An Optimal Method for Validating Satellite-derived Land Surface Phenology using in-situ Observations from National Phenology Networks

Yongchang Ye[a], Xiaoyang Zhang[a,*], Yu Shen[a], Jianmin Wang[a], Theresa Crimmins[b], Helfried Scheifinger[c]

[a] Geospatial Sciences Center of Excellence, Department of Geography, South Dakota State University, Brookings, SD 57007, USA

[b] School of Natural Resources and the Environment, University of Arizona, Tucson, AZ 85721, USA

[c]Zentralanstalt für Meteorologie und Geodynamik, 1190 Vienna, Austria

Abstract

Satellite-based land surface phenology (LSP) products play an important role in understanding atmosphere-vegetation carbon and energy exchanges. These products have been widely calculated from various satellite observations from local to global scales. However, the quality and accuracy of LSP products are often poorly quantified due to spatial mismatch between satellite observed pixels and in-situ observations. In the present study, we demonstrate an optimal algorithm leveraging the scalability, consistency, and representativeness of rich in-situ observations from national phenology networks to validate LSP products. Specifically, we demonstrate two approaches for validating the phenological timing of greenup onset in the operational Visible Infrared Imaging Radiometer Suite (VIIRS) LSP product developed at NASA using in-situ observations collected from the Pan European Phenological database (PEP725, 9664 site-years)

and the USA National Phenology Network (USA-NPN, 3144 site-years) spanning 2013–2020. The first approach assumes that in-situ data contain observations of phenological transitions (e.g., leaf-out) that are directly comparable with satellite detections. Accordingly, in-situ data were aggregated using four upscaling methods (mean, median, 30th percentile, and minimum bias) to directly compare with VIIRS LSP. The second approach assumes that species-specific phenological timing in in-situ data is basically impossible to spatially reconcile VIIRS LSP, but phenological events in a local area are driven by the same or very similar weather conditions. Therefore, interannual variations and long-term trends were applied to compare VIIRS LSP with in-situ data. The result shows first that the 30th percentile method is more promising in aggregating in-situ observations than the commonly used mean method. Second, direct comparison indicates that VIIRS greenup onset has a mean absolute difference of 13.9 ± 9.8 days with PEP725 in-situ observations and 12.3 ± 10.9 days with USA-NPN observations in well-selected deciduous forest sites. Third, the interannual comparison reveals that VIIRS greenup onset exhibits the same directions of multi-year anomalies and long-term trends as those of both PEP725 and USA-NPN observations in over 70% of sample sites. These findings improve our understanding of the scale mismatch and sample representativeness of species-specific phenology and the uncertainties of long-term LSP detections from remote sensing data.

Keywords: Phenology; In-situ observations; PEP725; USA-NPN; VIIRS; LSP validation

1. Introduction

Vegetation phenology refers to recurring plant life cycle events and their relationships with environmental conditions (Cleland et al., 2007). Phenology is a critical parameter for calculating photosynthetic activity and carbon sequestration (Richardson et al., 2009), quantifying plant-insect interactions (Senior et al., 2020), characterizing land cover/use changes (Nguyen et al., 2020), and forecasting crop yield (Sakamoto et al., 2013). In addition, long-term records of vegetation phenology are robust measures of climate change and ecosystem dynamics (Menzel et al., 2006; Zheng et al., 2016).

Vegetation phenology has been widely derived from satellite data over the past few decades. Satellite-derived phenology measures quantify the timing and magnitude of seasonal dynamics in vegetation communities over vegetated land surfaces, which is referred to as land surface phenology (LSP, de Beurs and Henebry, 2004). A large number of LSP products have been produced from local to global scales, as satellites are able to provide consistent wall-to-wall observations with a frequency of one to 16 days across the globe (Justice et al., 2013; Woodcock et al., 2008). The longest running LSP products are derived using the Advanced Very High Resolution Radiometer (AVHRR) data product, and including the start and end of vegetation growing season in Northern Hemisphere from 1982–2008 (de Jong et al., 2011) and 1982–2015 (Wu et al., 2021), North America from 1982–2006 (White et al., 2009), Europe 1982–2011 (Garonna et al., 2014), and globally from 1982–2013 (Julien and Sobrino, 2009). Similarly, LSP has been produced from SPOT-VEGETATION data in China from 1999–2013 (Wu et al., 2016), and Multi-temporal Medium Resolution Imaging Spectrometer (MERIS) from 2002 to 2012 (Rodriguez-Galiano et al., 2015).

The most widely used LSP products are the National Aeronautics and Space Administration (NASA) global 500m Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Dynamics Products (MCD12Q2, Ganguly et al., 2010). Unlike other LSP products, the MCD12Q2 is operationally produced every year from 2001 to 2019. Because MODIS sensors are expected to cease operation in the coming years, a NASA global 500 m LSP product is continuously produced from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor onboard the Suomi National Polar-orbiting Partnership (S-NPP) satellite, which launched in November 2011 and shows excellent continuity with MODIS records (Moon et al., 2019). These 500m global LSP products provide a long-term climate data record with scientific quality for various applications and are publicly accessible (https://ladsweb.nascom.nasa.gov/data/).

As with many satellite-derived products, evaluation and validation of LSP detections are critical for exploring reliable climate-ecosystem relationships, such as climate change and land disturbance. As a result, LSP validation is a significant mission in the CEOS (Committee on Earth Observation Satellites) Land Product Validation subgroup, focusing on the development of a validation good practices protocol (https://lpvs.gsfc.nasa.gov/Pheno/Pheno_home.html). However, LSP validation activities are limited because this objective is challenging due to limited datasets that are suitable for validation activities. Land surface phenology is currently evaluated using several independent ground-based datasets: (1) eddy flux tower measurements of Gross Primary Production (Huang et al., 2019; Huete et al., 2008; Wu et al., 2021; Xiao et al., 2019); (2) repeat landscape images collected through the PhenoCam Network (Klosterman et al., 2014; Moon et al., 2019; Richardson et al., 2018; Yan et al., 2019; Zhang et al., 2018a; Zhang et al., 2020b); (3) cryosphere and hydrology network records (White et al., 2009); and (4) observations of plant phenology collected across the landscape (Delbart et al., 2015; Donnelly et al., 2019; Khare et al.,

2019; Xie and Wilson, 2020). Because field-point observations are difficult to match spatially with satellite pixels, they frequently show poor agreement with LSP detections (Kowalski et al., 2020; Schwartz and Hanes, 2010; White et al., 2009).

In-situ phenology datasets best suited for use in LSP validation efforts are those collected through national phenology networks. These datasets offer several strengths for use in LSP validation efforts, including broad geographical and ecological coverage, observations from many species, and observations of seasonal events (phenophases) encompassing the entire growing season, from breaking leaf bud through leaf color change and leaf drop. Several such datasets exist in the Northern Hemisphere, including PlantWatch in Canada, the United States of America National Phenology Network (USA-NPN), the Pan European Phenology Project (PEP725) Network, and the Japan Phenological Eyes Network (PEN).

Ground-based phenology observations available through national-scale networks also present several challenges for use in LSP validation efforts. Observation frequency is variable and sometimes biased toward weekends (Courter et al., 2013), yielding variable precision in estimating when a transition such as leaf-out actually occurred on an individual plant. The period of observation record for a site is also variable: volunteer participants may regularly record observations for many years or for less than one season. Volunteer-contributed data are also subject to mis-identification of species and phenophases (Bison et al., 2019; Feldman et al., 2018; Fuccillo et al., 2015; MacKenzie et al., 2017; Tian et al., 2021). Finally, the spatial and temporal scale of ground-based observations does not match that of LSP products. In-situ phenology observations are collected on individual plants, whereas satellite pixels encompass a mosaic of species and vegetation types (Kowalski et al., 2020; Liang et al., 2011; Zhang et al., 2017). The discrete measurements collected on plants result in sharply defined life cycle events, such as the emergence

of first bloom, first leaf unfolding, and budburst, while LSP reflects transitions within fitted curves of remotely sensed greenness (Zhang et al., 2017). As a result, direct comparison between in-situ observations and satellite LSP detections tend to show large differences in the timing of seasonal transitions (Kowalski et al., 2020; Schwartz and Hanes, 2010; White et al., 2009).

The purpose of this study is to identify an optimal method for validation of satellite-derived LSP using in-situ observations from national phenology networks. Specifically, we investigated algorithms to aggregate species-specific spring phenological events from the PEP725 and USA-NPN datasets for the period 2013–2020. The aggregated in-situ observations were then applied to validate and evaluate the timing of vegetation greenup onset, or start of growing season (SOS), produced in VIIRS global land surface phenology product.

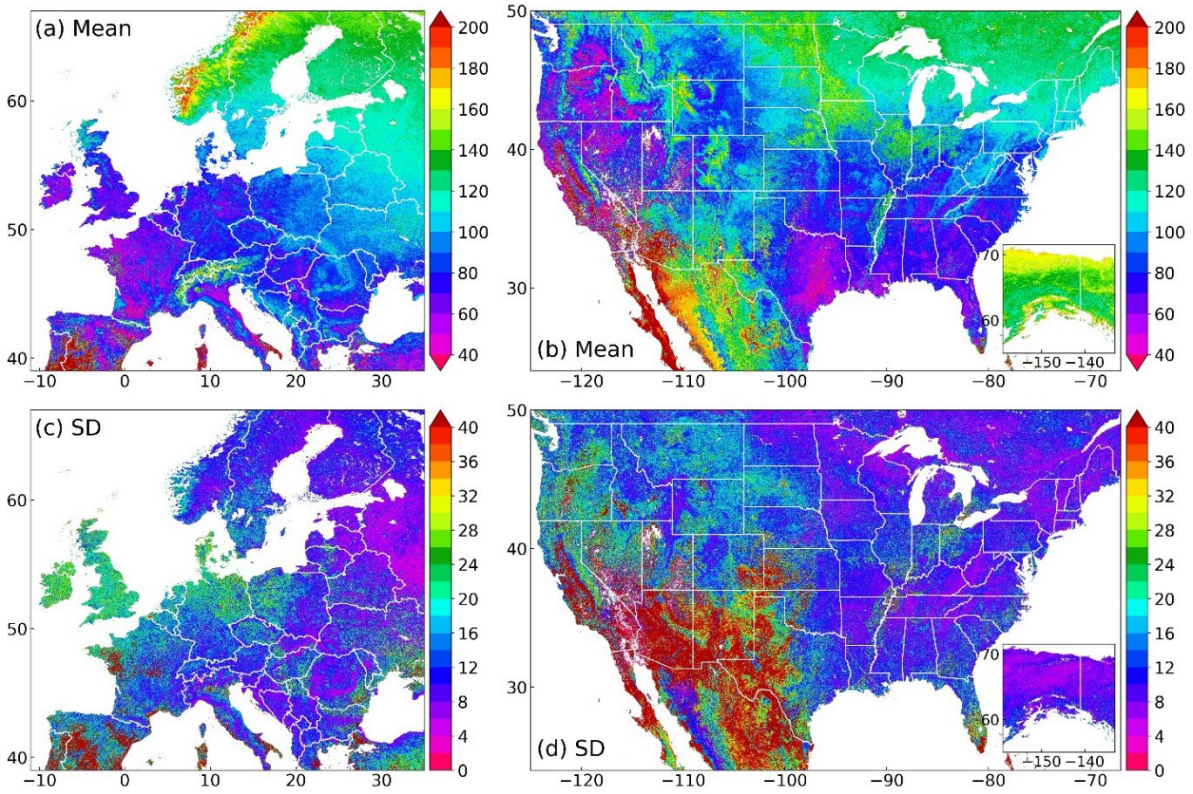
2. Methodology

2.1. VIIRS global land surface phenology product

The VIIRS global LSP product (VNP22Q2, 500 m, v001) is generated on an annual basis using VIIRS observations (Zhang et al., 2020a). It is suitable for characterizing and evaluating interannual-to-decadal scale changes in ecosystems in response to climate and environmental changes. The VIIRS LSP product indicates the timing of phenological transitions over the growing season. These timings include (1) greenup onset (the date of onset of greenness increase) that is equivalent to SOS, (2) date at the mid-points of the greenup phase, (3) maturity onset (the date of onset of greenness maximum), (4) senescence onset (the date of onset of greenness decrease), (5) date at the mid-points of the senescence phase, and (6) dormancy onset (the date of onset of greenness minimum).

The VIIRS LSP timings are derived from seasonal vegetation growing dynamics that are characterized using the daily two-band enhanced vegetation index (EVI2) calculated from VIIRS Nadir Bidirectional Reflectance Distribution Function (BRDF)-Adjusted Reflectance (NBAR). Briefly, the algorithm used to generate this LSP product consists of the following steps (Zhang et al., 2018b): (1) the daily EVI2 time series is composited to 3-day using the maximum value composite of best quality observations within a 3-day window in order to reduce the data size and still retain the fine temporal resolution; (2) a background EVI2 is calculated using the EVI2 values during the corresponding vegetation dormancy phase to remove the effect of snow and other abiotic factors; (3) a set of moving-window filters are used to fill the observation gaps associated with clouds and instrument issues; (4) a Savitzky-Golay filter and a running local median filter are applied to further reduce the noise; (5) the smoothed EVI2 time series are fitted using hybrid piecewise logistic functions; and (6) the phenological transition date is identified using local minimal or maximal rate of change in curvature in the fitted EV2 curve.

The VIIRS LSP product is available at Land Processes Distributed Active Archive Center (LP DAAC, Zhang et al., 2020a). In this study, we collected VIIRS LSP product across Europe and the USA for the period of 2013–2020 and focused on the greenup onset (or SOS) because it is the most important phenological timing during a vegetation growing season (**Fig. 1**).



152

153 **Fig. 1.** Spatial pattern of mean (a-b) and standard deviation (SD, c-d) of VIIRS SOS from 2013–
154 2020 in Europe (a and c) and the USA (b and d). The sub-windows embedded in (b) and (d) are
155 Alaska.

