www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

Global L-band equivalent AI-based vegetation optical depth dataset

Olya Skulovich[1]✉, Xiaojun Li[2], Jean-Pierre Wigneron[2] & Pierre Gentine[1]

The L-band vegetation optical depth data garners significant interest for its ability to effectively monitor vegetation, thanks to minimal saturation within this frequency range. However, the existing datasets have limited temporal coverage, constrained by the start of the respective satellite missions. Global L-band equivalent AI-Based Vegetation Optical Depth or GLAB-VOD is a global long-term consistent microwave vegetation optical depth dataset created using machine learning to expand the SMAP-IB VOD dataset temporal coverage from 2015-2020 to 2002-2020. The GLAB-VOD dataset has an 18-day temporal resolution and 25 km spatial resolution on the EASE2 grid and covers 2002-2020. An auxiliary consistent daily brightness temperature product, called GLAB-TB, is developed in parallel and ensures the consistency of the VOD product across time periods with different microwave satellites. As a result of its temporal consistency, this dataset can be used to study long-term global and regional trends in vegetation biomass and utilized in any other applications where long-term consistency is necessary. The GLAB-VOD dataset shows excellent spatial correlation globally when compared with biomass (up to $R = 0.92$) and canopy height ($R = 0.93$), outperforming its target dataset, SMAP-IB VOD.

Background & Summary

Observational datasets on vegetation biomass provide vital insights into the terrestrial carbon cycle and its potential dynamic under changing climate conditions. One measured indicator is vegetation optical depth (VOD), which represents the opacity of the vegetation determined through the extinction (including both attenuation and scattering effects) of electromagnetic waves emitted or reflected by the Earth's surface (usually measured in microwave band, typically L-band, C-band, or X-band). The extinction effects, described by radiative transfer theory, due to the vegetation is directly proportional to the vegetation water content. Hence, the information contained in VOD can be used for total biomass representation and its dynamic, as well as to estimate water-related vegetation stress[1–4].

VOD can be retrieved[5] from available passive microwave observations of the Earth's surface brightness temperature (TB). However, depending on the frequency, the retrieved values vary. That is due to the fact that higher frequency bands are mostly sensitive to the top of the canopy, whereas lower frequencies can penetrate deeper, comprising information from the full canopy, trunk and branches[6]. In addition to frequency, the disparity between different VOD retrievals can be related to varying sensors' footprints, radio frequency interference (RFI) contaminating the microwave signal, and the specifics of the retrieval model. For a comprehensive review of the different VOD products, retrieval models, and applications, we refer the readers to Frappart *et al.*[7].

There are currently three main VOD datasets available in C to K band (6.6 GHz to 19.35 GHz): the Global Land Parameter Data Record[8], Land Parameter Retrieval Model[9], and Global Long-term Microwave Vegetation Optical Depth Climate Archive (VODCA)[10]. While these products can be successfully implemented to study biomass dynamics[11], agricultural metrics[12], vegetation resilience[13], and sensitivity[14], there are concerns regarding these datasets. First, the data might lack consistency if derived from different sensors without explicitly addressing the sensors' compatibility[15]. For example, long-term trends in those VOD datasets vary significantly depending on the product[9]. Second, due to the lower penetration capability at higher frequencies, VOD derived from high-frequency sensors saturates for densely vegetated areas[3,16–19]. L-band microwaves penetrate deeper into canopies and can reach the soil. In fact, the two currently operating L-band sensors – the European Space Agency Soil Moisture and Ocean Salinity (SMOS)[20] and the National Aeronautics and Space Administration Soil Moisture Active Passive (SMAP)[21] – are primarily used to retrieve surface soil moisture from the measured Earth surface brightness temperatures, TBs. A soil moisture retrieval model needs to account for vegetation-related attenuation to accurately assess soil moisture. Hence, VOD products are available as side products of soil

[1]Columbia University, New York, NY, 10027, USA. [2]NRAE, UMR1391 ISPA, University of Bordeaux, F-33140, Villenave d'Ornon, France. ✉e-mail: os2328@columbia.edu

moisture retrievals. In particular, SMOS Level 2 (SMOS L2[20]) and Level 3 (SMOS L3[22]) provide VOD data from 2010, and SMAP L2_SM_P[23] and L2_SM_P_E[24] provide VOD data since the launch of the SMAP mission in 2015. The same sensor data can be used by other researchers to produce different VOD retrievals, for example, an alternative VOD dataset created from SMOS TB is SMOS-IC VOD[25]; the alternative VOD datasets based on SMAP data are MT-DCA VOD[1] and SMAP-IB VOD[26], and SMOSMAP-IB VOD[27] is based on the combination of SMOS and SMAP.

The SMAP-IB VOD dataset is created together with the soil moisture dataset from SMAP TB data based on the inversion of L-band Microwave Emission of the Biosphere model, covering the years from 2015 to the present. SMAP-IB VOD has good accuracy compared to other VOD products[28] while independent of auxiliary vegetation datasets by design. In particular, this feature allows SMAP-IB to demonstrate less saturation in dense forests compared to other SMAP products that incorporate optical vegetation data into their algorithms[26]. However, it should have a longer time span to maximize a dataset's potential. In this work, we aim to utilize AMSRE TB data to extend the SMAP-IB VOD dataset back to 2002 with a quality equivalent to the target dataset using machine learning. Following our recently developed methodology for seamless remote sensing data merging (used for soil moisture in Skulovich and Gentine, 2023[29]), we chose SMAP-IB VOD as our target dataset as a high-quality VOD dataset. To our knowledge, no studies have yet utilized machine learning tools to create VOD datasets, apart from being a complementary tool to estimate models' auxiliary parameters (e.g., soil parameters[30]). At the same time, machine learning is widely used to produce and improve soil moisture datasets[31–33], and soil moisture and VOD share common source data, so extending the machine learning tools to the VOD domain seems natural. By the nature of neural network (NN) training, a well-trained neural network produces output with a distribution that globally matches the target data distribution. Thus, we created a new VOD dataset, Global L-band equivalent AI-Based Vegetation Optical Depth or GLAB-VOD, that is consistent with SMAP-IB VOD and spans back to 2002. Such a VOD dataset is valuable for a wide range of scientific, environmental, and practical applications due to its consistency and temporal coverage that spans almost two decades. The two-decade-long dataset allows researchers to analyze trends in vegetation health, biomass, and moisture content more robustly, contributing to understanding how climate change affects global ecosystems and the carbon cycle. In particular, for evaluating the impact of climate on the trends of the vegetation biomass stocks[34], it is very important to validate models using long-term data sets. Extending current SMOS-based data set to 2002 is a very important achievement for such analyses. The GLAB-VOD dataset can help monitor forest conditions, detect deforestation, forest degradation, and recovery after disturbances, monitor crop conditions and grassland productivity, and capture gradual shifts that may not be evident in shorter timeframes. The dataset can be integrated into hydrological and carbon cycle models to improve their predictions. In addition, VOD data is essential for phenology studies to assess changes in the timing of vegetation life cycle events across different regions in recent years. Our approach allows a unique option of extending a dataset back into the past without losing in data quality. Past data can provide a more comprehensive baseline against which future changes can be measured. This dataset uniquely complements existing VOD datasets as it is the only long-term consistent global L-band equivalent VOD dataset.

Methods

The key strategy behind creating the GLAB-VOD dataset is based on the methodology presented in Skulovich and Gentine, 2023[29] and can be summarized as follows. We choose a target VOD dataset and train a neural network to predict this VOD directly from brightness temperatures from other sources with a longer observation history. In particular, our target VOD dataset is SMAP-IB VOD, retrieved from TB measurements collected by the SMAP satellite launched in 2015. Now, if we train NN with SMAP-IB VOD as a target but with SMOS TB as an input, such NN can output VOD for the whole period of SMOS observations, starting in 2010. In this example, we extended SMAP-IB VOD-like data back to 2010 in one step. In this study, we take several steps in this manner, contingent on the availability of TB data and their mutual compatibility, but in principle, the example above describes the essence of the approach. To substitute for numerous auxiliary datasets defining local conditions generally necessary for correct VOD retrievals from TB (for example, elevation and slope, soil texture, land cover, surface roughness, precipitation, vegetation parameters[35]), we use grid latitude and longitude as an input to the neural network. Further, the relationship between TBs and VOD is indelibly linked with soil moisture, hence, soil moisture data is added as NN input as well. CASM SM[29] is assumed to be consistent for the same period as this study and does not introduce any additional disparity in the input data. No other attenuation effects are considered, for example, due to rainfall since L-band frequencies are less sensitive to atmospheric conditions like clouds and rainfall, or to the effect of intercepted rain at the plants' surface, which was found to be of the order of a few Kelvins[16], especially at the temporal and spatial resolution of GLAB-VOD dataset. Next, let us briefly describe the three key features of the approach that allow us to achieve consistent and robust long-term data product.

