

Understanding the robustness of spectral-temporal metrics across the global Landsat archive from 1984 to 2019 – a quantitative evaluation

David Frantz^{a,b,*}, Philippe Rufin^{a,c}, Andreas Janz^a, Stefan Ernst^a, Dirk Pflugmacher^a, Franz Schug^d, Patrick Hostert^{a,e}

^a Geography Department, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

^b Geoinformatics – Spatial Data Science, Trier University, 54286 Trier, Germany

^c Earth and Life Institute, UCLouvain, Place Pasteur 3, 1348 Louvain-la-Neuve, Belgium

^d SILVIS Lab, Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, 1630 Linden Drive, Madison, WI 53706, USA

^e Integrated Research Institute on Transformations of Human Environment Systems (IRI THESys), Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

ARTICLE INFO

Keywords:

Landsat

MODIS

Time series

Temporal aggregation

Global

Data availability

ABSTRACT

The Landsat archive is one of the richest Earth observation datasets available and provides long-term data at fairly high temporal and spatial resolution globally. Temporal aggregation is frequently used to condense single observations into a more digestible feature space that provides spatially gap-free data to fulfill demands of many processing strategies that rely on homogeneous data coverage across a large area, e.g., machine learning-based land cover classification. Spectral Temporal Metrics (STMs) represent a conceptually simple feature space wherein multiple observations are temporally aggregated by statistical functions. The quality and inter-annual consistency of STMs is affected by data availability, including usable clear-sky observations that vary in time and space due to satellite lifecycles, sensor failures, changes of observation modes, climate regimes, orbital overlaps, as well as inter-annual variability of cloud cover. However, the relationship between data availability and STM consistency between years is still poorly understood, especially as differences in STMs between years can both result from inter-annual variability in data availability, as well as inter-annual variability of land surfaces. In this study, we systematically quantify the effect of inter-annual data availability on annual STMs for the years 1984–2019, while completely controlling for inter-annual land surface changes. Our results are expected to help assess where on Earth, and in what periods, specific metrics can be used or should be avoided when multi-annual consistency is required. We synthesized a global, nearly gap-free reference time series at daily temporal resolution from MODIS data. This “baseline” was subsequently degraded with actual annual Landsat mission observation scenarios resulting in synthetic annual time series that only differ with respect to data availability. Based on the differences between STMs generated from the baseline, and STMs computed from the degraded time series, we statistically quantified the accuracy, precision, and uncertainty (APU) for various STMs across the Landsat spectral bands. We compared the performance against a reasonable specification, i.e., a tolerated error. We aggregated APU metrics along climate zones annually to carve out regional and temporal differences. We found that huge regional differences exist, with the highest quality and consistency in arid climates (i.e., APU within specification). Errors in fully humid snow climates are high, yet systematic (biased but repeatable), whereas equatorial and temperate climates are characterized by unbiased but uncertain metrics, where accuracy or precision and uncertainty can exceed specification by a factor of three or more. Quality generally increased with time as a response to improved observation modes and data storage commitment, e.g., uncertainty improved from one sensor availability period to the next in >90% of all climate zones for the near infrared average – with the exception of the Landsat 7 scanline corrector failure in 2003 where quality decreased again in 62% of climate zones. We also derived and tested different measures of STM quality and found that the seasonal distribution of clear-sky observations is more important than the number of observations, e.g., the near infrared standard deviation's accuracy can be explained with an R^2 of 0.55, and 0.78 by the number of observations, and maximum time between subsequent observations in *Cfb* climates, respectively. Furthermore, our findings revealed how many observations, or how short the largest gap between consecutive observations must

* Corresponding author at: Geography Department, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

E-mail address: david.frantz@uni-trier.de (D. Frantz).

<https://doi.org/10.1016/j.rse.2023.113823>

Received 31 January 2023; Received in revised form 6 September 2023; Accepted 17 September 2023

Available online 21 September 2023

0034-4257/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

be to still produce reliable metrics (e.g., a maximum gaps of 42–45 days to obtain tolerated uncertainty of the near infrared average and standard deviation in *C/b* climates), which has substantial implications for the design of downstream applications relying on multi-annual STM. This study provides the tools for a global and systematic assessment of inter-annual STM consistency while controlling for land-surface dynamics and thereby paves the way for a systematic error quantification in Level 3 products.

1. Introduction

The Landsat data record is the longest, uninterrupted, cross-calibrated dataset (Czapla-Myers et al., 2015; Markham and Helder, 2012) that offers routine observations from space, and it is consequently one of the key Earth observation missions to monitor and understand global change processes on various spatio-temporal scales since the 1980s (Roy et al., 2014). Every day, 1200 new images were acquired during the Landsat 7 / 8 era (Wulder et al., 2019), which represented an approximate addition of 550 TB per year to the 6.65 PB archive.

Temporal aggregation is a means to condense time series into spatially gap-free data (i.e., Level 3 data as defined in Asrar and Greenstone (1995)) to fulfill demands of many processing strategies that rely on a homogeneous data coverage across a large area, e.g., machine learning-based classification of land cover (Pflugmacher et al., 2019) or regression-based estimation of tree height (Potapov et al., 2021). Different aggregation strategies have emerged over the years, e.g., a plethora of different compositing techniques (see Dennison et al., 2007; Lück and van Niekerk, 2016). In this study, we will focus on so-called *Spectral Temporal Metrics* (STMs), which we here define as “band-wise statistical description of all available high-quality observations within a defined temporal window”. STMs are conceptually simple metrics, wherein all observations of a single band are aggregated with a descriptive statistical function, e.g., mean, standard deviation, or quantiles (also: percentiles). STMs implicitly carry rich information on land surface phenology through temporal variability and data distribution (Frantz, 2019). As STMs represent a conceptually simple feature space that is relatively robust to noise (Zhu, 2017), they are highly suitable for many machine learning applications as they provide a spatially complete and usually gap-free dataset that still includes temporal information. Before applying the statistical function, observations need to be rigorously quality-screened using the per-pixel quality masks of the underlying Level 2 data. Ideally quality masks are non-cloud conservative for this purpose (Skakun et al., 2022), i.e., they minimize omission errors. The risk of remaining contamination with clouds, cloud shadows or snow depends on the choice of the statistical function. Omitted clouds or snow might pollute upper reflectance quantiles (incl. maximum), whereas omitted cloud shadows might affect lower quantiles (incl. minimum), and variability metrics (e.g., standard deviation) are usually more contamination-prone than state-representing metrics like quantiles or the average.

STMs have been extensively used in a broad variety of application domains, usually using classification- or regression-based machine learning techniques to produce qualitative or quantitative land cover maps, or properties of specific land cover classes at large spatial scales (Table 1). The quality of these maps is often driven by the quality of STMs, i.e., their ability to adequately represent land surface reflectance without data gaps or noise while preserving phenological information from a given set of available observations. However, few existing studies systematically evaluate the quality of the used STMs beyond the common approach to present available clear-sky observations (CSO), e.g., by relating the number of CSO (or *good observations*) to mapping quality (e.g., Hansen et al., 2016) or by testing STM robustness by performing a feature set selection or comparison (e.g., Hansen et al., 2005; Pflugmacher et al., 2019; Schug et al., 2020). Among the listed studies in Table 1, relatively few generate maps for multiple time steps, and only few of these relate data quality to mapping quality (e.g., Rufin et al., 2015). Some studies, however, implicitly account for intra-annual data

availability by using gap-filling techniques for STM creation to encounter missing or low-quality data (e.g. time series smoothing in Dara et al., 2018; Rufin et al., 2019, or gap-fill data from adjacent years in Potapov et al., 2015; Potapov et al., 2021). In any way, the quality of STMs can strongly vary spatially and temporally with the number and seasonal distribution of available CSO. In general, sufficient CSOs are required to produce reliable statistics. On an annual basis, the number of CSOs differ due to variability in data density resulting from (de-) commissioning of satellites, sensor failures (e.g., Landsat 7 scanline corrector), historic changes of the observation mode (i.e., acquisition plans, capacity, and storage commitment), and inter-annual variability of cloud cover (see e.g. Zhang et al., 2022). Spatially, CSOs also differ as a function of climate zones with different cloud prevalence. However, how many observations are needed to obtain a specific quality and is this quality consistent across years? Answering these question would be of high importance for users and their downstream applications.

Error quantification of Level 3 products is not as established as for other processing and analyses steps (but see Francini et al., 2023). Rigorous error quantification for Level 2 data, i.e., cloud detection and atmospheric correction has been progressively established in the past years (Doxani et al., 2018, 2023; Skakun et al., 2022), and error quantification in downstream analyses (Level 4) has long been established, e.g., for categorical (Olofsson et al., 2013, 2014) or continuous model output (Pineiro et al., 2008). For the intermediate Level 3 processing step, however, quantitative validation schemes are not used yet, and mostly constrained to simple sanity checks, e.g., visual examination, and/or the depiction of clear-sky observation maps. The latter, however, is only an indicative method as a proportional relationship between CSO quantity and Level 3 quality might not exist in every instance. As an example, more and seasonally well distributed observations are needed for dynamic, vegetated land surfaces (e.g., agriculture, pastures), while less and probably even seasonally clustered observations might suffice for barren surfaces – which are also most susceptible to systematic false positive cloud detections (Ernst et al., 2018).

The goal and scope of this study is to systematically quantify the

Table 1

Major application domains and published studies using Spectral Temporal Metrics for mapping land cover (LC), or properties of specific LC classes – note that this list is non-exhaustive and many other application areas exist.

Application domain	References
Multi-class LC and LC change	(Friedl et al., 2010; Gebhardt et al., 2014; Mack et al., 2017; Masiliūnas et al., 2021; Pflugmacher et al., 2019; Schug et al., 2020; Zhang et al., 2023)
Cropland (extent, change, land management, crop types)	(Deines et al., 2017; Ibrahim et al., 2021; Phalke and Özdoğan, 2018; Rufin et al., 2019, 2021; Waldner et al., 2017; Yin et al., 2018, 2020)
Grassland & pasture (extent, change, use intensity)	(Dara et al., 2018; Müller et al., 2015; Rufin et al., 2015)
Forest inventory (species, tree cover, biomass, canopy height)	(Baumann et al., 2018; Grabska et al., 2020; Hansen et al., 2005, 2016; Hemmerling et al., 2021; Higginbottom et al., 2018; Pasquarella et al., 2018; Potapov et al., 2021; Yin et al., 2017)
Forest change (deforestation, recovery)	(Griffiths et al., 2018; Hansen et al., 2013; Potapov et al., 2009, 2015; Viana-Soto et al., 2022)
Urban (built-up area and building features)	(Frantz et al., 2021; Schneider, 2012; Schneider et al., 2010; Schug et al., 2022)

effect of inter-annual data availability on the generation of annual STMs, while completely controlling for inter-annual changes of the land surface. To achieve this, we synthesize a *reference baseline* – a global, nearly gap-free time series at daily temporal resolution. This baseline is subsequently degraded with the annual observation scenarios of the Landsat mission, resulting in synthetic annual time series that only differ with respect to data availability. With established statistical concepts, we specifically address the following research questions:

- 1) How consistent are annual STMs derived from Landsat observation scenarios from 1984 to 2019?
- 2) Are there regional differences in STM quality across climate zones?
- 3) Which measure of clear-sky observation availability is suited to explain the quality of STMs?

2. Data and methods

2.1. Reference time series

To systematically investigate the research questions, we compiled a reference time series with daily resolution, hereafter denoted as the *baseline* (BL) regarding time series density (Fig. 1). Once established, the BL can be systematically degraded to reflect real-world and historic observation scenarios of medium-resolution sensors like Landsat. This ensures that differences between years can solely be attributed to data availability.

For this purpose, we generated the BL with daily MODIS (Moderate Resolution Imaging Spectroradiometer) data as the nature of this study puts temporal density well before spatial resolution and as these data permit to conduct this study at global scale. Thus, we acquired a complete year of daily MODIS surface reflectance data from both TERRA and AQUA satellites (collection 6 MOD09GA/MYD09GA (Vermote and Wolfe, 2015a, 2015b)), which provides a theoretical 730 observations per year. We selected the year 2015 as it marks the beginning of the warmest five-year period ever (World Meteorological Organization, 2019), and thus we consider it as the new normal climate and land surface status of today. The dataset is provided at 500 m spatial resolution in the sinusoidal tile grid with $10^\circ \times 10^\circ$ tiles. We downloaded near-global coverage from the Land Processes Distributed Active Archive Center (LP DAAC) Data Pool, i.e., 219 tiles with $>0.5\%$ of land coverage – excluding Antarctica.

The reflectance data were spectrally subset to the Landsat legacy bands, i.e., blue, green, red, near infrared and the 2nd shortwave infrared. The 1st shortwave infrared band was excluded from our analyses because it is malfunctioning on Aqua's MODIS (Salomonson and

Appel, 2006). All data were quality masked, i.e., all observations marked as any cloud type, cloud shadow or snow, as well as reflectance $\rho \leq 0$ or $\rho > 1$ were eliminated. To generate a temporally gap-free BL, the combined Aqua and Terra time series of each pixel were converted to daily temporal resolution using Radial Basis Function filter ensembles as described by Schwieder et al. (2016). The observations in the moving kernel are weighted (w) according to a Gaussian distribution:

$$w_t = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-0.5 \cdot \left(\frac{\Delta t}{\sigma}\right)^2\right) \quad (1)$$

where t is the time of the observation (i.e. day-of-year), Δt the observation time relative to the kernel center, and σ the full-width-at-half-maximum of the Gaussian bell-shaped curve. We used three kernels with $\sigma = 4, 8$, and 16 days, whose estimates are eventually aggregated using a weighted average based on the data availability within each kernel. This weighting strategy gives preference to kernels with a higher data density for the final value estimation. We employed a kernel cutoff that preserves 90% of the area under the Gaussian curve – this is done for computational reasons to allow for a fixed and finite kernel that is moved over the time series. The employed Radial Basis Function filtering has been proven suitable in past studies (Frantz et al., 2022) and was primarily chosen because of the advantage of smoothing filters for reconstruction (Zhang et al., 2021) and preservation of intra-seasonal variations (Jönsson et al., 2018).