156

157 2.2. Land cover and tree cover data

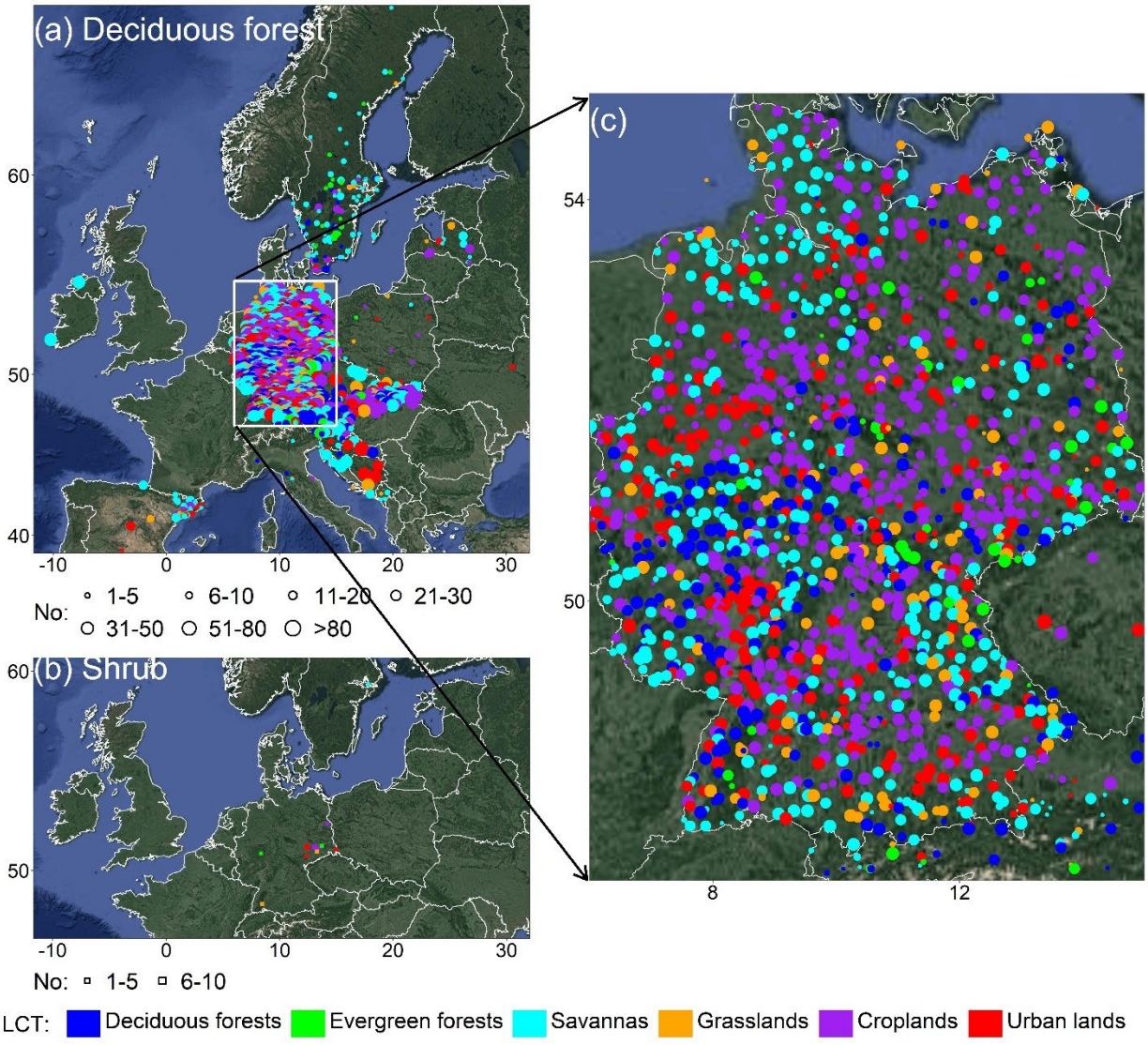
158 Land cover data were obtained from the MODIS land cover product (MCD12Q1, 500 m, v006).
159 The MCD12Q1 product contains land cover types (LCT) classified based on the International
160 Geosphere–Biosphere Programme (IGBP) scheme, which consists of 16 classes (Sulla-Menashe
161 et al., 2019). Tree cover data were extracted from the MODIS vegetation continuous fields product
162 (MOD44B, 250 m, v006). This product provides a continuous and quantitative portrayal of land
163 surface cover with a sub-pixel depiction of tree and non-tree cover using linear regression tree
164 models (DiMiceli et al., 2015). We collected MCD12Q1 and MOD44B products between 2013
165 and 2020, which are also available at LP DAAC.

166 Moreover, a fine resolution (30 m) land cover product in 2020 was also collected
167 (https://zenodo.org/record/4280923#.YUuU3bhKhaR). It is a continuation of a 2015 land cover
168 product that was generated by constructing a local adaptive random forest model using multi-
169 temporal Landsat-8 data from 2014–2016 (Zhang et al., 2021). This new product in 2020 was
170 produced based on the product in 2015 and combined the Landsat and Sentinel-1 data from 2019–
171 2020. The metric-composite method was used to minimize the effects of cloud and shadows
172 (Hansen et al., 2014). Specifically, the time series of individual bands and vegetation indices in
173 the study period was composited into different quantiles. The composite data were used as the
174 model inputs. Thus, this product is an integration of the land cover between 2014 and 2020 rather
175 than the land cover in a specific year.
176
177 2.3. In-situ observations of plant phenology
178 Species-specific in-situ measurements were accessed from national phenology networks in
179 Europe and the United States. The PEP725 Network is a joint effort among 32 European
180 meteorological services and project partners from across Europe (Templ et al., 2018). This project
181 coordinates phenological data collected by volunteers from 1868 to the present, and presently
182 holds nearly 12 million records encompassing 46 growing stages and 265 plant species. The USA-
183 NPN currently offers over 28 M phenological data records across the United States and spanning
184 2009-present. The phenology observations offered by the USA-NPN are contributed by volunteer
185 and professional scientists through the *Nature's No*tebook platform and following standardized
186 protocols (Denny et al., 2014; Rosemartin et al., 2014). Data were downloaded from the two
187 program's websites (www.pep725.eu and www.usanpn.org).

188 To evaluate the greenup onset (or SOS) in the VIIRS LSP product for the period between 2013
189 and 2020, we focused on the phenological timing of "first true leaf" in the PEP725 dataset and six
190 equivalent spring events for different vegetation types in the USA-NPN dataset. The PEP725
191 observations were concentrated in Germany (**Fig. 2**) and the USA-NPN observations were mainly
192 distributed in eastern U.S. and along the west coast of the U.S. (**Fig. 3**). We excluded any records
193 from the PEP725 dataset with a data quality label other than good (0). In the USA-NPN dataset,
194 we excluded records where multiple observers reporting on the same plant on the same day
195 submitted conflicting records of phenophase status and records when the date an observer first
196 reported "yes" to the phenophase (e.g., breaking leaf buds) was not preceded by a report of "no"
197 within the preceding 14 days. Further, we only retained the first date of reported "yes" for a
198 phenophase. After these data cleaning steps, our working data consisted of 69,222 records (88
199 species) from PEP725 and 29,871 (568 species) from USA-NPN (**Table 1**).
200
201 **Table 1.** Spring phenological events as well as number of records and species selected from the
202 PEP725 and USA-NPN between 2013 and 2020.

Data set	Phenological event	Records	Species
PEP725	First true leaf	69222	88
USA-NPN	Breaking leaf buds	21559	311
	Breaking leaf buds (lilac/honeysuckle)	1304	3
	Breaking needle buds (conifers)	948	11
	Breaking needle buds (deciduous)	138	3
	Initial growth (grasses/sedges)	3007	196
	Initial growth (forbs)	2106	34
	Emerging needles (pines)	809	12

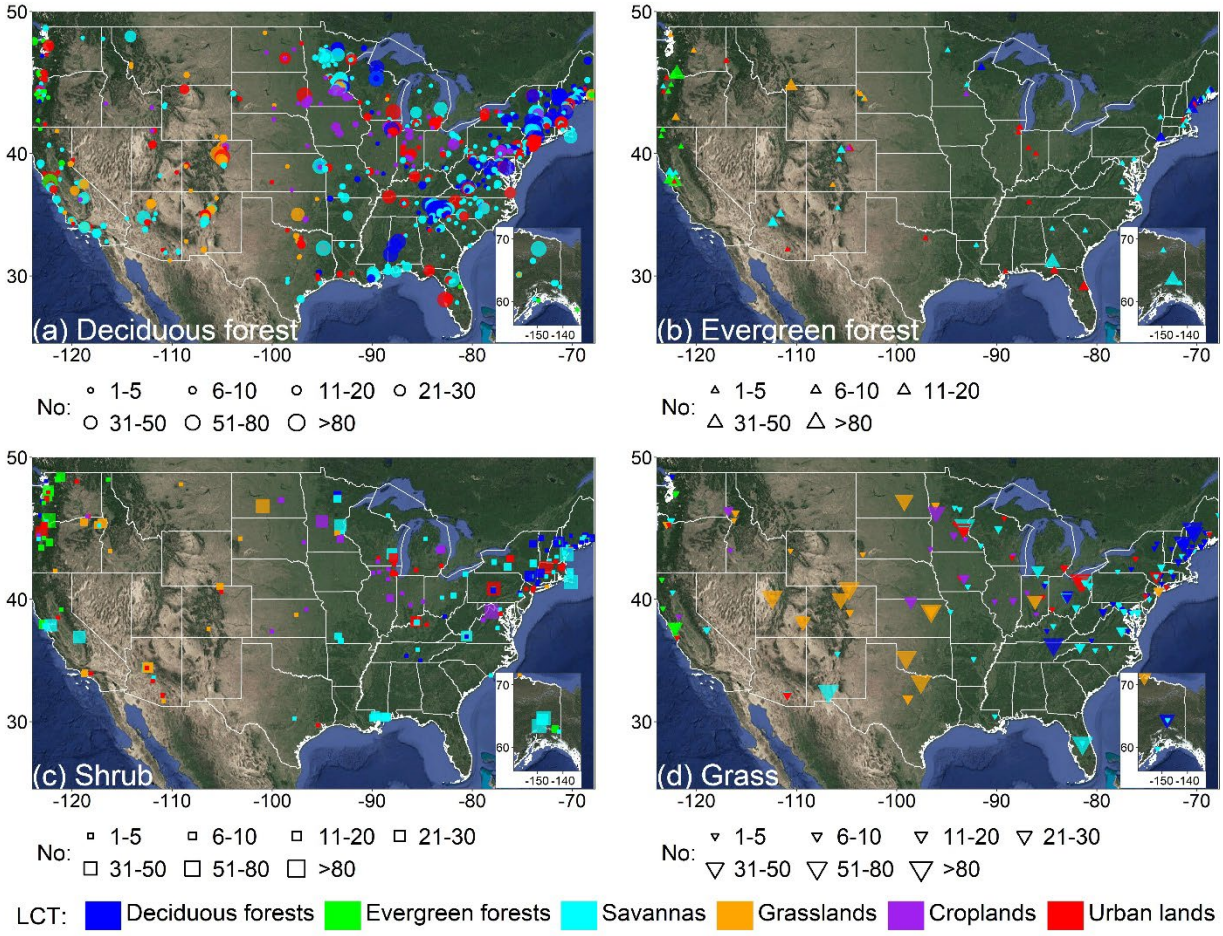
203



204

205 **Fig. 2.** Distribution of in-situ observations used in this analysis from the PEP725 dataset, 2013–
206 2020, including (a) deciduous forest and (b) shrub plant communities (N=1899). The distribution
207 of deciduous forest observations is enlarged presented in (c). Symbol size reflects the number of
208 observations (No). Color indicates the MODIS land cover type (LCT) at the sample location.
209 Background image was acquired from Google Map.

210



211

212 **Fig. 3.** Distribution of in-situ observations used in this analysis from the USA-NPN dataset, 2013–
213 2020, including (a) deciduous forest, (b) evergreen forest, (c) shrub, and (d) grass plant
214 communities (N=1477). Symbol size reflects the number of observations (No). Color indicates
215 the MODIS land cover type (LCT) at the sample location. Background image was acquired from
216 Google Map.

217

218 2.4. Generation of sample sites from in-situ observations and VIIRS LSP detections

219 To address the spatial mismatch between the in-situ phenology observations and the SOS
220 detections in the VIIRS LSP product and reduce the geolocation errors that could be as large as 2
221 km (Tian et al., 2021). we spatially aggregated in-situ observation records to 1.5×1.5 km^2 sample
222 sites, corresponding to 3×3 VIIRS LSP pixels. A sample site was established if it contained at least

one valid in-situ record and five valid VIIRS SOS observations. Next, to reduce the uncertainty, we applied the following filters to the in-situ observations: (1) the timing of spring leaf-out should be no later than 200 in day of year (DOY); (2) the record should have a difference less than 60 days in comparing to VIIRS SOS; and (3) the anomaly of spring leaf-out timing should be less than 30 days within eight years (2013–2020) for an individual plant. These criteria resulted in 1899 and 1477 sample sites from the PEP725 and USA-NPN datasets, respectively. Within a single sample site, the number of point observations varied from 1 to 87 in the PEP725 dataset and 1 to 610 in the USA-NPN dataset over 2013–2020.

2.5. Aggregation of in-situ point observations to sample sites

We aggregated the in-situ observations using four methods: mean, median, 30th percentile, and minimum bias. Specifically, the mean method calculated the mean of all in-situ observed records within an area of 3×3 VIIRS pixels, which is commonly used in the validation of satellite LSP (Donnelly et al., 2019; Xie and Wilson, 2020). The median method selected the median value from all records, which has the advantage of avoiding the effects from extreme values which could exist in in-situ observations collected by different volunteers. The 30th percentile method aggregated the in-situ observations by selecting the value at 30th percentile of all records. This algorithm is based on the assumption that the LSP SOS value in a satellite pixel reflects the timing at which SOS occurs in 30% of area (Zhang et al., 2017), which has been demonstrated in comparing 30m and 500m LSP across various ecosystems (Peng et al., 2021; Zhang et al., 2017). The minimum bias method selected the in-situ observation that had minimum absolute difference with VIIRS SOS. For sample sites with only one point observation, all four methods utilize this observation. For

245 sample sites with two observations, both the median and 30th percentile methods utilize the
246 observation occurring earlier.

247 The in-situ vegetation types in sample sites were also aggregated from individual species in
248 order to evaluate VIIRS SOS quality for various vegetation types. To do this, we first classified
249 each field-observed species into one of the five plant functional types (PFT): deciduous forest,
250 evergreen forest, shrub, grass, and crop. The PFT classification was based on species
251 characteristics retrieved from Wikipedia, Google Images, Missouri Botanical Garden
252 (https://www.missouribotanicalgarden.org/), the Gymnosperm Database
253 (https://www.conifers.org/pi/Abies_balsamea.php), and Global Biodiversity Information Facility
254 (https://www.gbif.org/). The species between shrubs and small trees were classified into either
255 shrub (mature height < 5 m) or tree (mature height ≥ 5 m). Because of the small number of in-
256 situ observations (< 0.1% of all the records), crop observations were excluded from the USA-NPN
257 dataset and evergreen forests, grasses, and crops observations were excluded from the PEP725
258 dataset. This resulted in 55869 (65 species) for deciduous forest and 7151 (12 species) for shrub
259 in the PEP725 dataset and 16205 (147 species) for deciduous forest, 1861 (34 species) for
260 evergreen forest, 4539 (152 species) for shrub, and 4390 (213 species) for grass in the USA-NPN
261 dataset. These PFT data in individual in-situ points were aggregated to sample sites based on the
262 SOS aggregation methods. For the mean aggregation method, the PFT of a sample site represented
263 the type with maximum quantity in all in-situ points within the site area. If there were multiple
264 types with an equal number of observations, the PFT of the sample site was assigned using the
265 following framework: evergreen forest > deciduous forest > shrub > grass. The selection of
266 evergreen forest first rather than deciduous forest was based on the fact that the evergreen forest
267 is likely to begin photosynthesis much earlier than deciduous forest in the spring (Richardson et

al., 2009). For the median, 30th percentile, and minimum bias methods, the PFT of a sample site was the individual plant corresponding to the chosen phenological timing. Thus, the representative PFT in a sample site varied with the four methods of SOS aggregations. This resulted in 9644 site-year samples in the PEP725 dataset and 3144 in the USA-NPN dataset, 2013– 2020 (**Table 2**). A high proportion of site-years were located in MODIS LCT that differed from in-situ PFT (**Table 3**).

