Remote Sensing of Environment 278 (2022) 113103

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Improving the characterization of global aquatic land cover types using multi-source earth observation data

Panpan Xu *, Nandin-Erdene Tsendbazar, Martin Herold, Jan G.P.W. Clevers, Linlin Li

Laboratory of Geo-Information Science and Remote Sensing, Department of Environmental Sciences, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB Wageningen, the Netherlands

ARTICLE INFO

Edited by Dr Marie Weiss

Keywords:
Global aquatic land cover mapping
Sentinel-1
Sentinel-2
ALOS/PALSAR
Multi-source earth observation data

ABSTRACT

The sustainable management of aquatic resources requires spatially explicit information on the water and vegetation presence of aquatic ecosystems. Previous Global Aquatic Land Cover (GALC) mapping has been focused on water bodies while lacking information on vegetation, and aquatic types have always been characterized by low accuracies in global land cover products, calling for specific attention to improve GALC mapping. The availability of a wealth of open Earth Observation (EO) data on cloud-computing platforms provides opportunities to map aquatic land cover globally. This study aims to evaluate the potential of multi-source freely available EO data, including optical, Synthetic Aperture Radar (SAR), and various ancillary datasets, for improving the characterization of aquatic land cover comprising both water and vegetation types on a global scale. Using different combinations of features derived from these data, the classification performance of five land cover classes (i.e., trees, shrubs, herbaceous cover, bare/sparsely vegetated lands, and water bodies) in aquatic areas was cross-validated. Results showed that Sentinel-2 data alone achieved similarly good overall accuracy as those combining multi-source data. However, the single-sensor Sentinel-2 data cannot discriminate highly mixed and spectrally similar types, such as shrubs, trees, and herbaceous vegetation. Integrating SAR features from the ALOS/PALSAR mosaic and Sentinel-1 data with optical features provided by Sentinel-2 data could help address this limitation to some extent. Although with a lower spatial and temporal resolution, the ALOS/PALSAR mosaic had a stronger impact on GALC classification than Sentinel-1 data when they were synergistically used. Features provided by ancillary datasets did not lead to significant improvement in the overall GALC classification. At class-level, topographic and soil features helped to reduce the commission error of shrubs, and climate variables were useful to improve the characterization of bare aquatic lands. The Global Ecosystem Dynamics Investigation (GEDI) forest canopy height dataset helped to characterize trees but also resulted in a decrease in accuracies of shrubs. By assessing multi-source earth observation data, this research represents an important step forward in the global mapping of comprehensive aquatic land cover types at high spatial resolution (i.e., 10 m).

1. Introduction

The aquatic ecosystem is one of the most productive ecosystems on the earth and provides essential ecosystem services for human beings, such as water retention and food security. Aquatic land cover is described as the land cover types that are significantly influenced by the presence of water over an extensive period of time around a year (Xu et al., 2020). Due to excessive human activities like cropland reclamation and urban development, global aquatic land cover (GALC) has undergone 35% losses since the 1970s (Ramsar Convention on Wetlands, 2018). The Sustainable Development Goal 6 puts emphasis on protecting and restoring water-related ecosystems, such as forests, wetlands, rivers, and lakes (United Nations, 2015). The monitoring of these aquatic ecosystems requires spatially explicit information on the water and also on the vegetation that affects the functioning of these systems (Mitsch and Gosselink, 2007).

Remote sensing provides an efficient way to monitor the spatial distribution of aquatic land cover, especially when applied to large scales. To date, a number of GALC datasets have been created, with a focus on providing information for water bodies (Pekel et al., 2016;

* Corresponding author at: Wageningen University & Research, Postbus 47, 6700 AA, Droevendaalsesteeg 3, 6708PB Wageningen, the Netherlands.
E-mail address: panpan.xu@wur.nl (P. Xu).

https://doi.org/10.1016/j.rse.2022.113103
Received 26 October 2021; Received in revised form 20 April 2022; Accepted 21 May 2022
Available online 2 June 2022

Verpoorter et al., 2014) or the possible existence of aquatic areas (Hu et al., 2017; Prigent et al., 2007). Although these datasets are useful to monitor the water in aquatic ecosystems, they fail to account for the vegetation that is required by a variety of applications, such as hydrological and climate modeling (Xu et al., 2020). The widely used Global Lakes and Wetlands Database (GLWD) (Lehner and Döll, 2004) provides a comprehensive description of different aquatic types. Sourcing from data in the 1980s, GLWD is outdated for monitoring the present status of GALC. Serving as the most fundamental geospatial data product, Global Land Cover (GLC) datasets also offer several classes to describe the vegetation for aquatic land cover. However, aquatic types have always been mapped with low accuracies in previous GLC products (Amler et al., 2015; Xu et al., 2021). With the increased availability of satellite data and improved computing capabilities, GLC mapping has progressed towards higher resolution, such as the FROM-GLC10 (Gong et al., 2019), WorldCover 2020 (Zanaga et al., 2021), and ESRI 2020 Land Cover (Karra et al., 2021) products with a 10 m-resolution. Despite the advantage of presenting more spatial details, aquatic land cover remains the most difficult to map (Gong et al., 2019) and is classified with relatively low accuracies in these high-resolution products (Gong et al., 2019; Tsendbazar et al., 2021b; Zanaga et al., 2021). Therefore, there is a critical need to update and improve the GALC mapping.

Unlike generic land cover, the characterization of aquatic land cover types is more challenging due to the interplay among vegetation and the underlying water and wet soils (Gallant, 2015). Optical imagery, providing the basic spectral information for land objects, has been widely used in characterizing water bodies (Pekel et al., 2016) and aquatic vegetation (Adam et al., 2010). Despite the advantage of multispectral images, cloud cover is a limiting factor for optical remote sensing systems. Synthetic Aperture Radar (SAR) data have the ability to penetrate through clouds and thus are less vulnerable to cloud contaminations. SAR can also penetrate into vegetation canopies, allowing them to collect information about the physical structure of vegetation (Mahdavi et al., 2018).

Synergistically using multi-source data is an effective way to overcome the limitations of a single sensor and improve the classification accuracy of aquatic land cover (Corcoran et al., 2013). Specifically, the combination of the high spatial and temporal resolution Sentinel-1 (S1) SAR and Sentinel-2 (S2) multispectral data has been employed to map the complex characteristics of aquatic land cover (Slagter et al., 2020). Although the integrated use of S1 and S2 data is often applied at local or regional scales (Mahdianpari et al., 2020; Slagter et al., 2020), it has not been explored for GALC mapping.

In general, SAR data with short wavelengths (e.g., C-band) are often used to map herbaceous vegetation (Mahdavi et al., 2018). Time-series of S1 data have been reported helpful for capturing flooded grasslands (Tsyganskaya et al., 2018). However, to penetrate further into the high vegetation canopy and sense the understory of water and vegetation, the longer-wavelength SAR (e.g., L-band) is required. The Advanced Land Observing Satellite Phased Arrayed L-band Synthetic Aperture Radar (ALOS/PALSAR) archives provide one of the most frequently used L-band SAR data (Rosenqvist et al., 2014). Due to the difficulty in acquiring and processing PALSAR-1 data globally and because PALSAR-2 data are not for free, the L-band SAR has not been applied for operational aquatic land cover mapping. However, the yearly updated ALOS/PALSAR mosaic (JAXA, 2016) provides a global SAR image, which offers an alternative for using individual L-band SAR images on a global scale. To date, it has not been assessed for GALC mapping.

Besides the spaceborne satellite data, various ancillary datasets such as topographic, climate, and soil data can provide complementary information for discriminating between different aquatic types, due to the fact that the occurrence of aquatic land is strongly influenced by the topography, climate conditions, and soil attributes (Mitsch and Gosselink, 2007). The confusion of the spectrally similar trees and shrubs has been an issue in GLC mapping (Xu et al., 2021). A recently published global forest canopy height product (Potapov et al., 2021), which provides continuous estimation of forest canopy height, could potentially be used for the classification of trees and shrubs that are characterized by different heights. Despite the wealth of ancillary datasets, it remains uncertain whether these data are effective for improving GALC mapping.

The development of cloud-computing platforms such as Google Earth Engine (GEE, Gorelick et al., 2017) allows users to access and analyse tremendous volumes of earth observation (EO) data efficiently. The S1 and S2 imagery as well as various ancillary datasets are readily available on the GEE platform, bringing new opportunities for the global-scale mapping of aquatic land cover types. In this study, our goal is to evaluate the potential of multi-source EO data including optical, SAR, and various ancillary datasets for improving the GALC characterization of both water and vegetation types. By exploring different combinations of data sources, we intend to find the important input variables that could improve the characterization of the water and vegetation for global aquatic ecosystems.

2. Materials and methods

2.1. Analysis overview

As this study focused on improving the land cover classification in aquatic areas, we used an existing map to predefine the baseline of global aquatic areas. For this purpose, the integrated GALC map from Xu et al. (2021) was taken. Furthermore, we conducted a point-level analysis in this study using the reference dataset (Section 2.2) provided by the Copernicus Global Land service GLC (CGLS-LC100) mapping project (Buchhorn et al., 2020; Tsendbazar et al., 2021a). This means we did not create a wall-to-wall map, but rather used the globally distributed reference aquatic sample sites to assess the value of multi-source EO data in global aquatic land cover mapping.

Following the United Nations Land Cover Classification System (LCCS)-based GALC characterization framework (Xu et al., 2020), land cover types mapped in this study include trees, shrubs, herbaceous cover, bare/sparsely vegetated lands (hereafter referred to as bare lands), and water bodies. Such a design allows users to adapt these basic land cover types to their own legends or combine them with other thematic maps (e.g., water seasonality) to derive their required information (Xu et al., 2020). Specific definitions of the five types can be found in Table S1 of the supplementary material.

The output map was intended for 2019–2020 with a 10 m spatial resolution. An overview of our analytical workflow can be seen in Fig. 1 and steps taken are detailed in the following sections. GEE was used for most of the data collection and preprocessing, and R (R Core Team, 2021) for modeling and analysis.

2.2. Reference dataset

In this study, we utilized the CGLS-LC100 global validation dataset (Tsendbazar et al., 2021a) as our reference data for both training and validation. It was developed for validating the annual (2015–2019) GLC maps of the CGLS-LC100 product (Buchhorn et al., 2020). For our purpose, we used the data from the year 2019. The reference dataset includes 21,752 sample locations across the globe (Tsendbazar et al., 2021a), which were created using a stratified random sampling design. Each sample location corresponded to a 100 m × 100 m area, and it was then divided into 10 × 10 subpixels (10 m × 10 m). The reference cover type was collected at the subpixel level by interpreting high-resolution satellite images on the GeoWiki platform (https://www.geo-wiki.org/). For more details about this reference dataset, please refer to Tsendbazar et al. (2018, 2021a).