We further downloaded the MODIS water mask product for 2015 (collection 6 MOD44W (Carroll et al., 2017)), which provides annual maps of surface water at 250 m spatial resolution. The water masks were spatially reduced to 500 m using mode resampling. The water mask was applied to all our analyses.

2.2. Observation scenarios

Real-world *Observation Scenarios* (OS) from the Landsat mission were used to degrade the baseline on an annual basis before computing STMs, i.e., using the same reflectance time series but with the real Landsat data density in the individual years. We downloaded Landsat Collection 1 quality assurance products from Google Earth Engine (Gorelick et al., 2017) and processed them to match the spatial characteristics of the MODIS baseline. Daily binary masks of Landsat CSO were generated between 01 January 1984 and 31 December 2019 for the complete globe – excluding Antarctica. We only used images with $<70\%$ cloud cover, which also is a common threshold applied in many studies (e.g. Frantz et al., 2022; Nill et al., 2022; Schug et al., 2022). We only allowed clear pixels according to the *pixel_qa* band, i.e., we excluded clouds, cloud

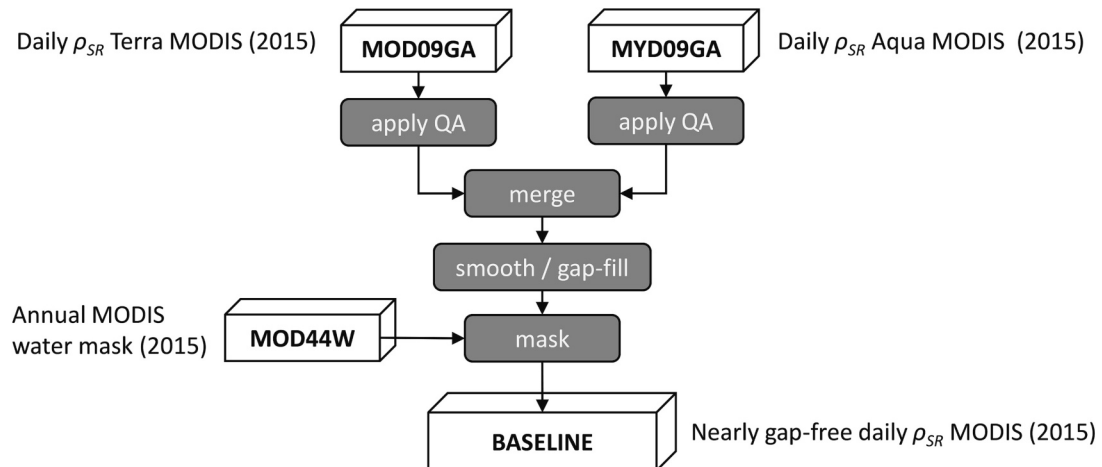


Fig. 1. Workflow for generating the temporally gap-free baseline time series with daily temporal resolution for the year 2015 from MODIS reflectance products. QA = Quality Assurance.

shadows, snow and nodata values. The masks were resampled to the ~500 m spatial resolution in a sinusoidal tile grid with $10^\circ \times 10^\circ$ tiles using nearest neighbor resampling. These binary masks were used with the baseline to reflect the actual Landsat data availability for each day.

2.3. Spectral temporal metrics

We generated a set of annual STMs solely based on the baseline (i.e., without applying any Landsat OS), denoted as STM_{BL} , which we use as reference for the following statistical analyses (Fig. 2). We computed following annual STMs: minimum (MIN), maximum (MAX), several quantiles (Q10, Q25, Q50, Q75, Q90), range (RNG), interquartile range (IRQ), average (AVG), standard deviation (STD), skewness (SKW), and kurtosis (KRT). Subsequently, we iterated over the years $y = [1984 \dots 2019]$ and degraded the baseline with the corresponding OS by applying the Landsat masks of year y , then computed STM_y (Fig. 2). The so-generated STMs, along with the underlying baseline time series, and year-specific OS for two example pixels are visualized in Fig. 3 for the near infrared band. Overall, this procedure generated 2405 images with global extent, i.e., (36 years +1 BL) $\times$ 5 bands $\times$ 13 aggregation functions.

2.4. Statistical analysis

Our statistical analysis centers around an evaluation in terms of accuracy, precision, and uncertainty (APU, Vermote and Kotchenova, 2008), which represents an established and comprehensive evaluation of surface reflectance estimates against a well characterized reference dataset. APU analyses have been extensively used to evaluate MODIS, Landsat, and Sentinel-2 surface reflectance products (e.g. Claverie et al., 2015; Doxani et al., 2018; Vermote et al., 2016; Vermote and Kotchenova, 2008). For surface reflectance, APU metrics are based on the difference between an estimation $\hat{\rho}$ and “truth” ρ for observation i :

$$\varepsilon_i = \hat{\rho}_i - \rho_i. \quad (2)$$

Accuracy A represents the mean error of the estimate versus the truth, thus an error metric that signals whether the estimate is systematically under- or overestimated, i.e., negatively or positively biased. The uncertainty U (in other words: RMSE) represents the statistical deviation of the estimate from the truth. It includes the bias, and hence is a mixture of random and systematic errors; if no bias is present, it only represents random errors. Precision P represents the statistical deviation (random error) around the true values corrected for the bias, i.e., without systematic errors. Thus, P is a measure of the repeatability of the estimate and signals whether the estimates are internally consistent –

even though they might be biased. The combination of APU allows for a separation of systematic and random errors (Fig. 4), wherein U is either driven by poor precision and a small bias (random), good precision and large bias (systematic), or any combination in between. As a rule, both P and A cannot be larger than U .

$$A = \frac{1}{n} \cdot \sum_{i=1}^n \varepsilon_i, \quad (3)$$

$$P = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (\varepsilon_i - A)^2}, \quad (4)$$

$$U = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n \varepsilon_i^2}. \quad (5)$$

Vermote and Kotchenova (2008) additionally derived a specification S for surface reflectance, which they derived from a conservative error budget analysis:

$$S_i = 0.005 + 0.05 \cdot \rho_i. \quad (6)$$

A surface reflectance estimate is considered “good” if ε is lower than its theoretical uncertainty for the true reflectance value ρ . This follows the assumption that a somewhat larger error is more acceptable for high reflectance values, whereas this same error is unreasonably large for low values.

STMs are annual aggregations of surface reflectance values, and therefore, the APU principle, as well as the specification shall be equivocally applicable for all STMs in units of reflectance, e.g., quantiles (i.e. selections of individual observations), as well as the average. For higher moments, we use the specification too, as we assume that a 5% deviation from the truth is a reasonable value. However, it is acknowledged that this assumption is not derived from an error budget analysis as for surface reflectance, thus results should be interpreted with more care as compared to the ones derived for STMs in units of reflectance.

Each STM_y (for any band) can be evaluated against the corresponding STM_{BL} (Fig. 2), which is considered as baseline. As a first step, we compute long-term APU metrics for each pixel individually to identify whether geographical patterns exist; hence n in eq. 3–5 corresponds to the number of years. For a more reliable interpretation with regards to the pixel-specific brightness of the land surface, we normalized the APU statistics in relation to their specification, such that A' , P' , U' values lower than 1 represent pixels that fall within the specification, and values larger than 1 indicate how much the specification was exceeded:

$$A' = |A|/S, \quad (7)$$

$$P' = P/S, \quad (8)$$

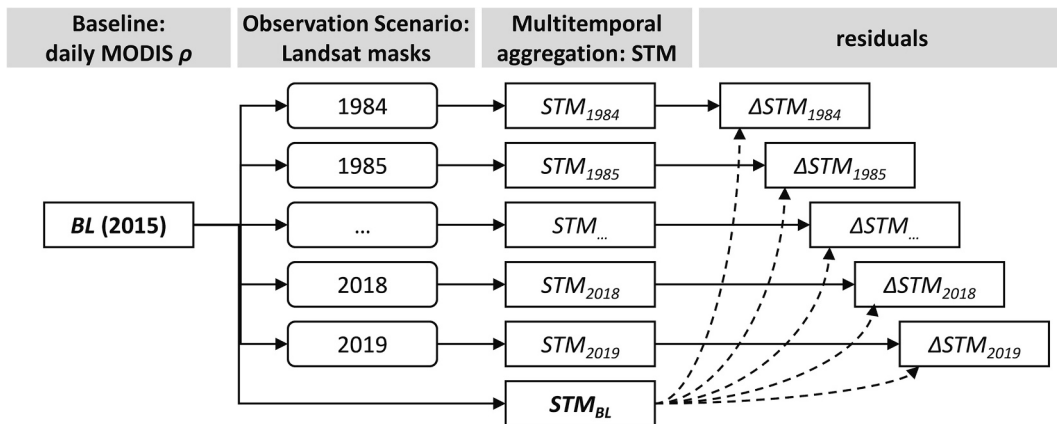


Fig. 2. Analysis workflow. The annual Landsat Observation Scenario is used to degrade the baseline (BL) time series, followed by computing annual sets of Spectral Temporal Metrics (STM_y). Annual STMs are subtracted from STMs based on the baseline, and differences used as basis for the statistical analysis; see accompanying text for a more in-depth description.

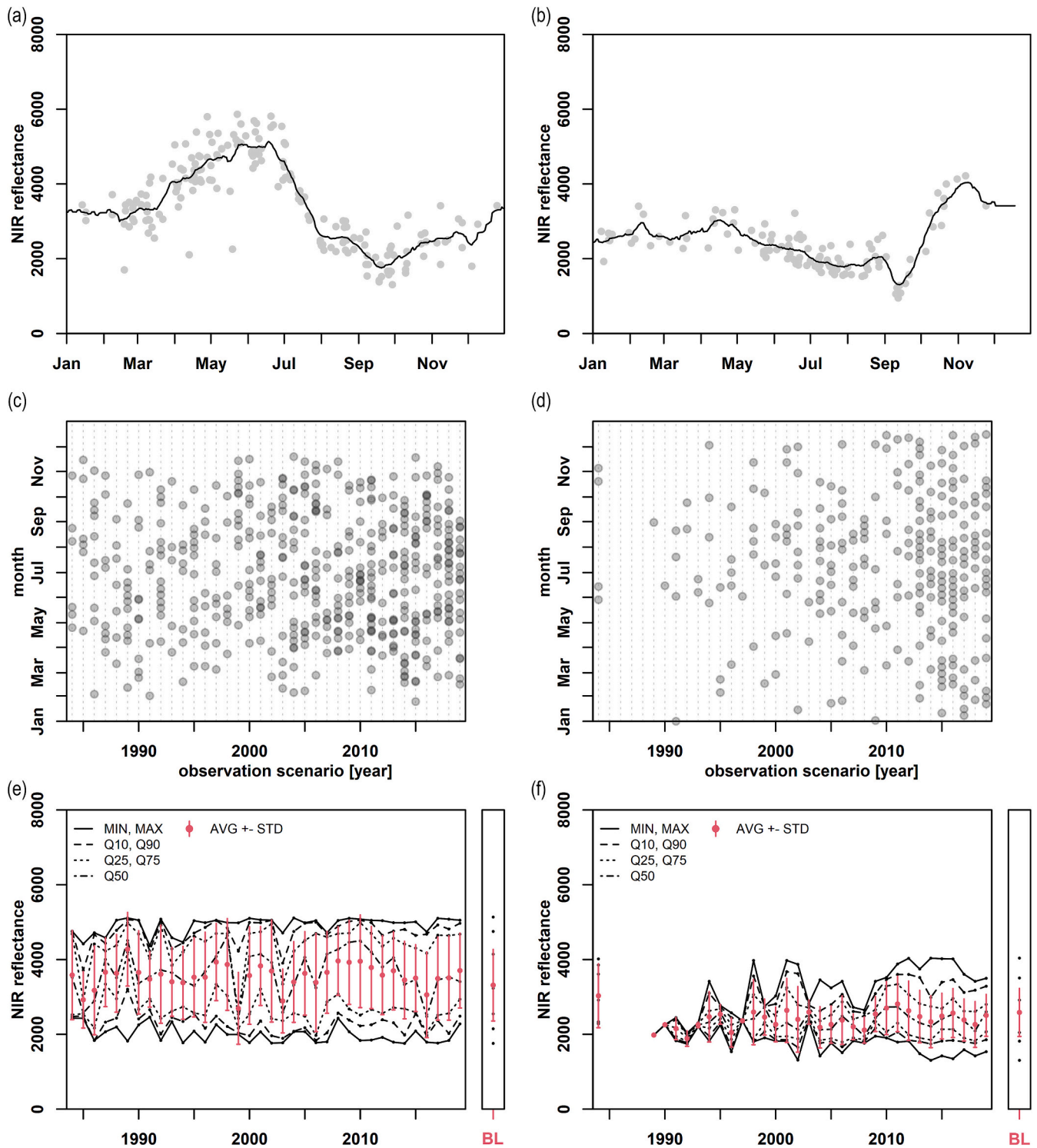


Fig. 3. Illustration of data preparation for two example grassland pixels in North Germany (left column) and the Democratic Republic of Congo (right column). (a-b): original MODIS time series (points) and Radial Basis Function-filtered time series (line); (c-d): Landsat Observation Scenario per year, i.e., date of clear-sky observations; (e-f): Spectral Temporal Metrics (STM) per year, computed from filtered time series (a-b) masked with annual observation scenarios (c-d); the baseline (BL) STM is shown at the far right of the sub-plot for reference.