The MODIS tree cover (250 m) was aggregated to sample sites by averaging all the best quality values within 6×6 pixel-units. The LCT from MODIS (500 m) and fine resolution (30 m) were first reclassified using a crosswalk method to deciduous forests, evergreen forests, grasslands, croplands, savannas (including shrublands), and urban lands. They were then aggregated by selecting the LCT of maximum quantity within a site area.

Table 2. Number of sample site-years (2013–2020) for each plant functional type (PFT) with four aggregation methods. "-" represents no samples.

Data sets	Methods	Deciduous forest	Evergreen forest	Shrub	Grass
PEP725	Mean	9598	-	66	-
	Median	9374	-	290	-
	30th percentile	9025	-	639	-
	Minimum bias	4443	-	5221	-
USA-NPN	Mean	2148	149	479	368
	Median	1946	148	602	448
	30th percentile	1952	139	620	433
	Minimum bias	2054	157	550	383

Table 3. Number of sample site-years (2013–2020) located at MODIS land cover types.

Data sets	Deciduous forests	Evergreen forests	Savannas	Grasslands	Croplands	Urban lands
PEP725	1328	285	2406	736	3049	1860
USA-NPN	743	149	936	291	211	814

2.6. Validation and evaluation of VIIRS SOS detections

2.6.1. Direct comparison between VIIRS SOS and in-situ observations

The direct comparison assumes that accurate satellite detections are temporal and spatial equivalent to the in-situ observations. Thus, the VIIRS SOS values were directly compared with in-situ observations that were aggregated from four different methods at different LCT and PFT. The analysis was also conducted on different tree cover categories that were stratified as 0–30%, 30–50%, and > 50% based on the MODIS vegetation continuous fields product (MOD44B). To further examine the difference between VIIRS SOS and in-situ observations, we explored the influences: (1) the proportion of different LCT within a sample site that was calculated from 30 m land cover product, and (2) the number of in-situ observations within a sample site. These validation and evaluation analyses were performed using the determination coefficient (R^2), mean absolute difference (MAD), and mean systematic bias (MSB) between in-situ observations and VIIRS SOS.

$$R^2 = \frac{N(\sum SOS_{VIIRS} \times SOS_{in-situ}) - (\sum SOS_{VIIRS})(\sum SOS_{in-situ})}{[N\sum SOS_{VIIRS}^2 - (\sum SOS_{VIIRS})^2][N\sum SOS_{in-situ}^2 - (\sum SOS_{in-situ})^2]} \quad (1)$$

$$MAD = \frac{\sum |SOS_{VIIRS} - SOS_{in-situ}|}{N} \quad (2)$$

$$MSB = \frac{\sum (SOS_{VIIRS} - SOS_{in-situ})}{N} \quad (3)$$

where SOS_{VIIRS} and $SOS_{in\text{-}situ}$ are start of growing season observed from VIIRS data and fields, respectively, and N is the number of sample sites.

2.6.2. Relative comparison of inter-annual variations and long-term trends

We also evaluated VIIRS SOS through comparing inter-annual variations between VIIRS and in-situ observations. This was based on the assumption that inter-annual variations and long-term trends of phenological events should be consistent for most species within a satellite pixel. Although VIIRS pixels (a mixture of multiple plant species and background) is impossible to spatially match with individual plant, phenological events in a local area are driven by the same or similar weather or climate conditions. The differences of the inter-annual and long-term variations between VIIRS and in-situ observations are mainly raised from systematic factors, such as the methods used to retrieve LSP and bias from phenological observers. The comparison of inter-annual variations and long-term trends was carried out as follows. First, the anomaly of a given year was calculated at a single sample site using the following formula:

$$Anomaly_i = SOS_i - \frac{\sum SOS_i}{n} \quad (4)$$

where SOS_i is the VIIRS SOS or in-situ SOS at the i^{th} year, n is the number of years with valid phenological values.

Second, the long-term trend was estimated by calculating the slope of linear regression from VIIRS SOS and in-situ observations during 2013–2020, separately. Although the time series was only eight years and therefore not sufficient for statistical analysis, the trend directions allowed us to reveal if the change of VIIRS and in-situ observations followed the same direction (early or late) in spring events. In this analysis, we excluded sample sites where valid observations were less than

eight years. This resulted in 557 sample sites in Europe and 30 in the USA, as the USA-NPN dataset held observations for far fewer sites during the earlier years of the study period.

3. Results

3.1. Variation of in-situ phenological metrics within a sample site

The phenological metrics calculated for a site using in-situ observations could vary greatly because of the diverse phenological cycles among different species or individual plants of the same species. This variation is illustrated using observations at three sample sites of deciduous forest in the PEP725 and USA-NPN datasets (**Fig. 4**). In the PEP725 example, each in-situ point had only one observation for a species. The in-situ SOS revealed considerable variation among different observations and different years. The smallest differences within a year were 16 days, 20 days, and 20 days for the three selected points, whereas the largest differences were 49 days, 53 days, and 42 days. In the USA-NPN example, each in-situ point could be comprised of observations for several individual plants for one given species. The in-situ SOS in the selected sample sites varied as much as 33 days for the same species, and 46 days among all the observations for a given year. In addition, a PFT within an in-situ point could be located at a different satellite-derived LCT. As demonstrated in the selected sites, the in-situ forest points were surrounded by croplands in the PEP725 sample sites (**Fig. 4**b and c) and mixed with grasslands in the USA-NPN sample site (**Fig. 4**i).

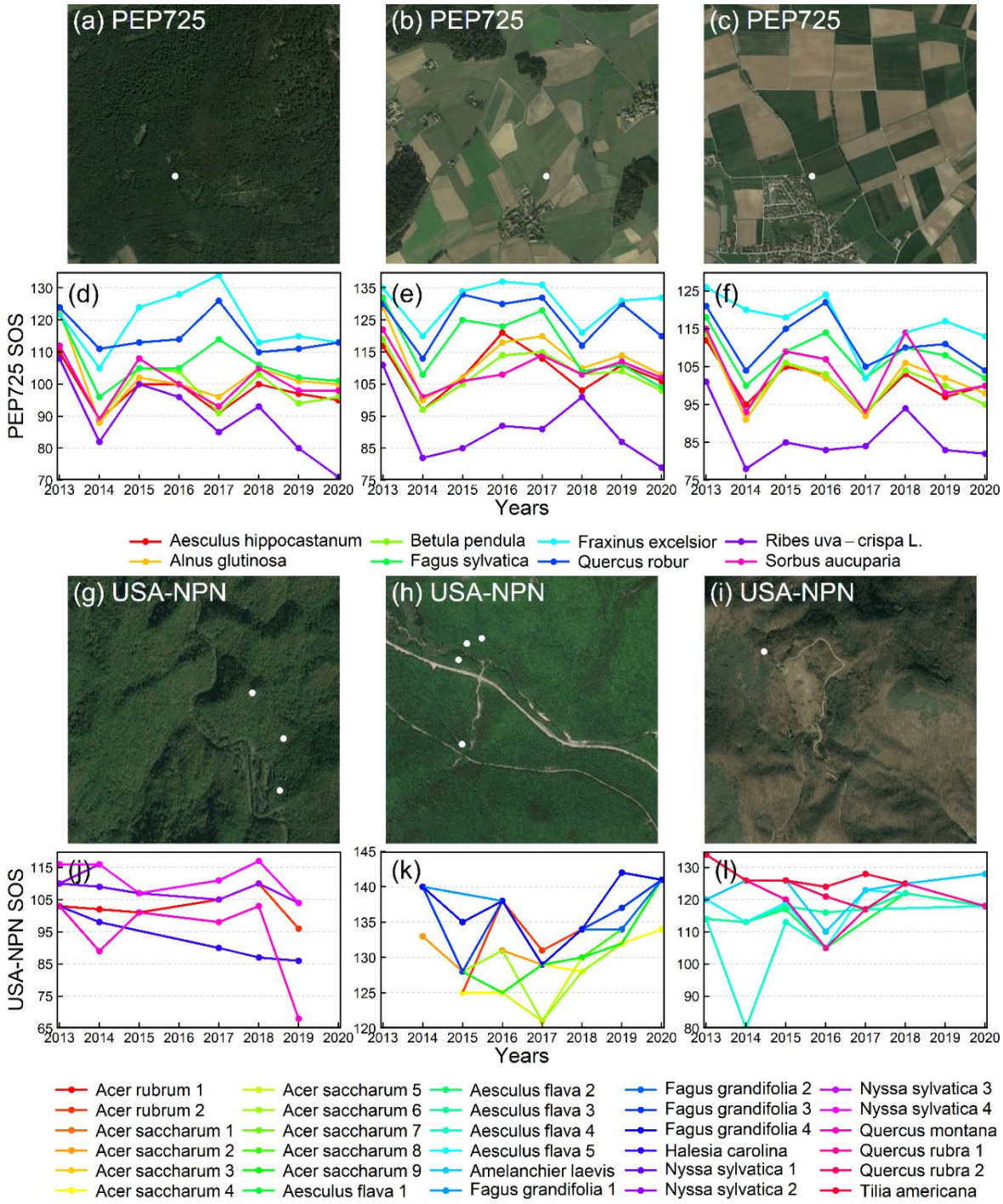


Fig. 4. In-situ phenological metrics for individual plants at three sample sites of deciduous forest in the PEP725 (a-f) and USA-NPN (g-l). (a-c) and (g-i) are the geolocations of in-situ points (white dots) in high resolution images acquired from Google Maps (2.5×1.5 km^2 area). (d-f) and (j-l) are their corresponding SOS observations for all individual plants, 2013–2020.

The range of in-situ SOS increased with the number of observed plants within a sample site (**Fig. 5**). In the PEP725 dataset, where the maximum number of sampled species was 14, the in-situ SOS range increased to ~37 days. This measure dipped slightly at observation numbers 11 and 12, which is likely due to the low proportion of site-years (total < 1%). In the USA-NPN dataset, the in-situ SOS range increased exponentially to ~30 days with observation number from 0 to 30. This measure became diverse with observation number larger than 30 where sampled site-years accounted for only 5.4% of all sites. Although the range could be over 60 days, the overall range tended to slightly increase.

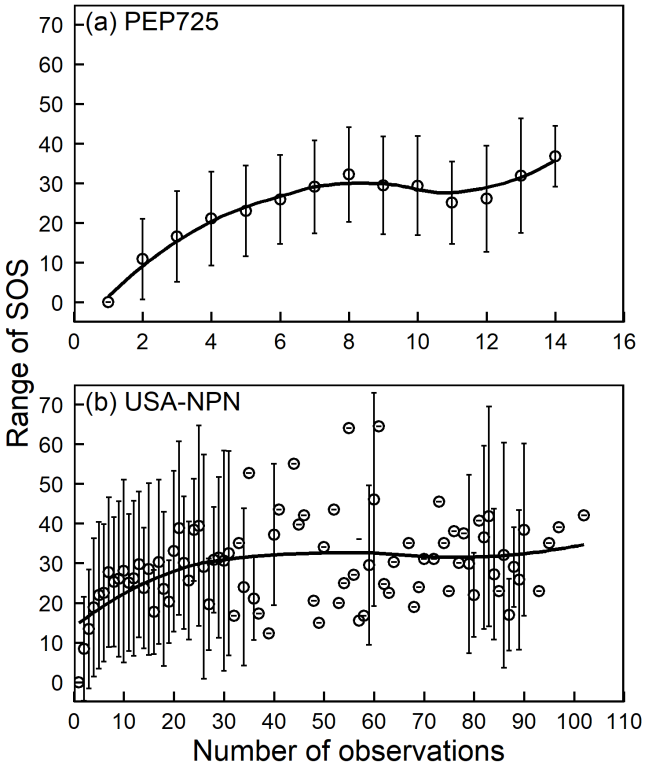


Fig. 5. Dynamics of SOS range among all observed plant individuals within a sampled site-year with the number of in-situ observations. Error bars show the standard deviation of all site-years with the same number of observations. The smoothed line was generated using a Loess regression (a method of local weighted regression).

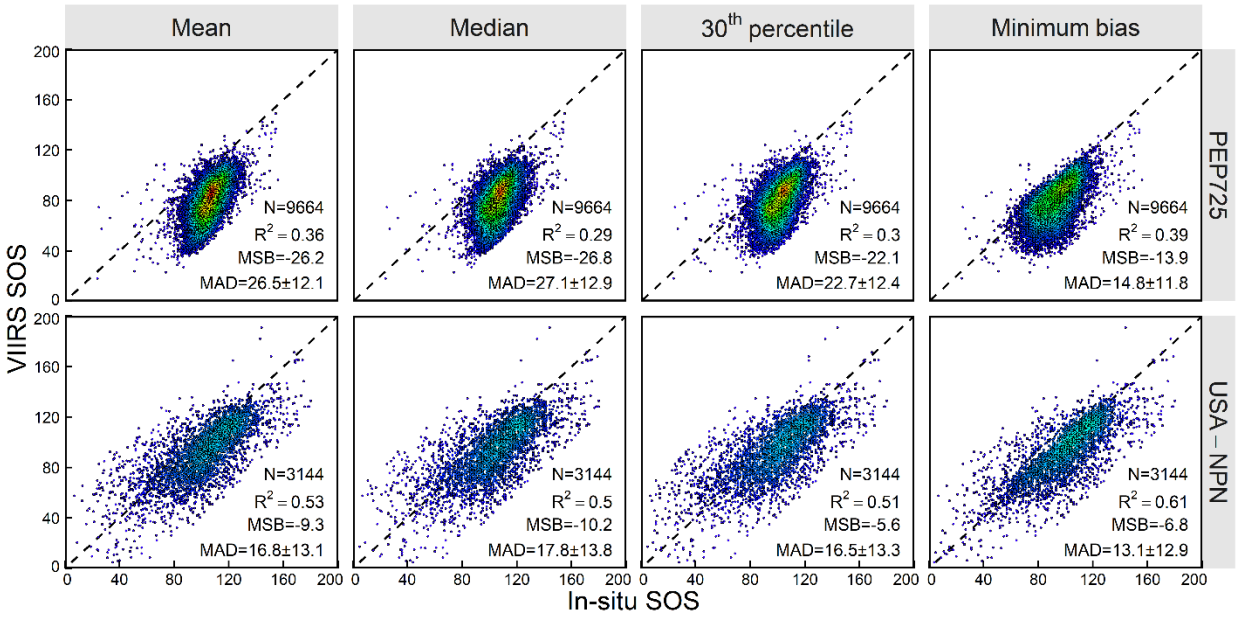
364

365 3.2. Direct comparisons between in-situ observations and VIIRS SOS

366 3.2.1. Difference for all site-year samples

367 Overall, MAD varied from 15–27 days in the PEP725 dataset and from 13–18 days in the
368 USA-NPN dataset (**Fig. 6**). MAD was smallest in the aggregations from the minimum bias method,
369 followed by 30th percentile, mean, and median methods. VIIRS SOS was more closely aligned
370 with in-situ observations in USA-NPN dataset than in PEP725 dataset across all aggregation
371 methods. MAD was similar in the USA-NPN and PEP725 datasets when in-situ values were
372 aggregated from the minimum bias method, however, it was six days and ten days smaller in the
373 USA-NPN dataset when in-situ values were aggregated using the 30th percentile method and both
374 mean and median methods, respectively. Moreover, VIIRS SOS showed systematic early biases
375 of 14–26 days in the PEP725 dataset, while the early biases were much smaller in the USA-NPN
376 dataset, with values of 6–11 days.

