AGU PUBLICATIONS
JGR
Journal of Geophysical Research: Biogeosciences

RESEARCH ARTICLE
10.1002/2016JG003576

Key Points:
- FLUXNET is unevenly representative across sites in terms of the measurement lengths and potentials of extrapolation in time
- Several drivers have trends or breakpoints in the baseline period and are underrepresented by FLUXNET measurement periods
- Justification of site temporal representativeness should consider the natural variability of climatological and biological conditions

Supporting Information:
- Supporting Information S1
- Data Set S1
- Data Set S2

Correspondence to:
H. Chu,
hchu@berkeley.edu

Citation:
Chu, H., D. D. Baldocchi, R. John, S. Wolf, and M. Reichstein (2017), Fluxes all of the time? A primer on the temporal representativeness of FLUXNET, *J. Geophys. Res. Biogeosci.*, *122*, 289–307, doi:10.1002/2016JG003576.

Received 5 AUG 2016
Accepted 13 JAN 2017
Accepted article online 24 JAN 2017
Published online 8 FEB 2017

Fluxes all of the time? A primer on the temporal representativeness of FLUXNET

Housen Chu[1], Dennis D. Baldocchi[1], Ranjeet John[2], Sebastian Wolf[3], and Markus Reichstein[4]

[1]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA, [2]Center for Global Change and Earth Observations, Michigan State University, East Lansing, Michigan, USA, [3]Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland, [4]Department Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany

Abstract FLUXNET, the global network of eddy covariance flux towers, provides the largest synthesized data set of CO_2, H_2O, and energy fluxes. To achieve the ultimate goal of providing flux information "everywhere and all of the time," studies have attempted to address the representativeness issue, i.e., whether measurements taken in a set of given locations and measurement periods can be extrapolated to a space- and time-explicit extent (e.g., terrestrial globe, 1982–2013 climatological baseline). This study focuses on the temporal representativeness of FLUXNET and tests whether site-specific measurement periods are sufficient to capture the natural variability of climatological and biological conditions. FLUXNET is unevenly representative across sites in terms of the measurement lengths and potentials of extrapolation in time. Similarity of driver conditions among years generally enables the extrapolation of flux information beyond measurement periods. Yet such extrapolation potentials are further constrained by site-specific variability of driver conditions. Several driver variables such as air temperature, diurnal temperature range, potential evapotranspiration, and normalized difference vegetation index had detectable trends and/or breakpoints within the baseline period, and flux measurements generally covered similar and biased conditions in those drivers. About 38% and 60% of FLUXNET sites adequately sampled the mean conditions and interannual variability of all driver conditions, respectively. For long-record sites (≥15 years) the percentages increased to 59% and 69%, respectively. However, the justification of temporal representativeness should not rely solely on the lengths of measurements. Whenever possible, site-specific consideration (e.g., trend, breakpoint, and interannual variability in drivers) should be taken into account.

1. Introduction

FLUXNET, the global network of eddy covariance flux towers, has grown rapidly over the last 25 years and largely enhanced our understanding of carbon, water, and energy cycles in terrestrial ecosystems [*Baldocchi*, 2014; *Papale et al.*, 2012]. To date, it is estimated that there are more than 900 active and historical tower sites worldwide and ~7000 site years (~60,000,000 site hours) of data that have been collected. By joining local towers and regional network teams globally, FLUXNET provides the largest synthesized data set of CO_2 (e.g., primary production and respiration), water vapor (e.g., evapotranspiration), and energy fluxes (e.g., sensible and latent heat fluxes). This includes the so-called "La Thuile data set" (initiated in 2007) containing ~1000 site years of data from 252 sites worldwide [*Agarwal et al.*, 2010; *Papale et al.*, 2012], which has been widely used (e.g., ~392 papers since 2007 with a topic of "FLUXNET" on Web of Science). To achieve the goal of providing flux information "everywhere and all of the time" [*Baldocchi*, 2008], machine-learning techniques (e.g., neural networks, regression trees, and kernel methods) were adopted to extrapolate the location- and time-constrained measurements to space- and time-explicit products, such as global long-term maps of CO_2, H_2O, and energy fluxes [*Beer et al.*, 2010; *Jung et al.*, 2009, 2011; *Papale and Valentini*, 2003; *Tramontana et al.*, 2015, 2016; *Xiao et al.*, 2008; *Yang et al.*, 2007, 2006]. Such gridded and harmonized products provide valuable information of terrestrial fluxes at the desired spatial and temporal scales, which have been used extensively in parameterizing, calibrating, and validating other models such as those from satellite-based remote sensing, land-atmosphere climate models, and global CO_2 flask measurements combined with atmospheric inversions [*Anav et al.*, 2015; *Beer et al.*, 2010; *Bonan et al.*, 2011; *Guanter et al.*, 2014; *Jiménez et al.*, 2011; *Kondo et al.*, 2015].

As the network of towers was never designed to maximize the spatial and temporal coverage, concerns were often raised in extrapolating flux information from measurement locations and periods to a space- and time-explicit extent, i.e., the so-called representativeness issue [*Schimel et al.*, 2015; *Sulkava et al.*, 2011; *Sundareshwar et al.*, 2007]. Adopting the definition proposed in *Nappo et al.* [1982], representativeness illustrates the extent to which a set of (flux) measurements taken in a given space-time domain (e.g., site locations and measurement periods) reflect the actual (flux) conditions in the different space-time domain taken on a scale appropriate for a specific application (e.g., terrestrial globe, 30 year climatological baseline). Over the years, several studies have attempted to evaluate the representativeness of FLUXNET in the context of such spatial and temporal extrapolations, of which the majority focused on the space domain. First, a few studies focused on the prediction uncertainties of fluxes by evaluating machine-learning model performance in extrapolation exercises [*Jung et al.*, 2009; *Papale et al.*, 2015]. As such approaches rely on the available data of fluxes for validation, the evaluation of representativeness was inherently constrained by data availability, in both space and time. Second, a few studies used proxies of fluxes derived from independent sources in evaluating network representativeness, such as model outputs from mechanistic models [*Carvalhais et al.*, 2010; *Sulkava et al.*, 2011], or data based on ground inventories and remote sensing [*Schimel et al.*, 2015]. Proxies enable to conduct evaluation in areas without direct flux measurements. Nonetheless, the use of proxies is still constrained by the data availability of proxies themselves (especially for those of ground inventories), and uncertainties in the estimation of proxies may be introduced to the evaluation process. Most importantly, models used to generate proxies often rely on flux data for model parameterization, calibration, or validation. Thus, it is disputed whether proxies are truly independent from flux data (i.e., still constrained by the information provided by flux data) and whether they are suitable for the evaluation of representativeness of flux measurements [*Sulkava et al.*, 2011]. Last, several studies proposed the evaluation of network representativeness based on the potential driver variables of fluxes [*Hargrove and Hoffman*, 2004; *Hargrove et al.*, 2003; *Sulkava et al.*, 2011; *Sundareshwar et al.*, 2007]. Those studies focused on multidimensional driver characteristics (e.g., climate, soil, and plant characteristics) and evaluated the similarity of driver conditions between the areas with and without measurements by using multivariate clustering techniques. In a sense, such approaches avoid using flux data directly but instead evaluate representativeness based on the extent to which the selected driver conditions in areas with flux measurements reflect the driver conditions in areas without measurements. As driver variables usually have much better data coverage than fluxes, the driver-based approaches have been proven to be powerful in evaluating network representativeness, especially in areas where the data or proxies of fluxes were unavailable [*Schimel et al.*, 2007].

By the time of writing, the new data synthesis FLUXNET2015/2017 is being carried out with the number of site years increasing substantially since the La Thuile data set (fluxnet.fluxdata.org) [*Pastorello et al.*, 2017]. Arguably, the newly established tower sites in the recent decades have largely improved the spatial representativeness of FLUXNET. For readers interested in spatial representativeness of flux networks, we refer them to other publications for detailed discussions [*Beringer et al.*, 2016; *Hargrove and Hoffman*, 2004; *Hargrove et al.*, 2003; *Kumar et al.*, 2016; *Sulkava et al.*, 2011; *Sundareshwar et al.*, 2007]. Yet new challenges emerge as many sites are only established after the late 2000s, and thus, the distribution of measurement lengths is highly skewed across sites. While the early established sites have reached more than 15–20 years of measurements, the majority (>50%) of sites only operate for 3–5 years or less. Thus, data users often have to make a compromise between spatial coverage (e.g., focusing on a short period with the maximal number of sites) and temporal coverage (e.g., focusing on a small subset of sites with long records). The objective of this study is to examine the representativeness of FLUXNET in the context of such uneven measurement lengths across the network. In other words, could sites with different measurement lengths provide flux information “all of the time,” such as that was intended to achieve by using machine-learning extrapolation? Specifically, we focused on the temporal representativeness of each tower site and tested whether the site-specific measurement period was sufficient to capture the natural variability of climatological and biological conditions at a given site location. We also proposed some guidelines and potential areas of research interests for future FLUXNET syntheses.

2. Materials and Methods

2.1. Overview

Here we propose an analysis framework to systematically evaluate the temporal representativeness of FLUXNET by focusing on the driver variables of target interests (e.g., CO_2, H_2O, latent, and sensible heat fluxes), similar to previous driver-based studies [*Hargrove and Hoffman*, 2004; *Hargrove et al.*, 2003; *Sulkava

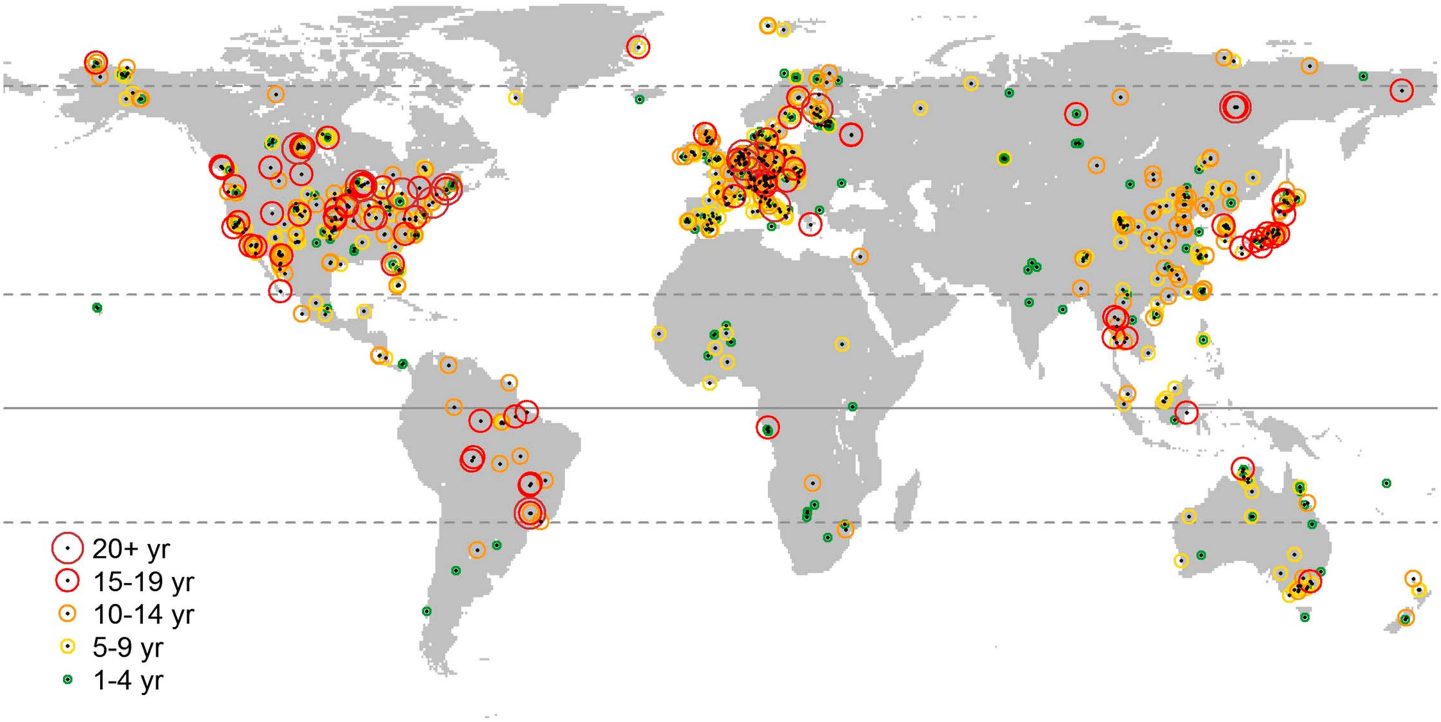

Figure 1. Map of active and historical FLUXNET tower sites used in the study. The color and size of the circle indicate the lengths of measurements as of December 2015. The solid and dashed lines denote equator, Tropic of Cancer/Capricorn, and the Arctic Circle, respectively. For data sources and details refer to Table S1.

