0.0007	<0.0001, 0.0066, 0.0220	<0.0001, 0.0046, 0.0220
	South America	152,549	75	0.10	<0.0001, <0.0001, <0.0001	<0.0001, <0.0001, <0.0001	<0.0001, <0.0001, <0.0001

Four null hypotheses were tested: (i) there is no overall trend (intercept $b_0 = 0$ in Eq. (3)), (ii) there is no difference among land-cover classes, (iii) there is no difference with latitude, and (iv) there is no land-cover class by latitude interaction. The number of pixels (n) differed according to the model, because models that included land-cover classes only used pixels with >50% of a single class. Partitions of 2000 pixels were used, giving different numbers of partitions among continents (n_p). The nugget gives the magnitude of non-spatially autocorrelated variation. Average values of the parameters governing the extent of spatial autocorrelation are $\widehat{r} = 285$, 316, 531, 287, 296, and 371 km, and $\widehat{g} = 0.42$, 0.52, 0.61, 0.47, 0.50, and 0.42 for Africa, Asia, Australia, Europe, North America, and South America. *P*-values were calculated using a partition LRT (P_{part}) and using Hochberg and FDR corrections for multiple comparisons (P_{hoch}, and P_{fdr}). Analyses were repeated three times, giving three *P*-values for each test.

getting the best statistical estimates of the time trends within pixels will improve the overall data analyses. Thus, we examined the challenges of both fitting pixel-scale time series and fitting map-scale patterns in the data.

4.1. Temporal patterns

Fitting a statistical model to time series from individual pixels gives both an estimate of the time trend and the statistical significance of the time trend. We investigated four statistical methods: least-squares regression (LS) (e.g., Fensholt and Proud, 2012; Myneni et al., 1997; Piao et al., 2011), the Mann-Kendall test combined with the Theil-Sen slope estimator (MK) (Fensholt et al., 2015; Zhu et al., 2016), the size-robust trend test (SR) (Fomby and Vogelsang, 2002; Vogelsang, 1998), and regression with autoregressive errors (AR) (Box et al., 1994; Ives et al., 2010). Of these, AR gave the most accurate estimates of the time trends in our simulations, that is, the estimates that have a combination of low bias and low standard errors. When there was temporal autocorrelation, LS and MK gave highly inflated type I error rates. While SR gave good type I errors, it had reduced statistical power to detect trends when they truly existed and gave estimates with relatively large standard errors (Table 2).

Despite the common practice of calculating *P*-values for trends in pixel-level time series, these *P*-values do not test hypotheses about the map-scale patterns in time trends. Nonetheless, an informal way to give a visual impression of patterns on a map is to use *P*-values as a filter: for example, only pixels with *P*-values <0.1 could be shown on the map. For this, *P*-values from AR are adequate, even though they can show mildly inflated type I errors when there is strong temporal autocorrelation and "short" time series of length 30. Nonetheless, the justification for using AR in the PARTS analysis is that it gives the most accurate estimates of the trends.

4.2. Spatial patterns

Although the problem of spatial autocorrelation in remote-sensing analyses is recognized (de Beurs et al., 2015; Tomaszewska et al., 2020; Zhou et al., 2001), currently there is no available statistical method that can account for spatial autocorrelation for maps of the size commonly obtained by remote sensing, which routinely exceed 10^5–10^6 pixels. Ignoring spatial autocorrelation incorrectly assumes that pixels are independent and can also fail to leverage information, for example, to identify differences in time trends among land-cover classes contained in nearby pixels. This problem can be seen visually. If there were no spatial autocorrelation, then pixels with high-magnitude trends would pepper a map (Fig. 4a), which is rarely seen in remote-sensing data. In contrast, random patterns in spatially autocorrelated data form clusters, and these random patterns can look deceptively "significant" (Fig. 4b,c).

We analyzed NDVI trends both in Alaska and among continents using a standard approach that ignores temporal and spatial autocorrelation: we estimated trends with LS and then performed comparisons among explanatory variables using ANOVA. With this approach, almost all time trend estimates in any statistical model were statistically significant (Figs. 6, 7). For example, the mean time trend averaged across all land-cover classes in all continents differed from zero, with the majority being positive (Fig. 7). In contrast, a PARTS analysis showed that only Asia and Europe contained land-cover classes that have significant time trends. Furthermore, the PARTS analysis showed less variability among land-cover classes than the ANOVA. The dataset that we analyzed (NDVI3g) has been used numerous times to analyze global greening and browning trends (e.g., Bi et al., 2013; Fensholt et al., 2015; Fensholt and Proud, 2012; Myneni et al., 1997; Piao et al., 2011; Xu et al., 2013; Zhou et al., 2001; Zhu et al., 2016), which is why we selected this dataset for our analyses. In this previous work, more-targeted analyses of specific regions were often made, in contrast to our continent-scale analyses that do not address regional patterns within continents. Furthermore, we only analyzed pixels with stable land-cover classes, thereby excluding