377



378

Fig. 6. VIIRS SOS and in-situ SOS at all site-years (2013–2020) in Europe (top) and the USA (bottom). Colors indicate density of sample number (low = low; red = high). MAD: mean absolute difference; MSB: mean systematic bias.

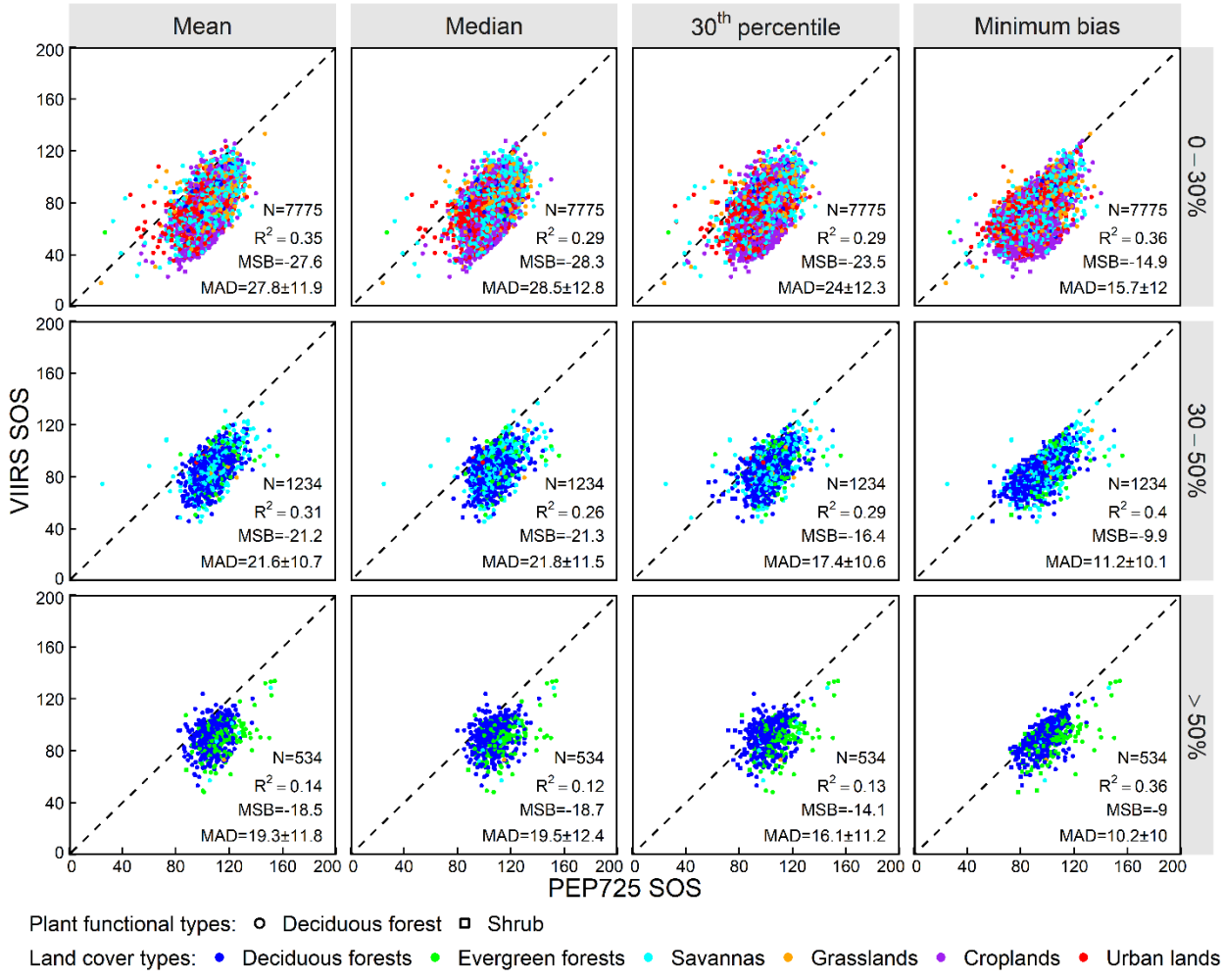
3.2.2. Difference with percent tree cover and land cover type

The vegetation purity in a VIIRS pixel was stratified by the tree cover and LCT. If tree cover was greater than 50% and LCT was classified as forests, the corresponding VIIRS SOS could represent in-suit observations from tree species. Otherwise, the VIIRS SOS mainly reflected the phenology from other vegetation species. As tree cover increased, MAD between VIIRS SOS and in-situ SOS decreased (**Fig. 7** and **Fig. 8**). MAD decreased the most in the PEP725 dataset, from 5–7 days when tree cover increased from 0–30% to 30–50%, and 1–2 days when tree cover increased > 50%. In contrast, the decrease in MAD was 1–4 days in the USA-NPN dataset when tree cover varied from 0–30% to 30–50% and remained similar with tree cover of 30–50% and > 50%. The pattern of MAD decrease with tree cover increase was similar for the in-situ data aggregated from the four different methods. However, MAD was smallest for the minimum bias method in different tree cover categories and followed by 30^{th} percentile method, which was similar for mean and median methods. The impacts were similar for systematic biases although the VIIRS SOS remained earlier than in-situ observations. The correlation between VIIRS SOS and in-situ SOS varied at different tree cover categories. In the PEP725 dataset, the R^2 for the three aggregation methods (except for minimum bias) decreased largely from tree cover of 0-30% to >50% because the reduction of samples resulted in the narrow range of SOS values. In the USA-NPN dataset, however, the R^2 for the four aggregation methods remained similar at three tree cover categories because the range of in-situ SOS was similarly large. Although the strength of the

relationship (R^2) varied, the VIIRS SOS measurements were all significantly correlated to in-situ SOS ($p<0.01$).

MAD also varied with LCT at three different categories of tree cover (**Table 4**). MAD was smaller when in-situ PFT was the same as MODIS LCT and increased when they were mismatched. The vegetation type mismatch occurred in majority sample sites in the PEP725 dataset, where 82% of PFT deciduous forest samples were located at sample sites with MODIS tree cover less than 30%, among which 21%, 32%, and 19% of PFT deciduous forest samples were corresponding to MODIS savannas+shrublands, croplands, and urban lands, respectively.

MAD was smallest when in-situ PFT and MODIS land cover matched well. For MODIS tree cover less than 30%, the land cover was unlikely related with forests but shrublands, croplands, and urban lands. Thus, the MAD in PFT shrub and MODIS savannas+shrublands was 18 days in the PEP725 dataset and 15 days in the USA-PNP dataset. For tree cover of 30–50%, the land cover was most likely savannas, so that MAD in PFT shrub and MODIS savannas+shrublands was 11.4 days in the USA-NPN dataset. For tree cover larger than 50%, the LCT should be forests. In that case, MAD in the deciduous forests identified in both PFT and MODIS data was 14 days in the PEP725 dataset and 12 days in the USA-NPN dataset (**Table 4**).



419

420 **Fig. 7.** Comparison between VIIRS SOS and in-situ SOS at all site-years (2013–2020) in Europe
421 under four aggregation methods and three tree cover categories (0–30%, 30–50%, and > 50%).
422 MAD: mean absolute difference; MSB: mean systematic bias.

423

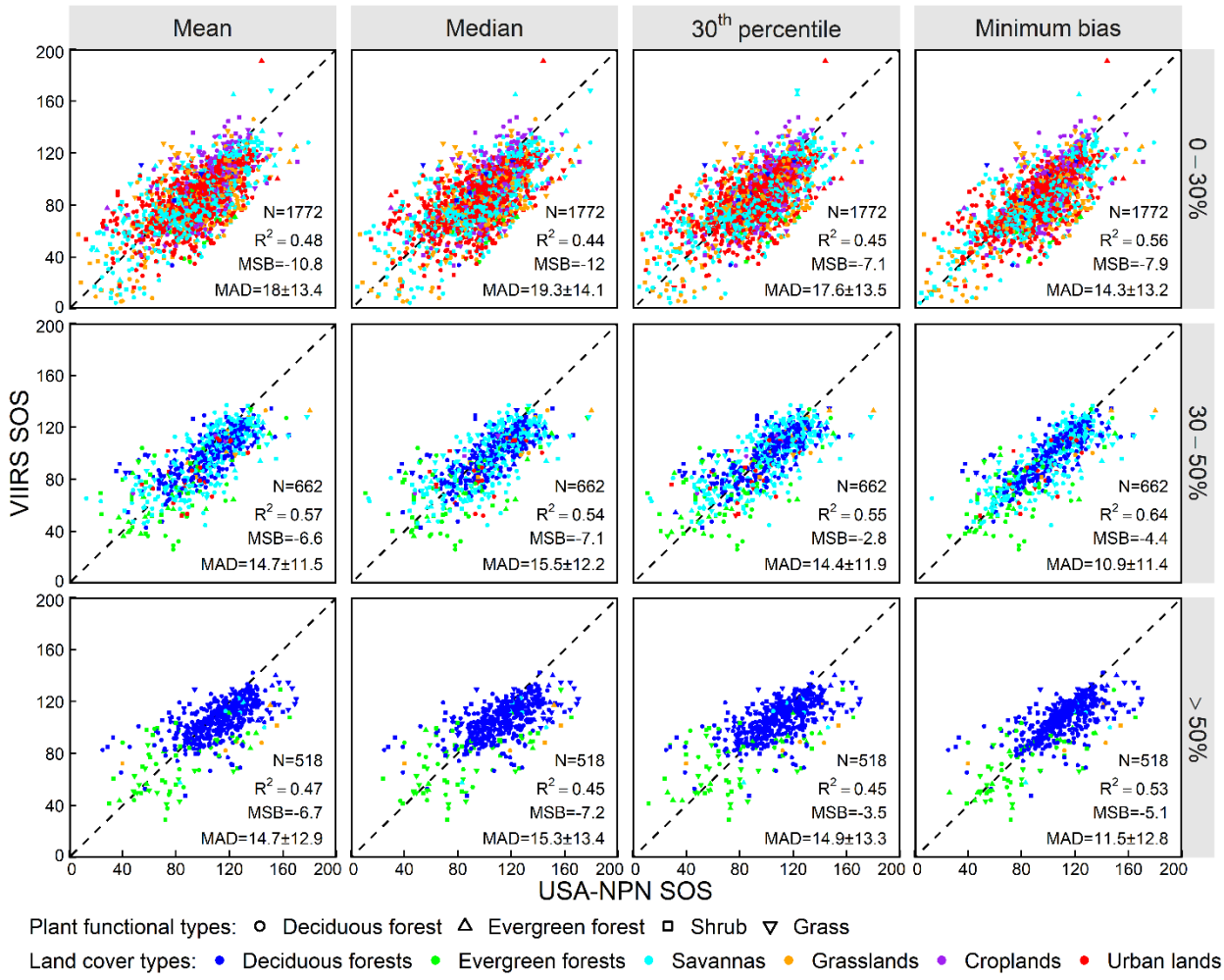


Fig. 8. Comparison between VIIRS SOS and in-situ SOS at all site-years (2013–2020) in the USA under four upscaling methods and three tree cover categories (0–30%, 30–50%, and > 50%). MAD: mean absolute difference; MSB: mean systematic bias.

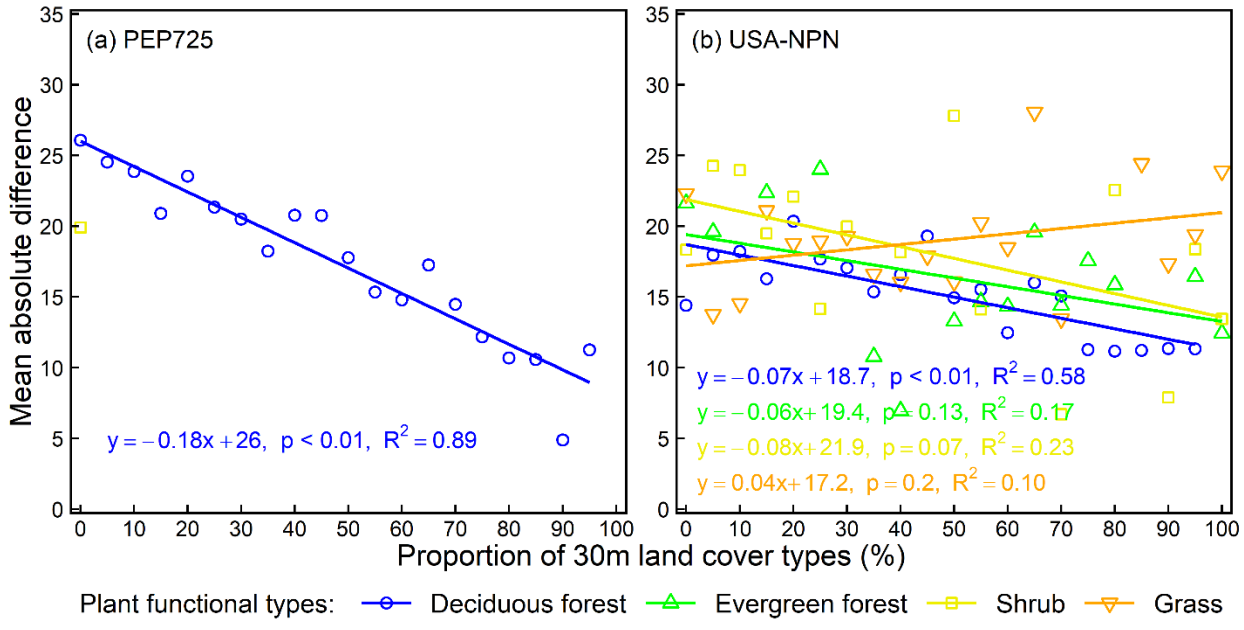
Table 4. Mean absolute difference (MAD) between VIIRS SOS and in-situ observations aggregated using 30th percentile method with the variation of MODIS tree cover categories, MODIS land cover types (LCT), and in-situ plant functional types (PFT) for sample sites from 2013–2020. Number of sampled site-years (< 20 was excluded) followed by its percent (%) in each tree cover category. "-" indicates no data.