et al., 2011; *Sundareshwar et al.*, 2007]. By focusing on independent driver data sets, we intend to infer the temporal representativeness beyond the periods with flux measurements, which were mostly bounded within 2005–2012. Also, by avoiding using flux data, we intend to expand our analysis works to a larger set of FLUXNET sites, where a standardized and openly shared flux data set was still not available [*Papale et al.*, 2012]. As our goal was to provide a generalized overview, we set the period of 1982–2013 as the long-term baseline (the longest possible) and included as many sites as possible by requiring the least amount of information from each tower site. The time series of preselected driver variables were obtained from independent gridded and harmonized data sets for the baseline period at each tower location. A series of statistical analyses were then applied to test whether the driver conditions (e.g., mean and variability) in years with flux measurements adequately represented those of the baseline period. In other words, we tested whether flux measurements (bounded within the measurement periods) had the possibility to "see" all the driver conditions of the baseline period. There were a few assumptions taken while conducting these analyses. (1) We assumed that there were no measurement errors and/or data gaps in flux measurements. That implied, flux information in years with measurements could be properly retrieved and used for extrapolation in years with no measurements. (2) The flux measurements at each tower site adequately reflected the area-average flux conditions over a given spatial extent, at which the area-average driver conditions were properly described by available driver data sets. (3) To provide flux information in years with no flux measurement, the years with measurements needed to experience similar driver conditions as those in years without measurements. Practically, these aforementioned assumptions may not always be valid at all sites. The violation of the assumptions generally reduced the potential representativeness of flux measurements, and thus, our analyses could be viewed as the optimistic estimates in the best case scenario. The implications of such assumptions for the interpretation of representativeness are discussed in section 3.4.

2.2. FLUXNET Registered Tower Sites

A full list of FLUXNET tower sites was compiled from the databases of FLUXNET-ORNL, FLUXNET-Fluxdata, AmeriFlux, European Fluxes Database, AsiaFlux, ChinaFlux, OzFlux, and U.S.-China Carbon Consortium (USCCC) (Table S1 in the supporting information). The site list and general site information were acquired,

Table 1. List of Selected Driver Variables

Abbreviation	Variable	Unit	Data Source
	Monthly (Time Series Analysis)		
tmp	Air temperature	°C	CRU TS 3.22
pre	Precipitation	mm	CRU TS 3.22
dtr	Diurnal air temperature range	°C	CRU TS 3.22
pet	Potential evapotranspiration	mm	CRU TS 3.22
ssrd	Incoming shortwave radiation	$W\ m^{-2}$	ERA-Interim
ndvi	Normalized difference vegetation index	Unitless	GIMMS NDVI3g
	Yearly (Standardized Anomaly, Equal Mean, Equal Variance, and Multivariate Similarity Analyses)		
tmp.mn	Mean annual air temperature	°C	CRU TS 3.22
tmp.hqtr	Mean air temperature in the warmest quarter[a]	°C	CRU TS 3.22
tmp.cqtr	Mean air temperature in the coolest quarter[a]	°C	CRU TS 3.22
pre.mn	Annual precipitation	mm	CRU TS 3.22
pre.hqtr	Precipitation in the warmest quarter[a]	mm	CRU TS 3.22
pre.cqtr	Precipitation in the coolest quarter[a]	mm	CRU TS 3.22
pptpet	Ratio of annual precipitation to annual potential evapotranspiration	Unitless	CRU TS 3.22
bioh	Potential growing degree[b]	°C	CRU TS 3.22
pet.mn	Annual potential evapotranspiration	mm	CRU TS 3.22
pet.hqtr	Potential evapotranspiration in the warmest quarter[a]	mm	CRU TS 3.22
dtr.mn	Mean diurnal air temperature range	°C	CRU TS 3.22
dtr.hqtr	Diurnal air temperature range in the warmest quarter[a]	°C	CRU TS 3.22
ssrd.mn	Mean incoming shortwave radiation	$W\ m^{-2}$	ERA-Interim
ssrd.hqtr	Incoming shortwave radiation in the warmest quarter[a]	$W\ m^{-2}$	ERA-Interim
ndvi.mn	Mean annual NDVI	Unitless	GIMMS NDVI3g
ndvi.hqtr	NDVI in the peak-NDVI quarter[a]	Unitless	GIMMS NDVI3g
ndvi.thwth.f	Length of greening period[c]	Month	GIMMS NDVI3g

[a] The warmest quarter is defined based on the 32 year mean seasonal variation of air temperature, where the three consecutive months have the highest mean air temperature. December and January are treated as consecutive months. Similar definition is also used to identify the coolest quarters and the peak-NDVI and trough-NDVI quarters.

[b] Sum of monthly air temperature when air temperature ≥5°C.

[c] The length of greening period is defined as the number of months in each year that monthly NDVI exceeds the site-specific NDVI threshold. Site-specific NDVI thresholds are defined as one of the following: (1) NDVI in the trough-NDVI quarter + 0.2 × (NDVI in the peak-NDVI quarter − NDVI in the trough-NDVI quarter) or 0.2 (whichever larger) for sites showing detectable seasonality in NDVI. (2) 0.2 for sites showing no detectable seasonality in NDVI. An example is presented in Figure 2b.

harmonized, and cross validated among databases. There were 916 sites in total registered across databases by the time of data query (December 2015), of which 875 sites had the basic general site information required for our analyses (i.e., site name, country, latitude, longitude, and measurement starting and ending years). Sites without specified ending years were treated as active. We further screened out sites that only started after our target period (1982–2013). Finally, 855 sites were retained and used for the following analyses unless specified otherwise (Figure 1 and Table S1). For easier interpretation, we grouped the sites into three main regions of South to North America, Africa-Europe, and Oceania-Asia while presenting the results.

2.3. Selected Driver Variables

We selected a series of potential driver variables based on literature survey and data availability (Table 1). The final list of selected drivers included climatic variables such as air temperature, precipitation, potential evapotranspiration, diurnal temperature range, incoming shortwave radiation, and vegetation indices such as satellite-based normalized difference vegetation index (NDVI). The selection of driver also considered the following factors: (1) Drivers were used previously in predicting long-term gridded fluxes via machine-learning techniques [*Jung et al.*, 2009, 2011; *Papale et al.*, 2015; *Papale and Valentini*, 2003; *Tramontana et al.*, 2016]. (2) Divers were used previously in evaluating spatial representativeness of flux tower networks [*Hargrove et al.*, 2003; *Sundareshwar et al.*, 2007]. (3) Drivers were available regarding the required spatial and temporal coverages (terrestrial globe, 1982–2013). (4) Certain variables were further excluded after preliminary tests such as plant functional type, topography, soil property, and nutrient content because their year-to-year variation was assumed negligible or their long-term gridded data were unavailable to the best of our knowledge.

Table 2. Schemata for Evaluating the Temporal Representativeness of Flux Measurement Periods

Statistics	Quantitative Criterion	Description	Interpretation
Trend + breakpoint	BFAST model, OLS-MOSUM test[a] ($p < 0.05$)	No trend or breakpoint	More likely, measurement period represents the entire baseline period.
		Detectable trend	Less likely
		Detectable breakpoint within the measurement period	Unknown
		Detectable breakpoint outside the measurement period	Less likely
Standardized anomaly	Student's *t* test ($p < 0.05$)[b]	Unevenly distributed within/outside the measurement period	Less likely
		Evenly distributed within/outside the measurement period	More likely
Equal mean	Wilcoxon Mann-Whitney test ($p < 0.05$)	Unequal means between the measurement period and the entire baseline period	Less likely
		Equal means between the measurement period and the entire baseline period	More likely
Equal variance	Levene's test ($p < 0.05$)	Unequal variance between the measurement period and the entire baseline period	Less likely
		Equal variance between the measurement period and the entire baseline period	More likely
Multivariate similarity	Least similar year at 5+, 10+, and 15+ year sites[c]	Low similarity between years with and without measurements	Less likely
		High similarity between years with and without measurements	More likely

[a] OLS-MOSUM test: ordinary least squares residual-based moving sum test [*Zeileis*, 2005].
[b] Student's *t* test is only conducted for pooled standardized anomalies.
[c] For implementation details refer to the main text (section 2.4).

For each tower site, the 32 year time series of the selected drivers were obtained from the gridded data sources (Table 1). No spatial downscaling was conducted, and the time series from the grid that covered the tower location were retrieved and used for all following analyses (section 2.1, assumption 2). Most climatic variables were obtained from the observation-based Climatic Research Unit (CRU) time series (TS) 3.22 data set [*Harris and Jones*, 2014; *Harris et al.*, 2014] (Table 1), which had a monthly temporal resolution and a 0.5° spatial resolution. Incoming shortwave radiation was obtained from the ERA-Interim reanalysis data collection [*Dee et al.*, 2011], which had a 6 h temporal resolution and a 0.75° spatial resolution. We further integrated the radiation data to a monthly scale. NDVI was obtained from the Global Inventory Modeling and Mapping Studies (GIMMS) NDVI3g data set [*Pinzon and Tucker*, 2014]. NDVI3g is the third generation of GIMMS NDVI derived from the advanced very high resolution radiometer (AVHRR) [*Anyamba et al.*, 2014] and has a bimonthly temporal resolution and a 0.083° spatial resolution. This latest version of GIMMS NDVI employs state-of-the-art techniques to remove artifacts in the time series, which are due to differences among sensors, solar zenith angle, and orbital drift. We used the embedded quality flags to account for the possibility of false positives from clouds or snow. We further aggregated the bimonthly NDVI to a monthly scale by using the maximum value composite method [*Guay et al.*, 2014; *Holben*, 1986]. In the preliminary tests, we found that such temporal aggregation helped filter the remaining problematic data points that were not captured by the embedded quality flags [*Forkel et al.*, 2013].

2.4. Statistical and Analysis Framework

Several groups of statistical analyses were carried out to test whether the driver conditions in years with flux measurements reflected those of the baseline period, including the univariate time series, equal mean, equal variance, and multivariate similarity analyses (Table 2). The software R was used for all statistical analyses [*R Development Core Team*, 2016]. Herein, we treated each year as the best resolved and inseparable unit in presenting our analyses ($n = 32$), as the FLUXNET synthesis data set was typically distributed with a minimum unit of one site year. The significance level was set to 0.05 in all our analyses.