1. **Deseasonalized signal**. Dividing the signal into fixed seasonal and varying residual components.
2. **Training scheme**. Developing a special NN training scheme to merge heterogeneous data sources using transfer learning.
3. **Uncertainty**. Assessing uncertainty of the final product using an ensemble of NN.

The following sections describe these steps in detail.

Key features of the approach.

Deseasonalized signal. In many parts of the globe, VOD and TB comprise a strong seasonal cycle. When an NN is trained on data with strong seasonality, good NN performance can be due

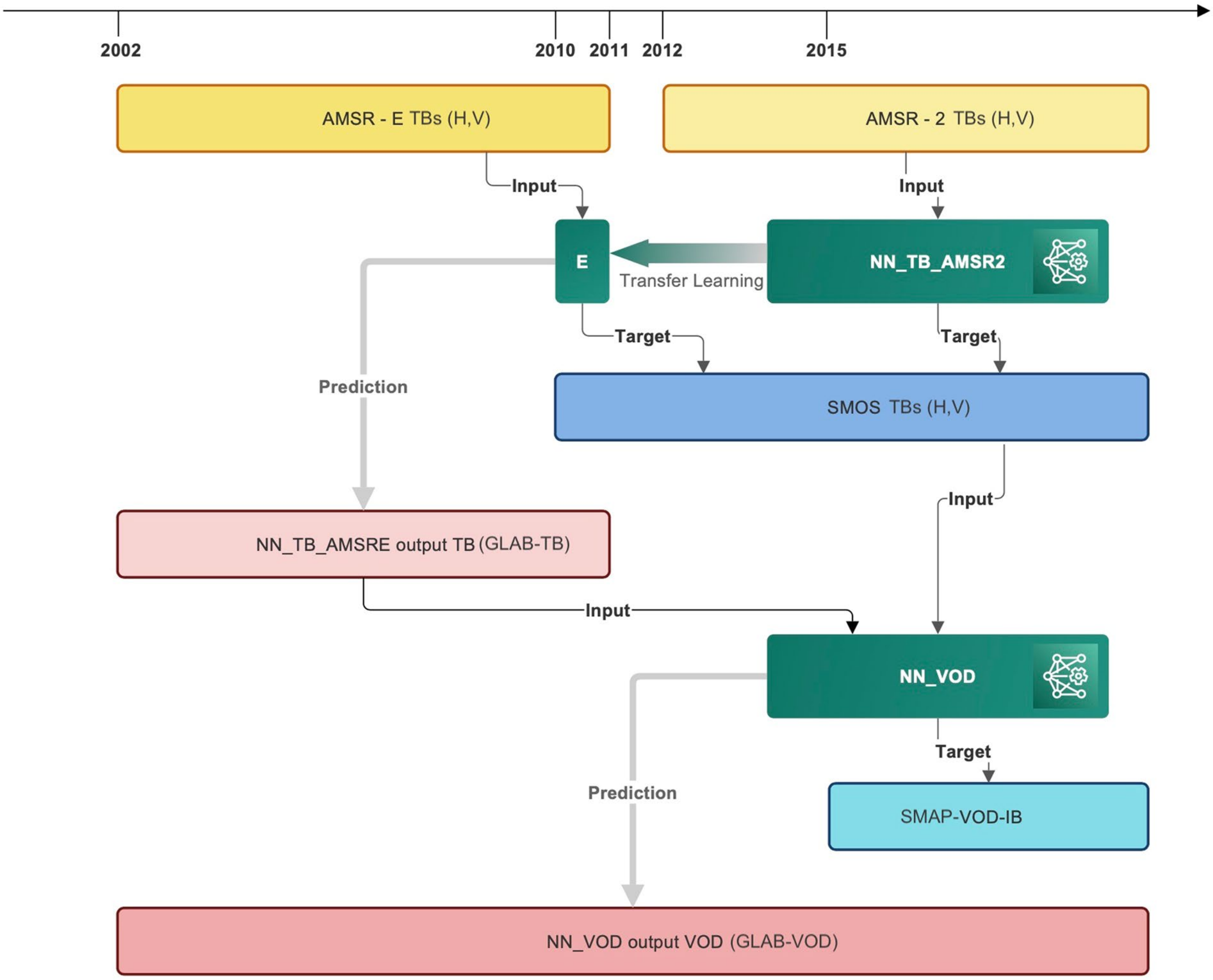

Fig. 1 Conceptual diagram of neural network training, input, target, and output dataset. Yellow blocks correspond to AMSR-family data; SMOS data is blue; SMAP-IB VOD is turquoise; NN predictions are pink; neural networks are denoted as dark green blocks with a network pictogram.

to NN capturing the seasonal amplitude rather than interannual variability or anomaly[36,37]. For that reason, we divide VOD and TB signal into fixed seasonal component and variable residual component, Eq. (1).

$$\widehat{VOD} = \underbrace{VOD_{fixed\_seas\_cycle}(lat,\ lon)}_{imposed} + \underbrace{VOD_{residuals}(lat,\ lon,\ TB_{residuals},\ SM)}_{targeted\ by\ NN} \tag{1}$$

The fixed seasonal component is defined per grid point as a sine wave with a period of one year fitted to the data[29]. The residual component in this configuration includes sub-seasonal periodic signals, trends, extremes, and noise. It is defined as a function of geographical location, residual component of TB signal, and full soil moisture signal. By this, we assume that the VOD-TB relationship is modulated by soil conditions defined through full soil moisture signal (rather than the relative residual component of the soil moisture signal).

It is the residual part of the signal that is targeted and predicted by NN in all cases.

Training scheme. The training scheme developed in this study is multi-stepped with the objective to combine different data sources into a consistent product seamlessly. Among all available TB data, we choose SMOS TB as NN input since the SMOS sensor is an L-band sensor, similar to SMAP. So, the first step is to obtain consistent TB data similar in quality to SMOS TB. Then, with these data as input, we can obtain VOD data for the whole period the input data exists. The conceptual scheme of the training is given in Fig. 1. Let us go over the depicted steps.

1. **NN_TB_AMSR2**. First, we train an NN, NN_TB_AMSR2 on data from 2012 to 2020, to reproduce SMOS-like TB from AMSR-2 TBs.

- **Inputs**. We take five AMSR-2 TBs in three frequencies – horizontal (H) and vertical (V) polarization TB measured at 10.7 GHz, H- and V-polarization TB measured at 18.7 GHz, and V-polarization TB measured at

36.5 GHz. We chose 10.7 and 18.7 GHz frequencies in AMSR-2 data as they penetrate deeper into the canopy and are hence more relevant for L-band (1.413 GHz) SMOS, less sensitive to artificial radio-frequency interference (RFI), and offers a longer period of satellite observations, which can be extended in the future to the earliest available data from 1998. We additionally include the measurements at 36.5 GHz in vertical polarization as it accounts for cloud liquid water[38] (that also attenuates the signal from the surface).

- **Target**. SMOS TB in H- and V-polarizations at the incidence angle of 40° are the NN_TB_AMSR2 targets. We train two separate NNs to target each of these TBs individually.