MODIS tree cover	In-situ PFT	MODIS LCT	PEP725		USA-NPN	
			Number (%)	MAD	Number (%)	MAD
0–30%	Deciduous forest	Deciduous forests	208 (2.7)	20.1±11.3	-	-
		Evergreen forests	39 (0.5)	24.2±11.8	-	-
		Savannas+shrublands	1859 (23.9)	23.2±11.6	381 (21.5)	17.5±13.4
		Grasslands	648 (8.3)	22.7±11.9	143 (8.1)	22.4±15.4
		Croplands	2789 (35.9)	26.6±12.8	121 (6.8)	16.6±12.2
		Urban lands	1687 (21.7)	22.2±11.2	483 (27.3)	16.9±12.9
	Evergreen forest	Savannas+shrublands	-	-	21 (1.2)	24.3±17.5
		Urban lands	-	-	30 (1.7)	22.2±16.2
	Shrub	Savannas+shrublands	119 (1.5)	18.2±11.5	111 (6.3)	15.1±11.5
		Grasslands	52 (0.7)	21.4±12.5	45 (2.5)	15.8±11.2
		Croplands	256 (3.3)	26.5±15.0	52 (2.9)	14.2±11.4
		Urban lands	98 (1.3)	20.1±12.9	127 (7.2)	16.5±14.3
	Grass	Savannas+shrublands	-	-	69 (3.9)	16.8±11.9
		Grasslands	-	-	62 (3.5)	19.8±13.4
		Croplands	-	-	32 (1.8)	20.3±13.4
		Urban lands	-	-	57 (3.2)	15.7±12.2
30–50%	Deciduous forest	Deciduous forests	657 (53.2)	17.1±9.9	181 (27.3)	11.3±8.8
		Evergreen forests	88 (7.1)	21.4±13.4	26 (3.9)	25.5±16.7
		Savannas	381 (30.9)	18±11	171 (25.8)	14.2±11.5
		Grasslands	25 (2)	18.8±10.5	-	-
		Urban lands	27 (2.2)	12.2±6.2	-	-
	Evergreen forest	Savannas	-	-	24 (3.6)	16.8±10.4
	Shrub	Deciduous forests	38 (3.1)	12.6±8.6	34 (5.1)	15.7±15.2
		Evergreen forests	-	-	23 (3.5)	19.1±14.5
		Savannas+shrublands	-	-	55 (8.3)	11.4±10.1
	Grass	Deciduous forests	-	-	36 (5.4)	12.1±10.8
		Savannas	-	-	54 (8.2)	17.7±12.3
> 50%	Deciduous forest	Deciduous forests	361 (67.6)	13.9±9.8	291 (56.2)	12.3±10.9
		Evergreen forests	132 (24.7)	22.2±12.7	-	-
	Shrub	Deciduous forests	23 (4.3)	11.5±7.6	75 (14.5)	13.7±14.4
		Evergreen forests	-	-	25 (4.8)	28.7±17
	Grass	Deciduous forests	-	-	64 (12.4)	15.7±12.5

436

437 To further reveal the impact of land cover mismatch on the VIIRS SOS validation. We
438 calculated the proportion of 30 m LCT which is the same as the in-situ PFT at a sample site (**Fig**.
439 **9**). For the in-situ PFT of deciduous forest in the PEP725 dataset, MAD linearly decreased ($p <$
440 0.01) with the proportion increase of deciduous forests calculated from 30 m LCT. The higher the

441 proportion of 30 m deciduous forests represents the better match of vegetation types. This pattern
442 is the same in the USA-NPN dataset, with the exception of the in-situ PFT of grass.
443



444

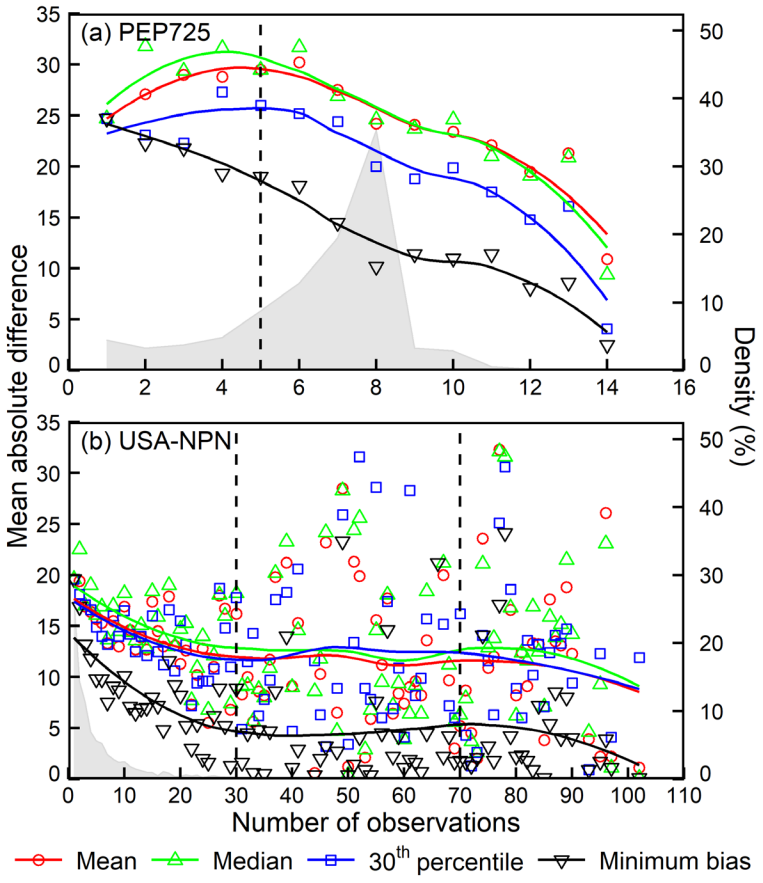
445 **Fig. 9.** Variation in mean absolute difference between VIIR SOS and in-situ SOS for different in-
446 situ plant functional types against the proportion of same vegetation types calculated from 30 m
447 land cover product at all site-years (2013–2020). The in-situ observations were aggregated using
448 the 30th percentile method.
449

450 3.2.3. Impact of the number of in-situ observations

451 As the number of in-situ observations increased, MAD between VIIRS SOS and in-situ SOS
452 decreased (**Fig. 10**). In the PEP725 dataset, there is only one observation for a species at a site-
453 year. The number of observations was less than 14 and peaked at eight (35% of sample sites). In
454 contrast, the number of observations was as high as 100 in the USA-NPN dataset, where one plant
455 species could contain multiple observations. The number of observations less than 5, 10, and 30
456 accounted for 59%, 79%, and 94% of sample sites, respectively.

In the PEP725 dataset, MAD was about 25 days for sites with one in-situ observation and reduced gradually with an increase in the number of observations (**Fig. 10**a). The reduction in MAD was linear for samples aggregated from the minimum bias method. For the other three methods of aggregation, MAD slightly increased with fewer than five observations and then steadily decreased. When the in-situ observations reached 14, MAD was as low as five days for the minimum bias method, about seven days for the 30^{th} percentile method, and around ten days for mean and median method.

In the USA-NPN dataset, MAD generally decreased non-monotonically with an increase in observations (**Fig. 10**b). MAD decreased substantially when observations were fewer than 30. Although the minimum bias method produced the smallest MAD, the 30^{th} percentile method performed slightly better than mean and median methods, especially with fewer than 30 observations.



469

470 **Fig. 10.** Variation in mean absolute difference between VIIR SOS and in-situ SOS in (a) Europe
471 and (b) the USA. The smoothed line was generated using a Loess regression (a method of local
472 weighted regression). The grey shadow indicates the density distribution of observation number
473 within a sample site.

474

475 3.2.4. Spatial variation in the difference between in-situ observations and VIIRS SOS

476 In Europe, MAD between VIIRS SOS and in-situ SOS was less than 20 days in northern and
477 eastern regions. In Germany, where most sites were distributed, MAD was 20–25 days at peak
478 frequency of sample sites and 0–25 days in 61% of sample sites (**Fig. 11**a). MAD of deciduous
479 forest and shrub were similar (~ 22 days). In the USA, MAD was lower than 20 days in
480 northeastern region, and greater than 20 days for many sites in the south and west (**Fig. 11**b).
481 Extremely large MAD (> 35) occurred where the number of in-situ records was limited (**Fig. 3**).

VIIRS SOS and in-situ SOS exhibited very good agreements at sample sites where in-situ PFT and satellite LCT matched well (**Fig. 11**c and d). MAD was 5–10 days at the peak frequency of sample sites in both Europe and the USA. The sample sites with MAD lower than 20 days accounted for about 68% in Europe and 82% in the USA. The PFT of deciduous forest exhibited smallest MAD (14.1±8.4 in Europe and 10.7±7.3 in the USA) compared to other vegetation types. The PFT of evergreen forest showed large MAD even in homogeneous sample sites.

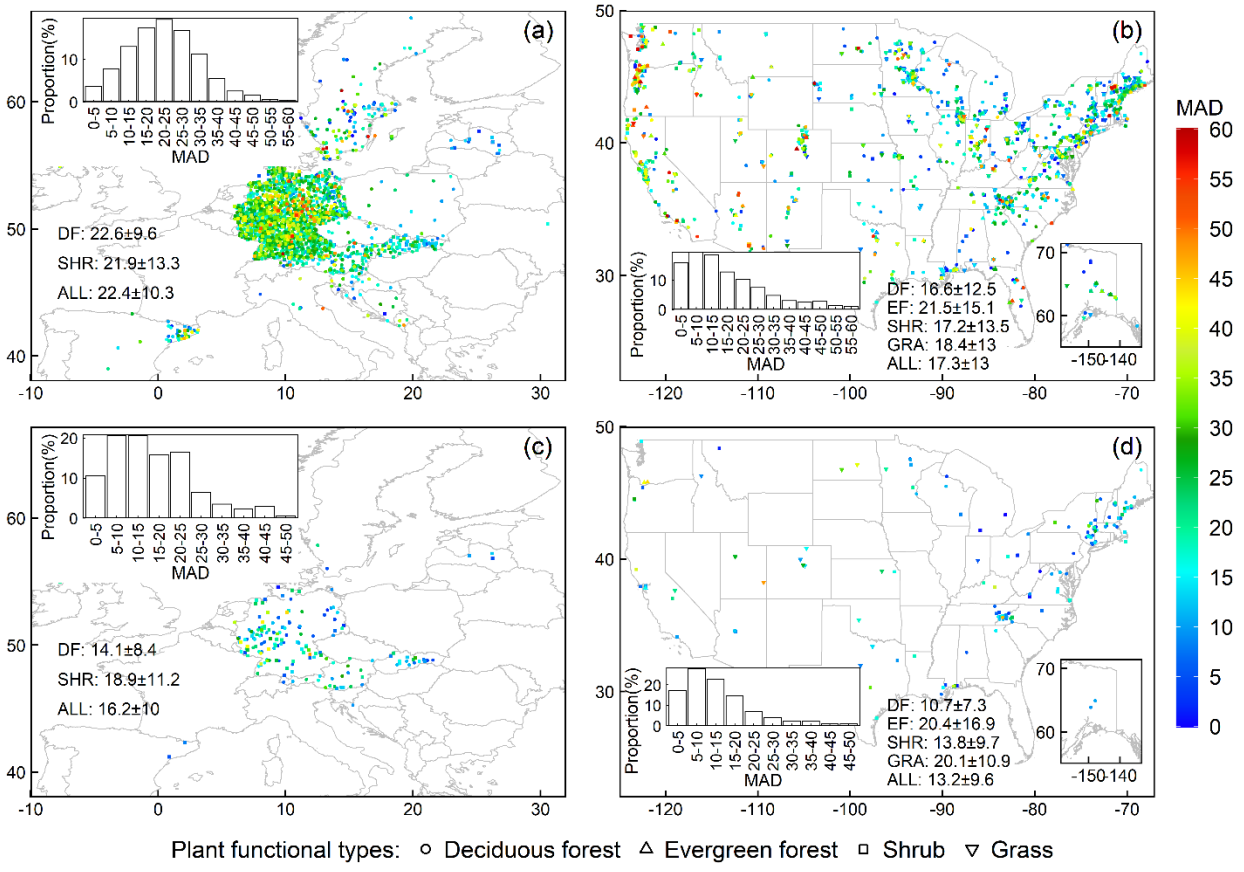


Fig. 11. Spatial variation of mean absolute difference (MAD) in eight years (2013–2020) between VIIRS SOS and in-situ SOS aggregated using the 30th percentile method in Europe (a and c) and the USA (b and d). (a) and (b) were derived from all the sample sites, while (c) and (d) were calculated from the sample sites where in-situ observations larger than 3 and in-situ plant functional types are the same as the MODIS land cover types. The mean values ± standard

deviations were calculated for different plant functional types (DF: deciduous forest; EF: evergreen forest; SHR: shrub; GRA: grass) and all vegetation types (ALL).

3.3. Comparison of inter-annual variations between VIIRS SOS and in-situ observations

3.3.1. Variation in multi-year anomaly

Inter-annual anomaly of VIIRS SOS and in-situ SOS were significantly correlated ($p < 0.01$, $R^2 = 0.3–0.41$ in Europe and 0.08–0.23 in the USA; **Fig. 12**). More importantly, the anomaly in both VIIRS SOS and in-situ observations exhibited the same positive and negative variations (in quadrant I and quadrant III in the scatterplots) in 70–76% of site-years although the proportion was slightly larger in the USA than Europe (**Table 5**). Further, the sample site-years with same anomaly direction were very similar for the four aggregation methods of in-situ data (**Table 5**), indicating similar abilities in acquiring inter-annual variations of SOS at sample sites.