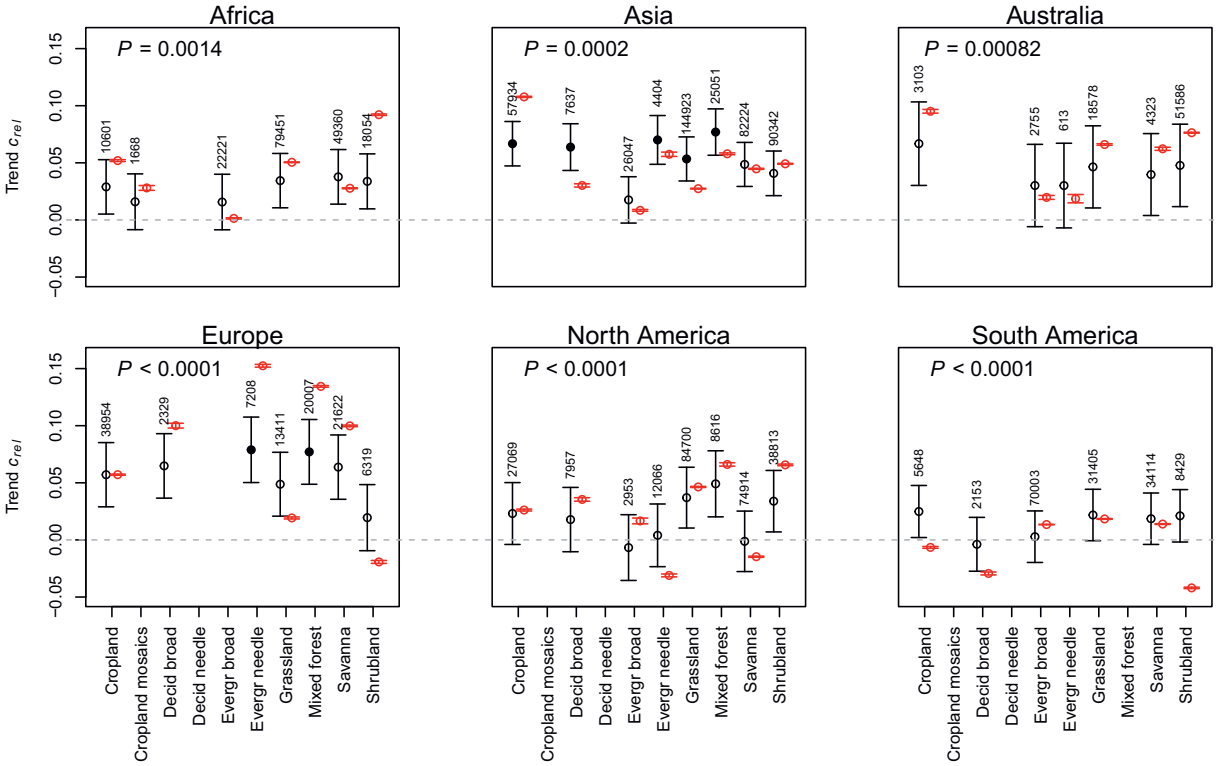

Fig. 7. Differences in time trends, c_{rel}, in mean annual NDVI among land-cover classes for six continents. Land-cover classes that have a statistically significant trend ($P < 0.01$) are shown with filled black dots, and non-significant classes are shown with black circles; bars represent unconditional standard errors. Analyses were performed after partitioning the dataset into subsets of 2000 pixels (Table 5), and for each continent land-cover classes represented at least 0.5% of the pixels. Outliers were removed. *P*-values at the top of each panel correspond to the statistical test that values of c_{rel} differ among land-cover classes. We also analyzed the pixel-level trends using LS regression and then performed an ANOVA to compare land-cover classes. The ANOVA estimates (±1 standard errors) are given in red. Land-cover classes are: Cropland, Cropland-vegetation mosaic (Cropland mosaics), Deciduous broadleaf forest (Decid broad), Deciduous needle leaf forest (Decid needle), Evergreen broadleaf forest (Evergr broad), Evergreen needle leaf forest (Evergr needle), Grassland, Mixed forest, Savanna, and Shrubland. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

trends caused by changing land cover and land use. Therefore, our conclusions are not directly comparable to previous analyses. Nonetheless, our results suggest that previous conclusions drawn about patterns of greening and browning throughout the world should be re-examined.