Eleven classes were originally included in the reference dataset following the LCCS classification scheme (Tsendbazar et al., 2021a), and we selected eight of them related to this study and merged these classes to obtain the five targeted aquatic types, namely trees (i.e.,

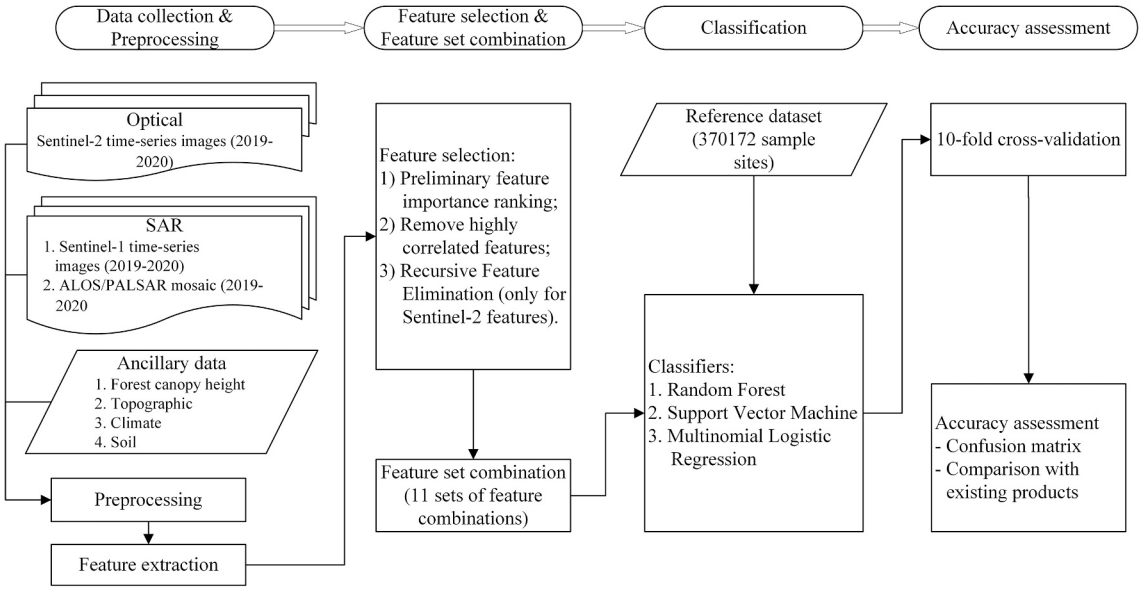

Fig. 1. Methodological workflow of this study.

corresponding to trees in the reference dataset), shrubs (i.e., corresponding to shrubs), herbaceous cover (i.e., corresponding to grassland, herbaceous wetland, moss/lichen, and cropland), bare lands (i.e., corresponding to bare/sparse vegetation), and water bodies (i.e., corresponding to water bodies).

As the initial reference dataset covered both aquatic and non-aquatic areas globally, we restrained the sample sites within aquatic areas using the integrated GALC Level-1 map from Xu et al. (2021). As a result, the data included 3801 sample locations comprising 370,172 subpixel sample sites (10 m) across the globe (80°N-56°S). The fact that not exactly 380,100 subpixel sample sites were included in these 3801 locations was because the integrated GALC Level-1 map could be partially covered by the sample locations. The spatial distribution of the reference dataset used in this study is shown in Fig. 2. The number of reference sample sites used for each class is shown in the legend.

2.3. Data collection and preprocessing

An overview of the data used in this study is given in Table 1. After preprocessing, a variety of features (Table 2) were obtained from the data sources.

2.3.1. Optical data

A total of 31,062,747 Level-2A S2 multispectral images at the reference sample sites during the observation period 2019-01-01 to 2020-12-31 with a cloud-cover of less than 20% were sourced from the GEE platform. The spatial distribution of valid S2 observations is shown in Fig. 3b. The images have already been atmospherically corrected, and

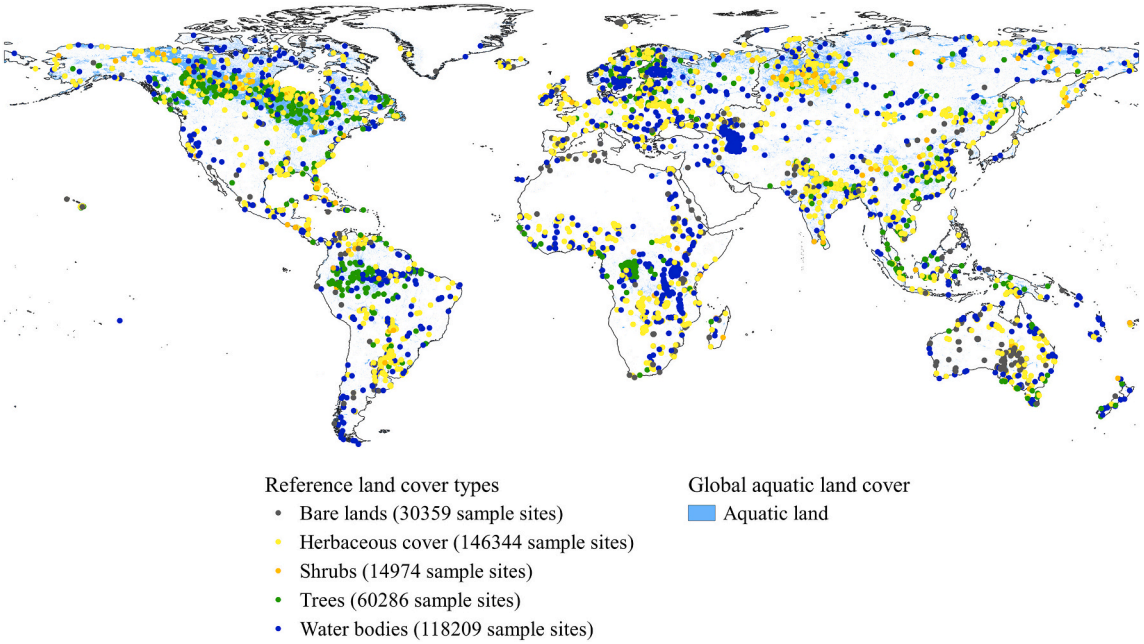

Fig. 2. Spatial distribution of the reference dataset used in this study. Numbers in brackets indicate the amount of reference sample sites used for each class. The global aquatic land cover map was sourced from Xu et al. (2021).

Table 1
Overview of the data used in this study.

Source	Product	Period of data	Spatial resolution	Temporal resolution	Coverage
Optical	Sentinel-2 MSI, Level-2A	2019–2020	10 m, 20 m	5-day	83°N-56°S
SAR	Sentinel-1 SAR GRD	2019–2020	10 m	6-day	85°N-60°S
	ALOS/PALSAR mosaic	2019–2020	25 m	Yearly	85°N-58°S
Topographic	SRTM DEM	2000	30 m	–	60°N-56°S
	Global Topographic Index by Marthews et al. (2015)	Circa 2000	15-arcsecond (~450 m at the Equator)	–	86°N-56°S
Forest canopy height	GEDI global forest canopy height	2019	30 m	–	52°N-52°S
Soil	OpenLandMap	1950–2017	250 m	–	84°N-56°S
Climate	WorldClim Version2 Bioclimatic variables	1970–2000	30-arcsecond (~1 km at the Equator)	–	84°N-56°S

Table 2
Overview of the features derived from the data sources. Detailed descriptions and calculation formulas of the features are presented in Tables S2-S4 (supplementary material).

Source	Product	Feature description	Features	Temporal metrics	Number of features
Optical	Sentinel-2	Spectral bands	B2; B3; B4; B8; B8A; B11; B12	Mean; Minimum; Maximum; Median; Standard deviation; 10th percentile; 90th percentile	119
		Water/vegetation indices	NDVI (Normalized Difference Vegetation Index); reNDVI (red-edge Normalized Difference Vegetation Index); mNDWI (modified Normalized Difference Water Index); SWI (Sentinel-2 Water Index); ND_NirSwir (Normalized Difference of NIR and SWIR2 bands); ARI (Anthocyanin Reflectance Index); NMDI (Normalized Multi-band Drought Index); AWEI (Automated Water Extraction Index); SAVI (Soil Adjusted Vegetation Index); REIP (Red Edge Inflection Point)		
SAR	Sentinel-1	SAR backscattering	VV; VH	Mean; Minimum; Maximum; Median; Standard deviation; 10th percentile; 90th percentile	36
		Polarization-derived features	Ratio; Normalized Difference		
		Texture features	Variance; Correlation; Contrast; Difference entropy (calculated for VV and VH, respectively)	Mean	
	ALOS/PAL-SAR mosaic	SAR backscattering	HH; HV		
		Polarization-derived features	Ratio; Normalized Difference	Mean	12
		Texture features	Variance; Correlation; Contrast; Difference entropy (calculated for HH and HV, respectively)		
Topogr-aphic	SRTM DEM	–	Elevation; Slope; Aspect; TPI (Topographic Position Index)	–	5
	Global Topographic Index	–	TWI (Topographic Wetness Index)	–	
Forest canopy height	GEDI global forest canopy height	–	Forest canopy height	–	1
Soil	OpenLandM-ap	–	Soil water content; Soil organic carbon content	–	2
Climate	WorldClim Version2 Bioclimatic variables	–	Temperature seasonality; Max temperature of warmest month; Min temperature of coldest month; Precipitation seasonality; Precipitation of wettest month; Precipitation of driest month	–	6

we further preprocessed them to mask clouds and cloud shadows, using the cloud masking algorithm developed by Braaten (2021). Spectral bands at a 20 m-resolution (i.e., B8A, B11, B12) were resampled to 10 m using the nearest neighbor resampling.

With the preprocessed time-series images in 2019–2020, two groups of features were extracted, namely the spectral bands and water/vegetation indices. The 17 optical features as shown in Table 2 have been used in various aquatic land cover classification research (DeLancey et al., 2019; Ludwig et al., 2019). From the time-series images, we further calculated seven temporal metrics over the entire 2019–2020 period, including the minimum, maximum, mean, median, 10th percentile, 90th percentile, and standard deviation. With the seven temporal metrics for each of the 17 optical features, a total of 119 variables (Table 2) were acquired from S2 data.

2.3.2. SAR data

A total of 38,554,729 S1 images at the reference sample sites from 2019-01-01 to 2020-12-31 were sourced from GEE, and the spatial distribution of S1 observations can be seen in Fig. 3a. The C-band S1 SAR data available on GEE are Ground Range Detected (GRD), acquired in Interferometric Wide swath mode with dual-polarization (VV and VH) images. They have been preprocessed as a Level-1 product after thermal noise removal, radiometric calibration, and terrain correction. Here, we further processed the data following a framework proposed by Mullissa et al. (2021) to obtain the analysis-ready SAR backscattering data. A border noise correction to remove border artifacts and a refined-Lee-filter for speckle filtering were implemented on the SAR time-series images within GEE.

The 25 m ALOS/PALSAR yearly mosaic is provided by the Japan Aerospace Exploration Agency (JAXA), created by mosaicking SAR images measured by PALSAR-1 or PALSAR-2 available each year (JAXA, 2016). In this study, mosaics were acquired for 2019 and 2020 (i.e., one mosaic per year). The data have been ortho-rectified and slope-corrected. We further applied a focal median filter (GEE, 2021) with a window size of 5 × 5 pixels to the image to reduce speckle effects. The data were in digital number (DN) and were converted to gamma-naught (γ^0) values in GEE using Eq. (1) (JAXA, 2016). The resulting data were resampled to 10 m before calculating features.