The BFAST (Breaks For Additive Seasonal and Trend) model was adopted to detect trends and breakpoints (i.e., change in slope of trend) in the monthly time series of six drivers for each tower site (Table 1) [*de Jong et al.*, 2013; *Verbesselt et al.*, 2010, 2012]. The model was adopted for its capability of detecting trends

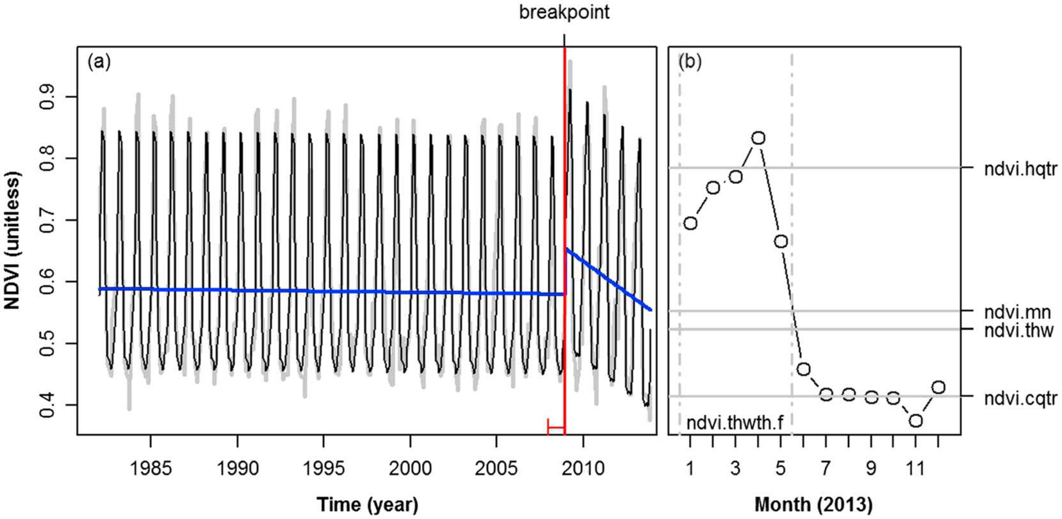

Figure 2. An example of (a) BFAST season-trend model fitting and (b) yearly aggregations of NDVI time series from the US-Ton (Tonzi Ranch) site. The grey lines in Figure 2a show the original monthly NDVI time series. The black, blue, and red lines denote the seasonality, trend, and breakpoint of the fitted BFAST model, respectively. The open circles in Figure 2b show the original monthly NDVI time series of one example year (2013). The horizontal grey lines indicate the yearly aggregations including mean NDVI in the peak-NDVI quarter (ndvi.hqtr), annual mean NDVI (ndvi.mn), site-specific NDVI threshold (ndvi.thw), and mean NDVI in the trough-NDVI quarter (ndvi.cqtr). The vertical dash-dotted lines show the start and end of the identified greening period (ndvi.thwth.f). Please note that ndvi.cqtr and ndvi.thw are not used directly in the analyses, so they are not listed in Table 1. Yet they are used to determine the greening periods, so they are kept here in the figure.

and breakpoints within time series that possessed periodic nature (Figure 2a). The model fitting was carried out only on monthly time series to ensure the model robustness [*Forkel et al.*, 2013]. The criterion of determining whether a breakpoint existed in the time series was based on the tests of ordinary least squares residual-based moving sum (OLS-MOSUM) [*Zeileis*, 2005]. The detection of a breakpoint indicated an abrupt or step change in time series (Figure 2a), which may associate with regime transition of a driver variable. If a breakpoint is identified outside the period of flux measurements, any inference or prediction that is tightly associated with the driver variable would be less likely to extrapolate to periods without flux measurements (Table 2).

Standardized anomaly, equal mean, equal variance, and multivariate similarity analyses were carried out on yearly aggregated data. Several aggregations such as annual integrals, means of the peak quarter, amplitudes, and periods of seasonal cycles were adopted similar to previous studies (Figure 2b) [*Hargrove et al.*, 2003; *Jung et al.*, 2009, 2011; *Sundareshwar et al.*, 2007]. Using these aggregations enabled our analyses at a yearly scale while still preserving within-year features of drivers in specific seasons (e.g., warm, growing seasons). In total, we kept 17 yearly aggregated divers for following analyses (Table 1).

First, we calculated the standardized anomaly for each yearly aggregated driver by normalizing the time series with its 32 year mean and standard deviation [*Wilks*, 2011]. For each tower site, the standardized anomaly illustrated the relative distribution of driver conditions within and outside the flux measurement periods (Table 2). We further pooled the standardized anomaly in the flux measurement periods from the three main regions (Oceania to Asia, Africa to Europe, and North to South America). The Student's *t* test was applied to test if the pooled anomaly was significantly different from zero. The test illustrated whether the flux measurement periods in each region adequately represented the baseline mean driver conditions.

Second, we adopted the Wilcoxon Mann-Whitney and Levene's tests to examine the equality of means and variances for each yearly aggregated variable at each tower site [*Hollander et al.*, 2014; *Levene*, 1960]. The Wilcoxon Mann-Whitney test is a nonparametric test for the equality of means between groups. Levene's test is a robust test for the equality of variances between groups. Both tests were adopted because they were less sensitive to departures from normality. The comparison was made only for sites having at least 5 years of flux measurements (i.e., 536 sites) and between the actual measured years and target baseline (1982–2013). The tests illustrated whether the driver conditions in years with flux measurements adequately represented the baseline mean conditions and interannual variability (Table 2).

Last, we calculated multivariate similarity metrics among each pair of the 32 years (1982–2013) at each site based on the multidimensional distances of 17 yearly aggregated drivers (Table 1). Both Manhattan and Euclidean distances were calculated in the preliminary tests, and we found that the choice of distance metrics had only minor effects on the interpretation of the final results. Thus, only the results using the Manhattan distances were presented here. For each tower site, the similarity was normalized to the full range of 0–100% such that the pair of years with the largest distance had zero similarity and the pairs of exact match (e.g., self-comparison) had 100% similarity. For each year without flux measurements, we then derived its maximum similarity to any of the years with flux measurements based on the cross-year similarity metrics. The maximum similarity represented how similar the years without flux measurements could be to any year with flux measurements in the multidimensional driver space and, thus, how likely the years not measured could be represented by the measured years (Table 2). To help interpret the similarity results across sites, we set a series of criteria based on the maximum similarity of the least similar year at each site and averaged across sites with 5+, 10+, and 15+ years of flux measurements. Such criteria express the extent to which the measurement period at each site could potentially represent when judging by the least similar years of the 5+, 10+, and 15+ year sites.

3. Results and Discussion

In the following analyses we showed the extent to which the selected driver conditions in years with flux measurements reflected those in years without flux measurements. We first focused on the time series characteristics (e.g., trend, breakpoint, and standardized anomaly) of each driver variable (section 3.1) and tested whether site-specific measurement periods were long enough to capture the baseline climatological and biological conditions (e.g., means and variance) at each tower site (section 3.2). We further examined the multivariate similarity of selected drivers and evaluated whether site-specific measurement periods were sufficient to adequately sample the multidimensional driver conditions at each tower site (section 3.3). We finally discussed the implications of our results and provided suggestions for future applications (section 3.4).

3.1. Univariate Time Series Analysis

Several drivers such as air temperature, diurnal temperature range, potential evapotranspiration, and NDVI had detectable trends and/or breakpoints within the baseline period. As the majority of tower sites started measurements in the late 2000s and operated in consecutive years, flux measurement periods tended to cover similar and biased driver conditions.

Air temperature showed increasing trends over the 32 year period across the majority of regions and sites (Figures 3h–3j). As most tower sites started in the late 2000s, the measured periods thus experienced generally above-average air temperature (Figures 3d–3g and S1a–S1f in the supporting information and Table 3). Several very warm years were identified in the periods with abundant site measurements, such as the summer of 2011 and the year of 2012 in North America, the summer of 2003 and the year of 2011 in Europe, and the summer of 2013 and the year of 2007 in Asia. On the other hand, we identified a few cool years with abundant site measurements such as the year of 2009 in North America, the winters of 2005 and 2010 in Europe, and the winters of 2011 and 2012 in Asia (Figures 3e–3g and S1a–S1f).

Precipitation showed no consistent trends across all regions and sites (Figures 4h–4j). Flux-measured periods tended to better sample the year-to-year variation in precipitation than air temperature (Figures 4d–4g and Table 3). There were regions and years that had abundant site measurements and experienced very high precipitation such as the summers of 2007 and 2012 in Europe, the year of 2010 in South Europe and Africa, the winters of 2007 and 2008 for 35°N–45°N in North America, and the years of 2010–2011 in Australia (Figures 4e–4g and S2a–S2f). On the other hand, very low precipitation was observed in the year of 2003 in Europe; the years of 2001, 2002, and 2012 in midlatitude North America; and the year of 2013 for 40°N–60°N in Asia.

Potential evapotranspiration, incoming shortwave radiation, and diurnal temperature range showed less consistent trends across all regions (Figures S3–S5 and Table 3). Yet there were local patterns of trends