$$U' = U/S,$$

(9)

2.5. Stratification

Year-specific APU statistics were stratified over spatial and temporal dimensions to further condense the information, hence we analyzed the

outcomes based on two main strata, i.e., (i) climate, and (ii) sensor availability periods:

(i) Climate. To learn about temporal patterns in different regions of the world, we stratified our data along the Köppen-Geiger climate classification after Kottek et al. (2006) using the 29 climate zones depicted in Fig. 5 and defined in Table 2. We used the climate classification of Rubel et al. (2017a, 2017b), which is available at a spatial

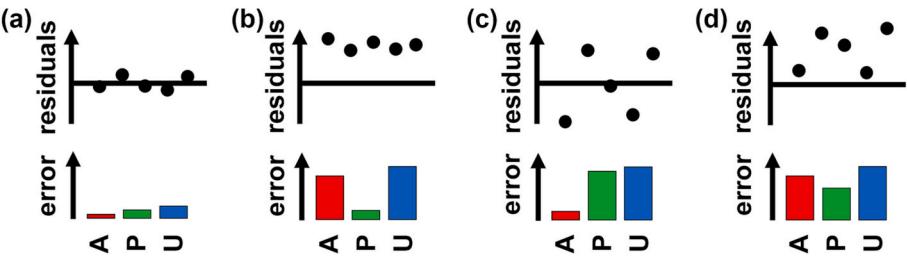


Fig. 4. Schematic illustration of Accuracy (A), Precision (P), and Uncertainty (U) for different types of errors. (a): high quality, (b): systematic, i.e., biased, but precise (repeatable); (c): random, i.e., unbiased, with poor precision; (d): mixture between systematic and random errors, i.e. biased and not precise.

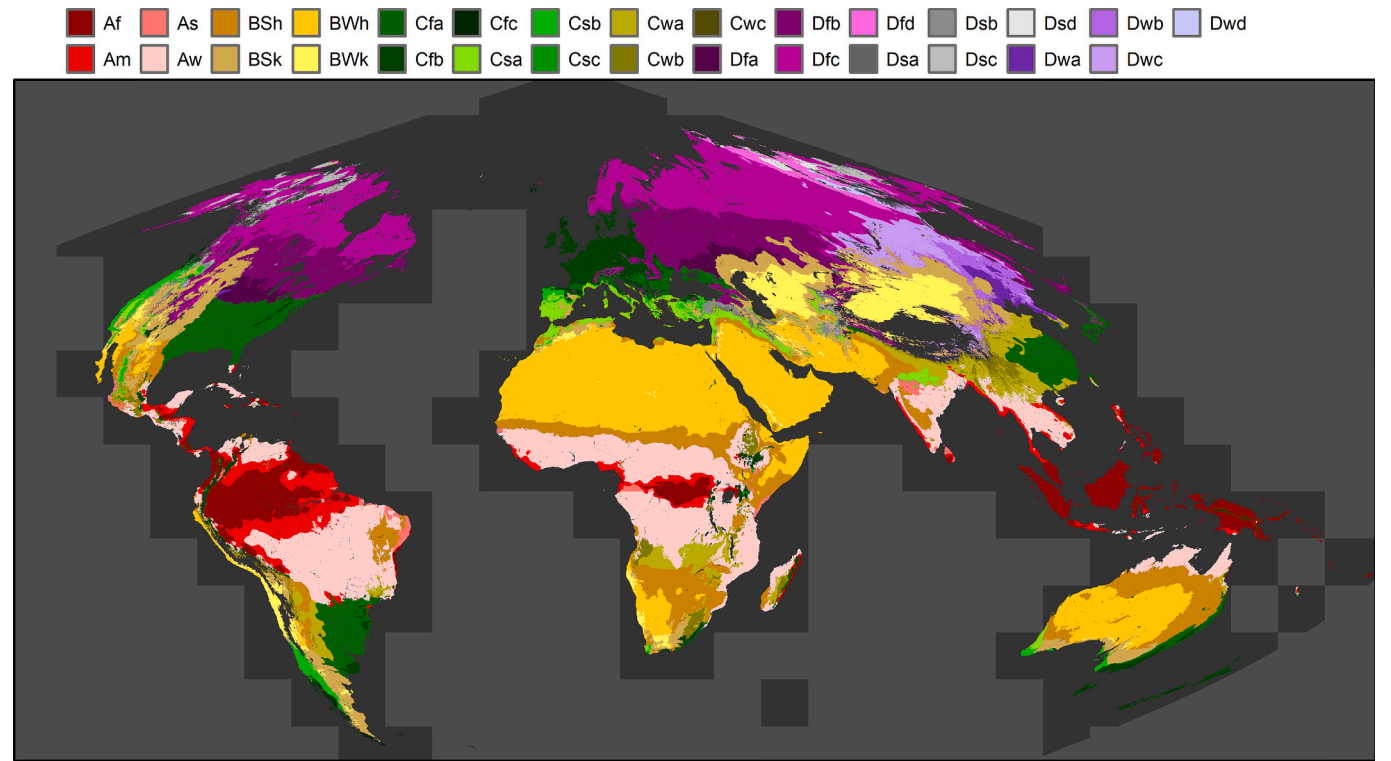


Fig. 5. Köppen-Geiger climate classification after Kottke et al. (2006), excluding polar climates. Dark grey background colors indicate MODIS tiles with >0.5% land coverage.

Table 2
Köppen-Geiger climate classification after Kottke et al. (2006), excluding polar climates.

Main climates	Precipitation	Temperature
A: equatorial	W: desert	h: hot arid
B: arid	S: steppe	k: cold arid
C: warm temperate	f: fully humid	a: hot summer
D: snow	s: summer dry	b: warm summer
	w: winter dry	c: cool summer
	m: monsoonal	d: extremely continental

resolution of 5 arc minutes and is representative for the observational period of 1986–2010. APU statistics for each year were computed considering all MODIS pixels within each climate zone.

(ii) Sensor availability periods. To provide a more synoptic view, we further aggregated APU statistics temporally according to the five main phases of the Landsat program that cover our study period, i.e., the phases where (1) both Landsat 4 and 5, (2) only Landsat 5, (3) both Landsat 5 and 7, (4) both Landsat 5 and SLC-off Landsat 7, (5) and both SLC-off Landsat 7 and Landsat 8 were active.

We further investigated how the annual APU statistics per climate zone correlate with annual measures of data density. We considered two measures, i.e., (1) the annual number of clear sky observations (# of CSO) as a measure of total data availability within a year, as well as (2) the maximum time (in days) between consecutive clear-sky observations within a year (max. dCSO) as a measure for seasonal data distribution.

3. Results

3.1. Spectral temporal metrics for Landsat observation scenarios

Fig. 6 (a,d) displays the AVG and STD metrics for the near infrared (NIR) band for the baseline, as well as the number of clear-sky observations for the baseline (g); a corresponding histogram is shown in Fig. 7. For large parts of the world (57%), the baseline includes 365 observations per pixel; in the remaining world, the baseline achieves a lower temporal density, which is either because of a complete unavailability of snow-free land observations in northern latitudes during the winter season, or persistent cloud coverage in highly convective tropical regions. The residuals of the AVG and STD metrics for the Landsat OS from 1995 and 2015 compared to the baseline (b-c, e-f), as well as the

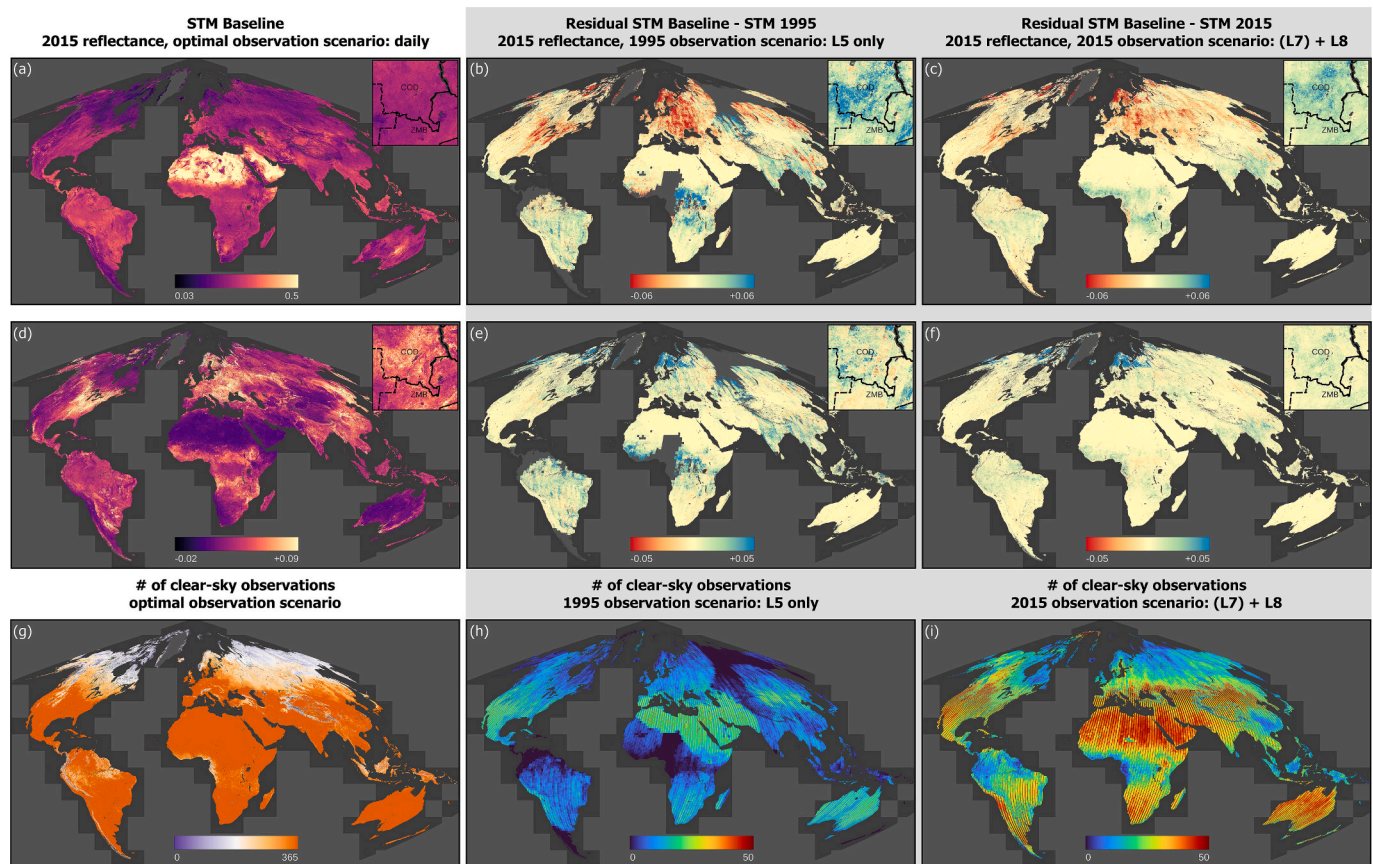


Fig. 6. Spectral Temporal Metrics (STM), annual STM residuals and number of clear sky observations (CSO) for two selected years. (a): near infrared average of the baseline time series; (b): residuals of the near infrared average of the 1995 Landsat observation scenario relative to the optimal observation scenario, i.e. the baseline; (c): same as (b) but for 2015; (d-f): same as (a-c) but for the near infrared standard deviation; (g): number of CSO of the optimal observation scenario, i.e. the baseline; (h): number of CSO for the 1995 Landsat observation scenario; (i) same as (h) but for 2015. Note: the same stretch is applied to (h-i), but a different stretch is applied to (g) because of large differences in values. L5 = Landsat 5, (L7) = Landsat 7 with SLC-off, L8 = Landsat 8. Panels (a-f) additionally depict a zoom-in on the border region of the Democratic Republic of Congo and Zambia in Southern Africa.

number of clear-sky observations are shown in Fig. 6 (h-i); histograms of CSO availability for all years are shown in Fig. 7. Data gaps are visible in the 1995 dataset due to inoperability of parts of the International Ground Stations Network (see <https://landsat.usgs.gov/historical-international-ground-stations/>), wherein the Argentinian, Gabonese, south Australian, and north-east Russian station did not receive data at this time. STM_{1995} was being computed from substantially fewer observations than STM_{2015} , and profound regional differences do exist, e. g., in the dry tropical forests in the Democratic Republic of the Congo, where AVG_{1995} indicates less vegetation vigor as opposed to AVG_{2015} , and even more so when compared to AVG_{BL} . Due to results' high dimensionality and space constraints, we are only able to present a selected subset of the STM within this manuscript.

3.2. Pixel-based, long-term APU

Fig. 8 displays long-term APU statistics relative to the specification for the AVG and STD metrics in the near infrared band; Table 3 gives the percentage of pixels in different performance classes. The average annual reflectance is in better agreement (over the entire time period) with the average baseline reflectance than the standard deviation of annual reflectance. Multiple areas are within specification for AVG (i.e., 71.15%, 61.19%, and 47.30% for A, P, and U, respectively). These include dry areas with comparatively high data density (cf. Fig. 6), but also more humid areas like the U.S. South or East Coast and large parts of South America. There is a regional pattern in the dominant error type, too. As an example, AVG in Eastern Europe is biased overall (Fig. 8 (a))–

but it is comparably precise (Fig. 8 (b)), which means that the error is similar in all years. Contrary, AVG is not biased in many tropical regions, but there is a poor precision (i.e., high scatter with both positive and negative residuals) between years, thus making the metric less repeatable. The STD metric usually exceeds the specification in all regions, i.e., in 72.4%, 98.75%, and 99.58% of all pixels in A, P, and U, respectively.