2. **Transfer learning**. Transfer learning[39] is an training technique where a model trained on one task is adapted for use on a different but related task. In the context of neural networks, transfer learning involves leveraging the knowledge gained from training a model on one dataset to accelerate learning or improve performance on another dataset or task. This is typically done by reusing the learned features or parameters of the pre-trained model and fine-tuning them on the new task or dataset. Indeed, AMSR-E TBs are not directly consistent with AMSR-2 data[40]. At this stage, we additionally train NN_TB_AMSR2, pre-trained at the previous step on 8 years of global data, with AMSR-E TBs as inputs. AMSR-E TBs are taken in the same frequencies and polarization as AMSR-2 at the previous step. The training set here comprises 22 months of data (January 2010 - October 2011), for which both AMSR-E and SMOS data exist and hence can be used for training.
   The NN_TB_AMSR2-E trained in this way can produce TBs consistent with SMOS TBs from AMSR-E data. We call this NN output GLAB-TB. The GLAB-TB dataset consists of daily records of SMOS-like TBs in H- and V-polarization starting in 2002.
3. **NN_VOD**. NN_VOD is trained with SMAP-IB VOD as a target and SMOS TB in H- and V- polarization and CASM soil moisture as input. Due to the noisiness of the VOD data, all datasets are brought to 18 days temporal resolution. When GLAB-TBs are used as an input, the NN_VOD output is GLAB-VOD.

The data for GLAB-VOD NN training comprises 21,716,611 data points, and the data for GLAB-TB NN comprises 47,800,254 data points. The data loss due to the necessary datasets' intersection is minimal for the GLAB-VOD NN (less than 0.2%) and 26.5% for the GLAB-TB NN (in comparison to the original AMSR-2 data).

During the training, the data is randomly divided into training, validation, and testing in proportions of 0.64 - 0.16 - 0.2. In the final version, the NNs have the following characteristics: 6 hidden layers, decreasing number of neurons from 1500 in the first layer to 200 neurons in the last hidden layer, leaky ReLU[41] activation function in the first layer, and ReLU in the subsequent, Huber loss function, and Adam optimizer[42] with learning rate 0.001. The input data is reprocessed with a robust scaler.

Uncertainty. In this work, we include the uncertainty of the final product associated with the model (structural, epistemic) uncertainty by training an ensemble of 7 NNs[43]. Then, we calculate the mean and standard deviation of the outputs of this sample, thus providing a Bayesian assessment of the uncertainty.

Data Records

The datasets created in this study[44] are available at https://zenodo.org/doi/10.5281/zenodo.10306094 publicly. The data is stored as yearly NetCDF files (for TBs and VOD separately) in the corresponding temporal resolutions (daily and 18 days) EASE-2 grid in 25 km spatial resolution.

The proposed approach aims to expand the SMAP-IB VOD dataset presented in Li *et al.*[26] back in time. SMAP-IB VOD is an L-band VOD product retrieved using the L-band Microwave Emission of the Biosphere model. We used 36 km spatial resolution data that covered 2015-2020 with daily temporal resolution. It showed excellent performance when compared to other vegetation indices[26]. The datset grid was changed to 25 km EASE2 grid for the compatibility purposes with other datasets. The dataset is publicly available at https://ib.remote-sensing.inrae.fr.

Brightness temperature data was taken from three different sources. The first is SMOS L3 TB product[22], which was filtered to eliminate the observations affected by RFI (where Root Mean Square Error for TB is higher than an 8K threshold[5]), similar to Li *et al.*[27]. Daily ascending orbit data in horizontal (H) and vertical (V) polarization with 25 km spatial resolution was taken from 2010 to 2020. SMOS L3 TB product is publicly available at https://www.catds.fr/sipad/. The second TB source is AMSR-2[45] daily data. We used brightness temperature data in H- and V- polarization measured at 10.7 GHz, H- and V-polarization TB measured at 18.7 GHz, and V-polarization TB measured at 36.5 GHz. The dataset is publicly available at https://nsidc.org/data/AU_Land/versions/1. AMSR-E[46] data is the third source of TBs. The same subset of frequencies as for AMSR-2 data are chosen. The dataset is publicly available at https://nsidc.org/data/ae_land3/versions/2.

Soil moisture data is taken from CASM datasets[29]. This dataset provides consistent soil moisture data from 2002 at 3 day temporal resolution and 25 km spatial resolution. The dataset is publicly available at https://zenodo.org/doi/10.5281/zenodo.7072511.

The summary of all data used in this study is given in Table 1.

Technical Validation

We consider the following metrics to assess the quality of the GLAB-TB and GLAB-VOD products.

- The trained NN performance.
- The products' ability to reproduce spatial patterns for TBs and VOD in the target datasets.
- The products' temporal consistency.
- The correlation between VOD and other vegetation-related indices.

Dataset	Variable	Reference	Temporal coverage used	Temporal resolution	Usage
SMAP-IB VOD	VOD	[26]	2015-2020	Daily	Target for NN_VOD
SMOS	TB-H, TB-V	[22]	2010-2020	Daily	Target for NN_TB_AMSR Input for NN_VOD
AMSR-2	10.7 GHz TB-H, TB-V, 18.7 GHz TB-H, TB-V, 36.5 GHz TB-V	[45]	2012-2020	Daily	Input for NN_TB_AMSR
AMSR-E	10.7 GHz TB-H, TB-V, 18.7 GHz TB-H, TB-V, 36.5 GHz TB-V	[46]	2002-2011	Daily	Input for NN_TB_AMSR transfer learning
CASM	Soil Moisture	[29]	2002-2020	3 days	Input for NN_TB_AMSR Input for NN_VOD
CCI Biomass	Biomass	[49]	2017	Annual	Auxiliary
Saatchi Biomass	Biomass	[48]	2015	Annual	Auxiliary
Canopy Height	Canopy Height	[50]	2019	Annual	Auxiliary
SMOS IC VOD	VOD	[25]	2015, 2017, 2019	Annual	Auxiliary
cSIF	Solar Induced Fluorescence	[51]	2019	Annual	Auxiliary
GLAB-TB	TB-H, TB-V	This work[44]	2002-2020	Daily	Output from NN_AMSR Input for NN_VOD
GLAB-VOD	VOD	This work[44]	2002-2020	18 days	Output from NN_VOD

Table 1. Datasets used in this study.

NN	Usage	Input	Target	Training period	R^2 residual	R^2 full signal
NN_VOD 3 days	Test the effect of averaging	residual SMOS TB (H and V), CASM	residual SMAP-IB VOD	2015-2020 (3 days)	0.10	0.98
NN_VOD no SM	Test the role of SM	residual SMOS TB (H and V), CASM	residual SMAP-IB VOD	2015-2020 (18 days)	0.24	0.99
NN_VOD full signal	Test seasonal cycle effect	SMOS TB (H and V), CASM	SMAP-IB VOD	2015-2020 (18 days)	-0.86	0.97
NN_TB_AMSR	Final version	residual AMSR-2 (5 TBs)	residual SMOS TB (H and V)	2012-2020 (daily)	0.58(H) 0.57(V)	0.82(H) 0.87(V)
NN_TB_AMSR transfer learning	Final version	residual AMSR-2 (5 TBs)	residual SMOS TB (H and V)	2010-2011 (daily)	0.59(H) 0.58(V)	0.82(H) 0.87(V)
NN_VOD	Final version	residual SMOS TB (H and V), CASM	residual SMAP-IB VOD	2015-2020 (18 days)	0.31	0.99

Table 2. Characteristics and performance of NN used in this study. Several sensitivity analysis results are included alongside the final versions of the NNs.

NN performance. Table 2 summarizes the NN performance measured as a coefficient of determination (R^2) for the NNs used to create GLAB-TB and GLAB-VOD and a few other NN configurations added for comparison. In the sensitivity analysis, many other NN configurations were tested, including NN with different number of neurons, hidden layers, loss function, learning rate, batch size, and scaling, as well as with different combinations of inputs. The chosen NN exhibited the best performance on an aggregate basis of several metrics, including correlation coefficient R, determination coefficient R^2, root mean squared error (RMSE) on training and test samples, and visual examination of the loss during the training for the training and validation samples.