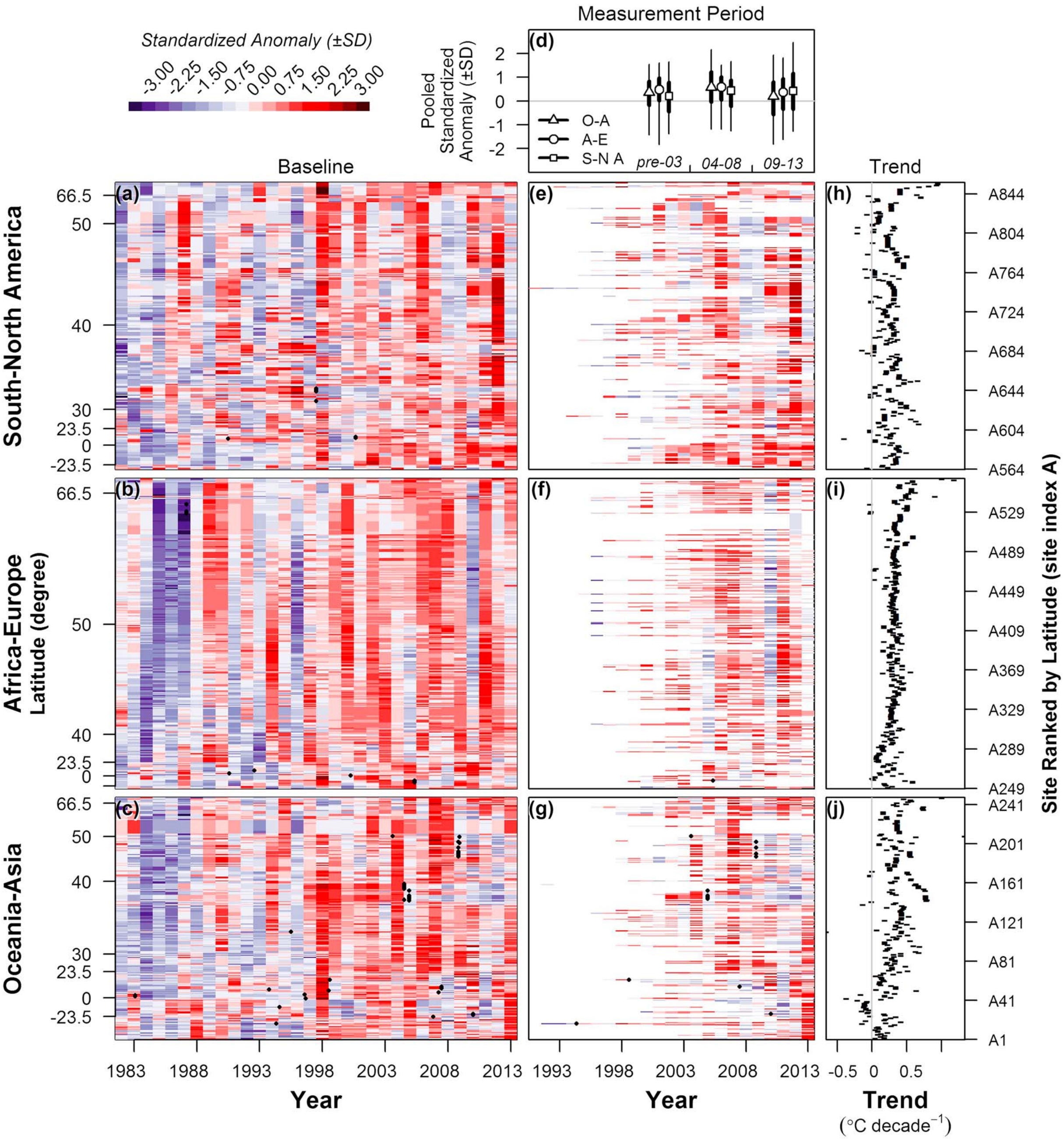

Figure 3. (a–g) Standardized anomaly and (h–j) slope of trend of air temperature from 1982 to 2013 at FLUXNET tower sites. In total, 855 sites are presented here. The figures are arranged as follows. The tower sites are grouped into the three main regions: South to North America (Figures 3a, 3e, and 3h), Africa to Europe (Figures 3b, 3f, and 3i), and Oceania to Asia (Figures 3c, 3g, and 3j). Within each region, the sites are ordered vertically by latitude and then equally spaced ranging from the most southern (bottom) to the most northern one (top). The corresponding latitudes and site index were labeled on the left and right *y* axes, respectively. For readers interested in identifying the sites in each panel, a full list of the site index is provided in Table S1 (site index A). Figures 3a–3c show the full record (1982–2013) of standardized anomaly, while Figures 3e–3g only show periods with tower measurements (i.e., the white color denotes no measurement). Each thin colored rectangle in Figures 3a–3c and 3e–3g denotes the standardized anomaly from one site year, and the colors denote the values of standardized anomaly. For a block of areas with similar colors (e.g., 2012 at midlatitude North America), it implies that there are a number of site years in adjacent latitudes that experience the similar standardized anomaly. Standardized anomaly is calculated based on the 32 year mean and standard deviation of air temperature at each tower site. Figure 3d shows the pooled standardized anomalies for the years with flux measurements. The data are grouped by three main regions and three main periods (i.e., pre-2003, 2004–2008, and 2009–2013). The open triangles, circles, and squares denote the medians of the pooled standardized anomalies in regions of Oceania to Asia, Africa to Europe, and North to South America, while the thick and thin lines denote the 25%–75% and 2.5%–97.5% quantile intervals, respectively. The closed circles in Figures 3a–3c and 3e–3g indicate the time and sites with identified breakpoints in the time series. Figures 3h–3j show the slopes of trends. For a site with an identified breakpoint, only trend in the relatively longer time period is shown.

Table 3. Means of the Pooled Standardized Anomalies (±SD) of Selected Drivers From Flux Measurement Periods[a]

	Oceania to Asia			Africa to Europe			North to South America		
Selected Diver	Pre-2003	2004–2008	2009–2013	Pre-2003	2004–2008	2009–2013	Pre-2003	2004–2008	2009–2013
tmp.mn	0.26	0.54	0.14	0.46	0.48	0.26	0.17	0.33	0.47
tmp.hqtr	0.15	0.26	0.48	0.54	0.25	0.41	0.13	0.20	0.52
tmp.cqtr	0.22	0.15	−0.14	0.21	0.34	−0.45	0.32	0.06	0.09
dtr.mn	ns	0.12	−0.11	0.25	0.14	0.18	ns	−0.19	−0.14
dtr.hqtr	ns	ns	ns	ns	0.27	0.21	ns	ns	ns
bioh	0.26	0.47	0.34	0.50	0.40	0.46	ns	0.30	0.54
pre.mn	ns	ns	0.24	ns	ns	0.17	−0.20	ns	ns
pre.hqtr	ns	−0.08	ns	ns	0.21	0.21	−0.12	ns	ns
pre.cqtr	ns	ns	0.10	ns	ns	0.14	−0.11	0.11	ns
pptpet	ns	−0.20	0.12	ns	ns	ns	−0.13	ns	−0.10
pet.mn	0.13	0.48	0.37	0.24	0.16	0.41	−0.12	0.07	0.33
pet.hqtr	ns	0.26	0.45	0.14	0.18	0.35	ns	ns	0.20
ssrd.mn	−0.26	0.38	0.10	0.24	0.12	−0.17	0.31	0.27	0.29
ssrd.hqtr	−0.22	0.22	ns	0.16	0.20	−0.26	0.26	0.28	0.10
ndvi.mn	0.19	ns	0.22	0.15	0.34	0.30	ns	ns	0.13
ndvi.hqtr	ns	ns	0.32	0.13	ns	0.60	−0.13	ns	0.53
ndvi.thwth.f	0.16	0.09	ns	ns	0.29	ns	ns	0.07	−0.12

[a]For better presentation, the means of pooled standardized anomalies are omitted if they are not significantly different from zero (ns: $p \geq 0.05$). The test illustrates whether the selected driver in the flux measurement periods in each region adequately represents the baseline mean driver conditions (i.e., zero). If the pooled standardized anomaly is significantly different from zero, then it expresses that the flux measurement periods experience biased driver conditions (i.e., higher or lower than baseline). The data are grouped by three main regions (Oceania to Asia, Africa to Europe, and North to South America) and three selected periods (pre-2003, 2004–2008, and 2009–2013).

appearing in certain regions, such as increasing potential evapotranspiration in Europe and Asia (Figures S3h–S3j) and increasing incoming shortwave radiation in North America (Figure S4h). Extreme years identified in air temperature or precipitation were also highlighted as extremes in potential evapotranspiration, incoming shortwave radiation, and diurnal temperature range. That included the warm and drought year of 2003 in Europe (Figures S3f, S4f, and S5f) and 2012 in North America (Figures S3e, S4e, and S5e). Noticeably, breakpoints within the time series of diurnal temperature range were detected at ~176 sites, of which the majority (93%) occurred before the periods with flux measurements (Figures S5a–S5c and S5e–S5g).

Most of the aforementioned region years with climatic anomalies and extremes were documented in previous studies, such as large-scale droughts (e.g., 2003 Europe, 2003–2004 western U.S., 2011 southern U.S., and 2012 Midwest U.S.) [*Ciais et al.*, 2005; *Coumou and Rahmstorf*, 2012; *Mallya et al.*, 2013; *Parazoo et al.*, 2015; *Schwalm et al.*, 2012; *Wolf et al.*, 2016], heat waves (e.g., 2003 summer Europe and 2012 North America) [*Ault et al.*, 2013; *Ciais et al.*, 2005; *Karl et al.*, 2012; *Wuebbles et al.*, 2014], and high precipitation causing flooding (e.g., 2010–2011 Australia and 2007 summer western Europe) [*Bastos et al.*, 2013; *Coumou and Rahmstorf*, 2012; *Haverd et al.*, 2013]. To the best of our knowledge, only a few of those extreme events have been studied extensively for their effects on CO_2 and energy fluxes across sites and regions [e.g., *Bastos et al.*, 2013; *Chu et al.*, 2016; *Ciais et al.*, 2005; *Schwalm et al.*, 2012; *Wolf et al.*, 2013, 2016; *Zscheischler et al.*, 2014]. These "hot spots" (region years) of climate extremes pointed out potential areas of future region-wide syntheses by using the new FLUXNET data set.

NDVI showed increasing trends in a large portion of sites and regions despite site-to-site variation (Figure 5). The widespread greening trends agreed with reports in previous studies, suggesting that the majority of sites were experiencing increased vegetation dynamics, possibly driven by long-term climate changes and anthropogenic influences (e.g., management and recovery from disturbance) [*de Jong et al.*, 2012, 2013; *Neigh et al.*, 2008; *Nemani et al.*, 2003; *Zhu et al.*, 2016]. Very green years with high NDVI were identified mostly in the last few years (2009–2013) corresponding to those with warmer air temperature and higher potential growing degrees (Figures 5d–5g and S6g–S6i and Table 3). The green years were generally more associated with higher NDVI in the peak quarter than longer greening periods (Figures S6a–S6f and Table 3). Yet as our analyses were conducted based on monthly time series, potential changes in the lengths of greening periods (e.g., days to weeks) may not be always distinguished.

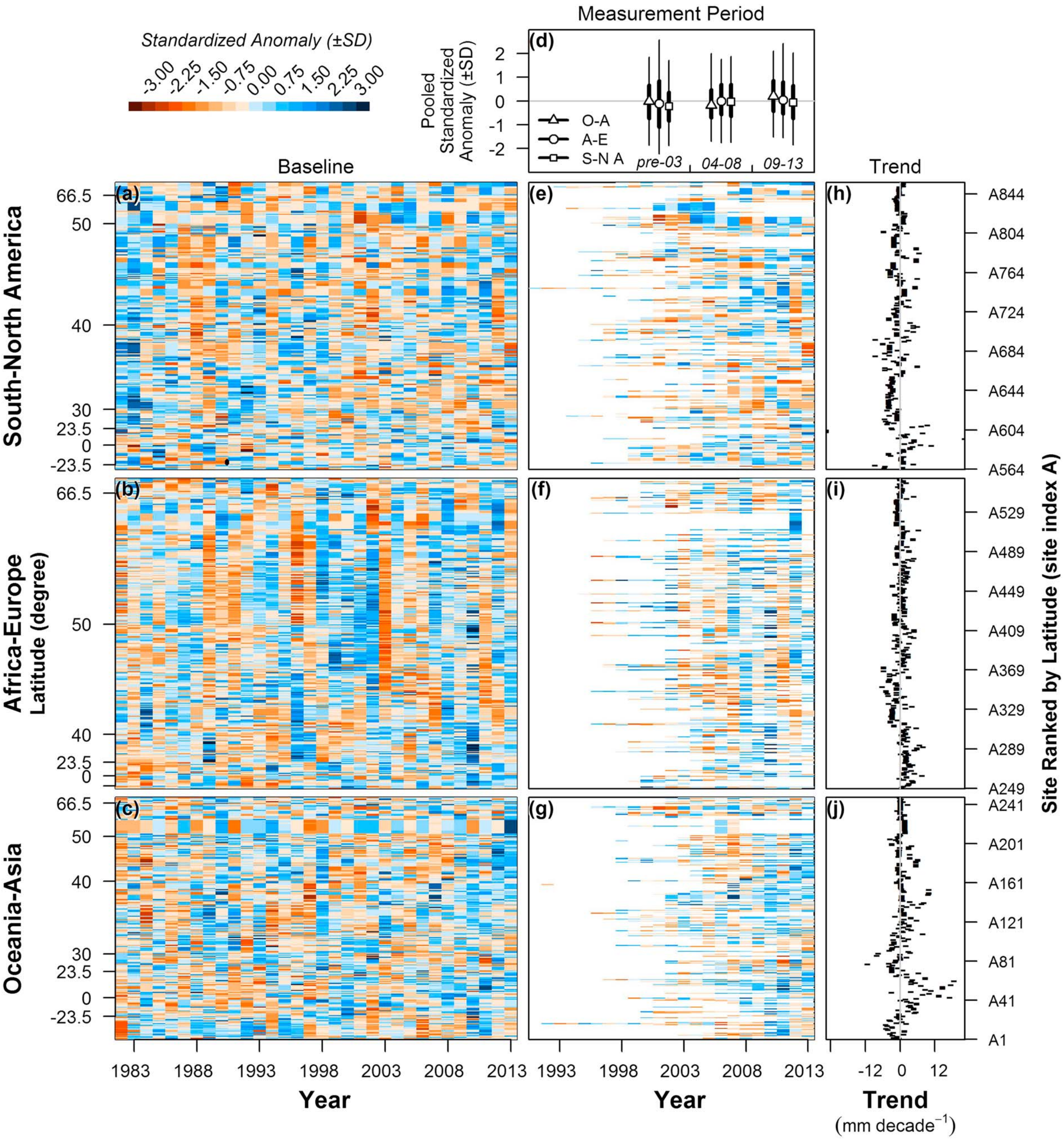

Figure 4. (a–g) Standardized anomaly and (h–j) slope of trend of precipitation from 1982 to 2013 at FLUXNET tower sites. In total, 855 sites are presented here. The tower sites are grouped into the three main regions: South to North America (Figures 4a, 4e, and 4h), Africa to Europe (Figures 4b, 4f, and 4i), and Oceania to Asia (Figures 4c, 4g, and 4j). Within each region, the sites are ordered vertically by latitude and then equally spaced ranging from the most southern (bottom) to the most northern one (top). The corresponding latitudes and site index were labeled on the left and right *y* axes, respectively. For readers interested in identifying the sites in each panel, a full list of the site index is provided in Table S1 (site index A). Figures 4a–4c show the full record (1982–2013) of standardized anomaly, while Figures 4e–4g only show periods with tower measurements (i.e., the white color denotes no measurement). Figure 4d shows the pooled standardized anomalies for the years with flux measurements. The data are grouped by three main regions and three main periods (pre-2003, 2004–2008, and 2009–2013). The open triangles, circles, and squares denote the medians of the pooled standardized anomalies in Oceania to Asia, Africa to Europe, and North to South America, while the thick and thin lines denote the 25%–75% and 2.5%–97.5% quantile intervals, respectively. The figure is organized similar to Figure 3.