3.3. Detailed APU for one climate zone

Fig. 9 depicts aggregated, but detailed results for the Cfb climate zone, i.e., warm temperate, fully humid, hot summer, which dominates in Central Europe. Only a subset of results is presented within the manuscript, for all plots, please see the extended supplemental material collection available at <https://zenodo.org/record/8096599> for all other metrics, bands and climate zones. The NIR AVG metric is characterized by fairly consistent APU metrics for the whole range of reflectance. In 1995 (a), accuracy is largely within specification, whereas precision and uncertainty are not. In 2015 (b), error metrics are smaller in general, although precision and uncertainty are still not within (or very close to the) specification for most part of the brightness gradient. Overall results vary between years (c), with larger errors in the earlier years and increasingly better quality towards the most recent period. Generally, accuracy is better than uncertainty and precision, which usually exceed the overall specification.

The NIR STD error metrics increase with the “true” values in both presented years. In 1995 (d), precision and uncertainty are out of specification for the entire range of “true” STD values, whereas accuracy

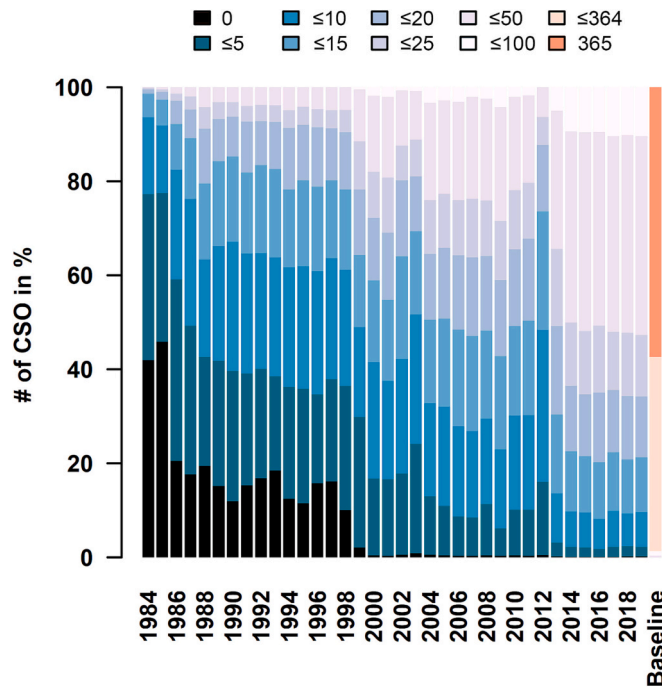


Fig. 7. Histogram of the number of clear-sky observations (CSO) per year, and in the baseline reference time series.

is very close to the threshold. In 2015 (e), accuracy is well within specification, and precision and uncertainty improved substantially compared to 1995, although still not within specification for most STD values. Overall results show a steady improvement in error metrics through time (f). At two distinct points in time, uncertainty and precision improve abruptly. These dates coincide with the launch of Landsat 7 and 8, respectively.

3.4. Synoptic view for all climate zones

To analyze error components over both space and time, Fig. 10 presents aggregations for all climate zones and different key periods of the Landsat mission. For all climate zones and metrics, early periods are

characterized by larger errors than more recent observation periods. In general, the STD metric exceeds specification more often than the AVG metric (see e.g., *BSk* zone), but even the latter is often beyond this threshold. There are large differences between climate zones. Errors in arid climates (*B**) are fairly low, whereas errors in fully humid snow climates (*Df**) are high, yet systematic (biased but precise). Especially equatorial (*A**), but also temperate climates (*C**) are characterized by unbiased, yet more uncertain STM metrics. Interestingly, the North American temperate climate (*Cfa*) is characterized by better APU performance than the *Cfb* climate dominating in Central Europe, even for contemporary data where acquisition capacity factors should be comparable.

3.5. Relationships between data availability measures and STM quality

We were also interested whether the performance of STMs could be explained by some measure of data availability. Fig. 11 (a,c) displays statistical relationships between annual APU metrics and the number of clear-sky observations (#CSO) used to produce the NIR AVG and STD metrics within the *Cfb* climate zone. Linear relationships do exist between each error component and #CSO. The variables are stronger correlated for the STD than for the AVG metrics. Accuracy and precision can be explained worst and best, respectively. In each case, however, relationships are stronger when describing data availability with the maximum time between subsequent clear-sky observations (max- Δ CSO, Fig. 11 (b,d)) as compared to the number of clear-sky observations. The strongest difference was found for the accuracy of the NIR STD, where R^2

Table 3

Summary of pixel-wise APU statistics relative to the pixel's specification for the near infrared average and standard deviation as shown in Fig. 8. The table shows the percentage of pixels in the specific performance class.

Performance class	STM	A	P	U
Within specification	AVG	71.15%	61.19%	47.30%
	STD	27.60%	1.25%	0.51%
Between 1 x and	AVG	17.80%	29.34%	30.61%
2 x specification	STD	18.83%	7.58%	5.30%
Between 2 x and	AVG	6.05%	6.57%	12.82%
3 x specification	STD	14.44%	13.94%	10.30%
Worse than	AVG	5.00%	2.59%	9.28%
3 x specification	STD	39.13%	77.23%	83.98%

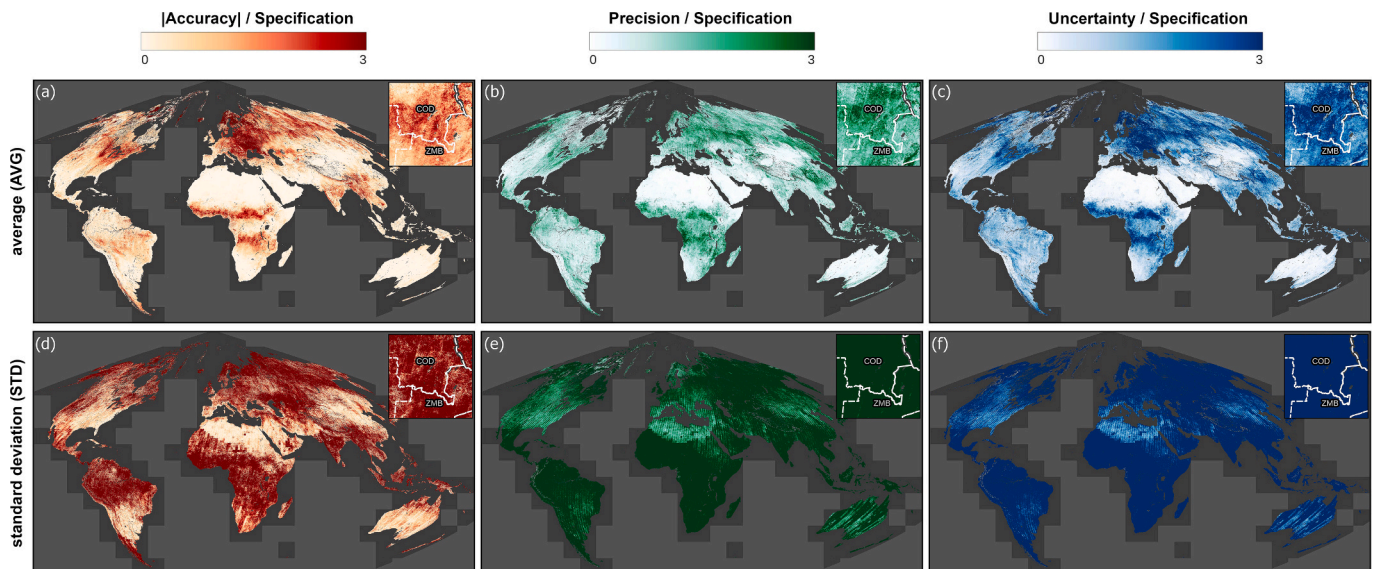


Fig. 8. Pixel-wise APU statistics relative to the pixel's specification for the near infrared average and standard deviation as of eq. 7–9. APU statistics for each pixel were computed along the temporal axis. The same stretch was applied to each plot; saturated colors exceed the specification by a factor 3 or more.

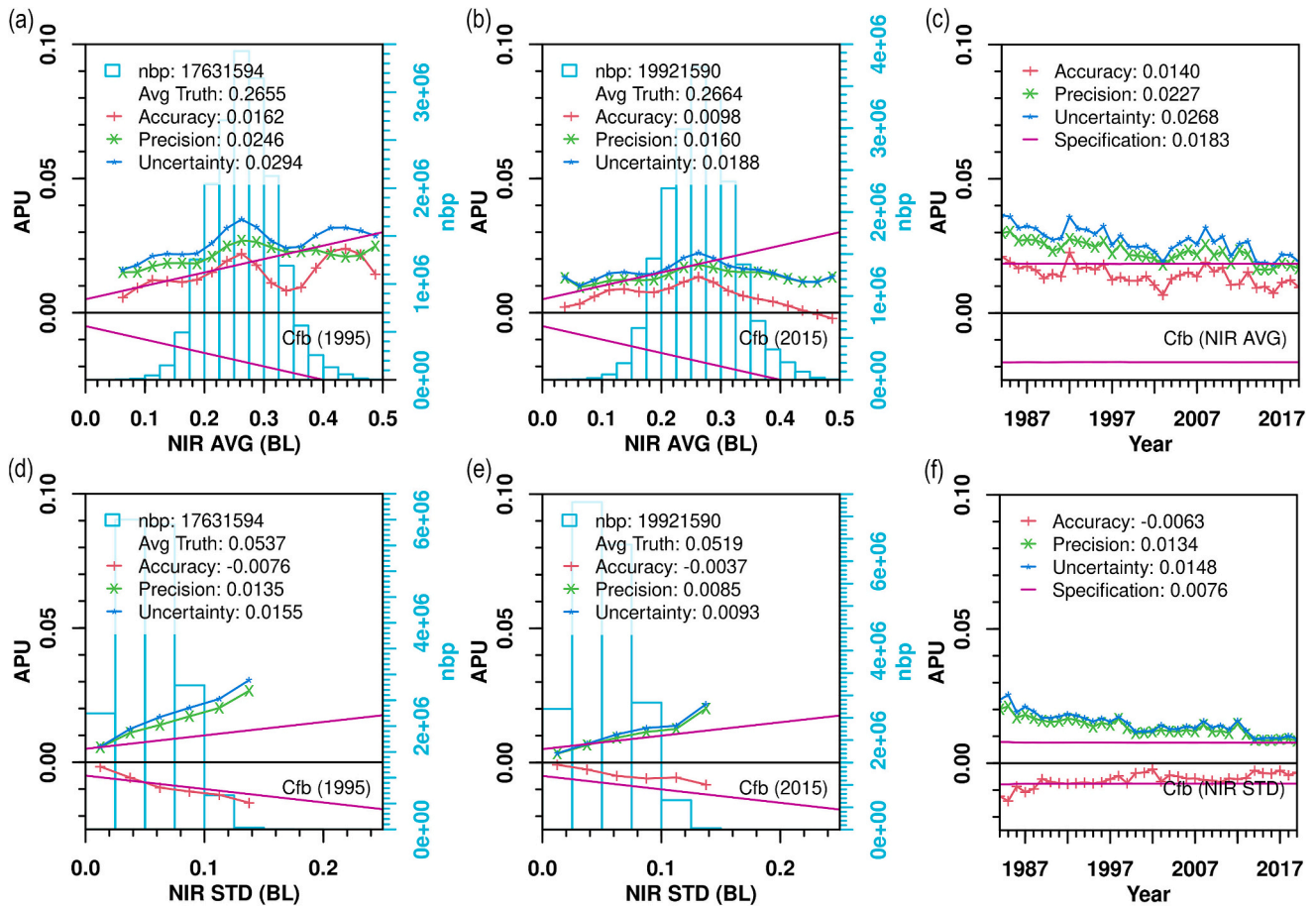


Fig. 9. APU analysis for near infrared STMs in climate zone Cfb. (a-c): average, (d-f): standard deviation. Columns 1–2: detailed APU plots for two selected years, i.e., 1995 and 2015. Column 3: Overall APU for all years (year-specific points represent the overall metrics in the legend of (a-b, d-e). BL = baseline).

increased from 0.55 to 0.78.

The mean specification was plotted as horizontal line in Fig. 11, and gives an indication of how many clear sky observations are required to generate an STM “within specification”, i.e., the point where the magenta line crosses the red, green, or blue regression lines. As an example, to generate a near infrared average STM in climate zone Cfb that has a tolerable uncertainty in relation to the baseline, about 26 clear sky observations (Fig. 11a) and <40 days between all consecutive observations (Fig. 11b) within a year are required. This measure is summarized in Table 4 for all climate zones and both data availability descriptors, i.e., #CSO, and max-ΔCSO. In general, a higher number of CSO and shorter maximum gaps are required to generate STMs with low uncertainty and precision, as compared to accuracy, where fewer CSO and larger gaps frequently produce tolerable metrics.