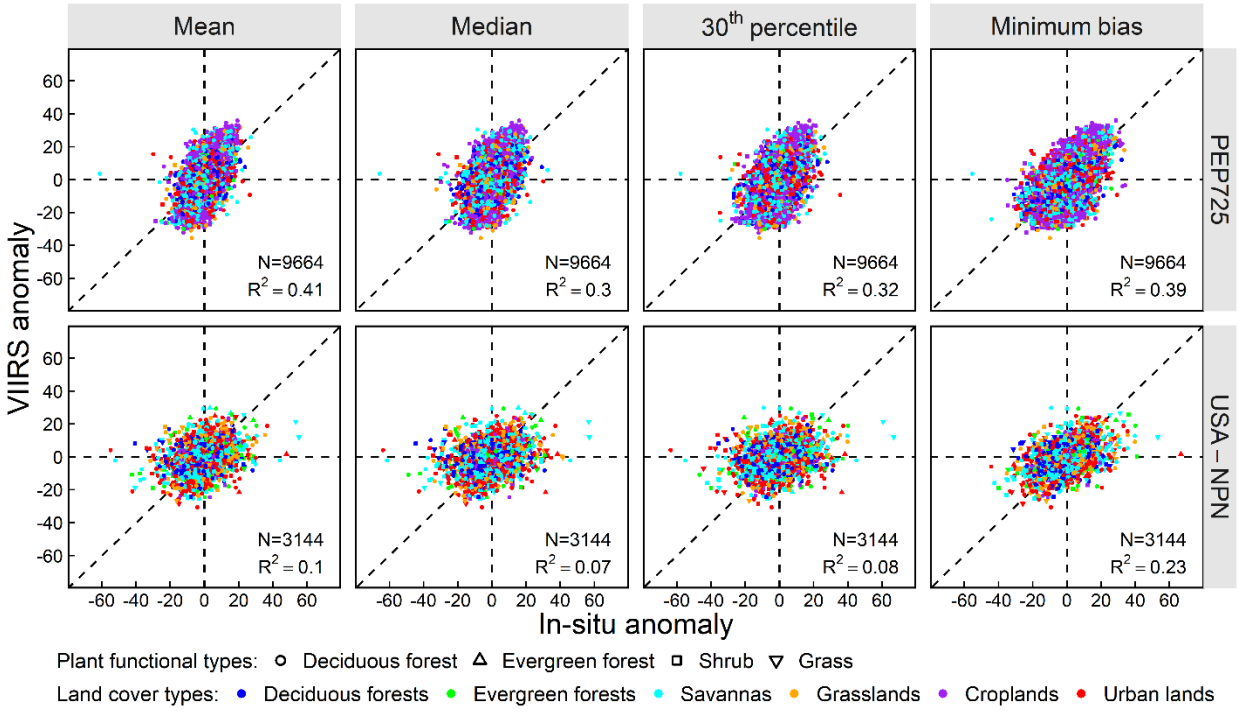


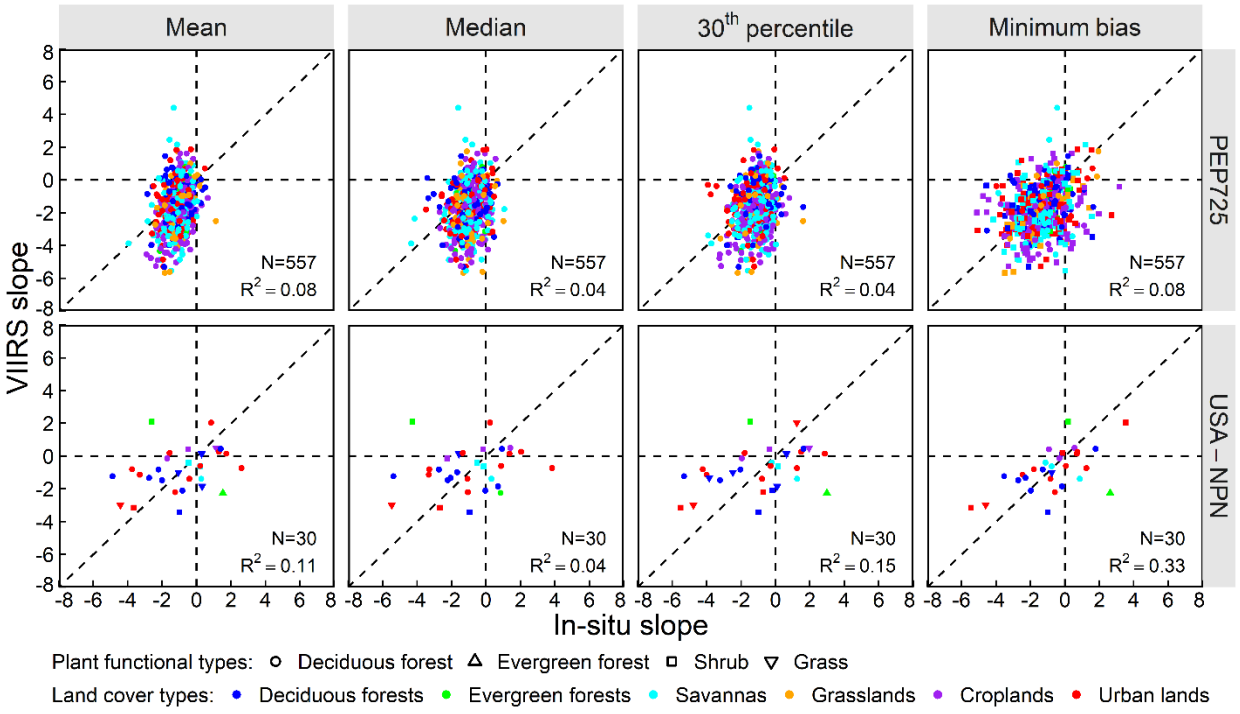
Fig. 12. Comparison of anomaly between VIIRS SOS and in-situ observations at all site-years (2013–2020) in Europe and the USA, where the in-situ observations were aggregated using four different methods.

Table 5. Proportion (%) of sample site-years (2013–2020) with positive or negative of anomaly in VIIRS SOS and in-situ observations. It corresponds to each of the 4 quadrants in **Fig. 12**.

Data sets	Methods	I (+, +)	II (-, +)	III (-, -)	IV (+, -)
PEP725	Mean	38.8	13.2	35.3	12.7
	Median	37.8	15.6	32.9	13.7
	30^{th} percentile	38.2	14.9	33.6	13.3
	Minimum bias	38.9	12.6	35.9	12.6
USA-NPN	Mean	50.2	14.3	21.1	14.3
	Median	50.5	14.7	20.8	13.9
	30^{th} percentile	50.2	14.5	21.1	14.3
	Minimum bias	52.4	11.9	23.6	12.1

3.3.2. Variation in long-term trends

The magnitude of linear slope was greater in VIIRS SOS than in-situ SOS aggregated from all four aggregation methods in Europe, but it demonstrated random differences in the USA (**Fig. 13**). The correlation between these measures was strongest for in-situ SOS aggregated using the minimum bias method, which was similar for other aggregation methods. The proportion of sample sites in same direction of linear trends (negative or positive) accounted for 82–87% and 70–77% in Europe and the USA, respectively (**Table 6**).



524

525 **Fig. 13.** Comparison of linear slopes between VIIRS SOS and in-situ SOS from 2013 to 2020 at
526 all sites in Europe and the USA using four aggregation methods.

527

528 **Table 6.** Proportion (%) of sample sites with positive or negative long-term slopes in VIIRS SOS
529 and in-situ observations. It corresponds to each of the 4 quadrants in **Fig. 13**.

Data sets	Methods	I (+, +)	II (-, +)	III (-, -)	IV (+, -)
PEP725	Mean	0.5	1.8	86.4	11.3
	Median	1.3	6.6	81.5	10.6
	30th percentile	0.4	2.9	85.3	11.5
	Minimum bias	3.4	9.3	78.8	8.4
USA-NPN	Mean	20.0	20.0	50.0	10.0
	Median	16.7	16.7	53.3	13.3
	30th percentile	20.0	16.7	53.3	10.0
	Minimum bias	20.0	13.3	56.7	10.0

4. Discussion

This study investigated the optimal method for validating satellite-derived land surface phenology using in-situ observations from national phenology networks. It is widely recognized that matching ground-based observations from national phenology networks with satellite LSP detections is challenging; as such, LSP validation efforts have historically relied on other sources of validation data such as PhenoCam imagery (Klosterman et al., 2014; Moon et al., 2019; Richardson et al., 2018; Zhang et al., 2018a). However, national phenology networks offer the greatest spatially extensive and taxonomically rich datasets for ground-truthing satellite-derived products. As a result, it is critical to understand the discrepancy and agreement between in-situ observations from the national phenology networks and satellite LSP detections. This study advances validation algorithms of LSP by fully using in-situ observations maintained by national phenology networks.

We developed and compared four different algorithms to aggregate in-situ species-specific phenology to sample sites that are compatible with satellite detections. Our study revealed that various approaches for aggregating in-situ observations have substantial impacts on the validation results. The conventionally used mean and median methods are limited for the aggregation of in-situ species-specific plant phenology, although the average of all the in-situ observations within a certain area were widely applied to compare with satellite detections (Berra et al., 2019; Delbart et al., 2015; Kowalski et al., 2020; Melaas et al., 2016; Peng et al., 2017a; Tian et al., 2021; White et al., 2009). The minimum bias method, on the other hand, yields the smallest difference between LSP and in-situ observations, revealing that in some plant species, LSP performs well at reflecting the timing of phenological transitions. This method, however, can only be a reference in the comparison of other methods because the manner of selecting ground truth is not independent from

553 the satellite derived LSP. In contrast, the 30th percentile method could be more practical because
554 it takes multiple in-situ observations into account. Although the optimal percentile could vary with
555 different landscapes and ecosystems, it is very close to 30th percentile at the national scale (Peng
556 et al., 2017b). Indeed, the 30th percentile is the optimal threshold that has been demonstrated by
557 investing scaling issues of SOS in the central U.S. (croplands, grasslands, and forests) and in the
558 semiarid west U.S. (shrublands and grasslands) (Peng et al., 2021; Zhang et al., 2017). The
559 advantage of the 30th percentile method was demonstrated in our comparisons between VIIRS SOS
560 and in-situ SOS, although the advantage in the USA is not as great as that in Europe (**Fig. 6-8**).
561 The difference in the 30th percentile method was 4–5 days and one day smaller than that in mean
562 and median methods in PEP725 and USA-NPN datasets, respectively, although it was 9 days and
563 3 days larger than that in minimum bias method. This pattern was found in the comparison with
564 all site-years, different tree cover categories, and LCT.

565 We showed quantitatively how to select compatible samples for comparing LSP detections
566 with in-situ observations. Phenological events are heterogeneous among plant species and even
567 within the same species (Richardson and O'Keefe, 2009). Generally, the timing of phenological
568 events in the same vegetation type is more comparable than those of different vegetation types.
569 However, the vegetated surface is commonly heterogeneous, with the mixture of various
570 vegetation types within a satellite pixel footprint. This is evident in **Fig. 4**, where the in-situ SOS
571 could be observed from a few deciduous trees within croplands or from plants with a wide range
572 of SOS values. Overall, 93–99% of in-situ observations in the PEP725 dataset were from
573 deciduous trees, while over 63% of them were located in the MODIS LCT of croplands, grasslands,
574 and savannas/shrublands with tree cover less than 50%. As a result, VIIRS SOS was overall much
575 earlier than in-situ observations in the PEP725 dataset, and the early bias increased with an

increasing proportion of croplands and grasslands (**Fig. 7**). In other words, the VIIRS SOS in these sample sites represented the phenological timings in crops or grasses instead of deciduous forests. However, most of the observations in the PEP725 dataset were European forest or garden tree species that leaf-out later than grasses and crops (Fu et al., 2014). Crop phenology is significantly influenced by crop management, while grasses respond more quickly to climate changes than forests and will dominate the satellite-derived phenology signals when forests are mixed with a certain proportion of grasses and crops (Donnelly et al., 2018). Other evaluations of LSP using in-situ observation data also reported an earlier pattern of satellite SOS (Cong et al., 2013; Donnelly et al., 2021; Ganguly et al., 2010; Pouliot et al., 2011; Soudani et al., 2008; Verma et al., 2016), especially among grasses and crops (Delbart et al., 2015; Peng et al., 2017a; Wu et al., 2016; Xie and Wilson, 2020). The greenness increase within a satellite pixel indicates a certain proportion of green leaves appear from early greenup plants. This has been demonstrated by comparing SOS at 500m VIIRS pixels with 30m Landsat or Harmonized Landsat and Sentinel-2 (HLS) pixels, indicating SOS at coarse pixels reflects time of 30% area becoming green (Peng et al., 2021; Zhang et al., 2017). On the other hand, in-situ data provide observations for pre-selected plant species which are not necessarily the early leaf-out species within a 500m VIIRS pixel.

In the USA-NPN dataset, the in-situ observations are more comparable to VIIRS SOS than those in the PEP725 dataset. In-situ observations in the USA-NPN dataset contain plants from more diverse vegetation species than those in PEP725 dataset (**Table 1**), where the number of observations in a sample site could be as large as 100 (**Fig. 10**) and 63% of in-situ deciduous forest samples (tree cover > 50%) matched with LCT in VIIRS SOS detections. Therefore, these observations better represented the timing of phenological events for a VIIRS pixel. To acquire reliable validation results, the selected samples should be comparable in vegetation types between

599 in-situ species and satellite pixels. At deciduous forest sites in both in-situ PFT and satellite LCT
600 with tree cover > 50%, the MAD was 14 days in Europe and 12 days in the USA (**Table 4**). This
601 suggests that VIIRS SOS produces similar accuracy in both Europe and the USA.
602 This study demonstrated that in-situ measurements and VIIRS detections were comparable
603 with small differences when obtained from the same vegetation types. This is further reflected in
604 the linear MAD decrease with the proportion increase of the 30 m LCT that was the same in-situ
605 PFT (**Fig. 9**). However, the MAD between VIIRS SOS and in-situ SOS in grasses showed no
606 significant trend (p=0.2) with the increase of grass proportion in a VIIRS pixel (**Fig. 9**b). The cause
607 of this pattern requires further investigations. It could be potentially associated to the fact that in-
608 situ grass observations, which originate primarly fromarid/semi-arid areas in the USA-NPN
609 dataset (**Fig. 3**), could experience multiple growing cycles with the controls of multiple rainy and
610 dry episodes. Thus, the phenological cycles between VIIRS detections and in-situ observations
611 could be inconsistent. It is challenging to match in-situ PFT and satellite LCT because the in-situ
612 observations within a VIIRS pixel always contains several plant functional types. The mismatch
613 could cause large MAD between in-situ SOS and VIIRS SOS.
614 This evaluation also revealed that multiple in-situ observations within a sample site are needed
615 to enhance the validation practice. A single in-situ observation is unlikely to represent the changes
616 of the entire vegetation community in a satellite pixel (Fisher and Mustard, 2007). Our results
617 show that MAD between VIIRS SOS and in-situ SOS decreases with an increase in the number of
618 observations (**Fig. 10**). For example, MAD dropped to ~15 days when the number of species
619 reached 13 in both the PEP725 and USA-NPN datasets (**Fig. 10**). MAD decreased to 10–15 days
620 when the sample size was between 20–80 and dropped almost linearly when the sample size was
621 greater than 80 in the USA-NPN (**Fig. 10**b). This finding is supported by a previous study that

622 suggests that the sample size should be larger than 20 and could be as many as 80 for optimal
623 validation of satellite-derived phenology (Liang et al., 2011). Therefore, it is suggested that the
624 sample pairs for validation should be selected from: (1) matched in-situ PFT and satellite LCT,
625 and (2) number of in-situ observations at least larger than three in a sample site, or an optimal
626 number of larger than six in the PEP725 dataset and ten in the USA-NPN dataset. Our results
627 display a MAD of 10–14 days for sites with the same type of in-situ PFT and MODIS LCT and
628 with the number of observations larger than three (**Fig. 11**c and d). The direct comparison suggests
629 that a satellite-derived LSP product is of high quality if its difference with current in-situ
630 observations is less than two weeks. The complexity and heterogeneity of plant species in a satellite
631 pixel make it impossible for a few in-situ observations to fully match satellite-derived LSP. Thus,
632 large differences in the SOS comparison are commonly obtained, even in deciduous forests
633 (Bornez et al., 2020; Delbart et al., 2015; Donnelly et al., 2019; Peng et al., 2017a; Soudani et al.,
634 2008; Xie and Wilson, 2020).