Breakpoints in NDVI time series were detected in ~175 sites (21%) across the network. There were certain periods and regions where breakpoints were mostly identified, such as the years of 2008–2010 in North America, 1986–1989 in Europe, 1994–1995 in Asia, and 2002–2010 in Oceania (Figures 5a–5c). Some of these region-wide breakpoints were reported in earlier studies [*de Jong et al.*, 2012, 2013; *Parazoo et al.*, 2015],

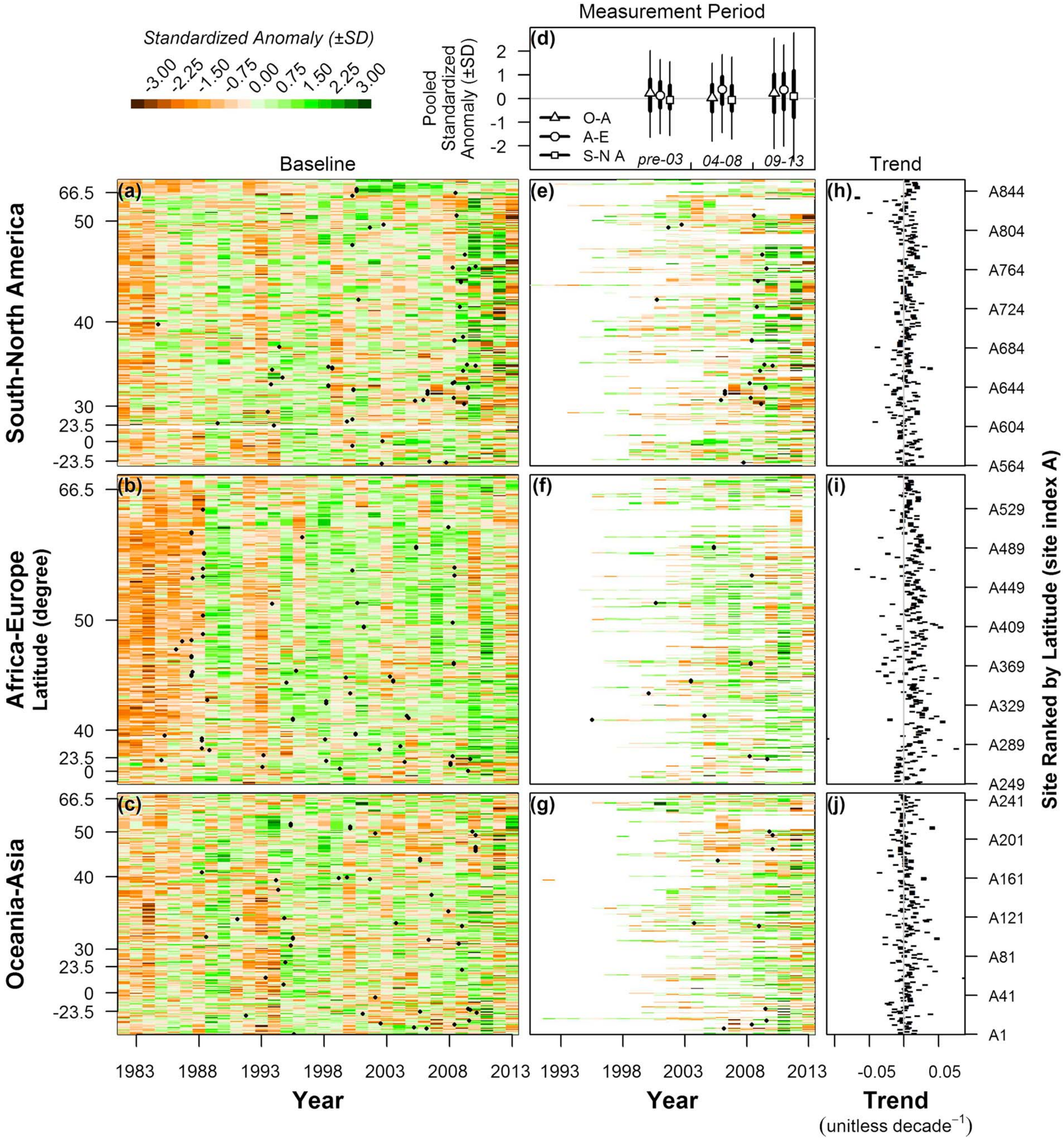

Figure 5. (a–g) Standardized anomaly and (h–j) slope of trend of normalized difference vegetation index (NDVI) from 1982 to 2013 at FLUXNET tower sites. In total, 855 sites are presented here. The tower sites are grouped into the three main regions: South to North America (Figures 5a, 5e, and 5h), Africa to Europe (Figures 5b, 5f, and 5i), and Oceania to Asia (Figures 5c, 5g, and 5j). Within each region, the sites are ordered vertically by latitude and then equally spaced ranging from the most southern (bottom) to the most northern one (top). The corresponding latitudes and site index were labeled on the left and right *y* axes, respectively. For readers interested in identifying the sites in each panel, a full list of the site index is provided in Table S1 (site index A). Figures 5a–5c show the full record (1982–2013) of standardized anomaly, while Figures 5e–5g only show periods with tower measurements (i.e., the white color denotes no measurement). Figure 5d shows the pooled standardized anomalies for the years with flux measurements. The data are grouped by three main regions and three main periods (pre-2003, 2004–2008, and 2009–2013). The open triangles, circles, and squares denote the medians of the pooled standardized anomalies in Oceania to Asia, Africa to Europe, and North to South America, while the thick and thin lines denote the 25%–75% and 2.5%–97.5% quantile intervals, respectively. The figure is organized similar to Figure 3.

where they were claimed to be associated with large-scale natural influences such as strong El Niño/La Niña, climate extremes. It is beyond our current scope to interpret the driving factors of these greening/browning trends and breakpoints. Nonetheless, the abundant site coverage in 2008–2010 in North America pointed out a potential area for future syntheses, where the focus could be on validating the breakpoints by using alternative remote sensing products with a finer spatial/temporal/spectral resolution (e.g., Moderate Resolution Imaging Spectroradiometer, Landsat, and Hyperion) [e.g., *Guay et al.*, 2014] or testing the robustness of models in extrapolating inferences or predictions across these breakpoints.

3.2. Mean Conditions and Interannual Variability of Selected Drivers

About 38% of FLUXNET sites (≥5 years only) adequately sampled the mean conditions of all selected driver variables, while ~60% of sites adequately sampled the interannual variability of all selected driver variables (Figure 6). The percentages of representative sites increased slightly to ~40% (mean condition) and ~65% (interannual variability) when counting only sites with 10+ years of measurements (Figures 6e–6g). Surprisingly, about 41% and 31% and of the long-record sites (≥15 years) still underrepresented the mean conditions and interannual variability in at least one of the driver variables. Air temperature, diurnal temperature range, and potential evapotranspiration were the variables mostly underrepresented by the measurement periods in terms of both their mean conditions and interannual variability, while NDVI was underrepresented mostly in terms of its mean conditions (Figure 6a). To a large extent, such underrepresented patterns could be attributed to the aforementioned trends and/or breakpoints existing in the time series of driver variables. As most tower sites operated in consecutive years, the measurement periods tended to sample similar and mostly higher-than-average conditions in these underrepresented driver variables. From a sampling perspective, it was crucial whether the measurement periods were able to see a few exceptional years with below-average driver conditions (e.g., cool years in Figures 3e–3g).

Similarly, the occurrence of breakpoints could also bias the mean conditions and interannual variability observed in the measurement periods. For driver variables where breakpoints occurred mostly before the measurement periods (e.g., diurnal temperature range), the measured years might not experience some of the far-end but crucial driver conditions (Figure S5). Thus, a substantial portion of sites underrepresented either mean conditions (22%) or interannual variability (16%) of diurnal temperature range (Figure 6a). On the other hand, NDVI had breakpoints occurring both within (e.g., 2008–2010 in North America) and before the measurement periods (e.g., 1986–1989 in Europe and 1994–1995 in Asia). NDVI tended to be less biased in areas where the breakpoints occurred within the measurement periods (Figures 6b–6g). Overall, FLUXNET still undersampled driver conditions in the early years of the baseline period, especially the mean conditions. Such undersampling was especially critical for variables that possessed significant trends and/or had breakpoints outside the measurement periods. Most underrepresented sites were located in the 30°N–40°N regions of North America and Asia and the tropical regions in South America (Figures 6b–6d).

3.3. Multivariate Similarity of Driver Conditions

FLUXNET is unevenly representative in terms of actual measurement lengths and also the extrapolation potentials in the time domain at each tower site. Our analyses suggest that at least 7–10 years of measurements are required to adequately sample the baseline multivariate driver conditions at most tower sites (Figure 7). Adopting the criteria of the maximum similarity of the least similar years from sites with 15+ years of measurements (Figure 7, leftmost dashed lines), sites with 7+ years of measurements generally had above-criteria similarity in most of the years without measurements (i.e., areas to the left of dashed lines). That means, by combining the years with flux measurements and years without measurements but having above-criteria similarity, sites with 7+ years of measurements generally represented driver conditions in around 28–30 years of the baseline, as sufficiently as those long-record sites. On the other hand, sites that had shorter measurement periods tended to have lower similarity in most of the years without flux measurements. Also, we found high site-to-site variation in the similarity metrics. Certain sites with 4–5 years of measurements had high similarity in up to ~20 of the years without flux measurements. Such similarity levels were comparable to some long-record sites (Figure 7). The site-to-site variation reflected partly the highly variable nature of interannual variability in the driver conditions across sites (i.e., some sites may experience higher interannual variability in the driver conditions than others). Additionally, it revealed the importance of sampling randomness (i.e., driver conditions in the years actually sampled) in determining the measurement representativeness. Thus, the justification of temporal representativeness should not rely solely on the lengths of measurements. Site-specific considerations (e.g., trend, breakpoint, and interannual variability in driver variables) should be taken whenever possible.