4.3. *Appropriate statistical models*

The statistical challenges we address all revolve around temporal and spatial autocorrelation, and this warrants discussion of what are temporal and spatial autocorrelation. From a statistical perspective, a model to test a specific hypothesis separates the variance in the response variable into that which can be explained by independent variables (the "mean" or "fixed" component of the model) and the unexplained variation that cannot (the "variable" or "random" component of the model). However, this does not mean that the unexplained variation is white noise. The variation treated as unexplained in the model has a cause, and a cause will often have temporal and spatial structure. Temporal and spatial autocorrelation can be interpreted as caused by environmental variables that are not measured (such as regional weather patterns) or by the ways in which biological systems respond to environmental variation. As an example of the latter, temporal autocorrelation will likely occur for any remote-sensing measure, such as NDVI, that reflects plant growth and succession following disturbances such as droughts (Fig. 2c,d). Spatial autocorrelation could be caused by the same factors that cause temporal autocorrelation, such as succession following a drought (temporal autocorrelation) if the drought occurred across many pixels (spatial autocorrelation). Similarly, areas of the same vegetation type (spatial autocorrelation) could respond to the same disturbance such as fire in the same way, recovering from the fire at similar rates (temporal autocorrelation).

Given the complexities of statistical hypothesis testing with temporal and spatial correlation, is it necessary to do statistical tests at all? This question is all the more germane when considering that Asia, despite greening in all land-cover classes, did not show a statistically significant increase in NDVI when combining all pixels (Figs. 1a, 7). Because we have measurements of NDVI for almost all of the land area of Asia, from a statistical perspective we have a complete census, not a sample, so why are statistics needed? The need for statistical tests can be illustrated with the following comparison. First, suppose foresters want to assess the growth of biomass in a 1-km^2 forest stand over 30 years. They measure the biomass of every tree in the stand every year, giving them the change in biomass without error. Because this is a complete sample, statistics are unnecessary. However, now suppose that their charge is not just to analyze the past, but to predict whether biomass will be higher in 10 years. To answer this question, spatial patterns of forest growth matter. If, for example, 1/3 of all locations exhibited a decrease in biomass, but these locations were scattered and could be attributed to deaths of large trees, then it would be reasonable to conclude that the treefalls are normal and to confidently predict a continued increase in biomass in the next 10 years. However, if instead biomass increased in the northern 2/3 of the stand but decreased in the southern 1/3, more caution is necessary. In the absence of an explanation for this spatial pattern, it is likely that a single stochastic event, maybe a fire 40 years ago in the northern 2/3 of the stand, is responsible for the increase in biomass there. Depending on the stage of succession of the northern 2/3, continued increase in biomass for the next 10 years might not occur. Thus, even though the foresters might have a complete sample of trees for 30 years,

considerations when answering the question about future change are different from considerations when answering the much narrower question of whether there has been a past change in a particular area.

This example of predicting change in a forest stand is analogous to the situation faced when testing hypotheses about global changes. When retrospectively analyzing a dataset, a significance test assesses the likelihood that a pattern observed in the sample that is analyzed will also occur in the entire population of all possible samples. If it is possible to obtain a complete sample (i.e., the entire population), statistics are not needed. However, when analyzing a dataset to make future predictions, a complete sample is impossible, because the appropriate collection of possible samples includes the future populations. We have not addressed the issue of long-term forecasting here, because this would require not only a model that fits past changes, but also justification for assuming that the processes driving past changes continue unabated into the future. Nonetheless, statistically assessing past changes in the context of the entire realm of possibilities is appropriate when using past changes to assess the realm of possible futures.

4.4. PARTS

PARTS can fit statistical models to very large datasets by partitioning them into subsets and then combining the separate statistical tests into one overall test. Our reason for subsampling pixels is only to reduce the computational burden of statistical model fitting. For example, a 100,000-pixel map results in 10^{10} pairwise spatial autocorrelations between pixels. For GLS, this autocorrelation matrix must be mathematically inverted, which is not computationally practical. Subsampling is sometimes used to avoid spatial autocorrelation altogether (Lahiri, 2003). For this, semi-variograms or other methods are used to identify the scale at which points can be assumed to be independent, and regular statistical models that do not account for spatial autocorrelation are then applied to the subsampled points. Not surprisingly, this approach loses a lot of information; for example, our analyses of Alaska (Table 4) would retain only 16 pixels if spatial autocorrelation was required to be less than 0.1. PARTS explicitly accounts for spatial autocorrelation and uses the correlated information in the analyses, which makes statistical tests more powerful than simply thinning data to the point of spatial independence.

