Journal of Agrometeorology

(A publication of Association of Agrometeorologists)

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 27 (4) : 454-463 (December - 2025)
https://doi.org/10.54386/jam.v27i4.3174
https://journal.agrimetassociation.org/index.php/jam

Research Paper

Clear-sky Land Surface Albedo (LSA) retrieval from OCM-3 data over diverse Indian landscapes using Machine learning approaches

ALIYA M. KURESHI[1], VISHAL N. PATHAK[2], DISHA B. KARDANI[1], JALPESH A. DAVE[3], DHIRAJ B. SHAH[1*], TEJAS P. TURAKHIA[4], ASHWIN GUJRATI[5], MEHUL R. PANDYA[5], and HIMANSHU J. TRIVEDI[3]

[1]*Sir P. T. Sarvajanik College of Science, Veer Narmad South Gujarat University, Surat 395007, Gujarat, India*
[2]*National Monsoon Mission, Indian Institute of Tropical Meteorology, Pune 411008, Maharashtra India*
[3]*N. V. Patel College of Pure and Applied Sciences, CVM University, Anand 388120, Gujarat, India*
[4]*Department of Instrumentation and Control Engineering, LDCE, GTU, Ahmedabad 380015, Gujarat, India*
[5]*Space Applications Centre, Indian Space Research Organisation, Ahmedabad 380015, Gujarat, India*
**Corresponding Author: dbs@ptscience.ac.in*

ABSTRACT

This study presents reliable methods for estimating clear-sky Land Surface Albedo (LSA) using Machine learning (ML) and satellite data, aiming to improve climate models and environmental monitoring. Top-of-atmosphere (TOA) radiance data from the Ocean Colour Monitor-3 (OCM-3) sensor aboard the Earth Observing Satellite (EOS-06) satellite containing 13 spectral bands were used, supported by 2.4 million synthetic simulations generated via the 6S (Second Simulation of a Satellite Signal in the Solar Spectrum) Radiative Transfer model (RTM). The simulations spanned diverse land covers, atmospheric states, sun and viewing geometries covering wavelengths from 0.4 to 2.5 µm. Three ML models Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multiple Linear Regression (MLR) were tested. Models were trained on 70% of the simulated data and tested on remaining 30% data. LSA retrieval was performed using OCM-3 TOA reflectance measurements in combination with MODIS-derived aerosol optical depth (AOD) and water-vapor inputs. This LSA derived from OCM-3 data then validated with MODIS LSA product (MCD43A3). Among the three models, RF achieved the best performance, with the lowest RMSE (0.00036) and strong agreement across various land types with MODIS data. The results confer the potential of ML models, especially RF, combined with radiative simulations, and can be used for operational estimation of LSA for OCM-3 data.

Keywords: Land surface albedo, Ocean colour monitor-3 (OCM-3), 6S (Second Simulation of a Satellite Signal in the Solar Spectrum), Machine learning, Random Forest.

Land Surface Albedo (LSA) is vital for Earth's energy balance, influencing climate and weather by regulating biophysical effects of natural and human activities (Lin *et al.,* 2022; Shuai *et al.,* 2014). It varies by surface type, with bright surfaces like snow reflecting more sunlight and darker ones like vegetation and water absorbing more (He *et al.,* 2018). Satellite-derived LSA products improve climate and weather models by aiding predictions of temperature, humidity, and ecosystem fluxes (Boussetta *et al.,* 2015; Kumar *et al.,* 2014). Various satellites provide LSA data at different resolutions, including Visible Infrared Imaging Radiometer Scale (VIIRS) (500m & 1km), Multi-angle Imaging Spectro Radiometer (MISR) (1 km), MODIS Albedo (500 m) and Sentinel-2 (10 m) (Wang *et al.,* 2013; Martonchik *et al.,* 1998; Chen *et al.,* 2023). This has increased interest in high-resolution, spatially accurate surface LSA mapping for climate studies.

LSA estimation often follows a traditional three-step algorithm. First, atmospheric correction preprocesses satellite radiance or reflectance data to remove atmospheric effects using Radiative Transfer Models (RTM) like MODTRAN (Dave *et al.,* 2023) and 6S (Schaaf *et al.,* 2002). Second, broadband albedo is derived by converting narrowband reflectance using sensor-specific empirical coefficients (Liang, 2001) for MODIS and Landsat. This translation ensures accurate LSA estimation despite varying sensor spectral responses. Finally, the anisotropic nature of surface reflectance, its dependence on illumination and viewing angles, is

Article info - DOI: *https://doi.org/10.54386/jam.v27i4.3174*
Received:31 August 2025; Accepted: 28 September 2025; Published online : 1 December 2025

corrected using Bidirectional Reflectance Distribution Function (BRDF) modelling, producing standardized LSA under black-sky (direct sunlight) and white-sky (diffuse light) conditions.

Traditional BRDF-based methods struggle with bright surfaces and assume constant atmospheric correction, causing retrieval errors. To overcome this, direct estimation methods were developed to link satellite observations to LSA using empirical or data-driven models, eliminating the need for detailed BRDF modelling or narrowband-to-broadband conversion (Liang, 2003; Qu *et al.,* 2014). Physical models generate training samples to improve accuracy (Wang *et al.,* 2013). This approach has been tested on various sensors, from low-resolution broadband to multispectral ones like MODIS and MISR, using inputs such as TOA reflectance, viewing geometry, and atmospheric parameters (AOD, water vapor, ozone). However, this method can produce noisy residuals, leading to occasional spurious LSA estimation.