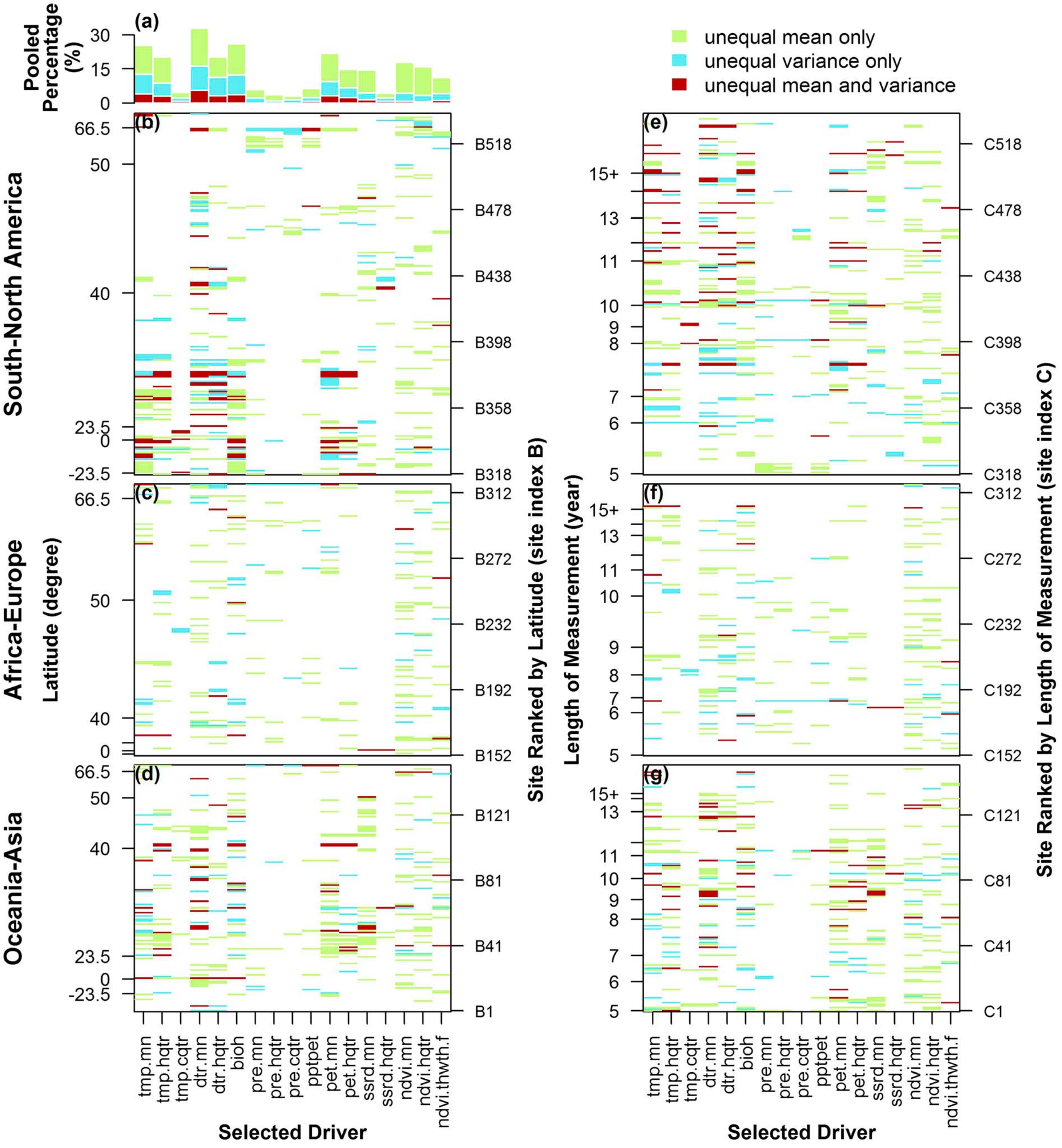

Figure 6. Comparison of mean and variance of selected drivers between the years with flux measurements and baseline period (1982–2013) at FLUXNET tower sites. The figures are arranged as follows. (a) The summed percentages of sites with unequal means and/or variance for each driver. The tower sites are grouped into the three main regions: (b and e) South to North America, (c and f) Africa to Europe, and (d and g) Oceania to Asia. In Figures 6b–6d, the sites are ordered vertically by latitude and then equally spaced ranging from the most southern (bottom) to the most northern one (top). In Figures 6e–6g, the sites are ordered vertically by length of tower measurements and then equally spaced ranging from the shortest (bottom) to the longest one (top). For readers interested in identifying the sites in each panel, a full list of the site index is provided in Table S1 (site indexes B and C). A list of driver abbreviations (*x* axis) can be found in Table 1. Each thin colored rectangle in Figures 6b–6g denotes the test results (i.e., unequal mean and/or variance) from one selected driver at one site. The brown colors indicate sites and drivers that have both unequal means and variance ($p < 0.05$) between the measured years and long-term baseline. The green and blue colors indicate sites and drivers that have unequal means and unequal variance, respectively. Sites with less than 5 years of measurements are not analyzed, and in total, 536 sites are presented here.

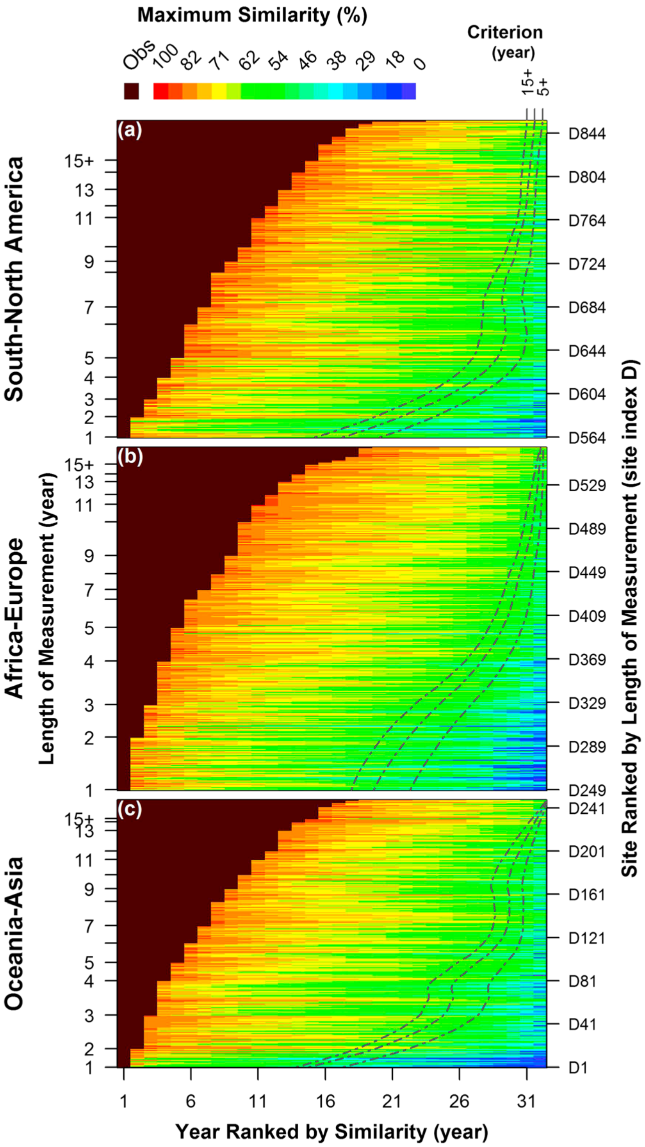

Figure 7. Multivariate similarity metrics between the years with and without flux measurements at FLUXNET tower sites. The tower sites are grouped into the three main regions: (a) South to North America, (b) Africa to Europe, and (c) Oceania to Asia. Within each region, the sites are ordered vertically by length of tower measurements and then equally spaced ranging from the shortest (bottom) to the longest one (top). The corresponding length of measurements and site index were labeled on the left and right *y* axes, respectively. For readers interested in identifying the sites in each panel, a full list of the site index is provided in Table S1 (site index D). At each site, similarity is calculated from the multivariate analysis of drivers across years. Each thin colored rectangle denotes the similarity from one site year. For each year without flux measurements, the maximum similarity is derived based on its cross-year similarity to any of the years with flux measurements. The years are ordered based on the maximum similarity and displayed subsequently (*x* axis) starting from the years with actual measurements (left, dark brown color (Obs)) and the years without flux measurements but most similar (red color) to the years least similar (right, blue color). The dashed lines show the smoothing contours at which the similarity is equivalent to the criteria defined by the averages of maximum similarity in the least similar year from sites with 5+, 10+, and 15+ years of measurements. The areas to the left of the dashed lines imply the number of site years that are adequately represented by the flux measurement periods.

3.4. Applications and Implications

Our study provided an initial exploratory analysis on the representativeness of FLUXNET in the time domain (esp., year-to-year variation). Despite the importance of temporal representativeness, this subject was less addressed in earlier studies because of the initial focus on maximizing the spatial representativeness of flux tower networks [*Hargrove et al.*, 2003; *Papale et al.*, 2015; *Sulkava et al.*, 2011; *Sundareshwar et al.*, 2007].

Previously, only a small number of sites had sufficiently long data (e.g., 10+ years), which further limited the feasibility of conducting such analyses. A few available modeling studies showed that the uncertainties in extrapolation were usually lower in time (e.g., among years) than in space (e.g., among sites) [*Jung et al.*, 2011; *Papale et al.*, 2015]. Our analyses emphasize the potentials of temporal extrapolation for relatively short and consecutive years, also because driver conditions tend to be similar across consecutive years [*Papale et al.*, 2015]. Yet we argue that extra caution should be taken on how well extrapolation works further back in time (e.g., 1980s). As we point out earlier, most of these modeling studies rely on the available data of fluxes for validation. Thus, the evaluation of extrapolation potentials is inherently constrained by data availability in time. Given that the flux measurements did not experience the constellations of drivers in those earlier years, an information transfer from spatial or seasonal variability must occur, which typically assumes that the underlying functional relationship between drivers and responses remains stationary. In addition, the secular trend in CO_2 concentration that may not be fully sampled by FLUXNET measurements certainly poses extra uncertainties on model extrapolation. Thus, we argue that subtle changes such as trends identified with machine-learning model extrapolation should be treated with caution.

Further research is needed to examine the spatiotemporal interchangeability in model extrapolation. Most previous machine-learning upscaling studies assumed certain interchangeability between spatial and temporal extrapolations [*Beer et al.*, 2010; *Jung et al.*, 2011; *Papale et al.*, 2015]. For example, model training was often carried out by pooling data from multiple sites and years of similar ecosystem types (e.g., climate and plant functional groups). In a sense, the cross-site difference in the driver conditions provides additional information that enables model extrapolation to years with different and unobserved responses (e.g., a cool and unseen year at a typically warm site can be predicted by using the data/model from a typically cool site with otherwise similar state-space combinations). Such assumption seemed inevitable in earlier studies because only a few sites had long-enough measurements to allow machine-learning approaches to sample all possible driver conditions at a given site. Recently, a few studies attempted to test such assumption of interchangeability [e.g., *Biederman et al.*, 2016]. It is beyond our scope to justify such interchangeability. However, we stress that different types of model are likely to have different capabilities in extracting the information from data and, thus, enabling the model extrapolation. To a large extent, the wide spectrum of model structures and selected drivers in previous studies has implicated the challenges and uncertainties lying within the spatiotemporal extrapolation [e.g., *Beer et al.*, 2010; *Jung et al.*, 2009, 2011; *Papale and Valentini*, 2003; *Tramontana et al.*, 2015, 2016; *Xiao et al.*, 2008; *Yang et al.*, 2007, 2006]. We urge future studies to target this topic, especially when more sites and longer records of data become available. In sum, we admit that our analysis framework may be adjusted depending on the extent to which the information of cross-site differences could be used in replacement of that of cross-year differences. In that case, one may consider to revise our current evaluation framework and to carry out the analyses on pooled data from multiple sites and years of similar ecosystem types.

Our analyses were confined by data availability of gridded drivers, which may not be comprehensive in all possible aspects. We advocate that future syntheses could use our findings as a primer and tailor their evaluation frameworks accordingly based on specific research questions. Certain limitations and implications of our study are discussed as follows.

First, the selection of baseline periods largely depends on the scope of research and also the availability of potential drivers. While we confine the current analysis within the last few decades where remote sensing data are available, one could certainly extend the baseline further to earlier periods (e.g., 1900s and 1950s) if only climatological drivers are considered or other alternative data sources of vegetation or soil properties become available. Similarly, the evaluation could be carried out to future periods (e.g., 2020s and 2050s) if the projection of future climatological conditions is available (e.g., from Coupled Model Intercomparison Project Phase 5). We also emphasize that the evaluation of temporal representativeness needs to be interpreted in the context of target baseline. Several drivers have showed substantial trends in our 32 year period. If one extends the target baseline to an even longer period (e.g., 100 years), then it is possible that the changes of driver conditions may exceed the ranges of interannual variability observed by the flux measurement periods (at least in certain times and regions).