$$\gamma^0 = 10 \times log_{10}\left(DN^2\right) - 83 \quad (1)$$

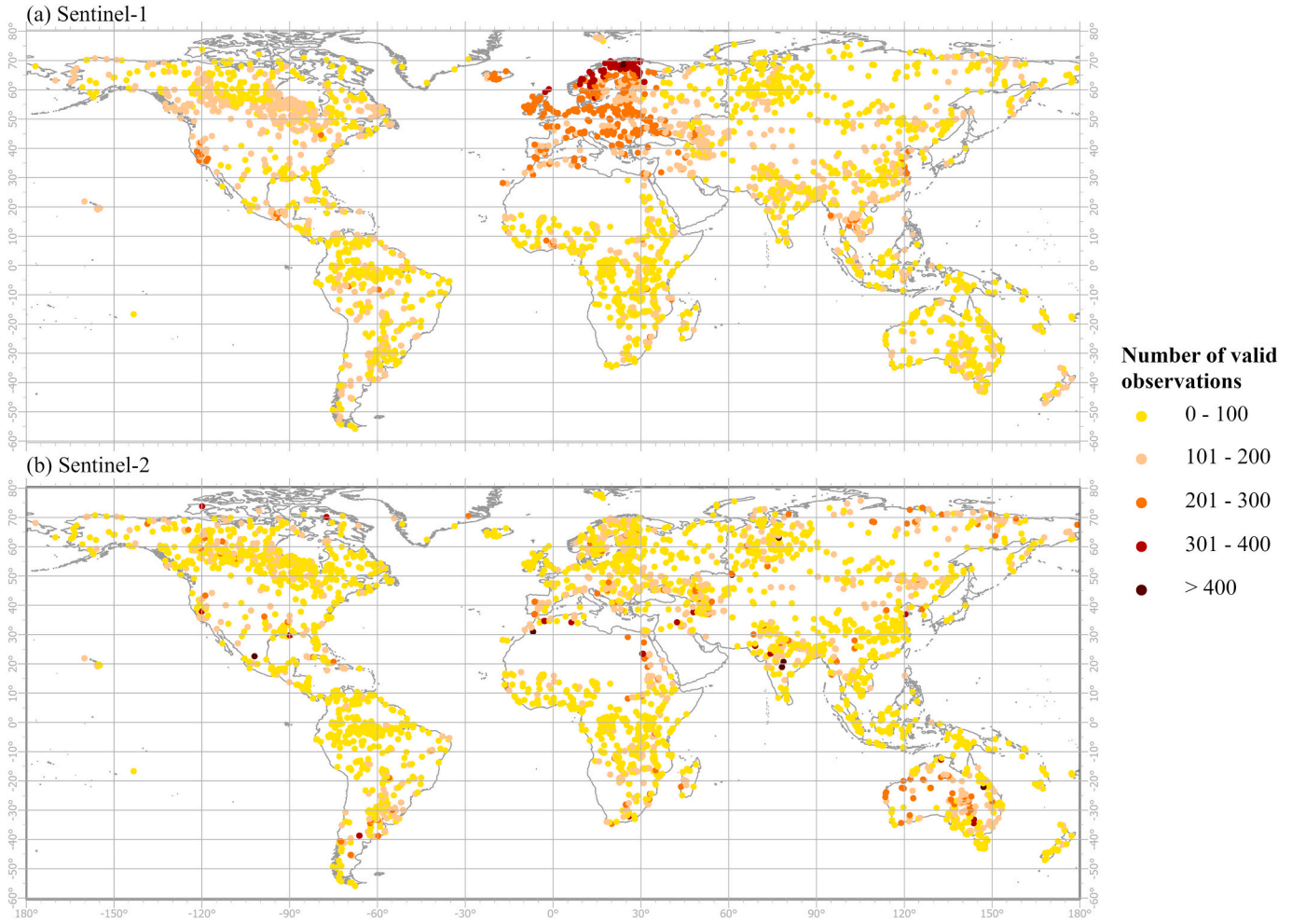

Fig. 3. Spatial distribution of the number of valid (a) Sentinel-1 and (b) Sentinel-2 observations during 2019–2020 at the reference sample sites.

Three groups of SAR features were extracted, including the SAR backscattering, polarization-derived features, and texture features (Table 2). The SAR backscattering was referred to VV, VH for S1, and HH, HV for PALSAR. Polarization-derived features comprised the ratio (e.g., $\frac{VV}{VH}$) and normalized difference (e.g., $\frac{HH-HV}{HH+HV}$) of the SAR bands. For S1 data, seven temporal metrics (Table 2) of the SAR backscattering and polarization-derived features were calculated over the entire time period. For PALSAR data, the mean-composite of the two images in 2019–2020 was used to derive the aforementioned two groups of features.

Texture measures from the Gray Level Co-occurrence Matrix (GLCM) provide valuable information on the pixel spatial relationships in a SAR image (Haralick et al., 1973). Based on GLCM, four texture features were used in this study, namely the variance, correlation, contrast, and difference entropy (Conners et al., 1984; Haralick et al., 1973). These texture features were calculated based on the mean-composite of the time-series SAR images for the VV, VH, HH, and HV bands. The GLCM neighborhood size was set to 5 for the calculation.

Finally, a total of 36 and 12 features were obtained from S1 and PALSAR data (Table 2), respectively.

2.3.3. Ancillary data

The topographic data used in this study originated from two sources. The first was taken from NASA's Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM), which describes global elevation in meters above sea level in the year 2000 with a 30 m-resolution (Farr et al., 2007). From the SRTM DEM, the elevation, slope, and aspect were extracted within GEE at 10 m-resolution. As GEE does not provide algorithms to calculate the Topographic Position Index (TPI, Wilson and Gallant, 2000) and the Topographic Wetness Index (TWI, Beven and Kirkby, 1979), two measures were taken to obtain the two features: 1) We exported the SRTM DEM out of GEE to calculate the TPI in GDAL (GDAL/OGR contributors, 2021); and 2) An analysis-ready TWI dataset generated by Marthews et al. (2015), serving as our second source of topographic data, was adopted to provide the global TWI estimate. This TWI dataset is calculated from the HydroSHEDS data (Lehner et al., 2008) and processed at a 15-arcsecond resolution under the WGS84 reference system. We resampled the TPI and TWI datasets to 10 m in GDAL, and then extracted the TWI and TPI values at the reference sample sites using the point sampling tool in QGIS.

The 30 m global forest canopy height dataset (Potapov et al., 2021) is developed by integrating Landsat time-series with the Global Ecosystem Dynamics Investigation (GEDI) lidar forest structure measurements. The data describe forest canopy height in meters within the latitude range 52°N ~ 52°S for the year 2019. On GEE, the data are split into seven continental mosaics, and we combined these mosaics to form a global image. The forest canopy height values at the reference sample sites were extracted at 10 m-resolution within GEE.

The climate data were adopted from the bioclimatic variables of the WorldClim Version2 product (Fick and Hijmans, 2017), which is available at Hijmans et al. (2020). These variables represent annual trends, seasonality, and extreme or limiting environmental factors. The dataset is at 30-arcsecond resolution and we resampled it to 10 m in GDAL. Six bioclimatic variables about the temperature and precipitation were used

in this study (Table 2). Values of these variables at the reference sample sites were extracted using QGIS.

Soil features were sourced from the soil water content (at 33 kPa, Field Capacity) and the soil organic carbon content product of the OpenLandMap (OpenGeoHub Foundation, 2018). These maps are developed based on machine learning predictions from a global compilation of soil profiles and samples (Hengl and MacMillan, 2019). The soil water content and the soil organic carbon content are both predicted at 6 standard depths (0, 10, 30, 60, 100, and 200 cm) at 250 m-resolution. In this study, we chose the 30 cm depth to represent the near-surface soil properties. Values of the two soil features at the reference sample sites were extracted in GEE at 10 m-resolution.

2.4. Feature selection and feature set combination

Among the optical features acquired from the S2 data, some are highly correlated and may contain redundant information that would increase the computation time and diminish the classification performance (Stromann et al., 2020). To minimize the multicollinearity, variables that have a high correlation (Pearson's correlation coefficient $\geq$ 0.9) with any of the high-importance variables, which were determined based on the impurity metrics provided by the random forest algorithm (Breiman, 2001), were iteratively removed. After this process, 64 out of the original 119 optical features remained. We then implemented the Recursive Feature Elimination (RFE, Guyon et al., 2002), a widely used feature selection algorithm, to exclude redundant variables from the remaining features. By ranking-ordering the features iteratively, RFE removes the ones with the least importance until finding the optimal combination of features. Finally, 57 out of the 64 variables were selected by RFE and they were used as S2 features in the classification model (Section 2.5).

For the S1 SAR features, only the multicollinearity was assessed using the same method as applied to S2 features. After removing the highly correlated ones, 24 out of the original 36 S1 features were retained. As the number of features from PALSAR and ancillary datasets were limited (Table 2), we did not implement the feature selection for these datasets. A detailed list of the features selected from each data source is shown in Table S5 (supplementary material).

We constructed 11 feature sets (Table 3) to assess their performance for GALC characterization, starting from the single-sensor optical or SAR features. To evaluate the added value of ancillary data, the GEDI forest canopy height, topographic, climate, and soil datasets were iteratively added onto the satellite-data-based S1-S2-PALSAR (i.e., S1S2P) combination.

Table 3
Different feature sets used for GALC characterization. Specific features selected from each data source are listed in Table S5 (supplementary material).

Abbreviation of the feature set	Data source	Number of features
P	PALSAR	12
S1	Sentinel-1	24
S1P	Sentinel-1, PALSAR	36
S2	Sentinel-2	57
S2P	Sentinel-2, PALSAR	69
S1S2	Sentinel-1, Sentinel-2	81
S1S2P	Sentinel-1, Sentinel-2, PALSAR	93
S1S2PF	Sentinel-1, Sentinel-2, PALSAR, Forest canopy height	94
S1S2PFT	Sentinel-1, Sentinel-2, PALSAR, Forest canopy height, Topographic	99
S1S2PFTC	Sentinel-1, Sentinel-2, PALSAR, Forest canopy height, Topographic, Climate	105
S1S2PFTCS	Sentinel-1, Sentinel-2, PALSAR, Forest canopy height, Topographic, Climate, Soil	107

2.5. Classification and accuracy assessment

Three classifiers including the Random Forest, Support Vector Machine, and the Multinomial Logistic Regression classifier were used for training and prediction in this study to obtain a reliable evaluation on the contribution of multi-source EO data. Random Forest (RF; Breiman, 2001) is an ensemble machine learning model that combines a set of decision trees constructed using a random subset in the training data. When working in a classification environment, the final prediction is determined by taking the majority of the predictions made by each individual decision tree in the forest. Support Vector Machine (SVM; Vapnik, 1999) is a machine learning algorithm which tries to find an optimal hyperplane to categorize data into different classes. As SVM does not support multiclass classification in its most basic type, the multiclass classification is broken down into multiple binary classifications. Multinomial Logistic Regression (MLR; Theil, 1969) is a classification method that allows for more than two categories of the dependent variable. The final class is assigned by calculating the probability of category membership on a dependent variable based on multiple independent variables.

The h2o package (Landry, 2016) in R was used to run RF and MLR models, and the liquidSVM package (Steinwart and Thomann, 2017) was used for implementing SVM. Both packages allow parallel processing. The setting of some key parameters of the RF, SVM, and MLR models are listed in Table S6 of the supplementary material, and we kept them as a constant while training the classification models for the 11 feature sets. Some reference sample sites may have missing values of some features caused by the cloud removal or the missing coverage of the source data (see Fig. S1 for the percentage of NA in each data source). For instance, the GEDI forest canopy height dataset only covers 52°N ~ 52°S. Thus, samples located outside this range have a null value. The RF model in h2o internally deals with missing values without imputation by minimizing the loss function when making a split decision for every node (H2O.ai, 2021a). As the MLR and SVM model cannot deal with missing values, we used the mean of valid observations to impute missing values.

A 10-fold cross-validation was applied to train the classification models as well as evaluating the classification accuracy. The reference dataset was partitioned into 10 random subsets based on the 100 m × 100 m sample locations (i.e., location IDs). Nine of them were used for training and the remaining one was used to validate the prediction accuracy. This was repeated 10 times, each time reserving a different subset for validation. Partitioning based on sample location level instead of subpixel level was chosen mainly to limit the influence of possible autocorrelation, thus preventing a subpixel from being selected for validation while the neighboring subpixels were selected for training. The number of reference samples used per class was shown in Fig. 2.

The overall accuracy (OA), user's accuracy (UA), and producer's accuracy (PA) were calculated as the median values based on confusion matrices produced from the 10-fold cross-validation to evaluate the performance of the predictions. Additionally, a Welch's *t*-test (Welch, 1947) in R was implemented to determine whether a statistically significant difference exists between different classification scenarios.