In the final configuration, NN_VOD achieves R^2 of 0.31 on the residual component of the VOD signal with $R^2 = 0.99$ for the full VOD signal. This suggests that variability in VOD mainly comes from seasonal variability. Note the change in the performance metrics for the NN trained on the full signals without seasonal cycle removal. Specifically, while the performance on the full signal deteriorates only slightly (0.97 vs. 0.99), this NN loses any ability to capture the relationship beyond the seasonal variation with $R^2_{residual} = -0.86$. Yet, we assume that the low signal-to-noise ratio of the residuals, even in the best-performing version, is the main reason for NN only capturing about 30% of the variability of the residual part of the VOD signal in the final version of the NN. To confirm this hypothesis, we compared the final version of the NN_VOD with an NN trained on data with a 3-day temporal resolution. This NN's $R^2_{residual} = 0.10$ suggests an even higher noise level compared to the data averaged to 18-days resolution. Finally, we see that SM data is also beneficial for capturing the signal beyond the seasonal variability: while R^2_{full} doesn't change for a NN trained without SM in the input, $R^2_{residual}$ for this NN drops to 0.24.

For NN_TB_AMSR, this NN can capture more information from the residual component of the TB signal, rightfully suggesting that TB variability is less subdued to seasonal cycle than VOD with $R^2_{residual}$ for NN_TB_AMSR reaching 0.58. However, the performance on the full signal is lower, implying that TBs from AMSR

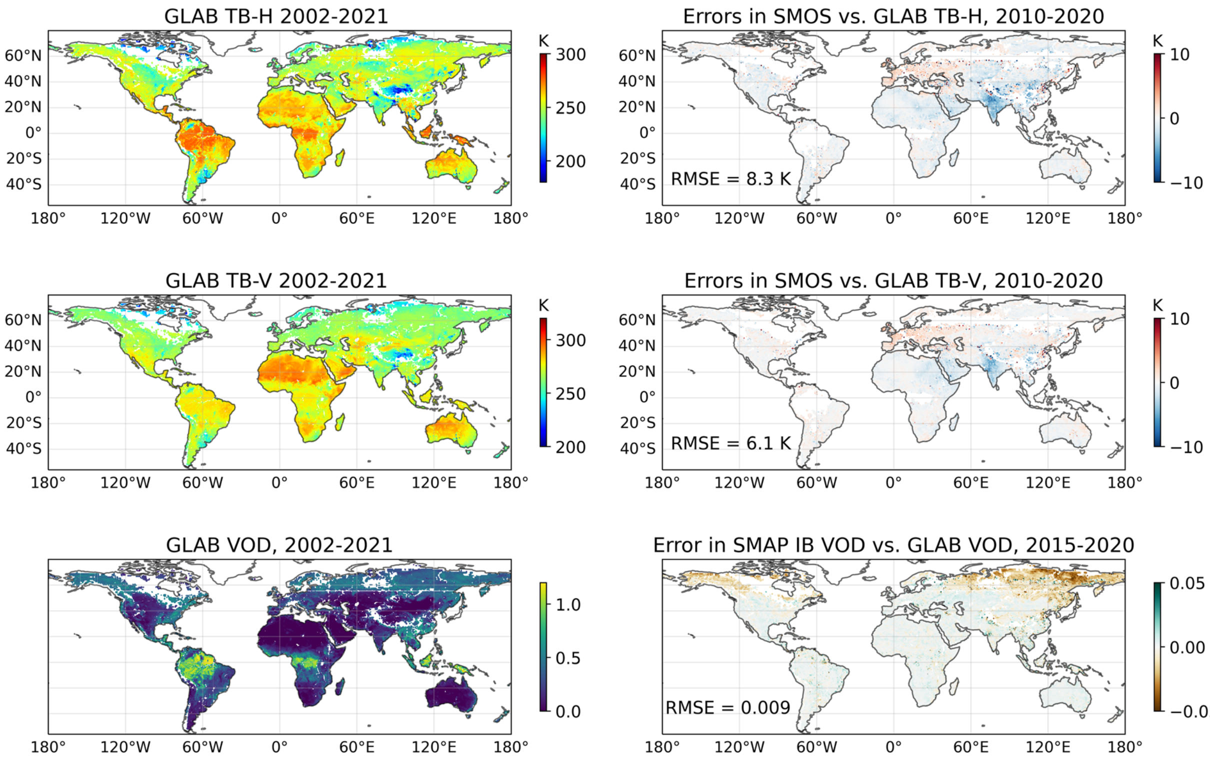

Fig. 2 Global spatial patterns in GLAB-TB (H and V) and GLAB-VOD averaged over 2002-2020 alongside with spatial distribution of errors in GLAB-TB vs. SMOS TB in 2010-2020 and GLAB-VOD vs. SMAP-IB VOD in 2015-2020. The errors are defined as temporally averaged target dataset minus GLAB product, i.e., negative error means the GLAB product overestimates the value in comparison to the target dataset.

cannot fully explain the variability in SMOS TB, which is to be expected given the difference in the frequencies at which they are measured. The transfer learning further slightly improves $R^2_{residual}$.

Supplementary Fig. S1 illustrates how well the NN output TB and VOD match the target dataset distribution. While GLAB-TB can capture the shape of the distribution well, including bi-modal distribution for TB in vertical polarization, it slightly underestimates the distribution tails while overestimating and skewing the peak. It is interesting to compare the TB full distributions with the TB seasonal cycle. TB variability cannot be reduced to seasonal variability, and NN significantly improves the match, capturing the dynamic from other sources of variability. GLAB-VOD distribution matches SMAP-IB VOD distribution almost perfectly, only slightly overestimating VOD equal to zero.

Spatial consistency. Figure 2 illustrates spatial patterns in GLAB-TB and VOD products. For comparison with the target products, the spatial distributions of the errors are given in the same figure. Overall, all GLAB products capture the spatial distribution of the targeted variables very well. For brightness temperatures, GLAB-TB seems to slightly overestimate the TB value over India. We hypothesize this behavior is related to RFI contaminating the TB data. SMOS data used in this study has been filtered to eliminate data affected by RFI with an RMSE threshold equal to 8 K. This already drastically affects the data availability, as illustrated in Supplementary Fig. S2. Further filtering the data can reduce the bias, however, in this tradeoff, we lean towards keeping the larger dataset. For the VOD product, the errors are negligible for most of the globe, except for high-latitude regions. However, even there, the errors are of 0.01 order of magnitude (global RMSE = 0.009, also as to be expected keeping in mind 0.99 R^2 achieved for the full VOD signal, Table 2). The deterioration in VOD performance in the high latitudes can be attributed to the overall complexity of microwave remote sensing in these regions related to orbital coverage and freeze/thaw conditions, numerous water bodies and very high organic content that affect both VOD and SM retrievals.

Temporal consistency. Figure 3 illustrates global averaged GLAB-VOD as a time series. For convenience, the date of SMOS data start and AMSR-E mission end are marked on the graph. To quantify our observations that GLAB-VOD has no bias related to the used input data, we can calculate the global mean VOD in 2002-2010 and compare it to the global mean VOD in 2012-2020. The corresponding values are 0.2364 with the model-related mean uncertainty of 0.046 in 2002-2010 vs. 0.2359 with 0.0038 uncertainty in 2012-2020. We can further confirm that GLAB-VOD is consistent in these two periods by comparing their global distributions (Supplementary Fig. S3). The two distributions are indeed almost identical. Note that VOD is a dynamic variable that can represent the change in biomass due to land use land cover changes, shifting phenology, water stress, or other disturbances. Hence, some changes in VOD between 2002-2010 and 2012-2020 can be related to the actual occurrence

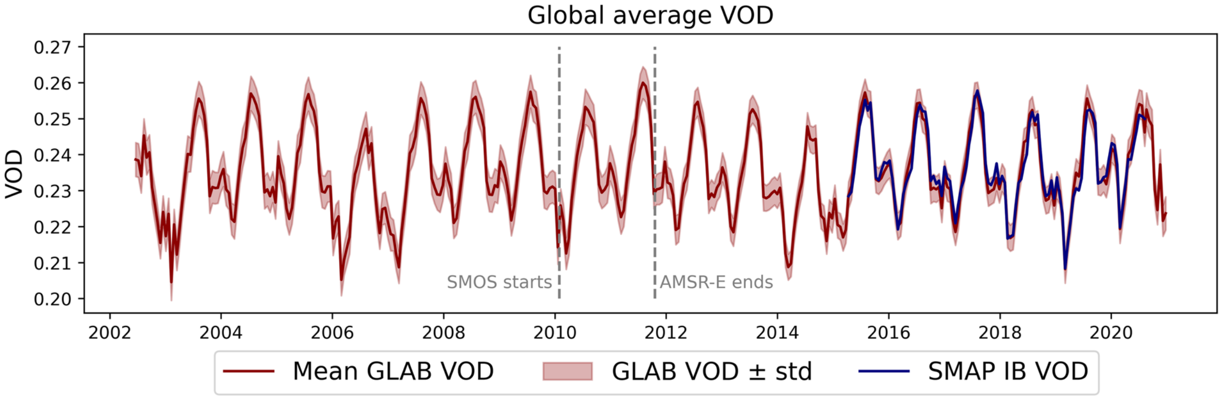

Fig. 3 Global averaged GLAB-VOD time series with uncertainty defined as one standard deviation from NN ensemble mean with SMAP-IB VOD in 2015-2020 for comparison. The grey vertical lines indicate the beginning of SMOS observations and the end of AMSR-E observations to illustrate the consistency of the dataset regardless of the input source.

of changes in biomass in the first decade in comparison to the second decade[47], rather the changes being a result of the model deficiency.