Second, the robustness and uncertainties of model extrapolation largely depend on the selected modeling approaches, explanatory variables, and also the interested fluxes [*Papale et al.*, 2015]. Thus, the selection of driver variables and their relative weights (assumed equal and independent in the study) may vary among studies and models [*Sulkava et al.*, 2011; *Tramontana et al.*, 2016]. Our current analyses were not tailored to target one specific type of fluxes, but the selection of drivers based on previous studies implicates that our conclusions should hold, in general, for CO_2, water vapor, and energy fluxes. A few studies proposed the use of joint products of multiple drivers as model predictors (e.g., radiation × vegetation index and soil water availability) [*Jung et al.*, 2011; *Tramontana et al.*, 2016], which may be introduced when necessary. Similarly, if one would extend the evaluation to other fluxes (e.g., methane fluxes), then other currently unaccounted potential drivers may need to be considered (e.g., water table and salinity) [e.g., *Ouyang et al.*, 2014; *Zhang et al.*, 2015].

Third, there are potential uncertainties in our analyses due to the mismatch of source areas between the flux towers and gridded drivers (i.e., assumption 2 in section 2.1). While we were aware and cautious about such mismatch, it was challenging to eliminate the uncertainties because both the network-wise tower footprint information and fine-resolution grid data that covered our target baseline were still unavailable. *Chen et al.* [2011] conducted a cross-site evaluation on the spatial representativeness of 12 Canadian flux tower sites and assessed the potential bias of flux footprints in representing the surrounding areas (e.g., in terms of replicating NDVI). They found that most of the tower sites well represented the surrounding ecosystems up to an area of ~10 km^2. Even in a few worst cases, the bias of NDVI between tower footprints and surrounding areas (e.g., 3 km radius) was still within ±10%. In a recent upscaling study, *Xu et al.* [2017] showed that the tower footprint bias may lead to up to ±27% of difference in the expected mean fluxes when upscaling tower observation to a 20×20 km^2 target grid. The magnitudes of footprint bias could vary from site to site and largely depend on the variability of flux footprints and also the surface heterogeneity in the surrounding areas. *Vuichard and Papale* [2015] conducted an extensive test across 250 FLUXNET tower sites and tested the robustness of downscaling meteorological variables (e.g., air temperature, incoming shortwave radiation, and precipitation) from gridded ERA-Interim reanalysis data to the site level. In their study, they adopted linear regression for downscaling and argued that such linear relation was generally sufficient for most tower sites. Yet they also stressed that caution need to be taken for sites that in the mountain or coastal areas, where the gridded data may vary the most from tower measurements. In sum, we argue that a certain portion of the potential bias from the mismatch in source areas should be relatively constant in time (i.e., systematic bias). Such systematic bias would have only minor effects on our results, as most of the analyses were focused on the variability of time series at each site (i.e., eliminating the potentially biased mean). However, extra caution should be exercised for sites with heterogeneous surrounding areas and/or relatively varying flux footprints. In those cases, a more sophisticated upscaling/downscaling approach [e.g., *Xu et al.*, 2017] may be needed to properly evaluate the representativeness of tower footprints.

Fourth, crucial but currently unavailable driver variables such as CO_2 concentration, disturbance, land use history, and management practice should be included when possible [*Jung et al.*, 2011; *Papale et al.*, 2015]. Probably, the time series of GIMMS NDVI may have provided a portion of information to the currently missing variables of disturbance and land use history (e.g., breakpoints and anomalously high/low NDVI). Several studies showed that remote sensing products along with time series-based methods (e.g., BFAST) could be used to detect disturbances [e.g., *Forkel et al.*, 2013; *Verbesselt et al.*, 2010, 2012]. Future syntheses should explore the potentials of utilizing other remote sensing products (e.g., those with a finer spatial, temporal, or spectral resolution) in deriving the missing but important information.

4. Conclusions

The global network FLUXNET is unevenly representative in terms of measurement lengths and, thus, flux information possessed by flux data collected at each tower site. Extra caution should be exercised when attempting to draw inference or make prediction beyond the extent of data coverage. In general, the similarity of driver conditions among years enables the extrapolation of flux information beyond the physically limited measurement periods. By combining years with flux measurements and years without measurements but having similar driver conditions, most decadal sites have shown to adequately represent the driver conditions in most years of the baseline period 1982–2013. To a large extent, the capability of extrapolation is determined by the lengths of actual measurements, i.e., the longer, the more representative. Yet we argue that the justification of temporal representativeness should not rely solely on the lengths of measurements, as our analyses showed substantial variation across sites, regions, and also the drivers selected. There is no single universal criterion (e.g., number of years) for determining the temporal representativeness. Site- or region-specific temporal characteristics in the driver conditions such as interannual variability, trends, and breakpoints often pose additional effects that could dampen or enhance the representativeness of measured periods. Thus, we suggest to perform initial tests on the temporal representativeness of study sites before conducting cross-site syntheses, especially for sites having only short measurement periods (e.g., $<$5–7 years).

We also strongly advocate continuing the data sharing, standardization, and harmonization in an open and timely manner within FLUXNET. This global network has grown rapidly since the La Thuile synthesis in 2007. However, the fraction of sites sharing data did not grow accordingly. By the time of writing, it is estimated that less than ~50% of registered sites and ~35% of collected data (i.e., site years) will participate in the ongoing FLUXNET2015/2017 data synthesis. We have emphasized the potentials and challenges of FLUXNET in providing flux information all of the time, and this goal can be further approached with collaborations among the teams of local towers, regional, and global networks.

Acknowledgments
The authors declare no conflict of interest. This work used eddy covariance data acquired and shared by the FLUXNET community, including but not limiting to these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The FLUXNET eddy covariance data processing and harmonization were carried out by the ICOS Ecosystem Thematic Center, AmeriFlux Management Project and Fluxdata project of FLUXNET, with the support of CDIAC, and the OzFlux, ChinaFlux, and AsiaFlux offices. We thank Leiming Zhang and Changliang Shao for validating the site information in China. This study is supported by the U.S. Department of Energy's Office of Science (DE-SC0012456). Sebastian Wolf acknowledges support from a Marie Curie International Outgoing Fellowship by the European Commission (grant 300083). We also thank Dario Papale, Gilberto Pastorello, and UC Berkeley Biomet Lab members for their constructive advices. Ankur Desai and two anonymous reviewers provided valuable suggestions for the quality of the study. The data used in this work are available in the following databases: Centre for Environmental Data Archival: CRU TS 3.22 (https://doi.org/10.5285/18BE23F8-D252-482D-8AF9-5D6A2D40990C), European Centre for Medium-Range Weather Forecasts: ERA-Interim (http://apps.ecmwf.int/datasets/data/interim-full-daily), and Global Inventory Monitoring and Modeling System (GIMMS): NDVI3g (https://eco-cast.arc.nasa.gov/data/pub/gimms/). A list of FLUXNET sites used in this study is provided in the supporting information (Table S1).

References

Agarwal, D. A., M. Humphrey, N. F. Beekwilder, K. R. Jackson, M. M. Goode, and C. van Ingen (2010), A data-centered collaboration portal to support global carbon-flux analysis, *Concurrency Comput.: Pract. Exp.*, *22*(17), 2323–2334.

Anav, A., et al. (2015), Spatiotemporal patterns of terrestrial gross primary production: A review, *Rev. Geophys.*, *53*, 785–818, doi:10.1002/2015RG000483.

Anyamba, A., J. Small, C. Tucker, and E. Pak (2014), Thirty-two years of Sahelian zone growing season non-stationary NDVI3g patterns and trends, *Remote Sens.*, *6*(4), 3101–3122.

Ault, T., G. Henebry, K. de Beurs, M. Schwartz, J. Betancourt, and D. Moore (2013), The false spring of 2012, earliest in North American record, *Eos Trans. AGU*, *94*(20), 181–182, doi:10.1002/2013EO200001.

Baldocchi, D. D. (2008), Turner review no. 15."Breathing" of the terrestrial biosphere: Lessons learned from a global network of carbon dioxide flux measurement systems, *Aust. J. Bot.*, *56*(1), 1–26, doi:10.1071/BT07151.

Baldocchi, D. D. (2014), Measuring fluxes of trace gases and energy between ecosystems and the atmosphere—The state and future of the eddy covariance method, *Global Change Biol.*, *20*, 3600–3609, doi:10.1111/gcb.12649.

Bastos, A., S. W. Running, C. Gouveia, and R. M. Trigo (2013), The global NPP dependence on ENSO: La Niña and the extraordinary year of 2011, *J. Geophys. Res. Biogeosci.*, *118*, 1247–1255, doi:10.1002/jgrg.20100.

Beer, C., M. Reichstein, E. Tomelleri, P. Ciais, M. Jung, N. Carvalhais, C. Rödenbeck, M. A. Arain, D. Baldocchi, and G. B. Bonan (2010), Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate, *Science*, *329*(5993), 834–838.

Beringer, J., et al. (2016), An introduction to the Australian and New Zealand flux tower network—OzFlux, *Biogeosciences*, *2016*(21), 5895–5916, doi:10.5194/bg-13-5895-2016.

Biederman, J. A., et al. (2016), Terrestrial carbon balance in a drier world: The effects of water availability in southwestern North America, *Global Change Biol.*, *22*(5), 1867–1879, doi:10.1111/gcb.13222.

Bonan, G. B., P. J. Lawrence, K. W. Oleson, S. Levis, M. Jung, M. Reichstein, D. M. Lawrence, and S. C. Swenson (2011), Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data, *J. Geophys. Res.*, *116*, G02014, doi:10.1029/2010JG001593.

Carvalhais, N., M. Reichstein, G. J. Collatz, M. D. Mahecha, M. Migliavacca, C. S. R. Neigh, E. Tomelleri, A. A. Benali, D. Papale, and J. Seixas (2010), Deciphering the components of regional net ecosystem fluxes following a bottom-up approach for the Iberian Peninsula, *Biogeosciences*, *7*(11), 3707–3729, doi:10.5194/bg-7-3707-2010.

Chen, B., N. C. Coops, D. Fu, H. A. Margolis, B. D. Amiro, A. G. Barr, T. A. Black, M. A. Arain, C. P.-A. Bourque, and L. B. Flanagan (2011), Assessing eddy-covariance flux tower location bias across the Fluxnet-Canada Research Network based on remote sensing and footprint modelling, *Agric. For. Meteorol.*, *151*(1), 87–100.

Chu, H., J. Chen, J. F. Gottgens, A. R. Desai, Z. Ouyang, and S. S. Qian (2016), Response and biophysical regulation of carbon dioxide fluxes to climate variability and anomaly in contrasting ecosystems in northwestern Ohio, USA, *Agric. For. Meteorol.*, *220*, 50–68, doi:10.1016/j.agrformet.2016.01.008.

Ciais, P., et al. (2005), Europe-wide reduction in primary productivity caused by the heat and drought in 2003, *Nature*, *437*(7058), 529–533.

Coumou, D., and S. Rahmstorf (2012), A decade of weather extremes, *Nat. Clim. Change*, *2*(7), 491–496.

de Jong, R., J. Verbesselt, M. E. Schaepman, and S. Bruin (2012), Trend changes in global greening and browning: Contribution of short-term trends to longer-term change, *Global Change Biol.*, *18*(2), 642–655.

de Jong, R., J. Verbesselt, A. Zeileis, and M. Schaepman (2013), Shifts in global vegetation activity trends, *Remote Sens.*, *5*(3), 1117–1133.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*(656), 553–597, doi:10.1002/qj.828.