To gain insight into how well predictions of this study compare with currently available global products, we selected four existing datasets and assessed their accuracies using the same reference dataset applied to our study. We used the CGLS-LC100 discrete map (Buchhorn et al., 2020) for 2019, WorldCover 2020 (Zanaga et al., 2021), the Hansen Global Forest Change (GFC) dataset version 1.7 (Hansen et al., 2013) describing forest changes for 2000–2019, and the Global Surface Water (GSW) yearly history for 2019 (Pekel et al., 2016) that identifies seasonal water and permanent water. All these datasets are available on GEE and we clipped them to the same aquatic areas using the integrated GALC Level-1 map (Xu et al., 2021) applied to our study. From the CGLS-LC100 discrete map and the WorldCover 2020 dataset, 19 out of 23 and 9 out of 11 land cover classes (Table S7 in the supplementary material),

respectively, were selected and reclassified into the five basic land cover types mapped in this study. To obtain trees in the year 2019 from the Hansen GFC data, we used the map in 2000 with a threshold of >50% tree cover as the basic tree cover extent, and excluded all forest losses from 2000 to 2019 from the basic map. As the seasonal water of the GSW dataset could be a mixture of water and temporarily flooded bare or vegetated types, we used only the permanent water to represent pure water bodies.

The variable importance produced by the RF and MLR model were used to assess the importance of features. In h2o, RF evaluates feature importance using the Gini impurity measure and MLR uses the regression coefficient to represent the feature importance (H2O.ai, 2021b). In this study, importance scores were calculated based on the trained classification model for the feature set combining all data sources. Standard deviations of the importance scores were also calculated based on the 10-fold outputs to assess the variability of these scores.

3. Results

3.1. Classification accuracy analysis

The cross-validated overall accuracies of the three classifiers are presented in Fig. 4. According to the classification results of the three classifiers, the highest OA was obtained with the feature set combining all available data sources based on RF (83.2%) and MLR (82.3%), whereas SVM achieved the highest OA (78.3%) when combining S1, S2, PALSAR, and GEDI forest canopy height. The classification based on the PALSAR-only data resulted in the lowest OA, which was 67.3%, 67.7%, and 65.5% based on MLR, RF, and SVM, respectively. In terms of statistical significance (i.e., Welch's *t*-test based on the 10-fold OA), the single-sensor S2 data and all multi-source feature sets significantly improved the OA compared to using the single-sensor S1 or PALSAR data. However, there is no significant difference among using multi-source feature sets compared to the single-sensor S2 data tested at the 95% confidence level.

The cross-validated class-specific accuracies derived from the three classifiers are shown in Fig. 5. We chose the best-performing RF prediction derived from the feature set combining all data sources to visualize the GALC prediction map (Fig. 6). Both the aquatic land cover types and correctness of the prediction (i.e., the commission, omission, and correct prediction) are visualized. The confusion matrix of this best prediction is presented in Table 4. The confusion matrices of RF predictions for the other feature sets can be found in Tables S8-S17 (supplementary material). As an illustration of the high-resolution map, we applied the 11 RF models that have been trained using the global reference dataset to a local area, i.e., the St. Lucia wetland in South Africa (32° 26′ 34.8"E, 28° 15' 32.4"S), to predict images of the 11 feature sets. From Fig. 7, it could be observed that at the local scale, the classification map derived from the S2-only feature set (Fig. 7e) was quite similar with those (Fig. 7f-l) integrating multi-source data.

Herbaceous cover had more errors of commission (100% - UA) than omission (100% - PA) and it tended to be overestimated at the cost of shrubs, trees, and bare aquatic lands (Table 4). Spatially, the commission error was higher in river floodplains (e.g., Paraguay River in South America, Indo-Gangetic Plain of India) and high northern latitudes (50°N ~ 70°N, Fig. 6a). Among all the feature sets, S2 data alone achieved a good performance in identifying herbaceous cover (Fig. 5a). The S1S2, S2P, and S1S2P feature set increased the UA of herbaceous cover by 0.6% ~ 1.4%, 1.1% ~ 1.9%, and 1.4% ~ 3.2% (based on the three classifiers) compared to the S2-only feature set, respectively. Incorporating the four ancillary datasets did not bring much improvement (Fig. 5a). Note that the accuracy used for comparison in this section was the median accuracy acquired from the 10-fold cross-validation.

Shrubs were predicted with the least accuracies and suffered

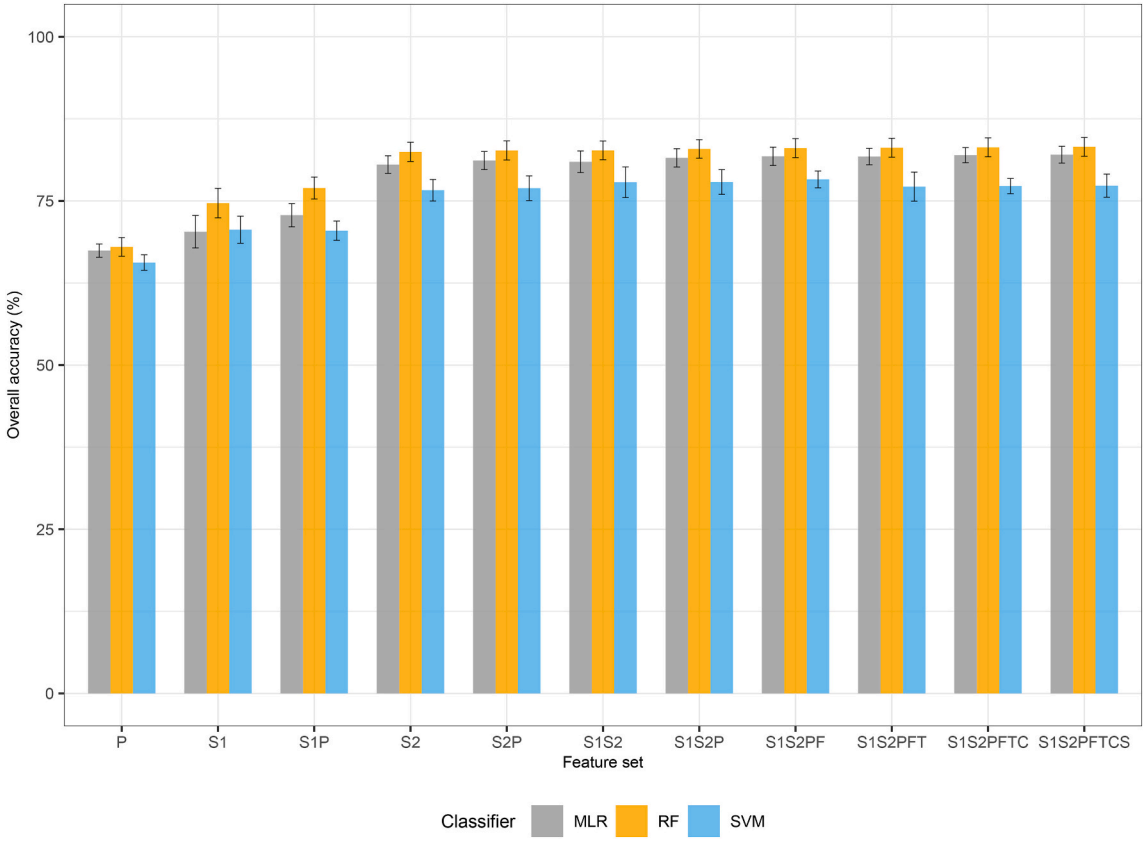

Fig. 4. The median overall accuracies derived from the 10-fold cross-validation based on three classifiers for the 11 feature sets. The error bar denotes the standard deviation.

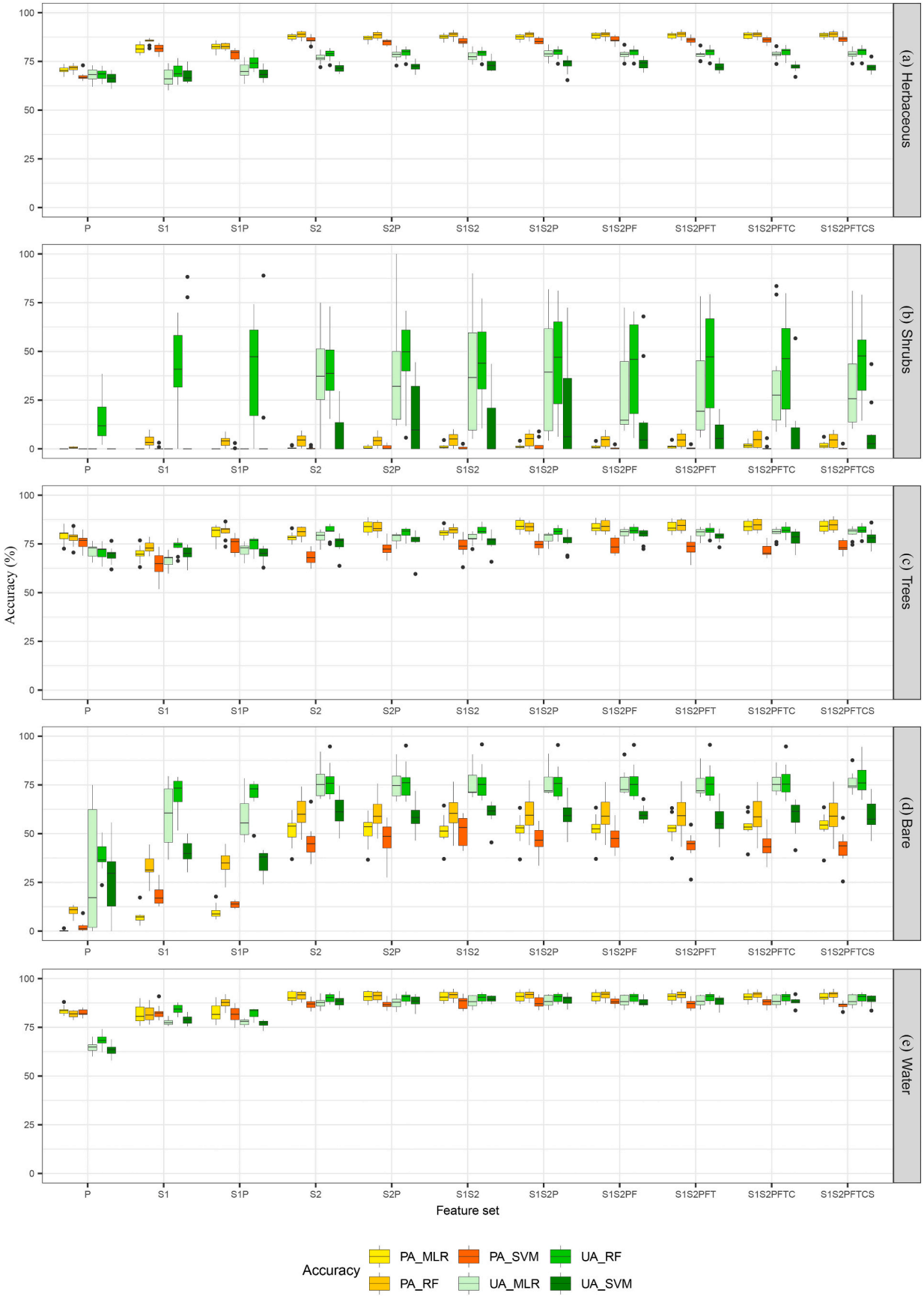

Fig. 5. Class-specific accuracies derived from the 10-fold cross-validation based on three classifiers for the 11 feature sets.