4. Discussion

4.1. Study design

Our study design provides the first attempt to systematically evaluate the effect of annual data availability on the quality and consistency of higher level EO datasets (i.e., STMs in this study) while completely controlling for inter-annual changes of the land surface. This was enabled by generating a reference time series, i.e., a synthetic baseline with daily temporal resolution based on a single year of hyper-temporal data. This dataset was subsequently degraded by applying the real-life annual observation scenarios from the Landsat mission, after which the time series were aggregated using statistical functions (STMs). As such, we could analyze how consistent the STM feature space is between

years without the vagaries of potential changes of the land surface.

We used 500 m MODIS surface reflectance data from 2015 to evaluate the consistency of STMs globally. This dataset was primarily chosen because of the high temporal resolution that we deemed most important for the sake of this study. Due to the wide swath with viewing angles of up to $\pm 55^\circ$ (Justice et al., 1998), as well as the usage of both Terra (morning) and Aqua (afternoon) observations, strong Bidirectional Reflectance Distribution Function (BRDF) effects exist in the employed data product. However, low-pass filtering (we used Radial Basis Function filter ensembles) is able to smooth the time series while still retaining the annual cycle (Bréon and Vermote, 2012); c.f. Fig. 3 a,b). This reduces the risk that a systematic bias is introduced when sampling the baseline with the acquisition patterns from different Landsat sensors. However, Bréon and Vermote (2012) also noted that such a bias might still be introduced if the underlying observations favor some directions during a given period. Nevertheless, if no smoothing of any kind would have been applied, sampling Landsat 7 observations would usually coincide with near-Nadir observations as Terra MODIS and Landsat 7 are in the same orbit with only 30 min between them (Gao et al., 2006); sampling Landsat 5 or Landsat 8 observations, however, would coincide with more oblique viewing angles and a different overpass time. Hence, we argue that our processing strategy reduces this effect to a large degree, while acknowledging that a specific correction of the daily surface reflectance for anisotropy would further improve consistency and should be considered in future research.

It is further noted that due to using both Terra and Aqua platforms, as well as the temporal smoothing filter, the baseline time series is characterized by an averaged local solar angle that is different from the local solar angle from which an actual Landsat time series is observed. Thus,

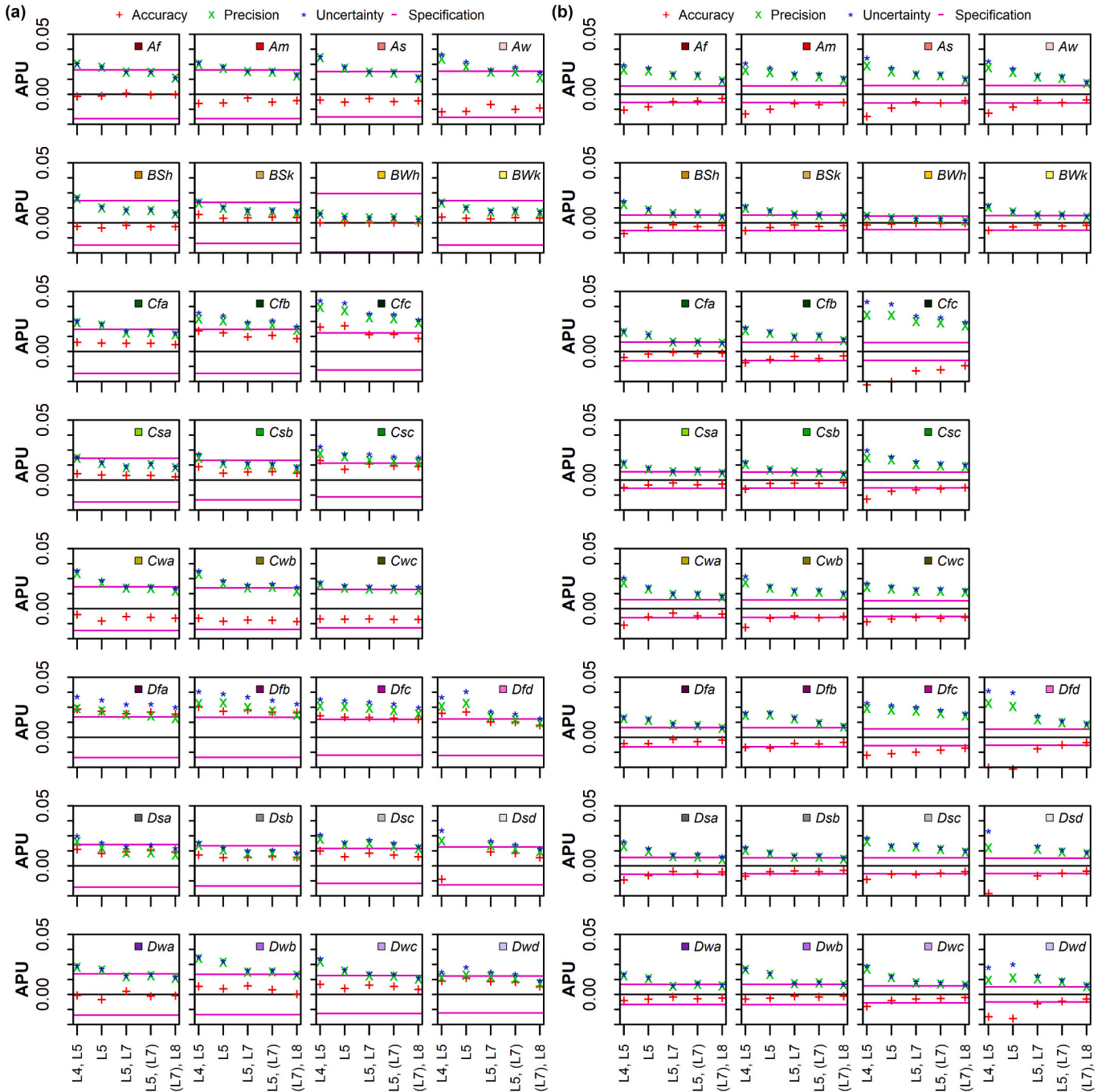


Fig. 10. Synoptic overview of the APU analysis of the near infrared band for the average (a) and standard deviation (b) metrics. The data were analyzed based on the strata defined in section 2.5. Each sub-plot represents one climate zone (see Fig. 5 and Table 2), and results are temporally grouped by sensor availability periods (x-axis), where L5 = Landsat 5, L7 = Landsat 7, (L7) = Landsat 7 with scanline corrector failure, L8 = Landsat 8. A detailed table with all numeric values of all spectral bands and STMs can be found in the supplemental material.

there might be a BRDF-induced absolute bias in reflectance between our simulated time series and an actual Landsat time series. However, our main consideration was to produce a baseline, which is inherently consistent. We assume that data availability-induced residuals would be similar within both individual time series, but a different study design would be needed to verify this.

We here evaluated a commonly used feature space in large-scale EO applications, i.e., annually binned STMs. However, our analysis and baseline data are also suitable to assess the performance of other statistical aggregation functions (note that we only presented results for the average and standard deviation – other aggregation functions are

available in the comprehensive supplemental material), other aggregation windows (e.g., seasonally-binned STMs), or methods (e.g., various pixel-based compositing methods (Qiu et al., 2023), or land surface phenology metrics (e.g. Brooks et al., 2020)) with respect to their inter-annual consistency. We believe that our proposed analysis strategy could also be instrumental in “measuring” the success of commonly used data augmentation or restoration techniques in the data themselves – instead of indirect approaches that compare and validate specific follow-up products. As an example, gap-filling techniques (e.g. Yan and Roy, 2018) are often used to presumably restore the correct data distribution before computing higher level metrics. With our methodology, it could

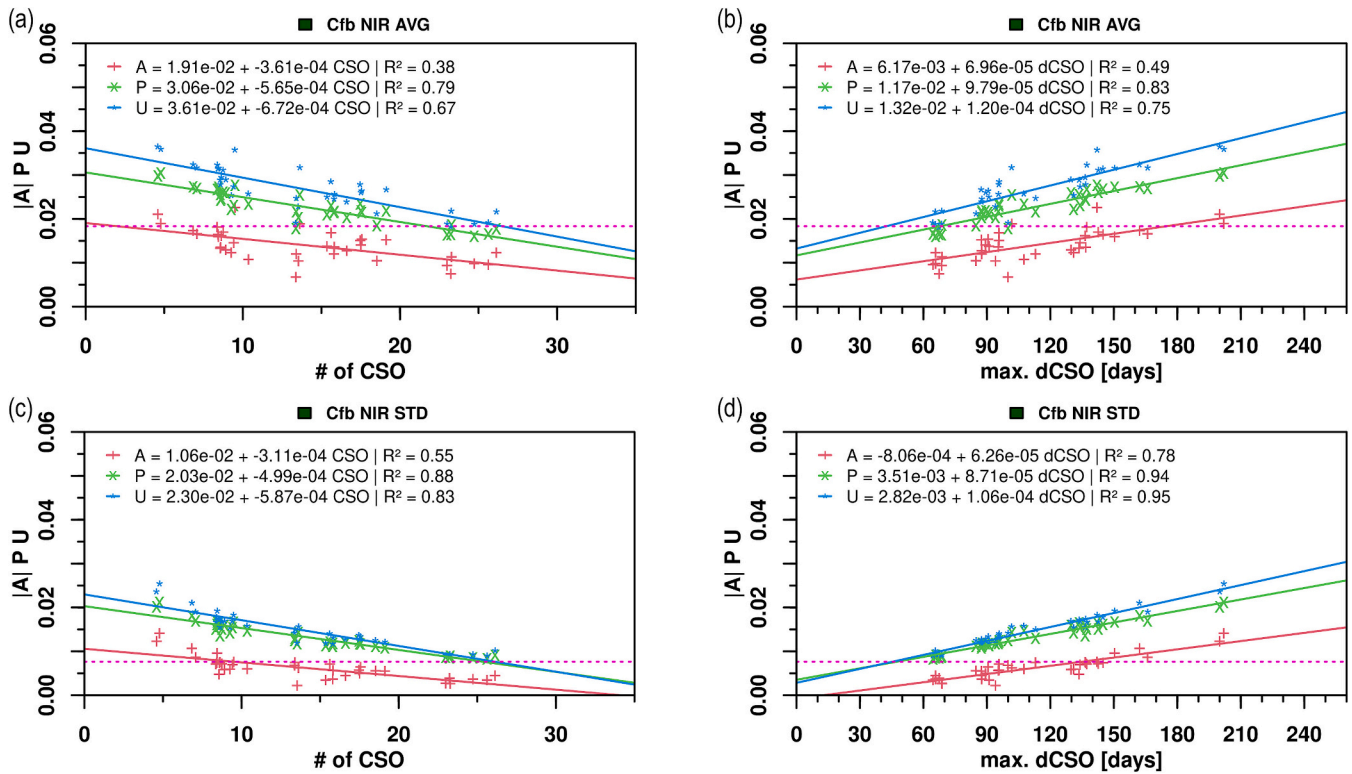


Fig. 11. Left: mean APU vs mean number of clear-sky observations (#CSO) for the NIR AVG (a) and NIR STD (c) metrics. Right: Mean APU vs the mean maximum time between subsequent clear-sky observations (max Δ CSO) for the NIR AVG (b) and NIR STD (d) metrics. All results are for the Cfb climate zone and each data point represents one year, i.e. a year-specific pair of error metrics and data availability. The mean specification for the average reflectance in this climate zone is drawn in magenta. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

be tested whether the feature space converges with that of the baseline, and hence method development could become more data-driven and evidence-based with shorter development cycles.

4.2. Spatio-temporal STM quality

4.2.1. Temporal patterns

The quality of the STMs has substantially improved with time. In some cases, the quality abruptly increased across all climate zones when key historic moments occurred (c.f. Fig. 9, summarized in Table 5), e.g., the commissioning of Landsat 7 along with the instatement of the Long-Term Acquisition Plan (LTAP) and increased downlink capacities (Arvidson et al., 2001, 2006). The commissioning of Landsat 8 marked another milestone, after which nearly all possible images were actually acquired for much of the globe (Roy et al., 2014). Although not investigated in this study, virtual satellite constellations (Wulder et al., 2015), e.g., Landsat plus Sentinel-2 (and also inclusion of Landsat 9 since the end of 2021) will further increase the inter-annual consistency and reduce errors for the most recent past. This is especially important, since Table 5 also reveals that a partial data loss as observed after the failure of Landsat 7's scanline corrector in 2003 (Wulder et al., 2011) resulted in a repeated deterioration of the STM quality in many climate zones. This highlights the fragility of a two-satellite observation setup that produces sufficiently frequent data, but a sensor loss can still push back observational capabilities. To this end, trans-continental partnerships with compatible satellite fleets (like the well-calibrated Landsat and Sentinel-2 missions) should further be strengthened as even newer satellite generations might become dysfunctional at any point in time (e.g., unexpected Sentinel-1B loss (Potin et al., 2022)).

We presume that our assessment is rather conservative with respect to the temporal evolution of STM quality. In our study design, we used the same baseline time series for each year, and only the data

availability was varied. However, data quality and characteristics of sensors have changed over time. Newer sensor generations come with better radiometric and spectral resolution. As an example, the atmospheric correction of Landsat 8 data is expected to be of better quality than that of Landsat 4, as e.g., bit depth was increased to 12 and new bands like the deep-blue band help with aerosol retrieval (Roy et al., 2014; Vermote et al., 2016). Similarly, undetected clouds are expected to occur less frequently in newer data (e.g., due to the addition of the cirrus band (Zhu et al., 2015)). Both aspects might result in larger random errors for older data as compared to contemporary data. In addition, systematic differences between sensors exist, due to different spectral response functions (Markham and Helder, 2012), as well as variability over time within a single sensor that results from orbit change induced solar zenith variations combined with surface reflectance anisotropy (Qiu et al., 2021; Zhang and Roy, 2016). Although both effects can be reduced with bandpass adjustment (Flood, 2014; Roy et al., 2016a) and Bidirectional Reflectance Distribution Function normalization approaches (Roy et al., 2016b), respectively, we assume that they add a temporally variable bias component towards the real-world error budget. The baseline, however, is mostly void of such errors. This is because the baseline time series was generated based on a temporal filtering that resulted in fairly smooth and clean time series (c.f. Fig. 3a, b). As such, our error assessment based on the baseline is overall conservative; if STM are computed from actual Landsat observations, somewhat larger random errors are to be expected. Nevertheless, general conclusions about the temporal data density effect can still be drawn from our study.