Forkel, M., N. Carvalhais, J. Verbesselt, M. D. Mahecha, C. S. Neigh, and M. Reichstein (2013), Trend change detection in NDVI time series: Effects of inter-annual variability and methodology, *Remote Sens.*, *5*(5), 2113–2144.

Guanter, L., Y. Zhang, M. Jung, J. Joiner, M. Voigt, J. A. Berry, C. Frankenberg, A. R. Huete, P. Zarco-Tejada, and J.-E. Lee (2014), Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence, *Proc. Natl. Acad. Sci. U.S.A.*, *111*(14), E1327–E1333.

Guay, K. C., P. S. A. Beck, L. T. Berner, S. J. Goetz, A. Baccini, and W. Buermann (2014), Vegetation productivity patterns at high northern latitudes: A multi-sensor satellite data assessment, *Global Change Biol.*, *20*(10), 3147–3158, doi:10.1111/gcb.12647.

Hargrove, W., and F. Hoffman (2004), Potential of multivariate quantitative methods for delineation and visualization of ecoregions, *Environ. Manage.*, *34*, S39–S60, doi:10.1007/s00267-003-1084-0.

Hargrove, W. W., F. M. Hoffman, and B. E. Law (2003), New analysis reveals representativeness of the AmeriFlux network, *Eos Trans. AGU*, *84*(48), 529–535, doi:10.1029/2003EO480001.

Harris, I., and P. D. Jones (2014), CRU TS3.22: Climatic Research Unit (CRU) time-series (TS) version 3.22 of high resolution gridded data of month-by-month variation in climate (Jan. 1901–Dec. 2013), produced by NCAS British Atmospheric Data Centre, doi:10.5285/18BE23F8-D252-482D-8AF9-5D6A2D40990C.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister (2014), Updated high-resolution grids of monthly climatic observations—The CRU TS3.10 dataset, *Int. J. Climatol.*, *34*(3), 623–642, doi:10.1002/joc.3711.

Haverd, V., M. R. Raupach, P. R. Briggs, J. G. Canadell, S. J. Davis, R. M. Law, C. P. Meyer, G. P. Peters, C. Pickett-Heaps, and B. Sherman (2013), The Australian terrestrial carbon budget, *Biogeosciences*, *10*(2), 851–869, doi:10.5194/bg-10-851-2013.

Holben, B. N. (1986), Characteristics of maximum-value composite images from temporal AVHRR data, *Int. J. Remote Sens.*, *7*(11), 1417–1434.

Hollander, M., D. A. Wolfe, and E. Chicken (2014), *Nonparametric Statistical Methods*, John Wiley, Hoboken, N. J.

Jiménez, C., C. Prigent, B. Mueller, S. Seneviratne, M. McCabe, E. Wood, W. Rossow, G. Balsamo, A. Betts, and P. Dirmeyer (2011), Global intercomparison of 12 land surface heat flux estimates, *J. Geophys. Res.*, *116*, D02102, doi:10.1029/2010JD014545.

Jung, M., M. Reichstein, and A. Bondeau (2009), Towards global empirical upscaling of FLUXNET eddy covariance observations: Validation of a model tree ensemble approach using a biosphere model, *Biogeosciences*, *6*(10), 2001–2013.

Jung, M., et al. (2011), Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, *J. Geophys. Res.*, *116*, G00J07, doi:10.1029/2010JG001566.

Karl, T. R., et al. (2012), U.S. temperature and drought: Recent anomalies and trends, *Eos Trans. AGU*, *93*(47), 473, doi:10.1029/2012EO470001.

Kondo, M., K. Ichii, H. Takagi, and M. Sasakawa (2015), Comparison of the data-driven top-down and bottom-up global terrestrial CO_2 exchanges: GOSAT CO_2 inversion and empirical eddy flux upscaling, *J. Geophys. Res. Biogeosci.*, *120*, 1226–1245, doi:10.1002/2014JG002866.

Kumar, J., F. M. Hoffman, W. W. Hargrove, and N. Collier (2016), Understanding the representativeness of FLUXNET for upscaling carbon flux from eddy covariance measurements, *Earth Syst. Sci. Data Discuss.*, *2016*, 1–25, doi:10.5194/essd-2016-36.

Levene, H. (1960), Robust tests for equality of variances, in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, edited by I. Olkin, pp. 278–292, Stanford Univ. Press, Stanford, Calif.

Mallya, G., L. Zhao, X. Song, D. Niyogi, and R. Govindaraju (2013), 2012 Midwest drought in the United States, *J. Hydrol. Eng.*, *18*(7), 737–745.

Nappo, C., J. Caneill, R. Furman, F. Gifford, J. Kaimal, M. Kramer, T. Lockhart, M. Pendergast, R. Pielke, and D. Randerson (1982), Workshop on the representativeness of meteorological observations, June 1981, Boulder, Colo, *Bull. Am. Meteorol. Soc.*, *63*(7), 761–764.

Neigh, C. S., C. J. Tucker, and J. R. Townshend (2008), North American vegetation dynamics observed with multi-resolution satellite data, *Remote Sens. Environ.*, *112*(4), 1749–1772.

Nemani, R. R., C. D. Keeling, H. Hashimoto, W. M. Jolly, S. C. Piper, C. J. Tucker, R. B. Myneni, and S. W. Running (2003), Climate-driven increases in global terrestrial net primary production from 1982 to 1999, *Science*, *300*(5625), 1560–1563.

Ouyang, Z., R. Becker, W. Shaver, and J. Chen (2014), Evaluating the sensitivity of wetlands to climate change with remote sensing techniques, *Hydrol. Process.*, *28*(4), 1703–1712.

Papale, D., and A. Valentini (2003), A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization, *Global Change Biol.*, *9*(4), 525–535.

Papale, D., D. Agarwal, D. Baldocchi, R. Cook, J. Fisher, and C. van Ingen (2012), Database maintenance, data sharing policy, collaboration, in *Eddy Covariance*, edited by M. Aubinet, T. Vesala and D. Papale, pp. 399–424, Springer, Netherlands, doi:10.1007/978-94-007-2351-1_17.

Papale, D., et al. (2015), Effect of spatial sampling from European flux towers for estimating carbon and water fluxes with artificial neural networks, *J. Geophys. Res. Biogeosci.*, *120*, 1941–1957, doi:10.1002/2015JG002997.

Parazoo, N. C., E. Barnes, J. Worden, A. B. Harper, K. B. Bowman, C. Frankenberg, S. Wolf, M. Litvak, and T. F. Keenan (2015), Influence of ENSO and the NAO on terrestrial carbon uptake in the Texas-northern Mexico region, *Global Biogeochem. Cycles*, *29*, 1247–1265, doi:10.1002/2015GB005125.

Pastorello, G. Z., D. Papale, H. Chu, C. Trotta, D. A. Agarwal, E. Canfora, D. D. Baldocchi, and M. S. Torn (2017), The FLUXNET2015 dataset: The longest record of global carbon, water, and energy fluxes is updated, *Eos Trans. AGU*, in press.

Pinzon, J., and C. Tucker (2014), A non-stationary 1981–2012 AVHRR NDVI3g time series, *Remote Sens.*, *6*(8), 6929–6960.

R Development Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Schimel, D., W. Hargrove, F. Hoffman, and J. MacMahon (2007), NEON: A hierarchically designed national ecological network, *Front. Ecol. Environ.*, *5*(2), 59–59.

Schimel, D., R. Pavlick, J. B. Fisher, G. P. Asner, S. Saatchi, P. Townsend, C. Miller, C. Frankenberg, K. Hibbard, and P. Cox (2015), Observing terrestrial ecosystems and the carbon cycle from space, *Global Change Biol.*, *21*(5), 1762–1776.

Schwalm, C. R., C. A. Williams, K. Schaefer, D. Baldocchi, T. A. Black, A. H. Goldstein, B. E. Law, W. C. Oechel, K. T. Paw U, and R. L. Scott (2012), Reduction in carbon uptake during turn of the century drought in western North America, *Nat. Geosci.*, *5*(8), 551–556.

Sulkava, M., S. Luyssaert, S. Zaehle, and D. Papale (2011), Assessing and improving the representativeness of monitoring networks: The European flux tower network example, *J. Geophys. Res.*, *116*, G00J04, doi:10.1029/2010JG001562.

Sundareshwar, P. V., et al. (2007), Environmental monitoring network for India, *Science*, *316*(5822), 204–205, doi:10.1126/science.1137417.

Tramontana, G., K. Ichii, G. Camps-Valls, E. Tomelleri, and D. Papale (2015), Uncertainty analysis of gross primary production upscaling using random forests, remote sensing and eddy covariance data, *Remote Sens. Environ.*, *168*, 360–373.

Tramontana, G., et al. (2016), Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, *Biogeosciences*, *2016*(13), 4291–4313, doi:10.5194/bg-13-4291-2016.

Verbesselt, J., R. Hyndman, G. Newnham, and D. Culvenor (2010), Detecting trend and seasonal changes in satellite image time series, *Remote Sens. Environ.*, *114*(1), 106–115.

Verbesselt, J., A. Zeileis, and M. Herold (2012), Near real-time disturbance detection using satellite image time series, *Remote Sens. Environ.*, *123*, 98–108.

Vuichard, N., and D. Papale (2015), Filling the gaps in meteorological continuous data measured at FLUXNET sites with ERA-Interim reanalysis, *Earth Syst. Sci. Data*, *7*(2), 157–171, doi:10.5194/essd-7-157-2015.

Wilks, D. S. (2011), *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, Calif.

Wolf, S., W. Eugster, C. Ammann, M. Häni, S. Zielis, R. Hiller, J. Stieger, D. Imer, L. Merbold, and N. Buchmann (2013), Contrasting response of grassland versus forest carbon and water fluxes to spring drought in Switzerland, *Environ. Res. Lett.*, *8*(3), 035007, doi:10.1088/1748-9326/9/8/089501.

Wolf, S., et al. (2016), Warm spring reduced carbon cycle impact of the 2012 US summer drought, *Proc. Natl. Acad. Sci. U.S.A.*, *130*(21), 5880–5885, doi:10.1073/pnas.1519620113.

Wuebbles, D. J., K. Kunkel, M. Wehner, and Z. Zobel (2014), Severe weather in United States under a changing climate, *Eos Trans. AGU*, *95*(18), 149–150, doi:10.1002/2014eo180001.

Xiao, J., et al. (2008), Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data, *Agric. For. Meteorol.*, *148*(11), 1827–1847.

Xu, K., S. Metzger, and A. R. Desai (2017), Upscaling tower-observed turbulent exchange at fine spatio-temporal resolution using environmental response functions, *Agric. For. Meteorol.*, *232*, 10–22, doi:10.1016/j.agrformet.2016.07.019.

Yang, F., M. A. White, A. R. Michaelis, K. Ichii, H. Hashimoto, P. Votava, A. Zhu, and R. R. Nemani (2006), Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine, *IEEE Trans. Geosci. Remote Sens.*, *44*(11), 3452–3461.

Yang, F., K. Ichii, M. A. White, H. Hashimoto, A. R. Michaelis, P. Votava, A.-X. Zhu, A. Huete, S. W. Running, and R. R. Nemani (2007), Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach, *Remote Sens. Environ.*, *110*(1), 109–122.

Zeileis, A. (2005), A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals, *Econometric Rev.*, *24*(4), 445–466.

Zhang, T.-T., J.-G. Qi, Y. Gao, Z.-T. Ouyang, S.-L. Zeng, and B. Zhao (2015), Detecting soil salinity with MODIS time series VI data, *Ecol. Indicators*, *52*, 480–489.

Zhu, Z., et al. (2016), Greening of the Earth and its drivers, *Nat. Clim. Change*, *6*, 791–795, doi:10.1038/nclimate3004.

Zscheischler, J., M. Reichstein, S. Harmeling, A. Rammig, E. Tomelleri, and M. D. Mahecha (2014), Extreme events in gross primary production: A characterization across continents, *Biogeosciences*, *11*, 2909–2924.