Further, we look at individual time series for five latitudinal bands ranging from 60 to 90° North and 30 to 60° South, in Fig. 4. GLAB-VOD can reproduce different dynamics inherent to different regions. At all latitude bands, seasonal cycle is evident in the data with different amplitude of the changes. The departures from the seasonal cycle represent year-to-year variability and irregularity of the signal due to stresses, extreme events, and in response to other forcing. For example, seasonal amplitude is the highest in mid latitudes, dominated by temperate forests and reflecting on the forests' spring and summer growing seasons and fall and winter dormant seasons. This is the leading mode of variability for this latitudinal band, with year-to-year variability reflected in the height of the summer peak and variability of the Fall dynamic (probably also related to crop yield and harvesting). These times are also characterized by the wider uncertainty band in our product in comparison to the rest of the year. This behavior is emergent, and, for example, for the tropical region, the VOD uncertainty is more uniform throughout the year. At the same time, NN is trained to generalize the TB-SM-VOD relationship globally, resulting in GLAB-VOD being more smoothed out compared to SMAP-IB VOD. For example, GLAB-VOD does not capture the peak present in SMAP-IB VOD data in high latitudes in 2019. However, the difference between the multiyear average peak VOD equal to 0.61 vs. 0.88 in 2019 requires further investigation into the cause of this event. Similarly, the SMAP-IB VOD data exhibits an apparent trend in the Southern Temperate Zone (30-60° South). GLAB-VOD does not reproduce this trend and is more stable year-to-year for the whole period between 2002 and 2021. Further studies can validate the presence or absence of this trend by comparing it to other vegetation-related observations, such as biomass, leaf area index, and photosynthesis.

We look into regional behavior and compare the target and the output VOD datasets even more in depth in Supplementary Figs. S4 and S5. Supplementary Fig. S4 illustrates the VOD timeseries at three individual locations – Santa Cruz region, Argentina (-50.05, -69.89), Nenaka region, AK, USA (64.91, -152.89), and Bordeaux region, France (44.58, 0.39), illustrating the behavior at one randomly chosen grid point where GLAB-VOD and SMAP-IB VOD do not fully correspond on the regional scale (according to the Fig. 4 described above). Supplementary Fig. S5 shows time series at 10 different regions around the globe (their location and extent are given in Supplementary Fig. S6, note the clear star-shaped RFI effects, for example, in Africa). From both examples, we can conclude that GLAB-VOD is less noisy than SMAP-IB VOD yet retains interannual variability. Some perceived noisiness in the data also comes from irregular data count (Supplementary Figs. S4 and S5), related to RFI data contamination (especially in China) and frozen conditions (Siberia, Northeast US, and Alaska). Decreasing trends in North-West Amazon region, Europe, and Australia, and increasing trend in China over the last years call attention to themselves in view of carbon sink dynamic. Long-term consistent datasets like GLAB-VOD allow to study this dynamic robustly.

Comparison to other vegetation metrics. To further validate the quality of our dataset, we compare how well the GLAB-VOD data correlates with other vegetation-related variables, in particular, with two biomass datasets[48,49], https://ceos.org/gst/jpl-biomass.html, https://climate.esa.int/en/projects/biomass/, canopy height data[50]https://zenodo.org/doi/10.5281/zenodo.5112903, and solar-induced fluorescence (SIF)[51]https://doi.org/10.17605/OSF.IO/8XQY6 datasets. For comparison, we also present the performance of another two VOD datasets: the target dataset that our product is based on – SMAP-IB VOD – and SMOS-based VOD dataset – SMOS IC VOD[25]. In all cases, one year of VOD data is compared to biomass and canopy height in the same year (See Table 1), both averaged to annual means.

Since the comparison is carried out for validation purposes, we assess the relative performance of GLAB-VOD via other VOD datasets. The assessment is based on the following assumptions:

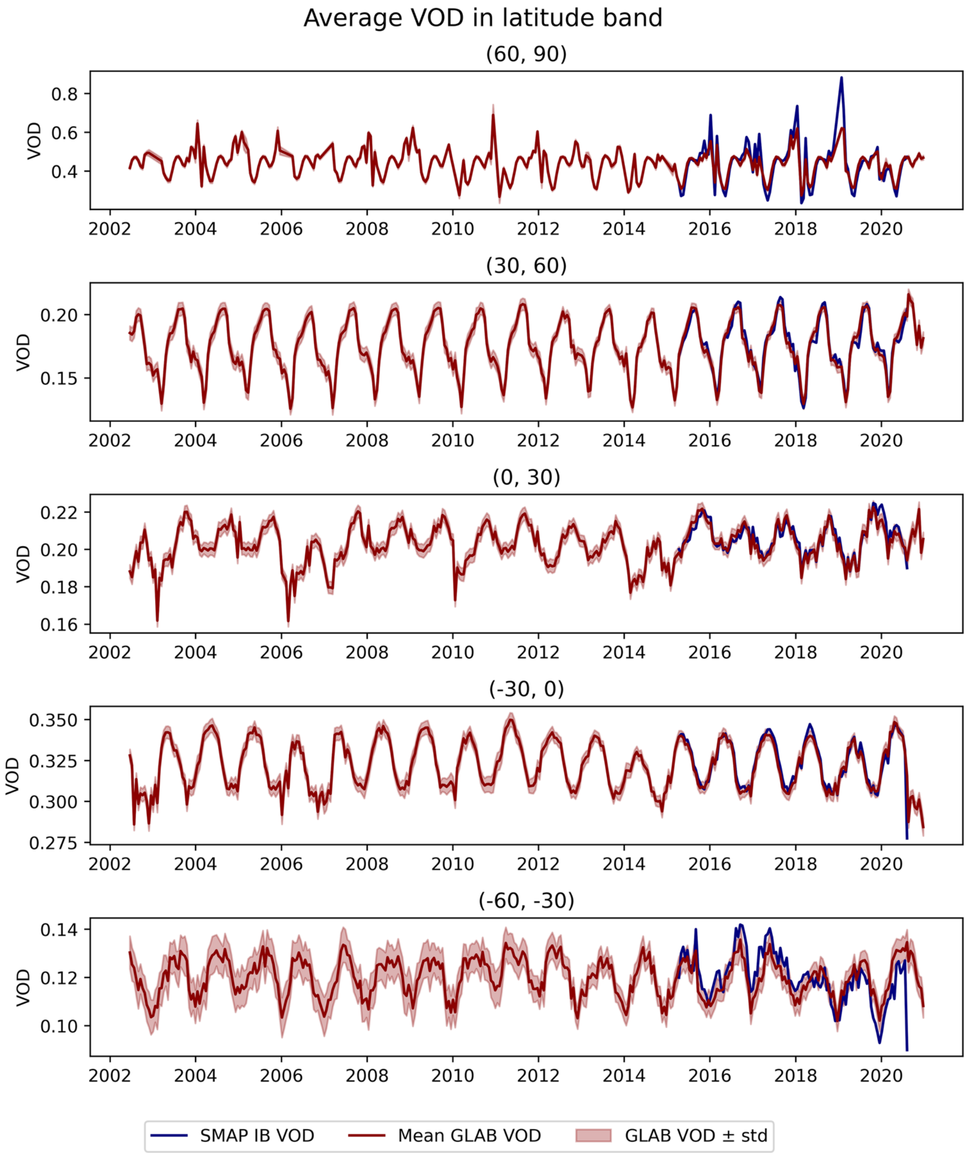

Fig. 4 GLAB-VOD time series with uncertainty defined as one standard deviation from NN ensemble mean with SMAP-IB VOD in 2015-2020 for comparison averaged for five latitudinal bands.