We expect that our results have important implications on the temporal transferability of machine learning models, i.e., when a model is trained with reference data from a specific year, and then applied to other years for which no training data is available. Zhang et al. (2023) recently showed that applying models trained on STMs from one year to

Table 4

Synoptic overview of the required data density to produce STMs (AVG and STD) of the near infrared band with tolerable errors for all climate zones; the data density is measured as the crossing point of the mean specification with a linear regression between the mean APU and a descriptor of annual data availability (see Fig. 11). In the left part of the table, the descriptor of data availability is the number of clear-sky observations (#CSO); in the right part, it is the maximum time between subsequent clear-sky observations (max-ΔCSO). “–” indicates instances where the crossing point could not have been derived as the metric was within specification for all years. Bold numbers mark instances where the required #CSO exceeds 45 observations, i.e., more observations than possible with a revisit frequency of 8 days (nominal Nadir temporal resolution of a two-satellite Landsat constellation), or where maximum data gaps (max-ΔCSO) shorter than 30 days are required, i.e., at least one valid observation per month must be available. A more detailed table with all numeric values of all spectral bands and STMs, as well as the R^2 of the regressions can be found in the supplemental material.

Climate	#CSO						max-ΔCSO in days					
	AVG			STD			AVG			STD		
	A	P	U	A	P	U	A	P	U	A	P	U
Af	–	6	6	8	18	17	–	182	182	156	1	22
Am	–	8	10	14	24	23	–	177	163	122	6	31
As	–	13	14	16	29	27	–	147	136	126	2	28
Aw	–	15	21	16	27	27	–	129	88	121	42	45
BSh	–	4	5	9	28	29	–	186	179	151	58	57
BSk	–	2	6	3	27	29	–	185	164	169	68	60
BWh	–	–	–	–	4	6	–	–	–	–	163	152
BWk	–	1	1	4	31	32	–	215	203	164	53	48
Cfa	–	19	21	1	24	24	–	84	72	178	59	58
Cfb	2	22	26	9	25	26	175	68	42	135	47	45
Cfc	8	22	22	17	32	26	187	2	3	78	1	1
Csa	–	6	9	1	31	34	–	135	125	158	51	41
Csb	–	10	18	5	25	27	–	143	111	161	79	71
Csc	1	27	39	24	39	37	270	103	29	126	9	24
Cwa	1	17	20	15	27	27	365	101	82	120	30	36
Cwb	1	18	23	18	31	31	365	98	64	103	1	9
Cwc	–	23	37	28	64	59	–	88	1	68	1	1
Dfa	51	24	47	1	25	26	6	78	13	151	74	72
Dfb	43	24	38	9	21	22	9	77	18	140	86	81
Dfc	15	24	33	19	35	33	105	24	1	82	1	1
Dfd	7	10	14	11	16	15	232	182	128	169	98	110
Dsa	1	8	23	23	28	33	1	156	91	92	69	46
Dsb	–	6	12	13	25	28	–	173	142	138	81	66
Dsc	1	17	22	12	28	28	365	110	67	156	11	15
Dsd	1	7	12	10	25	19	273	206	147	172	1	90
Dwa	1	22	23	1	25	27	243	72	68	157	63	60
Dwb	1	23	24	1	23	23	300	72	68	233	75	74
Dwc	1	14	16	7	18	18	306	123	113	184	95	91
Dwd	1	1	9	10	15	14	365	346	169	171	4	88

Table 5

Percentage of climate zones in which near infrared STM quality improved from one sensor availability period to the next; this analysis is based on summarizing the temporal differences of the data shown in Fig. 10.

–	L4 + L5 → L5		L5 → L5 + L7		L5 + L7 → L5 + (L7)		L5 + (L7) → (L7) + L8	
	AVG	STD	AVG	STD	AVG	STD	AVG	STD
A	64.29	85.71	67.86	96.43	31.03	31.03	89.66	100.00
P	89.29	92.86	96.43	96.43	62.07	79.31	100.00	100.00
U	92.86	92.86	96.43	96.43	37.93	58.62	100.00	100.00

another yields lower classification accuracy than in the base year, and they demonstrated that the drop in accuracy is dependent on the employed metrics. In this regard, before attempting a classification, our study design helps to identify which metrics are temporally robust with respect to a variable data density, while also including a geographic component. Our assessment quantifies the error along with error type. For temporal transfer, systematic errors might be less problematic when the estimate is precise, i.e., repeatable. However, if random errors dominate, i.e., if the uncertainty is high, inter-annual model transferability might be challenged and computing long-term trends might be less reliable. Our approach might be helpful in identifying when, where, and for what metrics inter-annual robustness is high or at least sufficient. As our dataset is void of inter-annual land modifications, i.e., STM_{BL} and STM_y , it would also be beneficial to test this hypothesis directly, e.g., by training classification or regression models that are to be applied to unseen years. This, however, was deemed out of scope of this study but might be worthwhile to investigate in detail in the future.

4.2.2. Geographic patterns

In addition to temporal patterns, distinct geographic patterns were found, too. Some patterns are related to data acquisition strategies, e.g., the LTAP, as well as the effectiveness of international ground stations to downlink and store the data. As an example, the circular-shaped pattern in the Sahara Desert in Fig. 8d-f coincides with the reach of the European ground stations, whereas the Gabonese ground station only captured data for a limited amount of time (see <https://landsat.usgs.gov/historical-international-ground-stations/>).

There are also differences across climate zones. Unsurprisingly, quality in arid climates was rather high, which resembles the overall data distribution patterns displayed in Fig. 6h,i and also average image quality reported by (Remelgado et al., 2023). This is because in arid desert and steppe climates, cloud coverage is generally low, and hence not only a large number of observations are available, but also a seasonally well-distributed data coverage, as well as a fairly stable reflectance trajectory throughout the year due to the lower abundance

of vegetation. As such, the full data distribution can be reasonably well approximated when temporally sampling with the nominal revisit frequency of the Landsat system, allowing for consistent long-term mapping, particularly with narrower seasonal aggregation windows (Rufin et al., 2021, 2022). In fully humid snow climates, errors were rather systematic. We presume that this is an effect of seasonal data distribution. In the baseline time series, there still is a substantial number of usable observations during the snow-free shoulder seasons. With sparser Landsat data, however, the probability of obtaining an image in these seasons is much lower than in the summer. In addition, vegetation development in such climates only happens in summer periods, hence the acquired observations are skewed towards higher reflectance (especially the near infrared reflectance). Although not being evaluated in this study, it might be beneficial to expand our analysis in the future to restrict temporal aggregation to the vegetation period of each pixel to gain additional insight into the interplay of phenology, intra-annual data aggregation windows and resulting metrics quality and consistency.

In equatorial climates, we observed the opposite behavior, i.e., unbiased but uncertain STMs. This is likely related to the relative intra-annual stability of vegetation reflectance around the equator. In addition, the timing of the acquisition is more random than in other climate zones, as cloud coverage is high throughout the year. As such, similar STM estimates can be obtained when having different CSO scenarios with similar acquisition frequency but variable acquisition dates. In temperate zones, there is a bias, which is sensible as data tends to be concentrated in summer months with reduced cloud probabilities – although not as exclusive as in the long snow phases in Df^* climates. We are also observing higher uncertainty, which we attribute to more unpredictable weather conditions in the temperate climates (as compared to more continental climates), where it is rather frequent that increased cloud coverage results in clustered observations in random parts of the growing season (e.g., less data in spring in one year, less data at peak vegetation in another year). An interesting effect was observed when comparing the temperate climates of Central Europe (*Cfb*) and Northern America (*Cfa*): The latter was characterized by better STM quality, which (at least for the near infrared AVG) are within the specification since the 2000s. Presumably, this is because the hot summer climate is less susceptible to the random cluster observations described before, and as such, we note that the application of methods developed in the U.S. might not work equally well in Europe. The negative effects of sparsely and randomly distributed observations in cloud-prone seasons on data quality and inter-annual consistency are in parts inherent to the annual aggregation windows analyzed in this work. Aggregating observations in seasonally constrained temporal windows with a more reliable quantity and distribution of observations represents a common strategy for mitigating such effects (e.g., Ibrahim et al., 2021; Nill et al., 2022). We did not evaluate strategies for seasonal aggregation in this study as the definition of suitable temporal windows is highly context- and application-specific, but suggest users to explore the use of narrower time windows for time series aggregation if appropriate. We underline that particular care about seasonal data density should be taken when STMs are used for large-area, e.g., continental, land cover mapping (e.g., Hansen et al., 2016; Pflugmacher et al., 2019). Here, we offer a quantitative approach to increase transparency.

4.3. How to assess STM quality?

The employed statistical error quantification in terms of an APU analysis allowed us to disentangle systematic and random errors in relation to a tolerated error, i.e., the specification. For the first time, our results quantify these effects, which were previously only qualitatively described by depicting clear-sky observation maps in terms of the number of usable observations to provide an overview of data availability, and to provide an indication for derived data quality (Frantz et al., 2021; Gebhardt et al., 2014; Müller et al., 2015; Rufin et al.,

2019). Our results confirm that a relationship exists between short-comings in STMs and the number of clear-sky observations. However, our study design also allowed us to quantify this relationship. The correlation with uncertainty and precision is comparably high, while the relationship with accuracy is substantially lower. Nevertheless, the intra-annual data distribution is of high relevance for all error components. As such, the relationship between error terms and CSO becomes much stronger when a suitable descriptor of data distribution is used instead of the number of CSOs, e.g., the maximum time between consecutive CSOs within a year. Seasonally well distributed observations are associated with lower errors when compared to a high data density just within a seasonal cluster – especially when vegetation is in focus. As such, the data distribution in seasonally dynamic vegetation surfaces can be much better approximated, which in turn generates statistical predictors that are much closer to the baseline – although derived from a smaller sample of observations. Our results give an indication about how many observations are necessary, and how short the largest gap between consecutive observations are allowed to be to still produce reliable metrics. These results might serve as a basis to evaluate how close a given CSO scenario is to the optimum to justify the choice of years to be analyzed, as well as to inform the choice of metrics in different climate zones to pre-assess whether it is feasible to include higher order features or whether analyses should be restricted to more robust metrics that can be reliably deduced from fewer observations.

5. Conclusion

We systematically quantified the effect of annual data availability within the Landsat archive when temporally aggregating imagery with statistical functions between years. This study is the first attempt to quantify effects of annual data availability while fully controlling for climate and land surface change. The presented results are expected to help assessing where on Earth, and in what periods, specific metrics can be used or should be avoided, as well as to contribute in assessing the quality and reliability of global and multitemporal downstream products. Generalized take home messages include:

- Types of metrics
 - Metrics of central tendency (e.g., average) are of higher quality and consistency than variability metrics (e.g., standard deviation) and require fewer observations to be reliably computed.
- Geographic patterns
 - Arid regions are characterized by the highest quality and consistency between years.
 - Climates with seasonal vegetation development are characterized by biased, but comparatively precise metrics.
 - Tropical climates are characterized by unbiased, but fairly unprecise metrics.
- Temporal patterns
 - The quality of metrics increases with time, especially when satellite constellations change.
 - As an exception, quality slightly deteriorated again when Landsat 7's scanline corrector failure occurred.
- Prediction of quality
 - The quality of metrics can be quantitatively predicted based on data availability.
 - Seasonal data distribution is more important than the number of observations.

We also see a potential in assessing whether temporal transfer of machine learning models to unseen years may be feasible, and where more consideration in model calibration is required. We conclude that the commonly used presentation of the number of clear-sky observation should be accompanied or replaced by other descriptors of data availability that have a stronger relationship to the quality of generated metrics, as e.g., the maximum time between consecutive clear-sky

observations.

CRedit authorship contribution statement

David Frantz: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Philippe Rufin:** Conceptualization, Methodology, Writing – review & editing. **Andreas Janz:** Conceptualization, Methodology, Writing – review & editing. **Stefan Ernst:** Conceptualization, Data curation, Writing – review & editing. **Dirk Pflugmacher:** Writing – review & editing. **Franz Schug:** Writing – review & editing. **Patrick Hostert:** Conceptualization, Project administration, Funding acquisition, Resources, Writing – review & editing.

Declaration of Competing Interest

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

Data availability

The raw satellite data used cannot be shared due to size limitations. All steps to reproduce the analysis will be documented at <https://github.com/davidfrantz/stm-consistency>.

Acknowledgements

We thank NASA EOSDIS Land Processes DAAC for providing MODIS data. We thank USGS, as well as Google Earth Engine for providing Landsat data. We thank the editors and two anonymous reviewers for their constructive feedback that substantially increased the quality of this manuscript. This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 414984028-SFB 1404: FONDA – Foundations of Workflows for Large-Scale Scientific Data Analysis. This study contributes to the Landsat Science Team.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.rse.2023.113823>.