Because the partitions used by PARTS are constructed randomly, the resulting *P*-values can differ among repeated analyses (Tables 4, 6). Nonetheless, in our applications this variation was not large enough to change any conclusions. Furthermore, even with this variability in *P*-values, PARTS gives correct type I errors and has good statistical power. PARTS is computationally efficient, making it possible to analyze even large datasets multiple times, in which case the median *P*-values can be used for hypothesis testing. As with any statistical model, diagnostics of model fit should be performed, and the R package remotePARTS contains the tools to perform the appropriate diagnostics.

By formulating a null hypothesis and statistical model to test it, PARTS forces an explicit statement of the problem in question, making it possible to compare different patterns. For example, for Alaska west of –141° longitude, we found a positive effect of latitude on greening (Table 4), but there was no effect of latitude on greening for North America as a whole (Table 6). PARTS could be used to analyze regions separately and then to statistically compare regions. For example, the strategy of using latitudinal bands as employed by Zhou et al. (2001) could be used to determine whether latitudinal trends in greening are only observed in the highest latitudinal bands. Similar analyses could investigate land-cover class by latitude interactions within different latitudinal bands. PARTS is flexible to address many hypotheses.

5. Conclusion

PARTS analyses make it possible to identify underlying patterns and test hypotheses at the scale of entire maps. By accounting for both temporal and spatial autocorrelation, PARTS both provides more power to detect underlying patterns and avoids false conclusions from the data.

The strength of PARTS is its ability to analyze patterns over an entire map, but the requirement for a single model to analyze the entire map can sometimes be a limitation. For example, in the map of greening trends throughout the world (Fig. 1a) there are many regions on the scale of several hundred kilometers in diameter that appear to differ from surrounding areas. PARTS treats these patterns as spatial autocorrelation in the random variation of the statistical model. PARTS can be used to test specific, a priori hypotheses about regional patterns, such as whether regions (e.g., countries, ecoregions, or tiles on a regularly spaced grid over the map) differ in the mean value of a response variable; this can be done by treating counties, ecoregions, or tiles as categorical variables in the same way as we treated land-cover classes for our analyses of NDVI. A key point here is that these are pre-defined regions that are treated as independent variables. PARTS is not designed to find regional patterns that are not initially specified in a model. This limits the ability of PARTS to be used for data exploration. As remote-sensing datasets continue to increase in temporal resolution (e.g., daily records) and spatial resolution (e.g., 3 m), additional statistical methods will be needed to investigate regional-scale patterns nested within map-scale patterns.

From a statistical perspective, PARTS requires little to be specified in a model. The time-series analyses of each pixel give estimates not only of time trends, but also of the intercept and strength of temporal autocorrelation (a_i and β_i in Eq. (1)), and uncertainty in the estimates. However, PARTS does not use this additional information when performing the spatial analysis. This setting aside of time-series information adds flexibility to the PARTS approach; different types of time-series models beyond those we considered here could be used in PARTS when needed to account for different types of trends, such as nonlinear time trends. Setting aside time-series information also adds robustness to the PARTS analyses, in the sense of being relatively insensitive to misspecification of the model. For example, unlike full spatiotemporal statistical models (Krainski et al., 2019; Wikle et al., 2019), PARTS does not require estimating how the values of the intercept from the pixel-level time-series (a_i in Eq. (1)) vary through space, because these values are not used in the spatial analysis. However, because PARTS sets aside information that is not needed to address a specific map-scale hypothesis, PARTS is not designed for predicting values at specific points in time and space.

The strength of PARTS is its ability to pose and test hypotheses at the scale of entire maps, even for maps with millions of pixels. PARTS will hopefully be a useful tool for making the most of remote-sensing data. The need for statistical methods to identify and test map-scale patterns is growing as time series of satellite imagery are becoming increasingly available. Statistical methods will make it possible to use these data to capture different facets of global change.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank the editor and three anonymous reviewers for their insightful comments that greatly improved our manuscript. This work was supported by a NASA AIST program grant [80NSSC20K0282] to ARI, VCR, FW and JZ; by an NSF grant [DEB-1556208] to ARI; and by a NASA LCLUC grant [80NSSC18K0316] and a NASA MuSLI grant [80NSSC18K0343] to VCR.