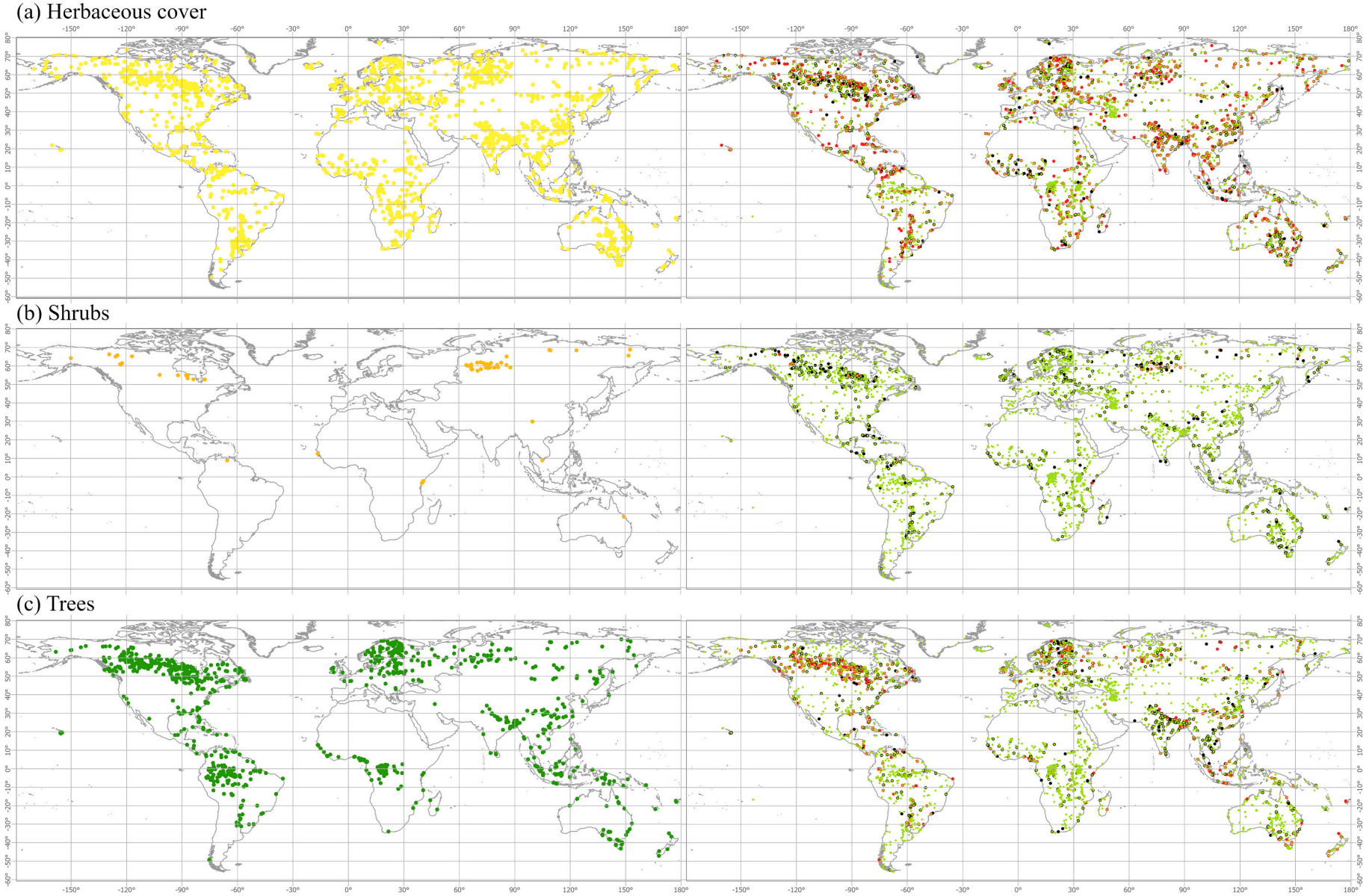

Fig. 6. Global aquatic land cover predictions derived from the best-performing RF prediction based on the feature set combining all data sources. In the correctness map, the "correct" class includes both the correctly predicted presence and the correctly predicted non-existence of the class; "commission" refers to the misclassification from other classes, and "omission" represents reference sample sites that are wrongly attributed to other types.

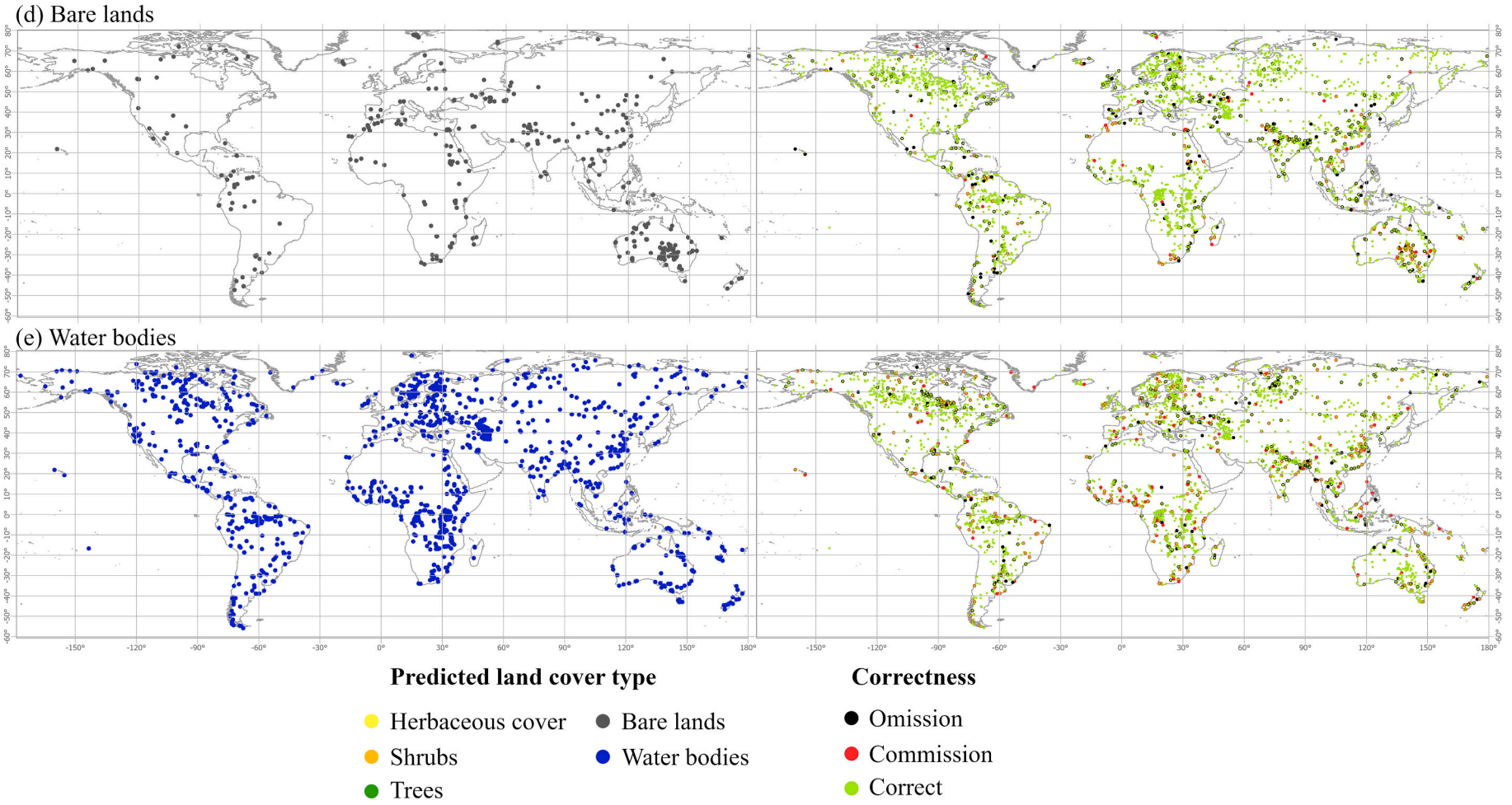

Fig. 6. (*continued*).

Table 4
The overall confusion matrix for the best-performing RF prediction derived from the feature set combining all available data (S1S2PFTCS).

		Reference						
		Herbaceous cover	Shrubs	Trees	Bare lands	Water bodies	Sum	User's accuracy (%)
Prediction	Herbaceous cover	129,959	9340	8325	7637	7468	162,729	79.9
	Shrubs	365	660	259	0	3	1287	51.3
	Trees	6048	4765	51,136	176	354	62,479	81.9
	Bare lands	3067	99	71	18,025	2049	23,311	77.3
	Water bodies	6905	110	495	4521	108,335	120,366	90.0
	Sum	146,344	14,974	60,286	30,359	118,209	370,172	
	Producer's accuracy (%)	88.8	4.4	84.8	59.4	91.7		83.2

considerable omissions (Table 4). Besides herbaceous cover, the majority of reference shrub sample sites were misclassified as trees. This was apparent in the boreal peatland regions (Fig. 6b), where dwarf shrubs in non-forested bogs can be misclassified as herbaceous vegetation (Fig. 6a), while shrubs in forested bogs are confused with short-stemmed trees (Fig. 6c) (Matthews and Fung, 1987). S2 data alone did not perform that well in characterizing shrubs (Fig. 5b). Instead, the S2P, S1S2, and S1S2P feature set improved the UA of shrubs by 11.1%, 5.3%, and 8.3% (based on RF) compared to using the single-sensor S2 data, respectively. Among the ancillary datasets, topographic features (i.e., S1S2PFT) increased the UA of shrubs by 1.3% and 4.6% based on RF and MLR compared to the S1S2PF feature set, respectively. The inclusion of soil features also brought 1.4% (RF) and 2.6% (SVM) increase in the UA. Integrating multi-source datasets did not improve the PA of shrubs, thus underestimation of shrubs remains critical.

Trees had considerable commissions in the abovementioned boreal peatland regions (Fig. 6c), resulting from its confusion with herbaceous vegetation and shrubs. The classification based on S1-only data had the lowest PA, i.e., the most omissions, of trees among the 11 feature sets, while the single-sensor PALSAR data performed better than S1 data in characterizing trees in aquatic areas (Fig. 5c). Although the single-sensor S2 data achieved relatively high accuracies for trees, there is still room for improvement. The S2P, S1S2, and S1S2P feature set increased the PA of trees by 1.5%, 0.9%, and 2.5% (based on RF