- At the annual scale and globally, VOD is directly related to biomass, and this relationship is close to linear for L-band retrievals[3,17,52].
- Canopy height corresponds to the total amount of vegetation[52]. At the annual scale and globally, VOD is closely related to canopy height, and this relationship is linear[26,53].
- SIF is a measure of photosynthesis[51]. Photosynthesis and biomass are not directly proportional, and the relationship depends on plan type, plant age, water, and nutrient availability[54]. We expect VOD products to be able to pick up these dependencies.

As evident from Fig. 5, the correlation coefficient R (calculated as Pearson's coefficient for the spatial correlation between VOD and corresponding proxies over the annual data for both) is close when comparing all three VOD datasets but is consistently higher for GLAB-VOD. The improvement seems to come from the partial dispersion of spurious associations outside the main correlation relationship. R is even higher for GLAB-VOD in comparison to SMAP-IB VOD, the target dataset (e.g., $R=0.919$ for GLAB-VOD vs. $R=0.897$ for SMAP-IB VOD when compared to Saatchi biomass, and $R=0.926$ vs. $R=0.911$, correspondingly, when compared to canopy height). That is notable considering that GLAB-VOD should inherit SMAP-IB VOD qualities through NN training. Since the comparison is made for one year only, this improvement comes from the features of the GLAB-VOD dataset rather than its augmented temporal coverage. GLAB-VOD is not only less noisy than SMAP-IB VOD, but it also seems to filter out the values not backed up by VOD-biomass dependence.

Finally, in Fig. 6 we consider a global spatially contiguous solar-induced fluorescence (cSIF)[51] as a function of VOD. Unlike biomass, we do not expect this relationship to be linear with respect to VOD, since for some

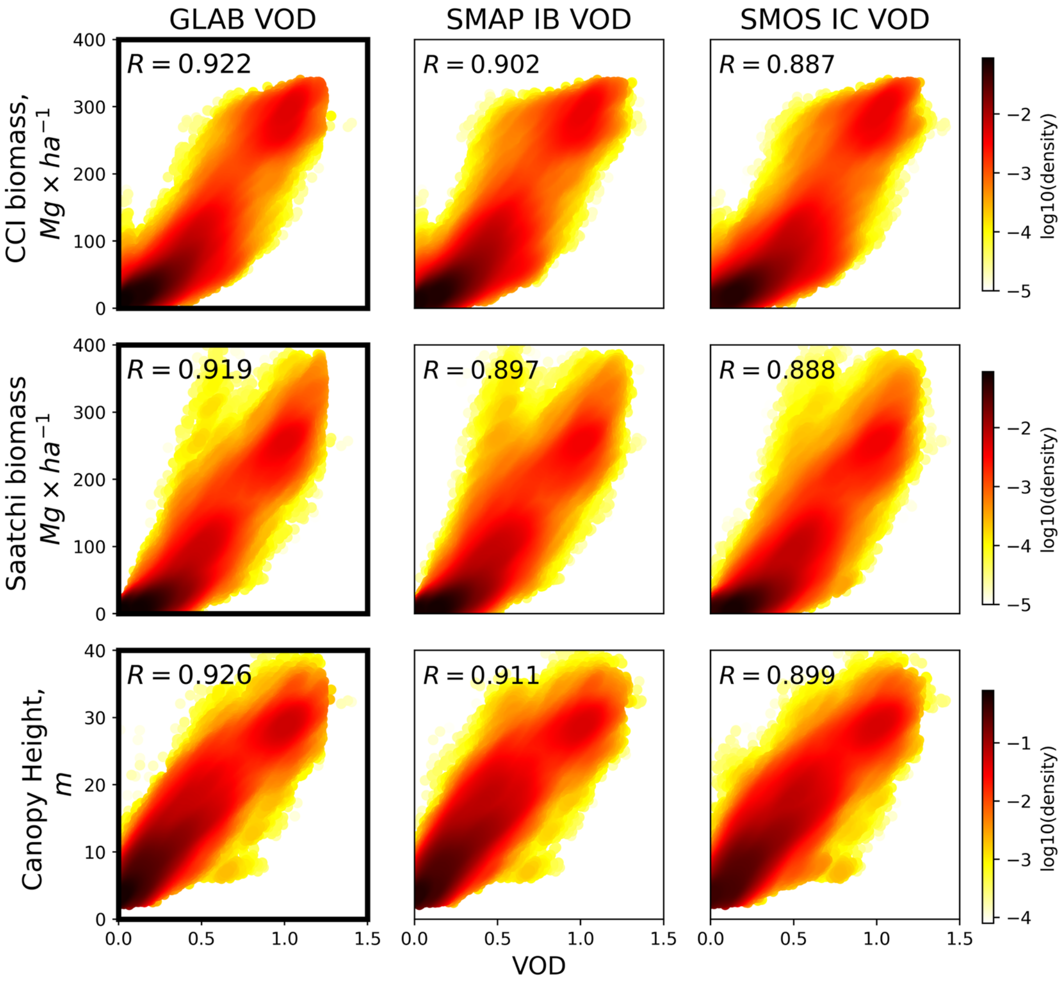

Fig. 5 Kernel density plot for biomass and canopy height in relation to VOD. GLAB-VOD (left column, highlighted by the black box) is compared to the SMAP-IB VOD (middle column) and SMAP IC VOD (right column). Correlation coefficient R is given per plot.

biomes, standing biomass is not necessarily actively photosynthesising. Indeed, from the shape of the curve we see cSIF saturation for evergreen forests. We further illustrate that by providing 2D density plots for crops, grasslands, deciduous broadleaf, and evergreen broadleaf forests (Fig. 6). The linear relationship between cSIF and VOD is more pronounced for the first three environments, and saturates for the evergreen forests.

Usage Notes

The dataset is open to public use without limitation. The permanent storage is at https://zenodo.org/doi/10.5281/zenodo.10306094, the data is stored as yearly data files in NetCDF format separately for brightness temperatures and VOD.

GLAB-TB data is archived and stored as “GLAB_TB_HV_yyyy.nc.zip” where *yyyy* is the corresponding year. The data has daily temporal and 25 km spatial resolution on EASE2 grid. Each data file contains geographical coordinates and date, with the corresponding variables:

- *TB_H* and
- *TB_V* – corresponding to the brightness temperatures in horizontal and vertical polarization, in degrees [*K*].

GLAB-VOD data is stored as “GLAB_VOD_yyyy.nc” where *yyyy* is the corresponding year. The data has 18-days temporal and 25 km spatial resolution on EASE2 grid. Each data file contains geographical coordinates and date, with the corresponding variables:

- *VOD* – the product of this study, the mean VOD data in the ensemble of NN;
- *VOD_residual* – the residual component of the VOD data, equal to the full VOD signal minus seasonal cycle, this is the mean VOD residual based on the ensemble of NN;
- *VOD_std* – standard deviation of the VOD based on the ensemble of NN;
- *VOD_residual_std* – standard deviation of the residual part of the VOD signal based on the ensemble of NN.