References

- Arvidson, T., Gasch, J., Goward, S.N., 2001. Landsat 7's long-term acquisition plan - an innovative approach to building a global imagery archive. *Remote Sens. Environ.* 78, 13–26. [https://doi.org/10.1016/S0034-4257\(01\)00263-2](https://doi.org/10.1016/S0034-4257(01)00263-2).
- Arvidson, T., Goward, S., Gasch, J., Williams, D., 2006. Landsat-7 long-term acquisition plan: development and validation. *Photogramm. Eng. Remote. Sens.* 72, 1137.
- Asrar, G., Greenstone, R., 1995. *MTPE EOS Reference Handbook*. NASA/Goddard Space Flight Center, Greenbelt, MD, USA.
- Baumann, M., Levers, C., Macchi, L., Bluhm, H., Waske, B., Gasparri, N.I., Kuemmerle, T., 2018. Mapping continuous fields of tree and shrub cover across the Gran Chaco using Landsat 8 and Sentinel-1 data. *Remote Sens. Environ.* 216, 201–211. <https://doi.org/10.1016/j.rse.2018.06.044>.
- Bréon, F.-M., Vermote, E., 2012. Correction of MODIS surface reflectance time series for BRDF effects. *Remote Sens. Environ.* 125, 1–9. <https://doi.org/10.1016/j.rse.2012.06.025>.
- Brooks, B.-G.J., Lee, D.C., Pomara, L.Y., Hargrove, W.W., 2020. Monitoring broadscale vegetational diversity and change across North American landscapes using land surface phenology. *Forests* 11. <https://doi.org/10.3390/f11060606>.
- Carroll, M., DiMiceli, C., Wooten, M., Hubbard, A., Sohlberg, R., Townshend, J., 2017. MOD44W MODIS/Terra Land Water Mask Derived from MODIS and SRTM L3 Global 250m SIN Grid V006. NASA EOSDIS Land Processes DAAC.
- Claverie, M., Vermote, E.F., Franch, B., Masek, J.G., 2015. Evaluation of the Landsat-5 TM and Landsat-7 ETM+ surface reflectance products. *Remote Sens. Environ.* 169, 390–403. <https://doi.org/10.1016/j.rse.2015.08.030>.
- Czapla-Myers, J., McCorkel, J., Anderson, N., Thome, K., Biggar, S., Helder, D., Aaron, D., Leigh, L., Mishra, N., 2015. The ground-based absolute radiometric calibration of Landsat 8 OLI. *Remote Sens.* 7, 600–626.
- Dara, A., Baumann, M., Kuemmerle, T., Pflugmacher, D., Rabe, A., Griffiths, P., Hölzel, N., Kamp, J., Freitag, M., Hostert, P., 2018. Mapping the timing of cropland abandonment and recultivation in northern Kazakhstan using annual landsat time series. *Remote Sens. Environ.* 213, 49–60. <https://doi.org/10.1016/j.rse.2018.05.005>.
- Deines, J.M., Kendall, A.D., Hyndman, D.W., 2017. Annual irrigation dynamics in the U. S. Northern High Plains derived from Landsat satellite data. *Geophys. Res. Lett.* 44, 9350–9360. <https://doi.org/10.1002/2017GL074071>.
- Dennison, P.E., Roberts, D.A., Peterson, S.H., 2007. Spectral shape-based temporal compositing algorithms for MODIS surface reflectance data. *Remote Sens. Environ.* 109, 510–522. <https://doi.org/10.1016/j.rse.2007.02.009>.
- Doxani, G., Vermote, E., Roger, J.-C., Gascon, F., Adriaensen, S., Frantz, D., Hagolle, O., Hollstein, A., Kirches, G., Li, F., Louis, J., Mangin, A., Pahlevan, N., Pflug, B., Vanhellemont, Q., 2018. Atmospheric correction inter-comparison exercise. *Remote Sens.* 10, 352. <https://doi.org/10.3390/rs10020352>.
- Doxani, G., Vermote, E.F., Roger, J.-C., Skakun, S., Gascon, F., Collison, A., Keukelaere, L.D., Desjardins, C., Frantz, D., Hagolle, O., Kim, M., Louis, J., Pacifici, F., Pflug, B., Poilvé, H., Ramon, D., Richter, R., Yin, F., 2023. Atmospheric correction inter-comparison exercise, ACIX-II land: an assessment of atmospheric correction processors for landsat 8 and Sentinel-2 over land. *Remote Sens. Environ.* 285, 113412. <https://doi.org/10.1016/j.rse.2022.113412>.
- Ernst, S., Lymburner, L., Sixsmith, J., 2018. Implications of pixel quality flags on the observation density of a continental Landsat archive. *Remote Sens.* 10, 1570. <https://doi.org/10.3390/rs10101570>.
- Flood, N., 2014. Continuity of reflectance data between Landsat-7 ETM+ and Landsat-8 OLI, for both top-of-atmosphere and surface reflectance: a study in the Australian Landscape. *Remote Sens.* 6, 7952–7970. <https://doi.org/10.3390/rs6097952>.
- Francini, S., Hermosilla, T., Coops, N.C., Wulder, M.A., White, J.C., Chirici, G., 2023. An assessment approach for pixel-based image composites. *ISPRS J. Photogramm. Remote Sens.* 202, 1–12. <https://doi.org/10.1016/j.isprsjprs.2023.06.002>.
- Frantz, D., 2019. FORCE—Landsat + Sentinel-2 analysis ready data and beyond. *Remote Sens.* 11, 1124. <https://doi.org/10.3390/rs11091124>.
- Frantz, D., Hostert, P., Rufin, P., Ernst, S., Röder, A., van der Linden, S., 2022. Revisiting the past: replicability of a historic long-term vegetation dynamics assessment in the era of big data analytics. *Remote Sens.* 14. <https://doi.org/10.3390/rs14030597>.
- Frantz, D., Schug, F., Okujeni, A., Navacchi, C., Wagner, W., van der Linden, S., Hostert, P., 2021. National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series. *Remote Sens. Environ.* 252, 112128. <https://doi.org/10.1016/j.rse.2020.112128>.
- Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., Huang, X., 2010. MODIS collection 5 global land cover: algorithm refinements and characterization of new datasets. *Remote Sens. Environ.* 114, 168–182. <https://doi.org/10.1016/j.rse.2009.08.016>.
- Gao, F., Masek, J., Schwaller, M., Hall, F., 2006. On the blending of the landsat and MODIS surface reflectance: predicting daily landsat surface reflectance. *IEEE Trans. Geosci. Remote Sens.* 44, 2207–2218. <https://doi.org/10.1109/tgrs.2006.872081>.
- Gebhardt, S., Wehrmann, T., Ruiz, M., Maeda, P., Bishop, J., Schramm, M., Kopeinig, R., Cartus, O., Kellndorfer, J., Ressl, R., Santos, L., Schmidt, M., 2014. MAD-MEX: Automatic wall-to-wall land cover monitoring for the Mexican REDD-MRV program using all Landsat data. *Remote Sens.* 6, 3923–3943.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202, 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>.
- Grabska, E., Frantz, D., Ostapowicz, K., 2020. Evaluation of machine learning algorithms for forest stand species mapping using Sentinel-2 imagery and environmental data in the polish Carpathians. *Remote Sens. Environ.* 251, 112103. <https://doi.org/10.1016/j.rse.2020.112103>.
- Griffiths, P., Jakimow, B., Hostert, P., 2018. Reconstructing long term annual deforestation dynamics in Pará and Mato Grosso using the Landsat archive. *Remote Sens. Environ.* 216, 497–513. <https://doi.org/10.1016/j.rse.2018.07.010>.
- Hansen, M.C., Potapov, P.V., Goetz, S.J., Turubanova, S., Tyukavina, A., Krylov, A., Kommareddy, A., Egorov, A., 2016. Mapping tree height distributions in Sub-Saharan Africa using Landsat 7 and 8 data. *Remote Sens. Environ.* 185, 221–232. <https://doi.org/10.1016/j.rse.2016.02.023>.
- Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. *Science* 342, 850–853. <https://doi.org/10.1126/science.1244693>.
- Hansen, M.C., Townshend, J.R.G., DeFries, R.S., Carroll, M., 2005. Estimation of tree cover using MODIS data at global, continental and regional/local scales. *Int. J. Remote Sens.* 26, 4359–4380. <https://doi.org/10.1080/01431160500113435>.
- Hemmerling, J., Pflugmacher, D., Hostert, P., 2021. Mapping temperate forest tree species using dense Sentinel-2 time series. *Remote Sens. Environ.* 267, 112743. <https://doi.org/10.1016/j.rse.2021.112743>.
- Higginbottom, T.P., Symeonakis, E., Meyer, H., van der Linden, S., 2018. Mapping fractional woody cover in semi-arid savannahs using multi-seasonal composites from Landsat data. *ISPRS J. Photogramm. Remote Sens.* 139, 88–102. <https://doi.org/10.1016/j.isprsjprs.2018.02.010>.
- Ibrahim, E.S., Rufin, P., Nill, L., Kamali, B., Nendel, C., Hostert, P., 2021. Mapping crop types and cropping systems in Nigeria with Sentinel-2 imagery. *Remote Sens.* 13. <https://doi.org/10.3390/rs13173523>.
- Jönsson, P., Cai, Z., Melaas, E., Friedl, M.A., Eklundh, L., 2018. A method for robust estimation of vegetation seasonality from landsat and Sentinel-2 time series data. *Remote Sens.* 10. <https://doi.org/10.3390/rs10040635>.