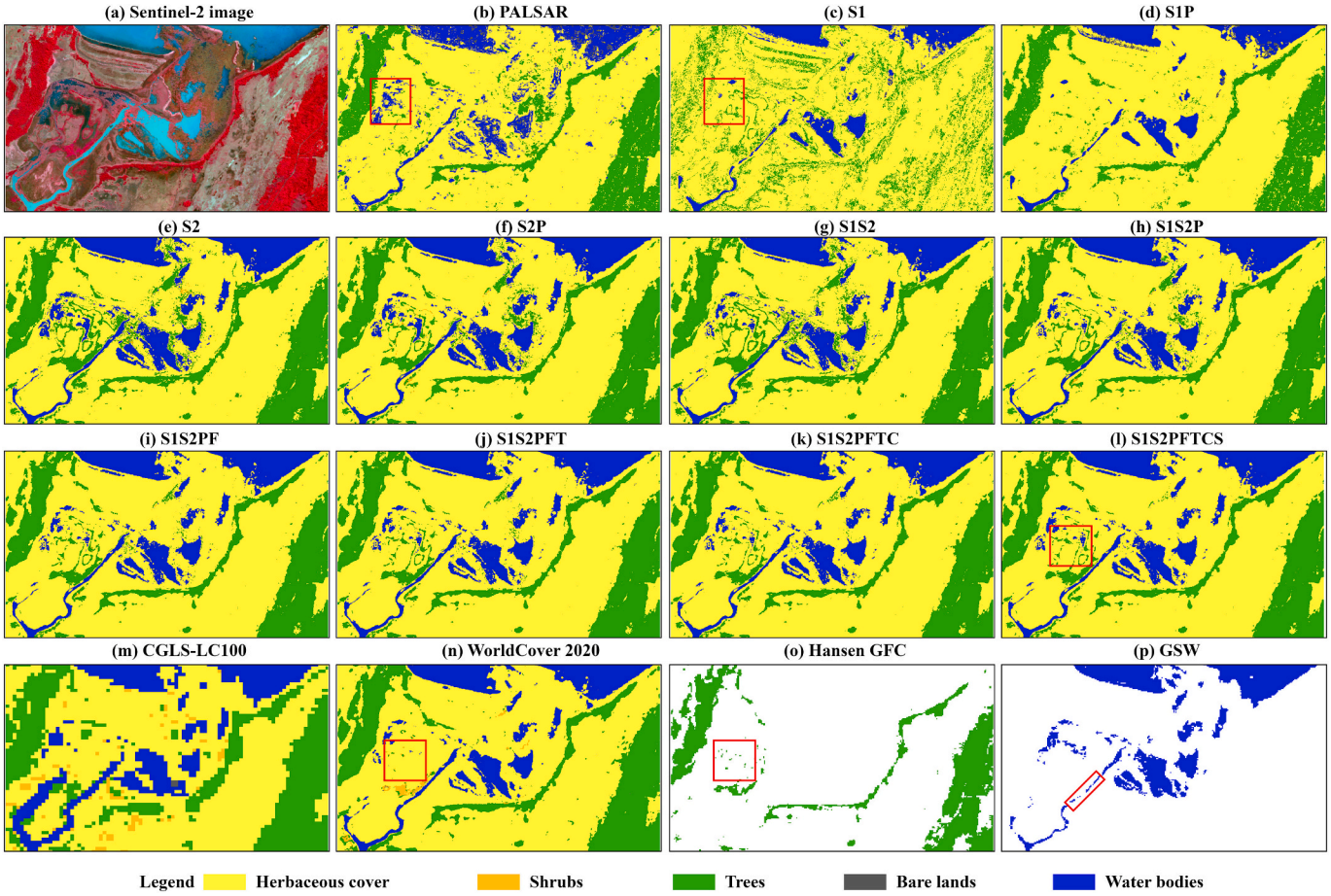

Fig. 7. Illustration of the prediction maps at local scale using a small area of St. Lucia wetland as an example. In this figure, (a) is a Sentinel-2 RGB image composited by the median B8, B4, and B3 bands; (b) ~ (l) are the RF prediction maps based on the 11 feature sets; (m) ~ (p) represent the CGLS-LC100, WorldCover 2020, Hansen GFC, and GSW dataset, respectively. Red boxes in (b) and (c) show a water area with submerged macrophytes. Red boxes in (l) and (n) show an area that is prone to cause misclassifications between herbaceous cover and trees. Red box areas in (o) and (p) show trees and water that were not detected by Hansen GFC and GSW dataset, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

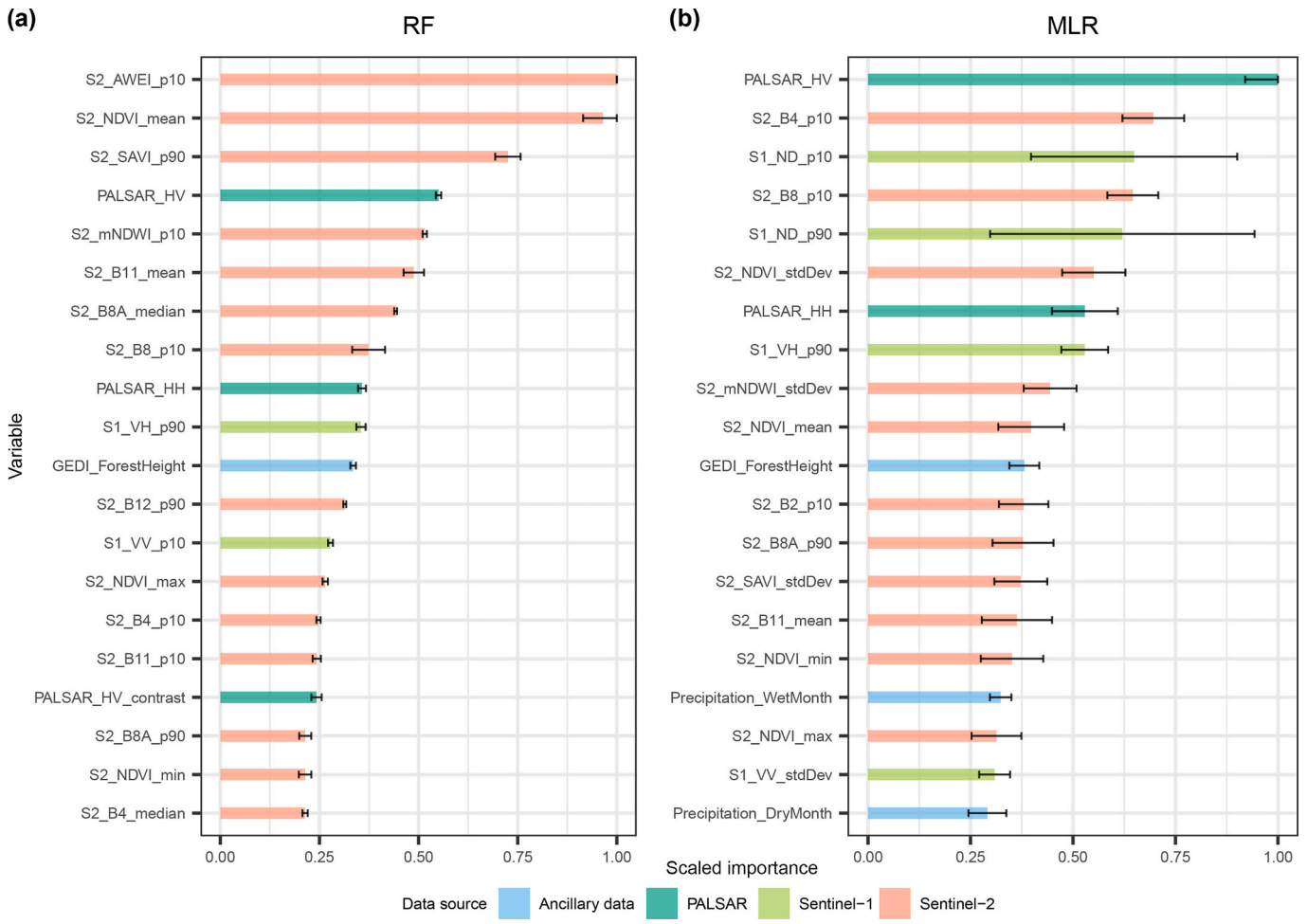

Fig. 8. Importance scores generated by the RF and MLR models for the top 20 features included in the feature set with all data sources. Scores shown in this figure are the median of the variable importance output from the 10-fold cross-validation. The error bar denotes the standard deviation.

Table 5
Comparison of class-specific accuracies between existing products and predictions (based on RF model) of this study.

Class	Best prediction of this study		S1S2 prediction of this study		CGLS-LC100		WorldCover 2020		Hansen GFC		GSW	
	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)	UA (%)	PA (%)
Herbaceous cover	79.9	88.8	78.9	88.7	72.4	80.0	81.1	79.0	–	–	–	–
Shrubs	51.3	4.4	51.6	4.5	22.0	17.7	24.8	16.8	–	–	–	–
Trees	81.9	84.8	81.6	81.9	68.4	83.6	74.5	86.3	75.8	68.2	–	–
Bare lands	77.3	59.4	76.2	60.2	76.7	42.6	66.4	42.2	–	–	–	–
Water bodies	90.0	91.7	90.1	91.4	89.4	79.7	83.1	90.2	–	–	92.7	82.5

predictions) compared to the S2-only feature set, respectively, but did not improve the UA. Adding the GEDI forest canopy height dataset onto the S1S2P feature set brought about 0.3% ~ 2.9% increase (based on the three classifiers) in the UA of trees, while resulted in a decrease of the accuracies of shrubs at the same time (S1S2PF in Fig. 5b). The other three ancillary datasets did not bring much improvement for trees.

For bare aquatic lands, the omission is more of an issue than commission. Bare lands were primarily misclassified as herbaceous cover and water bodies (Table 4). Besides the aforementioned Indo-Gangetic Plain of India, another hotspot area of their misclassifications was in Australia (Fig. 6d), where many sparsely vegetated dry lakes are seasonally flooded by water. The single-sensor S2 data obtained lower accuracies for bare aquatic lands (Fig. 5d) compared to its performance in characterizing water bodies, herbaceous cover, and trees. Integrating S1 and S2 data improved the PA by 0.5% and 8.3%, respectively, compared to the S2-only feature set in the RF and SVM prediction. Adding PALSAR, forest canopy height, and topographic features onto the S1S2 feature set did not result in much improvement in the accuracies. However, climate variables brought 3.2% and 6.5% increase in the UA based on MLR and SVM model, respectively. Soil features slightly improved the PA (0.4% ~ 1.0%) of bare aquatic lands.

Water bodies were the most correctly predicted class (Table 4). The main error in water classification was sourced from its confusion with herbaceous vegetation. Notably in African savannahs (10°N 0°E), water bodies had many misclassifications (Fig. 6e) from the seasonally flooded grasslands where herbaceous cover was omitted quite a lot (Fig. 6a). S2 features alone achieved high accuracies (Fig. 5e) in the water extraction. PALSAR-only data produced a lot of confusion among water, herbaceous cover, and bare aquatic lands (Table S8 in the supplementary material). The single-sensor S1 data outperformed PALSAR in characterizing water bodies globally (Fig. 5e), while it was less effective than PALSAR in identifying water bodies where vegetation was present, as shown in the

red box area of Fig. 7c. Integrating S2 data with the two sources of SAR or adding ancillary datasets did not bring much improvement compared with the S2-only feature set for water characterization.

3.2. Feature importance

Importance scores generated by the RF and MLR models for the top 20 features included in the feature set, which combined all data sources, are presented in Fig. 8. These scores represent the relative importance of features. The RF model and the MLR model generated different scores for specific features as they used different methods (Section 2.5) to evaluate the variable importance. Moreover, the way of dealing with correlated features by RF and MLR can explain the different importance they assign to different features. RF is good at dealing with correlated variables, while MLR tends to favor those uncorrelated because the correlated ones offer no information gain when building the model. Therefore, some features (e.g., the AWEI_p10 which was 72% correlated with mNDWI_p10) ranked higher by RF were not selected by MLR.

However, both models confirmed that optical features were driving the classification results (Fig. 8). Indices like the NDVI, mNDWI, SAVI, and spectral bands like B11 (i.e., SWIR1 band), B8A (i.e., narrow NIR band), B8 (broad NIR band), and B4 (red band) were important inputs for identifying aquatic land cover types. Concerning the temporal metrics, the 10th and 90th percentiles derived from the time-series data seemed to be more effective than other metrics (Fig. 8).

PALSAR features achieved higher importance scores than S1 features (e.g., HV and HH compared to VH and VV), indicating that PALSAR data had a stronger impact on GALC classification than S1 data when they were used synergistically. For both S1 and PALSAR data, the polarization-derived ratio and texture features had less influence compared to the backscattering data (e.g., HH, VH). The HV backscattering was of higher importance than HH, and the VH backscattering scored higher than VV (Fig. 8).

The GEDI forest canopy height scored the highest among features provided by ancillary datasets (Fig. 8). Topographic features were ranked lowest by both models. Climate variables, specifically the precipitation, were only favored by the MLR model.

3.3. Comparison with existing products

The class-specific accuracies of the CGLS-LC100, WorldCover 2020, Hansen GFC, and GSW datasets are shown in Table 5. For the GLC classification in aquatic areas, CGLS-LC100 and WorldCover 2020 obtained an OA of 74.9% and 78.3%, respectively (Table S18 in the supplementary material), which is lower than the OA of our best prediction (83.2%, Table 4). Except shrubs, all the class-specific accuracies of CGLS-LC100 were lower than those of our best prediction (Table 5). In addition, as can be seen from Fig. 7m, CGLS-LC100 presented less details of the aquatic land cover types than our maps. With similar input data, the OA of WorldCover 2020 was also lower than that of our S1S2-based prediction (82.7%, Table S13). WorldCover 2020 obtained a lower UA for shrubs and trees and a lower PA for herbaceous cover than our predictions (Table 5), because it had more confusion between herbaceous cover and trees or shrubs (Table S18), which is also visible from Fig. 7n (e.g., red box area). Accuracies of water bodies and bare aquatic lands were also lower in WorldCover 2020 (Table 5), caused by the confusion between the two classes and their misclassifications with herbaceous vegetation (Table S18). Despite that, CGLS-LC100 and WorldCover 2020 had a lower omission error for shrubs than that of our maps, possibly because they used more training samples of shrubs in the classification (Buchhorn et al., 2020; Zanaga et al., 2021). The Hansen GFC dataset obtained much lower accuracies for trees than our best prediction (Table 5). It could be observed that this dataset missed a lot of small patches of trees (e.g., red box area in Fig. 7o). The permanent water mapped by the GSW dataset omitted considerable water bodies compared with our best prediction, and this was especially apparent in small or narrow water areas (e.g., red box area in Fig. 7p).