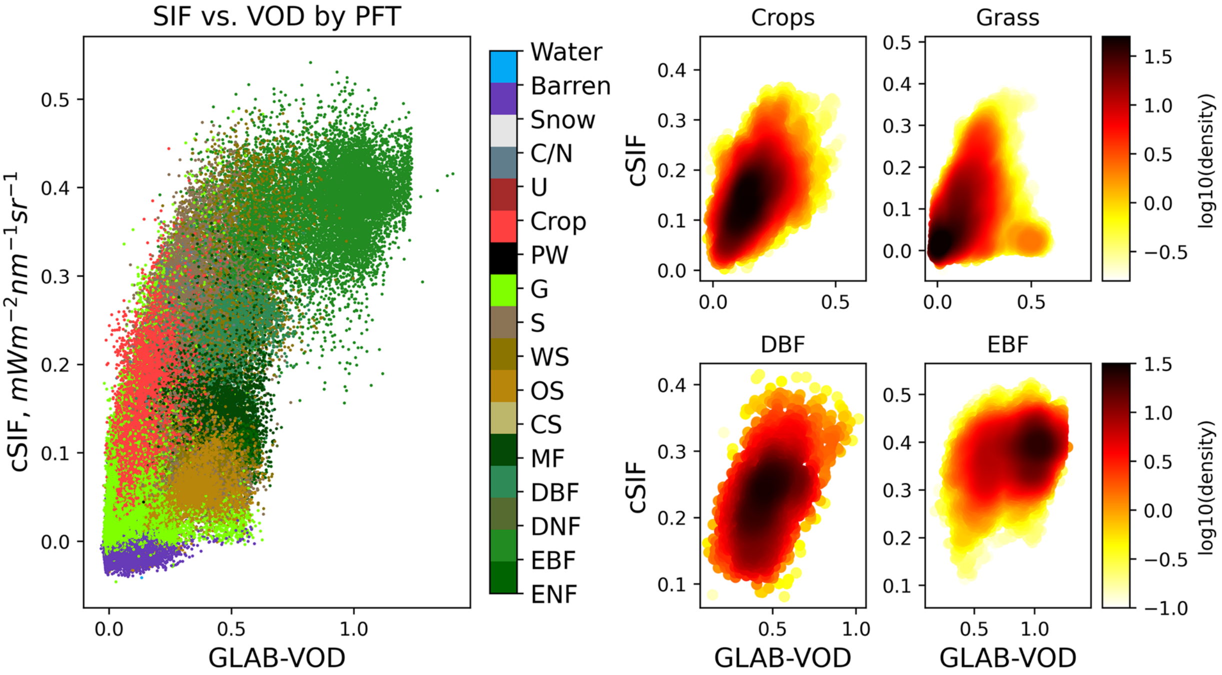

Fig. 6 Left panel: Scatter plot of cSIF as a function of VOD, colored by plant functional type (PFT). The PFTs are ENF: Evergreen Needleleaf Forests; EBF: Evergreen Broadleaf Forests; DNF: Deciduous Needleleaf Forests, DBF: Deciduous Broadleaf Forests, MF: Mixed Forests; CS: Closed Shrublands; OS: Open Shrublands: WS: Woody Savannas; S: Savannas; G: Grasslands; C: Croplands; PW: Permanent Wetlands; U: Urban and Built-up Lands; C/N: Cropland/Natural Vegetation Mosaics; Snow: Permanent Snow and Ice; Barren: Non-vegetated Barren (sand, rock, soil); Water: Water Bodies. Right panel: Kernel density plot for cSIF vs. VOD for Crops, Grass, Deciduous Broadleaf Forests, and Evergreen Broadleaf Forest.

Code availability

All code is written in Python, the analysis is conducted using Columbia University high performance computing clusters (Ginsburg), and is available at https://github.com/os2328/GLAB-dataset.

Received: 25 March 2024; Accepted: 20 August 2024;
Published online: 28 August 2024

References

1. Konings, A. G., Piles, M., Das, N. & Entekhabi, D. L-band vegetation optical depth and effective scattering albedo estimation from SMAP. *Remote Sensing of Environment* **198**, 460–470, https://doi.org/10.1016/j.rse.2017.06.037 (2017).
2. Tian, F. *et al.* Coupling of ecosystem-scale plant water storage and leaf phenology observed by satellite. *Nature ecology & evolution* **2**, 1428–1435, https://doi.org/10.1038/s41559-018-0630-3 (2018).
3. Konings, A. G., Holtzman, N. M., Rao, K., Xu, L. & Saatchi, S. S. Interannual variations of vegetation optical depth are due to both water stress and biomass changes. *Geophysical Research Letters* **48**, e2021GL095267, https://doi.org/10.1029/2021GL095267 (2021).
4. Dou, Y. *et al.* Reliability of using vegetation optical depth for estimating decadal and interannual carbon dynamics. *Remote Sensing of Environment* **285**, 113390, https://doi.org/10.1016/j.rse.2022.113390 (2023).
5. Wigneron, J.-P. *et al.* SMOS-IC data record of soil moisture and L-VOD: Historical development, applications and perspectives. *Remote Sensing of Environment* **254**, 112238, https://doi.org/10.1016/j.rse.2020.112238 (2021).
6. Forkel, M., Schmidt, L., Zotta, R.-M., Dorigo, W. & Yebra, M. Estimating leaf moisture content at global scale from passive microwave satellite observations of vegetation optical depth. *Hydrology and Earth System Sciences* **27**, 39–68, https://doi.org/10.5194/hess-27-39-2023 (2023).
7. Frappart, F. *et al.* Global monitoring of the vegetation dynamics from the vegetation optical depth (VOD): A review. *Remote Sensing* **12**, 2915, https://doi.org/10.3390/rs12182915 (2020).
8. Du, J. *et al.* A global satellite environmental data record derived from AMSR-E and AMSR2 microwave Earth observations. *Earth System Science Data* **9**, 791–808, https://doi.org/10.5194/essd-9-791-2017 (2017).
9. Liu, Y. Y., De Jeu, R. A., McCabe, M. F., Evans, J. P. & Van Dijk, A. I. Global long-term passive microwave satellite-based retrievals of vegetation optical depth. *Geophysical Research Letters* **38**, https://doi.org/10.1029/2011GL048684 (2011).
10. Moesinger, L. *et al.* The global long-term microwave vegetation optical depth climate archive (VODCA). *Earth System Science Data* **12**, 177–196, https://doi.org/10.5194/essd-12-177-2020 (2020).
11. Liu, Y. Y., van Dijk, A. I., McCabe, M. F., Evans, J. P. & de Jeu, R. A. Global vegetation biomass change (1988–2008) and attribution to environmental and human drivers. *Global Ecology and Biogeography* **22**, 692–705, https://doi.org/10.1111/geb.12024 (2013).
12. Karthikeyan, L., Chawla, I. & Mishra, A. K. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. *Journal of Hydrology* **586**, 124905, https://doi.org/10.1016/j.jhydrol.2020.124905 (2020).
13. Smith, T., Traxl, D. & Boers, N. Empirical evidence for recent global shifts in vegetation resilience. *Nature Climate Change* **12**, 477–484, https://doi.org/10.1038/s41558-022-01352-2 (2022).
14. Liu, L. *et al.* Tropical tall forests are more sensitive and vulnerable to drought than short forests. *Global Change Biology* **28**, 1583–1595, https://doi.org/10.1111/gcb.16017 (2022).