- Justice, C.O., Vermote, E., Townshend, J.R.G., DeFries, R., Roy, D.P., Hall, D.K., Salomonson, V.V., Privette, J.L., Riggs, G., Strahler, A., Lucht, W., Myneni, R.B., Knyazikhin, Y., Running, S.W., Nemani, R.R., Zhengming, W., Huete, A.R., Van Leeuwen, W., Wolfe, R.E., Giglio, L., Muller, J.P., Lewis, P., Barnsley, M.J., 1998. The moderate resolution imaging spectroradiometer (MODIS): land remote sensing for global change research. *IEEE Trans. Geosci. Remote Sens.* 36, 1228–1249. <https://doi.org/10.1109/36.701075>.
- Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World map of the Köppen-Geiger climate classification updated. *Meteorol. Z.* 15, 259–263. <https://doi.org/10.1127/0941-2948/2006/0130>.
- Lück, W., van Niekerk, A., 2016. Evaluation of a rule-based compositing technique for Landsat-5 TM and Landsat-7 ETM+ images. *Int. J. Appl. Earth Obs. and Geoinf.* 47, 1–14. <https://doi.org/10.1016/j.jag.2015.11.019>.
- Mack, B., Leinenkugel, P., Kuenzer, C., Dech, S., 2017. A semi-automated approach for the generation of a new land use and land cover product for Germany based on Landsat time-series and Lucas in-situ data. *Remote Sens. Lett.* 8, 244–253. <https://doi.org/10.1080/2150704X.2016.1249299>.
- Markham, B.L., Helder, D.L., 2012. Forty-year calibrated record of Earth-reflected radiance from Landsat: a review. *Remote Sens. Environ.* 122, 30–40. <https://doi.org/10.1016/j.rse.2011.06.026>.
- Masilūnas, D., Tsendbazar, N.-E., Herold, M., Lesiv, M., Buchhorn, M., Verbesselt, J., 2021. Global land characterisation using land cover fractions at 100 m resolution. *Remote Sens. Environ.* 259, 112409 <https://doi.org/10.1016/j.rse.2021.112409>.
- Müller, H., Rufin, P., Griffiths, P., Barros Siqueira, A.J., Hostert, P., 2015. Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian savanna landscape. *Remote Sens. Environ.* 156, 490–499. <https://doi.org/10.1016/j.rse.2014.10.014>.
- Nill, L., Grünberg, I., Ullmann, T., Gessner, M., Boike, J., Hostert, P., 2022. Arctic shrub expansion revealed by landsat-derived multitemporal vegetation cover fractions in the Western Canadian Arctic. *Remote Sens. Environ.* 281, 113228 <https://doi.org/10.1016/j.rse.2022.113228>.
- Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* 148, 42–57. <https://doi.org/10.1016/j.rse.2014.02.015>.
- Olofsson, P., Foody, G.M., Stehman, S.V., Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sens. Environ.* 129, 122–131. <https://doi.org/10.1016/j.rse.2012.10.031>.
- Pasquarella, V.J., Holden, C.E., Woodcock, C.E., 2018. Improved mapping of forest type using spectral-temporal Landsat features. *Remote Sens. Environ.* 210, 193–207. <https://doi.org/10.1016/j.rse.2018.02.064>.
- Pflugmacher, D., Rabe, A., Peters, M., Hostert, P., 2019. Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey. *Remote Sens. Environ.* 221, 583–595. <https://doi.org/10.1016/j.rse.2018.12.001>.
- Phalke, A.R., Özdoğan, M., 2018. Large area cropland extent mapping with Landsat data and a generalized classifier. *Remote Sens. Environ.* 219, 180–195. <https://doi.org/10.1016/j.rse.2018.09.025>.
- Piñeiro, G., Perelman, S., Guerschman, J.P., Paruelo, J.M., 2008. How to evaluate models: observed vs. predicted or predicted vs. observed? *Ecol. Model.* 216, 316–322. <https://doi.org/10.1016/j.ecolmodel.2008.05.006>.
- Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Armston, J., Dubayah, R., Blair, J. B., Hofton, M., 2021. Mapping global forest canopy height through integration of GEDI and landsat data. *Remote Sens. Environ.* 253, 112165 <https://doi.org/10.1016/j.rse.2020.112165>.
- Potapov, P., Hansen, Matthew, Stehman, Stephen V., Pittman, Kyle, Turubanova, Svetlana, 2009. Gross forest cover loss in temperate forests: biome-wide monitoring results using MODIS and Landsat data. *APPRES* 3, 1–23. <https://doi.org/10.1117/1.3283904>.
- Potapov, P., Turubanova, S.A., Tyukavina, A., Krylov, A.M., McCarty, J.L., Radeloff, V.C., Hansen, M.C., 2015. Eastern Europe's forest cover dynamics from 1985 to 2012 quantified from the full landsat archive. *Remote Sens. Environ.* 159, 28–43. <https://doi.org/10.1016/j.rse.2014.11.027>.
- Potin, P., Colin, O., Pinheiro, M., Rosich, B., O'Connell, A., Ormston, T., Gratadour, J.-B., Torres, R., 2022. Status and evolution of the Sentinel-1 mission. In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. Presented at the IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, pp. 4707–4710. <https://doi.org/10.1109/IGARSS46834.2022.9884753>.
- Qiu, S., Zhu, Z., Olofsson, P., Woodcock, C.E., Jin, S., 2023. Evaluation of Landsat image compositing algorithms. *Remote Sens. Environ.* 285, 113375 <https://doi.org/10.1016/j.rse.2022.113375>.
- Qiu, S., Zhu, Z., Shang, R., Crawford, C.J., 2021. Can Landsat 7 preserve its science capability with a drifting orbit? *Sci. Remote Sens.* 4, 100026 <https://doi.org/10.1016/j.srs.2021.100026>.
- Remelgado, R., Conrad, C., Meyer, C., 2023. Limitations in historical satellite archives bias SDG monitoring. *EarthArXiv* (pre-print). <https://doi.org/10.31223/X5QH37>.
- Roy, D.P., Kovalsky, V., Zhang, H.K., Vermote, E.F., Yan, L., Kumar, S.S., Egorov, A., 2016a. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. *Remote Sens. Environ.* 185, 57–70. <https://doi.org/10.1016/j.rse.2015.12.024>.
- Roy, D.P., Wulder, M.A., Loveland, T.R.C.E.W., Allen, R.G., Anderson, M.C., Helder, D., Irons, J.R., Johnson, D.M., Kennedy, R., Scambos, T.A., Schaaf, C.B., Schott, J.R., Sheng, Y., Vermote, E.F., Belward, A.S., Bindshadler, R., Cohen, W.B., Gao, F., Hipple, J.D., Hostert, P., Huntington, J., Justice, C.O., Kilic, A., Kovalsky, V., Lee, Z. P., Lymburner, L., Masek, J.G., McCorkel, J., Shuai, Y., Trezza, R., Vogelmann, J., Wynne, R.H., Zhu, Z., 2014. Landsat-8: science and product vision for terrestrial global change research. *Remote Sens. Environ.* 145, 154–172. <https://doi.org/10.1016/j.rse.2014.02.001>.
- Roy, D.P., Zhang, H.K., Ju, J., Gomez-Dans, J.L., Lewis, P.E., Schaaf, C.B., Sun, Q., Li, J., Huang, H., Kovalsky, V., 2016b. A general method to normalize Landsat reflectance data to nadir BRDF adjusted reflectance. *Remote Sens. Environ.* 176, 255–271. <https://doi.org/10.1016/j.rse.2016.01.023>.
- Rubel, F., Brugger, K., Haslinger, K., Auer, I., 2017a. The climate of the European Alps: shift of very high resolution Köppen-Geiger climate zones 1800–2100. *Meteorol. Z.* 26, 115–125. <https://doi.org/10.1127/metz/2016/0816>.
- Rubel, F., Brugger, K., Haslinger, K., Auer, I., 2017. Re-analyzed Köppen-Geiger high resolution map 1986–2010. Climate Change & Infectious Diseases Group, Institute for Veterinary Public Health, Vetmeduni Vienna, Austria.
- Rufin, P., Frantz, D., Ernst, S., Rabe, A., Griffiths, P., Özdoğan, M., Hostert, Patrick, 2019. Mapping cropping practices on a National Scale using Intra-Annual Landsat Time Series Binning. *Remote Sens.* 11, 232. <https://doi.org/10.3390/rs11030232>.
- Rufin, P., Müller, D., Schwieder, M., Pflugmacher, D., Hostert, P., 2021. Landsat time series reveal simultaneous expansion and intensification of irrigated dry season cropping in southeastern Turkey. *J. Land Use Sci.* 16, 94–110. <https://doi.org/10.1080/1747423X.2020.1858198>.
- Rufin, P., Müller, H., Pflugmacher, D., Hostert, P., 2015. Land use intensity trajectories on amazonian pastures derived from landsat time series. *Int. J. Appl. Earth Obs. and Geoinf.* 41, 1–10. <https://doi.org/10.1016/j.jag.2015.04.010>.
- Rufin, P., Peña-Guerrero, M.D., Umirbekov, A., Wei, Y., Müller, D., 2022. Post-soviet changes in cropping practices in the irrigated drylands of the Aral Sea basin. *Environ. Res. Lett.* 17, 095013 <https://doi.org/10.1088/1748-9326/ac8daa>.
- Salomonson, V., Appel, I., 2006. Development of the aqua MODIS NDSI fractional snow cover algorithm and validation results. *Geosci. Remote Sens. IEEE Trans.* 44, 1747–1756. <https://doi.org/10.1109/tgrs.2006.876029>.
- Schneider, A., 2012. Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. *Remote Sens. Environ.* 124, 689–704. <https://doi.org/10.1016/j.rse.2012.06.006>.
- Schneider, A., Friedl, M.A., Potere, D., 2010. Mapping global urban areas using MODIS 500-m data: new methods and datasets based on 'urban ecoregions'. *Remote Sens. Environ.* 114, 1733–1746. <https://doi.org/10.1016/j.rse.2010.03.003>.
- Schug, F., Frantz, D., Okujeni, A., Hostert, P., 2022. Sub-pixel building area mapping based on synthetic training data and regression-based unmixing using Sentinel-1 and -2 data. *Remote Sens. Lett.* 13, 822–832. <https://doi.org/10.1080/2150704X.2022.2088253>.
- Schug, F., Frantz, D., Okujeni, A., van der Linden, S., Hostert, P., 2020. Mapping urban-rural gradients of settlements and vegetation at national scale using Sentinel-2 spectral-temporal metrics and regression-based unmixing with synthetic training data. *Remote Sens. Environ.* 246, 111810 <https://doi.org/10.1016/j.rse.2020.111810>.
- Schwieder, M., Leitão, P.J., da Cunha Bustamante, M.M., Ferreira, L.G., Rabe, A., Hostert, P., 2016. Mapping Brazilian savanna vegetation gradients with Landsat time series. *Int. J. Appl. Earth Obs. and Geoinf.* 52, 361–370. <https://doi.org/10.1016/j.jag.2016.06.019>.
- Skakun, S., Wevers, J., Brockmann, C., Doxani, G., Aleksandrov, M., Batić, M., Frantz, D., Gascon, F., Gómez-Chova, L., Hagolle, O., López-Puigdollers, D., Louis, J., Lubej, M., Mateo-García, G., Osman, J., Peressutti, D., Pflug, B., Puc, J., Richter, R., Roger, J.-C., Scaramuzza, P., Vermote, E., Vesel, N., Zupanc, A., Žust, L., 2022. Cloud mask intercomparison eXercise (CMIX): an evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sens. Environ.* 274, 112990 <https://doi.org/10.1016/j.rse.2022.112990>.
- Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sens. Environ.* 185, 46–56. <https://doi.org/10.1016/j.rse.2016.04.008>.
- Vermote, E., Wolfe, R., 2015. MOD09GA MODIS/Terra Surface Reflectance Daily L2G Global 1km and 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC.
- Vermote, E., Wolfe, R., 2015. MYD09GA MODIS/Aqua Surface Reflectance Daily L2G Global 1km and 500m SIN Grid V006. NASA EOSDIS Land Processes DAAC.
- Vermote, E.F., Kotchenova, S., 2008. Atmospheric correction for the monitoring of land surfaces. *J. Geophys. Res. Atmos.* 113 <https://doi.org/10.1029/2007JD009662>.
- Viana-Soto, A., Okujeni, A., Pflugmacher, D., García, M., Aguado, I., Hostert, P., 2022. Quantifying post-fire shifts in woody-vegetation cover composition in Mediterranean pine forests using landsat time series and regression-based unmixing. *Remote Sens. Environ.* 281, 113239 <https://doi.org/10.1016/j.rse.2022.113239>.
- Waldner, F., Hansen, M.C., Potapov, P.V., Löw, F., Newby, T., Ferreira, S., Defourny, P., 2017. National-scale cropland mapping based on spectral-temporal features and outdated land cover information. *PLOS ONE* 12, e0181911. <https://doi.org/10.1371/journal.pone.0181911>.
- World Meteorological Organization, 2019. The Global Climate in 2015–2019. Chairperson, Publications Board World Meteorological Organization (WMO) 7 bis, avenue de la Paix Tel.: +41 (0) 22 730 84 03 P.O. Box 2300 Fax: +41 (0) 22 730 81 17 CH-1211 Geneva 2, Switzerland.
- Wulder, M.A., Hilker, T., White, J.C., Coops, N.C., Masek, J.G., Pflugmacher, D., Crevier, Y., 2015. Virtual constellations for global terrestrial monitoring. *Remote Sens. Environ.* 170, 62–76. <https://doi.org/10.1016/j.rse.2015.09.001>.
- Wulder, M.A., Loveland, T.R., Roy, D.P., Crawford, C.J., Masek, J.G., Woodcock, C.E., Allen, R.G., Anderson, M.C., Belward, A.S., Cohen, W.B., Dwyer, J., Erb, A., Gao, F., Griffiths, P., Helder, D., Hermosilla, T., Hipple, J.D., Hostert, P., Hughes, M.J., Huntington, J., Johnson, D.M., Kennedy, R., Kilic, A., Li, Z., Lymburner, L., McCorkel, J., Pahlevan, N., Scambos, T.A., Schaaf, C., Schott, J.R., Sheng, Y., Storey, J., Vermote, E., Vogelmann, J., White, J.C., Wynne, R.H., Zhu, Z., 2019.

- Current status of Landsat program, science, and applications. *Remote Sens. Environ.* 225, 127–147. <https://doi.org/10.1016/j.rse.2019.02.015>.
- Wulder, M.A., White, J.C., Masek, J.G., Dwyer, J., Roy, D.P., 2011. Continuity of landsat observations: short term considerations. *Remote Sens. Environ.* 115, 747–751. <https://doi.org/10.1016/j.rse.2010.11.002>.
- Yan, L., Roy, D.P., 2018. Large-area gap filling of Landsat reflectance time series by spectral-angle-mapper based spatio-temporal similarity (SAMSTS). *Remote Sens.* 10 <https://doi.org/10.3390/rs10040609>.
- Yin, H., Brandão, A., Buchner, J., Helmers, D., Iuliano, B.G., Kimambo, N.E., Lewińska, K.E., Razenkova, E., Rizayeva, A., Rogova, N., Spawn, S.A., Xie, Y., Radeloff, V.C., 2020. Monitoring cropland abandonment with landsat time series. *Remote Sens. Environ.* 246, 111873 <https://doi.org/10.1016/j.rse.2020.111873>.
- Yin, H., Khamzina, A., Pflugmacher, D., Martius, C., 2017. Forest cover mapping in post-soviet Central Asia using multi-resolution remote sensing imagery. *Sci. Rep.* 7, 1375. <https://doi.org/10.1038/s41598-017-01582-x>.
- Yin, H., Prishchepov, A.V., Kuemmerle, T., Bleyhl, B., Buchner, J., Radeloff, V.C., 2018. Mapping agricultural land abandonment from spatial and temporal segmentation of landsat time series. *Remote Sens. Environ.* 210, 12–24. <https://doi.org/10.1016/j.rse.2018.02.050>.
- Zhang, H.K., Roy, D.P., 2016. Landsat 5 thematic mapper reflectance and NDVI 27-year time series inconsistencies due to satellite orbit change. *Remote Sens. Environ.* 186, 217–233. <https://doi.org/10.1016/j.rse.2016.08.022>.
- Zhang, H.K., Roy, D.P., Luo, D., 2023. Demonstration of large area land cover classification with a one dimensional convolutional neural network applied to single pixel temporal metric percentiles. *Remote Sens. Environ.* 295, 113653 <https://doi.org/10.1016/j.rse.2023.113653>.
- Zhang, J., Shang, R., Rittenhouse, C., Witharana, C., Zhu, Z., 2021. Evaluating the impacts of models, data density and irregularity on reconstructing and forecasting dense Landsat time series. *Sci. Remote Sens.* 4, 100023 <https://doi.org/10.1016/j.srs.2021.100023>.
- Zhang, Y., Woodcock, C.E., Arévalo, P., Olofsson, P., Tang, X., Stanimirova, R., Bullock, E., Tarrio, K.R., Zhu, Z., Friedl, M.A., 2022. A global analysis of the spatial and temporal variability of usable Landsat observations at the pixel scale. *Frontiers in Remote Sensing* 3.
- Zhu, Z., 2017. Change detection using Landsat time series: a review of frequencies, preprocessing, algorithms, and applications. *ISPRS J. Photogramm. Remote Sens.* 130, 370–384. <https://doi.org/10.1016/j.isprsjprs.2017.06.013>.
- Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the fmask algorithm: cloud, cloud shadow, and snow detection for landsats 4–7, 8, and sentinel 2 images. *Remote Sens. Environ.* 159, 269–277. <https://doi.org/10.1016/j.rse.2014.12.014>.