635 We also demonstrate that comparisons of inter-annual variations and long-term trends provide
636 a robust evaluation of satellite LSP quality. Long-term LSP data records are critical to investigating
637 the impacts of environmental changes on ecosystems in a warming climate. This suggests that the
638 LSP detections meet the requirement for exploring ecosystem dynamics if inter-annual variations
639 and trends reflect well of the in-situ data variations. Although VIIRS SOS and in-situ SOS
640 presented a difference of 2–4 weeks for all the sample sites (**Fig. 6**), they exhibited inter-annual
641 anomalies and long-term trends in the same direction in over 70% of sample sites in four different
642 aggregation methods (**Tables 5** and **6**). The inconsistent direction in the remained samples may
643 result from (1) uncertainties in in-situ observations or satellite detections, (2) inherence of plant
644 species, and (3) unique microclimate at sample sites (Tian et al., 2020). It is interesting that the

sample sites with same direction of trends were larger in Europe than the USA, although MAD was overall larger in Europe. This indicates that the larger MAD was the result of mismatch between satellite LCT and in-situ PFT, instead of the bad quality of VIIRS SOS detections. The comparison results also suggest that the VIIRS SOS has the ability to monitor the phenological response to yearly environmental changes and long-term climate changes at the landscape scale.

Finally, the in-situ observations in national phenology networks are expected to play an increasingly important role in validating satellite-derived LSP in the future. Mobile applications have been developed for acquiring accurate measurements, including precise locations for observations. To motivate more participants in tracking plant and animal phenology, the USA-NPN released a mobile application (Nature's Notebook) appropriate for use by professional and citizen scientists to record phenology observations. With the availability of accurate geolocation and phenological events, in-situ observations could be well matched to the LSP detected from PlanetScope (~3m), Sentinel-2 (~10m), and Landsat (~30m). Compared to moderate or coarse LSP, the LSP from finer spatial resolution imagery generally derivate less from in-situ observations. In our study, the links between VIIRS SOS and in-situ SOS were only acceptable at homogenous sites (MAD: 10–14 days), while the difference between fine spatial resolution SOS and in-situ SOS could be less than 10 days in Europe (Kowalski et al., 2020) and the USA (Zhang et al., 2020b). This is reasonable because fine spatial resolution pixels are "purer" relative to moderate spatial resolution pixels. In homogenous area, the difference between the two kinds of resolution could be reduced. For example, a study in the semiarid region of the western United States dominated by shrublands did not show a significant difference between VIIRS SOS and Harmonized Landsat and Sentinel-2 (HLS) SOS (Peng et al., 2021). Additionally, our results show that VIIRS SOS tended to be earlier than in-situ SOS (as explained in the paragraphs 3 and 4 in

this section), because VIIRS captured the greenup onset when about 30% of the plants start leaf out within the pixels (Peng et al., 2017b; Zhang et al., 2017). This study is also supported by a previous study that suggested that the SOS of a coarser pixel is usually earlier than the average of inside fine pixels because it depends on those fine pixels with earlier green-up and higher growth speed (Liu et al., 2019). Although these finer spatial resolution LSP could be properly aggregated for substantially improving our knowledge of the quality in the long-term global LSP products at moderate resolutions (~500m), it is challenging to apply at large scale due to the low temporal resolution and cloud contaminations. The moderate or coarse satellite data remains the most useful data source in monitoring global and continental LSP. However, the SOS derived from moderate or coarse satellite data may represent the phenology of only earliest species (Fu et al., 2014), which are not necessarily the dominant species in a plant community.

Although we in this study investigated four methods to match observations from the national phenology networks with satellite detections, matching spatially in-situ observations and satellite detections remains a challenge. The geolocation of in-situ observations can exhibit large uncertainties, though this shortcoming is expected to improve greatly with the increase in the use of mobile applications to collect the observations. Further, individual plants observed from fields do not always represent the vegetation community present in moderate resolution satellite pixels accurately. To best link these two datasets, vegetation types should be classified using high resolution data, such as Worldview (< 0.5m) and PlanetScope (~3m), and the vegetation phenology could be detected from daily PlanetScope time series. Thus, the in-situ PTF could be spatially match well with the vegetation type in a PlanetScope pixel and then the phenology detected from PlanetScope could be calibrated using in-situ observations. Finally, the calibrated 3m PlanetScope

690 phenology could be aggregated to evaluate and validate the global LSP products produced from
691 VIIRS and MODIS observations.
692
693 **5. Conclusions**
694 We compared different methods for the validation of satellite-derived LSP products solely
695 using in-situ phenology observations from national phenology networks. Because of the limited
696 selections in plants, errors in geolocation, and the uncertainty in phenology observations, in-situ
697 observations are unlikely to be fully representative of satellite pixel footprints and content.
698 Therefore, a sample site for LSP validation requires more than three in-situ observations, or an
699 optimal number of larger than six in the PEP725 dataset and ten in the USA-NPN dataset. For
700 aggregating in-situ observations in a sample site, the 30^{th} percentile method is more practical than
701 the mean, median, and minimum bias methods. Further, direct comparison should be performed
702 by selecting the samples where observed in-situ plant species are consistent to satellite land cover
703 types, which is able to reveal the magnitude of discrepancy between in-situ observations and LSP
704 detections. It is suggested that a LSP product is of high quality if its difference with in-situ
705 observations is less than two weeks. Inter-annual anomaly and long-term trends provide an
706 alternative way to evaluate the quality of LSP products, which not only reduce the impacts of
707 spatial mismatches but also meet the requirement for investigating long-term changes of climate
708 and ecosystems. Overall, the innovative method in this study provides new knowledge in LSP
709 evaluation using in-situ observations and the result improves our understanding of the scale
710 mismatch and sample representativeness of species-specific phenology and the uncertainties of
711 long-term LSP detections from remote sensing data.
712

Acknowledgements

This work was supported by NASA contract 80NSSC18K0626. We would like to express our profound gratitude to the staff of the PEP725 project (particularly to Markus Ungersböck) and USA-NPN and to the many volunteer participants in these phenology monitoring programs for providing us with the in-situ databases. The authors also wish to thank the LP DAAC for providing the MODIS and VIIRS satellite data.

References

Berra, E.F., Gaulton, R., & Barr, S., 2019. Assessing spring phenology of a temperate woodland: A multiscale comparison of ground, unmanned aerial vehicle and Landsat satellite observations. Remote Sens. Environ. 223, 229-242. https://doi.org/10.1016/j.rse.2019.01.010.

Bison, M., Yoccoz, N.G., Carlson, B.Z., & Delestrade, A., 2019. Comparison of budburst phenology trends and precision among participants in a citizen science program. IJBm. 63, 61-72. https://doi.org/10.1007/s00484-018-1636-x.

Bornez, K., Descals, A., Verger, A., & Penuelas, J., 2020. Land surface phenology from VEGETATION and PROBA-V data. Assessment over deciduous forests. IJAEO. 84. https://doi.org/10.1016/j.jag.2019.101974.

Cleland, E.E., Chuine, I., Menzel, A., Mooney, H.A., & Schwartz, M.D., 2007. Shifting plant phenology in response to global change. Trends Ecol. Evol. 22, 357-365. https://doi.org/10.1016/j.tree.2007.04.003.

Cong, N., Wang, T., Nan, H.J., Ma, Y.C., Wang, X.H., Myneni, R.B., & Piao, S.L., 2013. Changes in satellite-derived spring vegetation green-up date and its linkage to climate in China from 1982 to 2010: a multimethod analysis. Global Change Biol. 19, 881-891. https://doi.org/10.1111/gcb.12077.

Courter, J.R., Johnson, R.J., Stuyck, C.M., Lang, B.A., & Kaiser, E.W., 2013. Weekend bias in Citizen Science data reporting: implications for phenology studies. IJBm. 57, 715-720. https://doi.org/10.1007/s00484-012-0598-7.

de Beurs, K.M., & Henebry, G.M., 2004. Land surface phenology, climatic variation, and institutional change: Analyzing agricultural land cover change in Kazakhstan. Remote Sens. Environ. 89, 497-509. https://doi.org/10.1016/j.rse.2003.11.006.

de Jong, R., de Bruin, S., de Wit, A., Schaepman, M.E., & Dent, D.L., 2011. Analysis of monotonic greening and browning trends from global NDVI time-series. Remote Sens. Environ. 115, 692-702. https://doi.org/10.1016/j.rse.2010.10.011.

Delbart, N., Beaubien, E., Kergoat, L., & Toan, T.L., 2015. Comparing land surface phenology with leafing and flowering observations from the PlantWatch citizen network. Remote Sens. Environ. 160, 273-280. https://doi.org/10.1016/j.rse.2015.01.012.

Denny, E.G., Gerst, K.L., Miller-Rushing, A.J., Tierney, G.L., Crimmins, T.M., Enquist, C.A.F., Guertin, P., Rosemartin, A.H., Schwartz, M.D., Thomas, K.A., & Weltzin, J.F., 2014. Standardized phenology monitoring methods to track plant and animal activity for science and resource management applications. IJBm. 58, 591-601. https://doi.org/10.1007/s00484-014-0789-5.

DiMiceli, C., M. Carroll, R. Sohlberg, D. Kim, & M. Kelly, J.T., 2015. MOD44B MODIS/Terra Vegetation Continuous Fields Yearly L3 Global 250m SIN Grid V006 [Data set]. https://doi.org/10.5067/MODIS/MOD44B.006 accessed on 7 September 2021.

Donnelly, A., Liu, L., Zhang, X., & Wingler, A., 2018. Autumn leaf phenology: discrepancies between in situ observations and satellite data at urban and rural sites. Int. J. Remote Sens. 39, 8129-8150. https://doi.org/10.1080/01431161.2018.1482021.

Donnelly, A., Yu, R., & Liu, L.L., 2021. Comparing in situ spring phenology and satellite-derived start of season at rural and urban sites in Ireland. Int. J. Remote Sens. 42, 7821-7841. https://doi.org/10.1080/01431161.2021.1969056.

Donnelly, A., Yu, R., Liu, L.L., Hanes, J.M., Liang, L., Schwartz, M.D., & Desai, A.R., 2019. Comparing in-situ leaf observations in early spring with flux tower CO2 exchange, MODIS EVI and modeled LAI in a northern mixed forest. Agric. For. Meteorol. 278. https://doi.org/10.1016/j.agrformet.2019.107673.

Feldman, R.E., Zemaite, I., & Miller-Rushing, A.J., 2018. How training citizen scientists affects the accuracy and precision of phenological data. IJBm. 62, 1421-1435. https://doi.org/10.1007/s00484-018-1540-4.

Fisher, J.I., & Mustard, J.F., 2007. Cross-scalar satellite phenology from ground, Landsat, and MODIS data. Remote Sens. Environ. 109, 261-273. https://doi.org/10.1016/j.rse.2007.01.004.

Fu, Y.S.H., Piao, S.L., Op de Beeck, M., Cong, N., Zhao, H.F., Zhang, Y., Menzel, A., & Janssens, I.A., 2014. Recent spring phenology shifts in western Central Europe based on multiscale observations. Global Ecol. Biogeogr. 23, 1255-1263. https://doi.org/10.1111/geb.12210.

Fuccillo, K.K., Crimmins, T.M., de Rivera, C.E., & Elder, T.S., 2015. Assessing accuracy in citizen science-based plant phenology monitoring. IJBm. 59, 917-926. https://doi.org/10.1007/s00484-014-0892-7.

Ganguly, S., Friedl, M.A., Tan, B., Zhang, X.Y., & Verma, M., 2010. Land surface phenology from MODIS: Characterization of the Collection 5 global land cover dynamics product. Remote Sens. Environ. 114, 1805-1816. https://doi.org/10.1016/j.rse.2010.04.005.

Garonna, I., De Jong, R., De Wit, A.J.W., Mucher, C.A., Schmid, B., & Schaepman, M.E., 2014. Strong contribution of autumn phenology to changes in satellite-derived growing season length estimates across Europe (1982-2011). Global Change Biol. 20, 3457-3470. https://doi.org/10.1111/gcb.12625.

Hansen, M.C., Egorov, A., Potapov, P.V., Stehman, S.V., Tyukavina, A., Turubanova, S.A., Roy, D.P., Goetz, S.J., Loveland, T.R., Ju, J., Kommareddy, A., Kovalskyy, V., Forsyth, C., & Bents, T., 2014. Monitoring conterminous United States (CONUS) land cover change with Web-Enabled Landsat Data (WELD). Remote Sens. Environ. 140, 466-484. https://doi.org/10.1016/j.rse.2013.08.014.

Huang, X.J., Xiao, J.F., & Ma, M.G., 2019. Evaluating the Performance of Satellite-Derived Vegetation Indices for Estimating Gross Primary Productivity Using FLUXNET Observations across the Globe. Remote Sens. 11. https://doi.org/10.3390/rs11151823.

Huete, A.R., Restrepo-Coupe, N., Ratana, P., Didan, K., Saleska, S.R., Ichii, K., Panuthai, S., & Gamo, M., 2008. Multiple site tower flux and remote sensing comparisons of tropical forest dynamics in Monsoon Asia. Agric. For. Meteorol. 148, 748-760. https://doi.org/10.1016/j.agrformet.2008.01.012.

Julien, Y., & Sobrino, J.A., 2009. Global land surface phenology trends from GIMMS database. Int. J. Remote Sens. 30, 3495-3513. https://doi.org/10.1080/01431160802562255.

Justice, C.O., Roman, M.O., Csiszar, I., Vermote, E.F., Wolfe, R.E., Hook, S.J., Friedl, M., Wang, Z.S., Schaaf, C.B., Miura, T., Tschudi, M., Riggs, G., Hall, D.K., Lyapustin, A.I., Devadiga,

S., Davidson, C., & Masuoka, E.J., 2013. Land and cryosphere products from Suomi NPP VIIRS: Overview and status. J. Geophys. Res. Atmos. 118, 9753-9765. https://doi.org/10.1002/jgrd.50771.