4. Discussion

Previous GALC products are limited by mainly delineating open surface water, while few datasets provide information on the presence of water and vegetation types together. This study represents an important step forward in characterizing these key components of aquatic ecosystems collectively, which is helpful for users that require multiple aquatic types in their applications. Compared with generic GLC mapping, the interaction among water, vegetation, and wet soils makes it more complex to discriminate between different aquatic land cover types. For instance, herbaceous vegetation in river floodplains is prone to cause confusion among water, bare lands, and herbaceous cover. As seen in the results of previous GLC products (e.g., CGLS-LC100) and even in the most contemporary products (e.g., WorldCover 2020), aquatic areas suffer low accuracies. This highlights the need for specific attention in improving the mapping of aquatic land cover. In this study, we focused on aquatic areas only and explored ways to improve the GALC classification using multi-source EO data. Findings of our research may provide some useful information for future GALC mapping initiatives.

Result of this study showed that with Sentinel-2 data alone, comparably good results could be achieved as those combining multi-source data in the overall GALC classification. Optical features also obtained higher importance scores compared to most SAR features and ancillary datasets (Fig. 8). With a similar high spatial and temporal resolution, the single-sensor Sentinel-1 data have been reported to outperform Sentinel-2 data in characterizing detailed aquatic types in a local-scale study (Slagter et al., 2020). However, in our case, Sentinel-2 was better than Sentinel-1 in GALC characterization. This could partially be attributed to the better class separability based on optical features compared to SAR features derived from Sentinel-1 data. From Fig. S2 in the supplementary material, we found that different aquatic land cover types were more distinguishable by indices like NDVI and mNDWI than by the VV and VH backscattering.

However, the sole use of optical data does not provide sufficient information to accurately discriminate highly mixed and spectrally similar types, such as shrubs, trees, and herbaceous vegetation. Adding SAR features which are able to penetrate into the canopy and sense the vegetation structure could address the inefficiency of optical data to some extent. For example, the S2P, S1S2, and S1S2P (Table S12, S13, and S14 in the supplementary material) feature sets have helped to reduce the misclassifications among the three types. However, adding PALSAR or Sentinel-1 data did not reduce the omission error of shrubs and the commission error of trees, indicating that integrating Sentinel-2 data with the ALOS/PALSAR mosaic and Sentinel-1 data was still limited at addressing the most prominent issues in characterizing shrubs and trees in aquatic areas.

The single-sensor S1 data outperformed PALSAR data in identifying herbaceous cover (Fig. 5a), bare aquatic lands (Fig. 5d), and water bodies (Fig. 5e). However, Sentinel-1 data are sometimes ineffective to characterize water bodies with the existence of vegetation. This is because the backscatter signal will be increased by vegetation, which would lower the contrast between water and the surrounding non-water classes (Tsyganskaya et al., 2018). Compared with the longer-wavelength L-band radar, C-band data are more vulnerable to such conditions (Fig. S3 in supplementary material). Sentinel-1 data also have a lower capability in identifying trees (Fig. 5c) in aquatic areas, especially when dealing with dense tree canopies (Fig. S3 in supplementary material). Longer wavelengths can penetrate deeper into tree canopies, whereas the shorter-wavelength C-band radar will be reflected by leaves and branches, resulting in the decreased polarization signals for trees (Udali et al., 2021), which may further cause confusion among trees, shrubs, and herbaceous vegetation. When integrated with Sentinel-2 data, the S2P feature set achieved similar performances as with the

S1S2 feature set in identifying the five aquatic land cover types (Fig. 5). The PALSAR HV and HH backscattering obtained higher importance scores than the Sentinel-1 VH and VV backscattering (Fig. 8) when they were used synergistically. It should be noted that the PALSAR data used in this study had a yearly temporal resolution; it would hold more potential in characterizing complex aquatic ecosystems if higher temporal information is made available.

Adding ancillary datasets did not significantly improve the overall performance of the GALC classification (Fig. 4). Among all variables derived from ancillary datasets, the GEDI forest canopy height, which was of the highest importance, was expected to help separating shrubs from trees. Although adding this dataset onto the satellite-data-based S1S2P feature set reduced the commission error of trees by 0.3% ~ 2.9%, it resulted in an decrease in the accuracies of shrubs as well (i.e., S1S2PF in Fig. 5b). According to Potapov et al. (2021), the GEDI forest canopy height dataset is less accurate in estimating <3 m forest canopies, indicating the need of improving the height estimation for shrubs. Moreover, supplementing the height information for herbaceous vegetation or using longer-wavelength SAR data (e.g., P-band) that have better penetrating capability holds considerable potential in reducing the confusion between shrubs, herbaceous cover, and trees.

Previous studies have demonstrated the advantage of using topographic variables to detect the aquatic land existence (Hird et al., 2017), while our study shows that topographic features were less effective in the detailed-level GALC characterization. Among these five aquatic land cover types, topographic features were only relevant to reduce the commission error of shrubs (i.e., S1S2PFT in Fig. 5b). Climate information was reported beneficial for the classification of shrubs (Masiliūnas et al., 2021), while in our case it only contributed to the identification of bare aquatic lands. According to Fig. S4 in the supplementary material, different aquatic types are hardly separable by climate variables. One possible explanation could be that the coarse spatial resolution (i.e., 1 km) of the WorldClim dataset makes it difficult to capture the detailed information in complex aquatic environments. Furthermore, this dataset represents a multi-year mosaic for 1970–2000, which cannot provide the temporal variations in the phenology. Incorporating soil data only slightly reduced the commission error of shrubs and the omission error of bare aquatic lands. This might be attributed to the built-in uncertainties in the source dataset, i.e., soil data from the OpenLandMap, which was derived from model simulations (Hengl and MacMillan, 2019) rather than direct observations.

In this study, shrubs were predicted with the least accuracies and the largest error range in GALC classification, which is also a known issue in general GLC mapping (Herold et al., 2008; Tsendbazar et al., 2021a). Besides that shrubs are difficult to be separated from trees and herbaceous vegetation, the availability of reliable reference data is another crucial issue affecting the classification accuracy of shrubs because fewer training data, which further cause imbalances among classes, may lead to unstable performances of the classification model. Increasing the shrub training data in aquatic areas is therefore necessary to improve the shrub characterization.

Results of this study were based on machine learning algorithms and pixel-based classification. It is possible that the contribution of various data sources might be different if deep learning and/or other object-based methods are used. The classification based on RF, SVM, and MLR suggests that synergistically using optical, SAR (i.e., L-band and C-band), and ancillary datasets could produce a better overall performance for GALC characterization. However, considering that high-quality (e.g., high accuracy and spatial resolution) ancillary datasets might not always be available, and land cover mapping on a global-scale is resource-intensive and time-consuming because all features have to be loaded, preprocessed, and trained over the whole globe, the decision to use either multi-source or single-sensor data should be considered depending on the cover type being investigated. In water- or herbaceous-dominated aquatic areas, the single-sensor Sentinel-2 data could achieve good results. When mapping more complex aquatic ecosystems, SAR data should be considered to be integrated with optical data.

An accurate GALC characterization will be influenced by multiple factors. This study explored possible solutions from a data/sensor perspective, and unavoidably have some limitations. Firstly, to reduce data volume and computing intensity, the two-year time-series Sentinel-1 and -2 data were compressed to seven temporal composites (i.e., mean, median, percentiles, etc.), which may not be able to detect different phases of the growth cycle of various vegetation types. To make use of the vegetation phenology, future studies could consider harmonic analysis (e.g., Fourier methods) to derive phenological features (e.g., start of season, end of season) from the time-series data. Furthermore, data availability across different regions of the globe as well as in different seasons over a year will also affect the classification performance (LaRocque et al., 2020) and may cause spatial differences in the correctness of predictions. This was not thoroughly investigated in our study and could be evaluated for an improved GALC mapping in the future. Leveraging on the high spatial and temporal resolutions, Sentinel data are increasingly being used to classify more detailed land cover types even at species level (Sun et al., 2021), although such detailed reference data are not available on a global scale yet.

5. Conclusions

The under-presentation of vegetative information in existing GALC products and the poor performance of GLC products in delineating aquatic land cover require specific attention to improve GALC mapping. This study evaluated the potential of freely available multi-source EO data in improving the GALC characterization. Multiple classification scenarios were implemented based on different combinations of features derived from optical, SAR, and various ancillary datasets. The cross-validated results showed that accuracies obtained from the single-sensor Sentinel-2 data were comparable to results derived from combining multi-source data in the overall GALC classification. Integrating Sentinel-2 data with SAR features from Sentinel-1 data and the ALOS/PALSAR mosaic reduced misclassifications among shrubs, trees, and herbaceous vegetation but was still limited at addressing the most prominent issues in characterizing shrubs (i.e., omission) and trees (i.e., commission) in aquatic areas. Although with a lower spatial and temporal resolution, the ALOS/PALSAR mosaic was better than Sentinel-1 data at identifying trees in aquatic areas as well as water bodies with vegetation presence. Among ancillary datasets, topographic and soil features were relevant for the identification of shrubs, whereas climate variables contributed to the characterization of bare aquatic lands, but they did not bring significant improvement in overall accuracies. The GEDI forest canopy height dataset improved the characterization of trees, while it also decreased the accuracy of shrubs. To improve the GALC mapping, more efforts are needed to supplement a sufficient amount of reliable reference data for the less accurate classes such as shrubs.

CRediT authorship contribution statement

Panpan Xu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Nandin-Erdene Tsendbazar:** Conceptualization, Methodology, Investigation, Resources, Supervision, Writing – review & editing. **Martin Herold:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Jan G.P.W. Clevers:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Linlin Li:** 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.

Acknowledgements

This study was funded by the China Scholarship Council (grant no. 201804910841). The authors are thankful to Johannes Reiche, Dainius Masiliūnas, Adugna Mullissa, Bart Slagter, and Yaqing Gou for their kind suggestions for our data preprocessing and analysis of the results. The authors are grateful to the anonymous reviewers for their constructive comments on our work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2022.113103.

References

Adam, E., Mutanga, O., Rugege, D., 2010. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. Wetl. Ecol. Manag. 18, 281–296. https://doi.org/10.1007/s11273-009-9169-z.

Amler, E., Schmidt, M., Menz, G., 2015. Definitions and mapping of East African wetlands: a review. Remote Sens. https://doi.org/10.3390/rs70505256.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69. https://doi.org/10.1080/02626667909491834.

Braaten, J., 2021. Sentinel-2 Cloud Masking with s2cloudless. https://github.com/google/earthengine-community/blob/master/tutorials/sentinel-2-s2cloudless/index.ipynb (accessed 14 December 2020).

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.-E., Li, L., Tarko, A., 2020. Copernicus Global Land Service: Land Cover 100m: Version 3 Globe 2015-2019: Product User Manual. Zenodo, Geneve, Switzerland. https://doi.org/10.5281/zenodo.3938963.

Conners, R.W., Trivedi, M.M., Harlow, C.A., 1984. Segmentation of a high-resolution urban scene using texture operators. Comput. Vision. Graph. Image Process. 25, 273–310. https://doi.org/10.1016/0734-189X(84)90197-X.

Corcoran, J.M., Knight, J.F., Gallant, A.L., 2013. Influence of multi-source and multi-temporal remotely sensed and ancillary data on the accuracy of random forest classification of wetlands in northern Minnesota. Remote Sens. 5, 3212–3238. https://doi.org/10.3390/rs5073212.