15. Tao, S. *et al*. Little evidence that Amazonian rainforests are approaching a tipping point. *Nature Climate Change* **13**, 1317–1320, https://doi.org/10.1038/s41558-023-01853-8 (2023).
16. Wigneron, J.-P. *et al*. Modelling the passive microwave signature from land surfaces: A review of recent results and application to the L-band SMOS & SMAP soil moisture retrieval algorithms. *Remote Sensing of Environment* **192**, 238–262, https://doi.org/10.1016/j.rse.2017.01.024 (2017).
17. Mialon, A. *et al*. Evaluation of the sensitivity of SMOS L-VOD to forest above-ground biomass at global scale. *Remote Sensing* **12**, 1450, https://doi.org/10.3390/rs12091450 (2020).
18. Vaglio Laurin, G. *et al*. Monitoring tropical forests under a functional perspective with satellite-based vegetation optical depth. *Global Change Biology* **26**, 3402–3416, https://doi.org/10.1111/gcb.15072 (2020).
19. Bousquet, E. *et al*. Influence of surface water variations on VOD and biomass estimates from passive microwave sensors. *Remote Sensing of Environment* **257**, 112345, https://doi.org/10.1016/j.rse.2021.112345 (2021).
20. Kerr, Y. H. *et al*. The SMOS soil moisture retrieval algorithm. *IEEE Transactions on Geoscience and Remote Sensing* **50**, 1384–1403, https://doi.org/10.1109/TGRS.2012.2184548 (2012).
21. Entekhabi, D. *et al*. SMAP handbook–soil moisture active passive: Mapping soil moisture and freeze/thaw from space. *SMAP Project* (2014).
22. Al Bitar, A. *et al*. The global SMOS level 3 daily soil moisture and brightness temperature maps. *Earth System Science Data* **9**, 293–315, https://doi.org/10.5194/essd-9-293-2017 (2017).
23. Chan, S. K. *et al*. Assessment of the SMAP passive soil moisture product. *IEEE Transactions on Geoscience and Remote Sensing* **54**, 4994–5007, https://doi.org/10.1109/TGRS.2016.2561938 (2016).
24. Chan, S. *et al*. Development and assessment of the SMAP enhanced passive soil moisture product. *Remote Sensing of Environment* **204**, 931–941, https://doi.org/10.1016/j.rse.2017.08.025 (2018).
25. Fernandez-Moran, R. *et al*. SMOS-IC: An alternative SMOS soil moisture and vegetation optical depth product. *Remote Sensing* **9**, 457, https://doi.org/10.3390/rs9050457 (2017).
26. Li, X. *et al*. A new SMAP soil moisture and vegetation optical depth product (SMAP-IB): Algorithm, assessment and inter-comparison. *Remote Sensing of Environment* **271**, 112921, https://doi.org/10.1016/j.rse.2022.112921 (2022).
27. Li, X. *et al*. The first global soil moisture and vegetation optical depth product retrieved from fused SMOS and SMAP L-band observations. *Remote Sensing of Environment* **282**, 113272, https://doi.org/10.1016/j.rse.2022.113272 (2022).
28. Li, X. *et al*. Alternate INRAE-Bordeaux Soil Moisture and L-Band Vegetation Optical Depth Products from SMOS and SMAP: Current status and overview. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, 2629–2632, https://doi.org/10.1109/IGARSS52108.2023.10283412 (IEEE, 2023).
29. Skulovich, O. & Gentine, P. A long-term consistent artificial intelligence and remote sensing-based soil moisture dataset. *Scientific Data* **10**, 154, https://doi.org/10.1038/s41597-023-02053-x (2023).
30. Liu, X. *et al*. A new global C-band vegetation optical depth product from ASCAT: Description, evaluation, and inter-comparison. *Remote Sensing of Environment* **299**, 113850, https://doi.org/10.1016/j.rse.2023.113850 (2023).
31. Rodríguez-Fernández, N. J. *et al*. Long term global surface soil moisture fields using an SMOS-trained neural network applied to AMSR-E data. *Remote Sensing* **8**, 959, https://doi.org/10.3390/rs8110959 (2016).
32. Kolassa, J., Gentine, P., Prigent, C. & Aires, F. Soil moisture retrieval from AMSR-E and ASCAT microwave observation synergy. part 1: Satellite data analysis. *Remote Sensing of Environment* **173**, 1–14, https://doi.org/10.1016/j.rse.2015.11.011 (2016).
33. Yao, P. *et al*. A long term global daily soil moisture dataset derived from AMSR-E and AMSR2 (2002–2019). *Scientific data* **8**, 143, https://doi.org/10.1038/s41597-021-00925-8 (2021).
34. Wigneron, J.-P. *et al*. Global carbon balance of the forest: satellite-based L-VOD results over the last decade. *Frontiers in Remote Sensing* **5**, 1338618, https://doi.org/10.3389/frsen.2024.1338618 (2024).
35. O'Neill, P. E. *et al*. SMAP algorithm theoretical basis document: Level 2 and 3 soil moisture (passive) data products (2021).
36. Nelson, M., Hill, T., Remus, W. & O'Connor, M. Time series forecasting using neural networks: Should the data be deseasonalized first? *Journal of forecasting* **18**, 359–367, (1999).
37. Zhang, G. P. & Qi, M. Neural network forecasting for seasonal and trend time series. *European journal of operational research* **160**, 501–514, https://doi.org/10.1016/j.ejor.2003.08.037 (2005).
38. Han, M. *et al*. A surface soil temperature retrieval algorithm based on AMSR-E multi-frequency brightness temperatures. *International Journal of Remote Sensing* **38**, 6735–6754, https://doi.org/10.1080/01431161.2017.1363438 (2017).
39. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**, 1345–1359, https://doi.org/10.1109/TKDE.2009.191 (2009).
40. Wang, M. *et al*. A consistent record of vegetation optical depth retrieved from the AMSR-E and AMSR2 X-band observations. *International Journal of Applied Earth Observation and Geoinformation* **105**, 102609, https://doi.org/10.1016/j.jag.2021.102609 (2021).
41. Maas, A. L. *et al*. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, **30-1**, 3 (Atlanta, GA, 2013).
42. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, https://doi.org/10.48550/arXiv.1412.6980 (San Diego, 2015).
43. Caldeira, J. & Nord, B. Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms. *Machine Learning: Science and Technology* **2**, 015002, https://doi.org/10.1088/2632-2153/aba6f3 (2020).
44. Skulovich, O., Gentine, P., Wigneron, J.-P. & Xiaojun, L. GLAB-VOD: Global L-band AI-Based Vegetation Optical Depth Dataset Based on Machine Learning and Remote Sensing. *Zenodo* https://doi.org/10.5281/zenodo.10306095 (2023).
45. Jackson, T., Chan, S. K., Bindlish, R. & Njoku, E. G. AMSR-E/AMSR2 Unified L2B Half-Orbit 25 km EASE-Grid Surface Soil Moisture, Version 1, https://doi.org/10.5067/IKQ0G7ODMLC7 (2018).
46. Njoku, E. G. AMSR-E/Aqua Daily L3 surface soil moisture, interpretive parameters & QC EASE-Grids, Version 2, https://doi.org/10.5067/AMSR-E/AE_LAND3.002 (2004).
47. Wang, M. *et al*. Satellite observed aboveground carbon dynamics in Africa during 2003–2021. *Remote Sensing of Environment* **301**, 113927, https://doi.org/10.1016/j.rse.2023.113927 (2024).
48. Saatchi, S. S. *et al*. Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the national academy of sciences* **108**, 9899–9904, https://doi.org/10.1073/pnas.1019576108 (2011).
49. Santoro, M. & Cartus, O. ESA biomass climate change initiative (biomass_CCI): Global datasets of forest above-ground biomass for the years 2010, 2017 and 2018, v2. *Centre for Environmental Data Analysis* https://doi.org/10.5285/5f331c418e9f4935b8eb1b836f8a91b8 (2021).
50. Lang, N. *et al*. Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. *Remote Sensing of Environment* **268**, 112760, https://doi.org/10.1016/j.rse.2021.112760 (2022).
51. Zhang, Y., Joiner, J., Alemohammad, S. H., Zhou, S. & Gentine, P. A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks. *Biogeosciences* **15**, 5779–5800, https://doi.org/10.5194/bg-15-5779-2018 (2018).
52. Li, X. *et al*. Global-scale assessment and inter-comparison of recently developed/reprocessed microwave satellite vegetation optical depth products. *Remote Sensing of Environment* **253**, 112208, https://doi.org/10.1016/j.rse.2020.112208 (2021).
53. Rodríguez-Fernández, N. J. *et al*. An evaluation of SMOS L-band vegetation optical depth (L-VOD) data sets: high sensitivity of L-VOD to above-ground biomass in Africa. *Biogeosciences* **15**, 4627–4645, https://doi.org/10.5194/bg-15-4627-2018 (2018).
54. Hu, H.-J., Xu, K., He, L.-C. & Wang, G.-X. A model for the relationship between plant biomass and photosynthetic rate based on nutrient effects. *Ecosphere* **12**, e03678, https://doi.org/10.1002/ecs2.3678 (2021).

Acknowledgements

O.S. and P.G. acknowledge support from the LEMONTREE (Land Ecosystem Models based on New Theory, obseRvations and ExperimEnts) project, funded through the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures programme. J.-P.W. acknowledges support from the CNES (Centre National d'Etudes Spatiales, France) TOSCA programme.

Author contributions

O.S. developed the methodology, created the datasets, analysed the results, and wrote the manuscript. X.L. provided SMOS TB, SMAP-IB VOD, biomass, and canopy height data, and advised on the methodology and analysis on the results. P.G. and J.-P.W. advised on the methodology and analysis on the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03810-2.

Correspondence and requests for materials should be addressed to O.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.