Khare, S., Drolet, G., Sylvain, J.D., Pare, M.C., & Rossi, S., 2019. Assessment of Spatio-Temporal Patterns of Black Spruce Bud Phenology across Quebec Based on MODIS-NDVI Time Series and Field Observations. Remote Sens. 11. https://doi.org/10.3390/rs11232745.

Klosterman, S.T., Hufkens, K., Gray, J.M., Melaas, E., Sonnentag, O., Lavine, I., Mitchell, L., Norman, R., Friedl, M.A., & Richardson, A.D., 2014. Evaluating remote sensing of deciduous forest phenology at multiple spatial scales using PhenoCam imagery. BGeo. 11, 4305-4320. https://doi.org/10.5194/bg-11-4305-2014.

Kowalski, K., Senf, C., Hostert, P., & Pflugmacher, D., 2020. Characterizing spring phenology of temperate broadleaf forests using Landsat and Sentinel-2 time series. IJAEO. 92. https://doi.org/10.1016/j.jag.2020.102172.

Liang, L.A., Schwartz, M.D., & Fei, S.L., 2011. Validating satellite phenology through intensive ground observation and landscape scaling in a mixed seasonal forest. Remote Sens. Environ. 115, 143-157. https://doi.org/10.1016/j.rse.2010.08.013.

Liu, L.C., Cao, R.Y., Shen, M.G., Chen, J., Wang, J.M., & Zhang, X.Y., 2019. How Does Scale Effect Influence Spring Vegetation Phenology Estimated from Satellite-Derived Vegetation Indexes? Remote Sens. 11. https://doi.org/10.3390/rs11182137.

MacKenzie, C.M., Murray, G., Primack, R., & Weihrauch, D., 2017. Lessons from citizen science: Assessing volunteer-collected plant phenology data with Mountain Watch. Biol. Conserv. 208, 121-126. https://doi.org/10.1016/j.biocon.2016.07.027.

Melaas, E.K., Sulla-Menashe, D., Gray, J.M., Black, T.A., Morin, T.H., Richardson, A.D., & Friedl, M.A., 2016. Multisite analysis of land surface phenology in North American temperate and boreal deciduous forests from Landsat. Remote Sens. Environ. 186, 452-464. https://doi.org/10.1016/j.rse.2016.09.014.

Menzel, A., Sparks, T.H., Estrella, N., Koch, E., Aasa, A., Ahas, R., Alm-Kubler, K., Bissolli, P., Braslavska, O., Briede, A., Chmielewski, F.M., Crepinsek, Z., Curnel, Y., Dahl, A., Defila, C., Donnelly, A., Filella, Y., Jatcza, K., Mage, F., Mestre, A., Nordli, O., Penuelas, J., Pirinen, P., Remisova, V., Scheifinger, H., Striz, M., Susnik, A., Van Vliet, A.J.H., Wielgolaski, F.E., Zach, S., & Zust, A., 2006. European phenological response to climate change matches the warming pattern. Global Change Biol. 12, 1969-1976. https://doi.org/10.1111/j.1365-2486.2006.01193.x.

Moon, M., Zhang, X.Y., Henebry, G.M., Liu, L.L., Gray, J.M., Melaas, E.K., & Friedl, M.A., 2019. Long-term continuity in land surface phenology measurements: A comparative assessment of the MODIS land cover dynamics and VIIRS land surface phenology products. Remote Sens. Environ. 226, 74-92. https://doi.org/10.1016/j.rse.2019.03.034.

Nguyen, L.H., Joshi, D.R., Clay, D.E., & Henebry, G.M., 2020. Characterizing land cover/land use from multiple years of Landsat and MODIS time series: A novel approach using land surface phenology modeling and random forest classifier. Remote Sens. Environ. 238. https://doi.org/10.1016/j.rse.2018.12.016.

Peng, D., Wang, Y., Xian, G., Huete, A.R., Huang, W., Shen, M., Wang, F., Yu, L., Liu, L., Xie, Q., Liu, L., & Zhang, X., 2021. Investigation of land surface phenology detections in shrublands using multiple scale satellite data. Remote Sens. Environ. 252, 112-133. https://doi.org/10.1016/j.rse.2020.112133.

Peng, D.L., Wu, C.Y., Li, C.J., Zhang, X.Y., Liu, Z.J., Ye, H.C., Luo, S.Z., Liu, X.J., Hug, Y., & Fang, B., 2017a. Spring green-up phenology products derived from MODIS NDVI and EVI: Intercomparison, interpretation and validation using National Phenology Network and AmeriFlux observations. Ecol. Indicators. 77, 323-336. https://doi.org/10.1016/j.ecolind.2017.02.024.

Peng, D.L., Zhang, X.Y., Zhang, B., Liu, L.Y., Liu, X.J., Huete, A.R., Huang, W.J., Wang, S.Y., Luo, S.Z., Zhang, X., & Zhang, H.L., 2017b. Scaling effects on spring phenology detections from MODIS data at multiple spatial resolutions over the contiguous United States. Isprs J Photogramm. 132, 185-198. https://doi.org/10.1016/j.isprsjprs.2017.09.002.

Pouliot, D., Latifovic, R., Fernandes, R., & Olthof, I., 2011. Evaluation of compositing period and AVHRR and MERIS combination for improvement of spring phenology detection in deciduous forests. Remote Sens. Environ. 115, 158-166. https://doi.org/10.1016/j.rse.2010.08.014.

Richardson, A.D., Hollinger, D.Y., Dail, D.B., Lee, J.T., Munger, J.W., & O'Keefe, J., 2009. Influence of spring phenology on seasonal and annual carbon balance in two contrasting New England forests. Tree Physiol. 29, 321-331. https://doi.org/10.1093/treephys/tpn040.

Richardson, A.D., Hufkens, K., Milliman, T., & Frolking, S., 2018. Intercomparison of phenological transition dates derived from the PhenoCam Dataset V1.0 and MODIS satellite remote sensing. Sci. Rep. 8. https://doi.org/10.1038/s41598-018-23804-6.

Richardson, A.D., & O'Keefe, J., 2009. Phenological Differences Between Understory and Overstory. In: A. Noormets (Eds.), Phenology of Ecosystem Processes: Applications in Global Change Research, pp. 87-117. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-0026-5_4.

Rodriguez-Galiano, V.F., Dash, J., & Atkinson, P.M., 2015. Intercomparison of satellite sensor land surface phenology and ground phenology in Europe. Geophys. Res. Lett. 42, 2253-2260. https://doi.org/10.1002/2015gl063586.

Rosemartin, A.H., Crimmins, T.M., Enquist, C.A.F., Gerst, K.L., Kellermann, J.L., Posthumus, E.E., Denny, E.G., Guertin, P., Marsh, L., & Weltzin, J.F., 2014. Organizing phenological data resources to inform natural resource conservation. Biol. Conserv. 173, 90-97. https://doi.org/10.1016/j.biocon.2013.07.003.

Sakamoto, T., Gitelson, A.A., & Arkebauer, T.J., 2013. MODIS-based corn grain yield estimation model incorporating crop phenology information. Remote Sens. Environ. 131, 215-231. https://doi.org/10.1016/j.rse.2012.12.017.

Schwartz, M.D., & Hanes, J.M., 2010. Intercomparing multiple measures of the onset of spring in eastern North America. IJCli. 30, 1614-1626. https://doi.org/10.1002/joc.2008.

Senior, V.L., Evans, L.C., Leather, S.R., Oliver, T.H., & Evans, K.L., 2020. Phenological responses in a sycamore-aphid-parasitoid system and consequences for aphid population

dynamics: A 20 year case study. Global Change Biol. 26, 2814-2828. https://doi.org/10.1111/gcb.15015.

Soudani, K., le Maire, G., Dufrene, E., Francois, C., Delpierre, N., Ulrich, E., & Cecchini, S., 2008. Evaluation of the onset of green-up in temperate deciduous broadleaf forests derived from Moderate Resolution Imaging Spectroradiometer (MODIS) data. Remote Sens. Environ. 112, 2643-2655. https://doi.org/10.1016/j.rse.2007.12.004.

Sulla-Menashe, D., Gray, J.M., Abercrombie, S.P., & Friedl, M.A., 2019. Hierarchical mapping of annual global land cover 2001 to present: The MODIS Collection 6 Land Cover product. Remote Sens. Environ. 222, 183-194. https://doi.org/10.1016/j.rse.2018.12.013.

Templ, B., Koch, E., Bolmgren, K., Ungersbock, M., Paul, A., Scheifinger, H., Rutishauser, T., Busto, M., Chmielewski, F.M., Hajkova, L., Hodzic, S., Kaspar, F., Pietragalla, B., Romero-Fresneda, R., Tolvanen, A., Vucetic, V., Zimmermann, K., & Zust, A., 2018. Pan European Phenological database (PEP725): a single point of access for European data. IJBm. 62, 1109-1113. https://doi.org/10.1007/s00484-018-1512-8. Data set accessed 2021-09-03.

Tian, F., Cai, Z., Jin, H., Hufkens, K., Scheifinger, H., Tagesson, T., Smets, B., Van Hoolst, R., Bonte, K., Ivits, E., Tong, X., Ardö, J., & Eklundh, L., 2021. Calibrating vegetation phenology from Sentinel-2 using eddy covariance, PhenoCam, and PEP725 networks across Europe. Remote Sens. Environ. 260, 112456. https://doi.org/10.1016/j.rse.2021.112456.

Tian, J.Q., Zhu, X.L., Shen, Z.Y., Wu, J., Xu, S., Liang, Z.C., & Wang, J.T., 2020. Investigating the urban-induced microclimate effects on winter wheat spring phenology using Sentinel-2 time series. Agric. For. Meteorol. 294. https://doi.org/10.1016/j.agrformet.2020.108153.

Verma, M., Friedl, M.A., Finzi, A.C., & Phillips, N., 2016. Multi-criteria evaluation of the suitability of growth functions for modeling remotely sensed phenology. Ecol. Model. 323, 123-132. https://doi.org/10.1016/j.ecolmodel.2015.12.005.

White, M.A., de Beurs, K.M., Didan, K., Inouye, D.W., Richardson, A.D., Jensen, O.P., O'Keefe, J., Zhang, G., Nemani, R.R., van Leeuwen, W.J.D., Brown, J.F., de Wit, A., Schaepman, M., Lin, X.M., Dettinger, M., Bailey, A.S., Kimball, J., Schwartz, M.D., Baldocchi, D.D., Lee, J.T., & Lauenroth, W.K., 2009. Intercomparison, interpretation, and assessment of spring phenology in North America estimated from remote sensing for 1982-2006. Global Change Biol. 15, 2335-2359. https://doi.org/10.1111/j.1365-2486.2009.01910.x.

Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P.S., Vermote, E.F., Vogelmann, J., Wulder, M.A., Wynne, R., & Team, L.S., 2008. Free access to Landsat imagery. Sci. 320, 1011-1011. https://doi.org/10.1126/science.320.5879.1011a.

Wu, C., Wang, J., Ciais, P., Peñuelas, J., Zhang, X., Sonnentag, O., Tian, F., Wang, X., Wang, H., Liu, R., Fu, Y.H., & Ge, Q., 2021. Widespread decline in winds delayed autumn foliar senescence over high latitudes. Proceedings of the National Academy of Sciences. 118, e2015821118. https://doi.org/10.1073/pnas.2015821118.

Wu, C.Y., Hou, X.H., Peng, D.L., Gonsamo, A., & Xu, S.G., 2016. Land surface phenology of China's temperate ecosystems over 1999-2013: Spatial-temporal patterns, interaction effects,

covariation with climate and implications for productivity. Agric. For. Meteorol. 216, 177-187. https://doi.org/10.1016/j.agrformet.2015.10.015.

Xiao, J.F., Chevallier, F., Gomez, C., Guanter, L., Hicke, J.A., Huete, A.R., Ichii, K., Ni, W.J., Pang, Y., Rahman, A.F., Sun, G.Q., Yuan, W.P., Zhang, L., & Zhang, X.Y., 2019. Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. Remote Sens. Environ. 233. https://doi.org/10.1016/j.rse.2019.111383.

Xie, Y.Y., & Wilson, A.M., 2020. Change point estimation of deciduous forest land surface phenology. Remote Sens. Environ. 240. https://doi.org/10.1016/j.rse.2020.111698.

Yan, D., Scott, R.L., Moore, D.J.P., Biederman, J.A., & Smith, W.K., 2019. Understanding the relationship between vegetation greenness and productivity across dryland ecosystems through the integration of PhenoCam, satellite, and eddy covariance data. Remote Sens. Environ. 223, 50-62. https://doi.org/https://doi.org/10.1016/j.rse.2018.12.029.

Zhang, X., Liu, L., Chen, X., Gao, Y., Xie, S., & Mi, J., 2021. GLC_FCS30: global land-cover product with fine classification system at 30m using time-series Landsat imagery. Earth Syst. Sci. Data. 13, 2753-2776. https://doi.org/10.5194/essd-13-2753-2021.

Zhang, X.Y., Friedl, M.A., & & Henebry, G., 2020a. VIIRS/NPP Land Cover Dynamics Yearly L3 Global 500m SIN Grid V001 [Data set]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/VIIRS/VNP22Q2.001 accessed on 14 June 2020.

Zhang, X.Y., Jayavelu, S., Liu, L.L., Friedl, M.A., Henebry, G.M., Liu, Y., Schaaf, C.B., Richardson, A.D., & Gray, J., 2018a. Evaluation of land surface phenology from VIIRS data using time series of PhenoCam imagery. Agric. For. Meteorol. 256-257, 137-149. https://doi.org/10.1016/j.agrformet.2018.03.003.

Zhang, X.Y., Liu, L.L., Liu, Y., Jayavelu, S., Wang, J.M., Moon, M., Henebry, G.M., Friedl, M.A., & Schaaf, C.B., 2018b. Generation and evaluation of the VIIRS land surface phenology product. Remote Sens. Environ. 216, 212-229. https://doi.org/10.1016/j.rse.2018.06.047.

Zhang, X.Y., Wang, J.M., Gao, F., Liu, Y., Schaaf, C., Friedl, M., Yu, Y.Y., Jayavelu, S., Gray, J., Liu, L.L., Yan, D., & Henebry, G.M., 2017. Exploration of scaling effects on coarse resolution land surface phenology. Remote Sens. Environ. 190, 318-330. https://doi.org/10.1016/j.rse.2017.01.001.

Zhang, X.Y., Wang, J.M., Henebry, G.M., & Gao, F., 2020b. Development and evaluation of a new algorithm for detecting 30 m land surface phenology from VIIRS and HLS time series. Isprs J Photogramm. 161, 37-51. https://doi.org/10.1016/j.isprsjprs.2020.01.012.

Zheng, Z.T., Zhu, W.Q., Chen, G.S., Jiang, N., Fan, D.Q., & Zhang, D.H., 2016. Continuous but diverse advancement of spring-summer phenology in response to climate warming across the Qinghai-Tibetan Plateau. Agric. For. Meteorol. 223, 194-202. https://doi.org/10.1016/j.agrformet.2016.04.012.