DeLancey, E.R., Kariyeva, J., Bried, J.T., Hird, J.N., 2019. Large-scale probabilistic identification of boreal peatlands using Google earth engine, open-access satellite data, and machine learning. PLoS One 14. https://doi.org/10.1371/journal.pone.0218165.

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., Alsdorf, D., 2007. The shuttle radar topography mission. Rev. Geophys. 45 https://doi.org/10.1029/2005RG000183.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315.

Gallant, A.L., 2015. The challenges of remote monitoring of wetlands. Remote Sens. 7, 10938–10950. https://doi.org/10.3390/rs70810938.

GEE, 2021. ee.Image.focal_median. https://developers.google.com/earth-engine/apidocs/ee-image-focal_median (accessed 27 May 2021).

Gong, P., Liu, H., Zhang, M., Li, C., Wang, J., Huang, H., Clinton, N., Ji, L., Li, Wenyu, Bai, Y., Chen, B., Xu, B., Zhu, Z., Yuan, C., Ping Suen, H., Guo, J., Xu, N., Li, Weijia, Zhao, Y., Yang, J., Yu, C., Wang, X., Fu, H., Yu, L., Dronova, I., Hui, F., Cheng, X., Shi, X., Xiao, F., Liu, Q., Song, L., 2019. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. Sci. Bull. 64, 370–373. https://doi.org/10.1016/J.SCIB.2019.03.002.

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.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.

H2O.ai, 2021a. Distributed Random Forest (DRF). https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html (accessed 7 October 2021).

H2O.ai, 2021b. Variable Importance. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/variable-importance.html (accessed 7 Otober 2021).

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., 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.

Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. IEEE Trans. Syst. Man. Cybern. SMC-3, 610–621. https://doi.org/10.1109/TSMC.1973.4309314.

Hengl, T., MacMillan, R.A., 2019. Predictive Soil Mapping with R. OpenGeoHub Foundation, Wageningen, the Netherlands.

Herold, M., Mayaux, P., Woodcock, C.E., Baccini, A., Schmullius, C., 2008. Some challenges in global land cover mapping: an assessment of agreement and accuracy in existing 1 km datasets. Remote Sens. Environ. 112, 2538–2556. https://doi.org/10.1016/j.rse.2007.11.013.

Hijmans, R.J., Cameron, S., Parra, J., 2020. Historical Climate Data. https://worldclim.org/data/worldclim21.html (accessed 1 January 2020).

Hird, J.N., DeLancey, E.R., McDermid, G.J., Kariyeva, J., 2017. Google Earth Engine, open-access satellite data, and machine learning in support of large-area probabilistic wetland mapping. Remote Sens. 9 https://doi.org/10.3390/rs9121315.

GDAL/OGR contributors, 2021. GDAL/OGR Geospatial Data Abstraction Software Library. Open Source Geospatial Foundation. https://gdal.org.

Hu, S., Niu, Z., Chen, Y., Li, L., Zhang, H., 2017. Global wetlands: potential distribution, wetland loss, and status. Sci. Total Environ. 586, 319–327.

JAXA, 2016. Global 25m Resolution PALSAR-2/PALSAR Mosaic and Forest/Non-Forest Map (FNF) Dataset Description. Japan Aerospace Exploration Agency, Earth Observation Research Center, Tsukuba, Ibaraki, Japan.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J.C., Mathis, M., Brumby, S.P., 2021. Global land use / land cover with sentinel 2 and deep learning. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp. 4704–4707. https://doi.org/10.1109/IGARSS47720.2021.9553499.

Landry, M., 2016. Machine Learning with R and H2O. Mountain View, California, USA.

LaRocque, A., Phiri, C., Leblon, B., Pirotti, F., Connor, K., Hanson, A., 2020. Wetland mapping with Landsat 8 OLI, Sentinel-1, ALOS-1 PALSAR, and LiDAR data in southern New Brunswick, Canada. Remote Sens. 12, 2095.

Lehner, B., Döll, P., 2004. Development and validation of a global database of lakes, reservoirs and wetlands. J. Hydrol. 296, 1–22. https://doi.org/10.1016/j.jhydrol.2004.03.028.

Lehner, B., Verdin, K., Jarvis, A., 2008. New global hydrography derived from spaceborne elevation data. Eos (Washington. DC) 89, 93–94. https://doi.org/10.1029/2008EO100001.

Ludwig, C., Walli, A., Schleicher, C., Weichselbaum, J., Riffler, M., 2019. A highly automated algorithm for wetland detection using multi-temporal optical satellite data. Remote Sens. Environ. 224, 333–351. https://doi.org/10.1016/j.rse.2019.01.017.

Mahdavi, S., Salehi, B., Granger, J., Amani, M., Brisco, B., Huang, W., 2018. Remote sensing for wetland classification: a comprehensive review. GIScience Remote Sens. 55, 623–658. https://doi.org/10.1080/15481603.2017.1419602.

Mahdianpari, M., Salehi, B., Mohammadimanesh, F., Brisco, B., Homayouni, S., Gill, E., DeLancey, E.R., Bourgeau-Chavez, L., 2020. Big Data for a Big Country: the first generation of Canadian Wetland inventory map at a spatial resolution of 10-m using Sentinel-1 and Sentinel-2 data on the Google Earth Engine cloud computing platform. Can. J. Remote. Sens. 46, 15–33. https://doi.org/10.1080/07038992.2019.1711366.

Marthews, T.R., Dadson, S.J., Lehner, B., Abele, S., Gedney, N., 2015. High-resolution global topographic index values for use in large-scale hydrological modelling. Hydrol. Earth Syst. Sci. 19, 91–104. https://doi.org/10.5194/hess-19-91-2015.

Masiliūnas, 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 https://doi.org/10.1016/j.rse.2021.112409.

Matthews, E., Fung, I., 1987. Methane emission from natural wetlands: global distribution, area, and environmental characteristics of sources. Glob. Biogeochem. Cycles 1, 61–86. https://doi.org/10.1029/GB001i001p00061.

Mitsch, W., Gosselink, J., 2007. Wetlands, 4th ed. Wiley, New York, USA.

Mullissa, A., Vollrath, A., Odongo-Braun, C., Slagter, B., Balling, J., Gou, Y., Gorelick, N., Reiche, J., 2021. Sentinel-1 SAR backscatter analysis ready data preparation in Google earth engine. Remote Sens. https://doi.org/10.3390/rs13101954.

OpenGeoHub Foundation, 2018. About OpenLandMap. https://openlandmap.org (accessed 15 November 2018).

Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422. https://doi.org/10.1038/nature20584.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Blair, J.B., Hofton, M., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sens. Environ. 253 https://doi.org/10.1016/j.rse.2020.112165.

Prigent, C., Papa, F., Aires, F., Rossow, W.B., Matthews, E., 2007. Global inundation dynamics inferred from multiple satellite observations, 1993-2000. J. Geophys. Res. Atmos. 112 https://doi.org/10.1029/2006JD007847.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramsar Convention on Wetlands, 2018. Global Wetland Outlook: State of the world's Wetlands and their Services to People. Ramsar Convention Secretariat, Gland, Switzerland.

Rosenqvist, A., Shimada, M., Suzuki, S., Ohgushi, F., Tadono, T., Watanabe, M., Tsuzuku, K., Watanabe, T., Kamijo, S., Aoki, E., 2014. Operational performance of the ALOS global systematic acquisition strategy and observation plans for ALOS-2 PALSAR-2. Remote Sens. Environ. 155, 3–12.

Slagter, B., Tsendbazar, N.E., Vollrath, A., Reiche, J., 2020. Mapping wetland characteristics using temporally dense Sentinel-1 and Sentinel-2 data: a case study in the St. Lucia wetlands, South Africa. Int. J. Appl. EARTH Obs. Geoinf. 86 https://doi.org/10.1016/j.jag.2019.102009.

Steinwart, I., Thomann, P., 2017. liquidSVM: A fast and versatile SVM package arXiv1702.06899.

Stromann, O., Nascetti, A., Yousif, O., Ban, Y., 2020. Dimensionality reduction and feature selection for object-based land cover classification based on Sentinel-1 and Sentinel-2 time series using Google Earth Engine. Remote Sens. https://doi.org/10.3390/rs12010076.

Sun, C., Li, J., Liu, Yongxue, Liu, Yongchao, Liu, R., 2021. Plant species classification in salt marshes using phenological parameters derived from Sentinel-2 pixel-differential time-series. Remote Sens. Environ. 256, 112320 https://doi.org/10.1016/J.RSE.2021.112320.

Theil, H., 1969. A multinomial extension of the linear logit model. Int. Econ. Rev. 10, 251–259. https://doi.org/10.2307/2525642.

Tsendbazar, N.-E., Herold, M., de Bruin, S., Lesiv, M., Fritz, S., Van De Kerchove, R., Buchhorn, M., Duerauer, M., Szantoi, Z., Pekel, J.-F., 2018. Developing and applying a multi-purpose land cover validation dataset for Africa. Remote Sens. Environ. 219, 298–309. https://doi.org/10.1016/j.rse.2018.10.025.

Tsendbazar, N.-E., Herold, M., Li, L., Tarko, A., de Bruin, S., Masiliunas, D., Lesiv, M., Fritz, S., Buchhorn, M., Smets, B., Van De Kerchove, R., Duerauer, M., 2021a. Towards operational validation of annual global land cover maps. Remote Sens. Environ. 266, 112686 https://doi.org/10.1016/J.RSE.2021.112686.

Tsendbazar, N.-E., Li, L., Koopman, M., Carter, S., Herold, M., Georgieva, I., Lesiv, M., 2021b. WorldCover Product Validation Report V1.1. https://esa-worldcover.s3.amazonaws.com/v100/2020/docs/WorldCover_PVR_V1.1.pdf (accessed 22 October 2021).

Tsyganskaya, V., Martinis, S., Marzahn, P., Ludwig, R., 2018. Detection of temporary flooded vegetation using Sentinel-1 time series data. Remote Sens. 10 https://doi.org/10.3390/rs10081286.

Udali, A., Lingua, E., Persson, H., 2021. Assessing forest type and tree species classification using Sentinel-1 C-band SAR data in southern Sweden. Remote Sens. 13, 3237. https://doi.org/10.3390/rs13163237.

United Nations, 2015. Transforming our World: The 2030 Agenda for Sustainable Development. Department of Economic and Social Affairs, United Nations.

Vapnik, V., 1999. The Nature of Statistical Learning Theory. Springer Science & Business Media.

Verpoorter, C., Kutser, T., Seekell, D.A., Tranvik, L.J., 2014. A global inventory of lakes based on high-resolution satellite imagery. Geophys. Res. Lett. 41, 6396–6402. https://doi.org/10.1002/2014GL060641.

Welch, B.L., 1947. The generalization of 'STUDENT'S' problem when several different population varlances are involved. Biometrika 34, 28–35.

Wilson, J.P., Gallant, J.C., 2000. Secondary topographic attributes. Terrain Anal. Princ. Appl. 87–131.

Xu, P., Herold, M., Tsendbazar, N.-E., Clevers, J.G.P.W., 2020. Towards a comprehensive and consistent global aquatic land cover characterization framework addressing multiple user needs. Remote Sens. Environ. 250 https://doi.org/10.1016/j.rse.2020.112034.

Xu, P., Tsendbazar, N.-E., Herold, M., Clevers, J.G.P.W., 2021. Assessing a prototype database for comprehensive global aquatic land cover mapping. Remote Sens. 13 https://doi.org/10.3390/rs13194012.

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., Cartus, O., Santoro, M., Fritz, S., Georgieva, I., Lesiv, M., Carter, S., Herold, M., Li, L., Tsendbazar, N.-E., Ramoino, F., Arino, O., 2021. ESA WorldCover 10 m 2020 v100. Zenodo. https://doi.org/10.5281/ZENODO.5571936.