Recently, ML methods have become powerful tools for handling nonlinear regression in remote sensing parameter retrieval (Belgiu and Drăguț, 2016). Models like Random Forest (RF), Artificial Neural Networks (ANN) (Extreme Gradient Boosting (XGBoost) (Bijlwan *et al.,* 2024) and K-Nearest Neighbors (KNN) effectively manage complex, nonlinear relationships with flexibility. Chen *et al.,* (2023) used synthetic datasets from MODIS and RTM to retrieve high-resolution Sentinel-2 LSA, validating performance with in-situ measurements across varied terrains. Similar work (Lin *et al.,* 2022) was carried out to developed a direct estimation framework for fine-scale LSA mapping using Sentinel-2, combining high-quality BRDF training data, ESA land cover data, and MODIS BRDF/LSA products under diverse viewing angles. Despite progress, challenges remain in extending LSA retrieval to diverse geographic regions with dynamic land cover changes.

To address existing gaps, a large-scale simulated dataset was created using the 6S RTM to train and test three machine learning models: RF, XGBoost, and MLR. These models, developed from theoretical data, were then applied to actual OCM-3 satellite reflectance. Their performance was evaluated by comparing estimated LSA with MODIS LSA products, demonstrating the models' ability to deliver reliable, high-resolution LSA across various landforms.

MATERIALS AND METHODS

Study area

India has a wide range of iconic landscapes dividing the nation into six physiographic regions mainly - Northern Mountains consisting of Himalayan ranges, Peninsular Plateaus, Indo-Gangetic Plains, Thar Desert, and Coastal plains, creating a unique and diversified background for remote sensing applications (NRSC, 2021). Therefore, to capture this diversity, the present study focuses on four prominent states of India namely Gujarat, Rajasthan, Haryana, Madhya Pradesh each with a unique combination of topography, climate and atmospheric influence that together represents the major landscapes of the country.

Gujarat with semi-arid zones along with the unique saline desert areas of Rann of Kutch, offering a highly reflective and dynamic surface environment (Mehta *et al.,* 2012). Major cities like Ahmedabad, Vadodara and Surat have led to rapid urbanization, further transforming the land covers over Gujarat state. Moreover, agriculture is a key sector in Gujarat, with crops like cotton, groundnut, maize, pearl millet and sorghum being prominently cultivated in its fertile plains (ICAR, 2025) which provides seasonal variation in LSA. Rajasthan can be divided into four major regions, the western desert along with barren hills, level rocky or sandy plains, the Aravalli hills and south-eastern plateau with varied climate from semi-arid to arid and categorized by dry, barren terrain with minimal vegetation and rainfall makes it suitable for evaluating model robustness over bright, high LSA surfaces (Sharma *et al.,* 2015). Haryana, is a landlock arid to semi-arid state in the North Western region of India with highly cultivable regions, it is a subsection of the Indo-Gangetic Plain lying with intensively managed croplands and seasonal vegetation. (Singh *et al.,* 2023). Madhya Pradesh covers a diverse terrain including dense forests, open shrubland, farmlands, and plateau regions a mix of natural and anthropogenic surface types, best suited for examining the model performance across diverse and heterogeneous landscapes (Lohare *et al.,* 2023).

Together, these regions capture the wide environmental and geographical variability of India which allows to check the robustness of the developed method and supports the creation of more reliable and high-resolution LSA products that can be applied across varying landscapes.

OCM TOA reflectance

This study utilizes the level-1C Top of Atmosphere (TOA) radiance (Level-1_RadianceGeoreferenced) ($mW \cdot cm^{-2} \cdot sr^{-1} \cdot \mu m^{-1}$) datasets of the Ocean Colour Monitor (OCM-3) sensor at a spatial resolution of 360m (Local Area Cover – LAC) with less than 10% cloud coverage, acquired from Bhoonidhi Portal (https://bhoonidhi.nrsc.gov.in/). The obtained TOA radiance was pre-processed and converted to TOA-reflectance equation (1)

$$R_{TOA}(\lambda) = \frac{\pi L_{TOA}(\lambda)}{E_0(\lambda) \cos\theta} \quad (1)$$

where, R_{TOA} - Calculated TOA-Reflectance, L_{TOA} - Level-1C TOA radiance of OCM-3 sensor, E_0 - Solar spectral irradiance, θ- Solar zenith angle (SZA), λ – Wavelength, d - Earth Sun distance.

The data comprises of 13 spectral bands of OCM-3 sensor as summarized in (Table 1) along with the associated band-wise solar spectral irradiance (E_0) (Pandya and Pathak, 2021) which served as fundamental inputs for converting TOA-radiance to TOA-reflectance.

MODIS land surface albedo (MCD43A3)

The Moderate Resolution Imaging Spectroradiometer (MODIS) product (MCD43A3, V061) provided by NASA land processes distributed active archive center (NASA, 2015) and accessed via Google Earth Engine (GEE) platform (https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MCD43A3) was used in this study as reference

Table 1: Spectral characteristics of OCM-3 bands and corresponding solar irradiance (E_0).

Band No	OCM3 band range (μm) Bandwidth(nm)	Solar spectral irradiance (E0) ($mW \cdot cm^{-2} \cdot \mu m^{-1}$)
1	0.402 - 0.422 (20)	171.60
2	0.438 - 0.448 (10)	188.05
3	0.485 - 0.495 (10)	192.87
4	0.505 - 0.515 (10)	191.20
5	0.550 - 0.560 (10)	186.17
6	0.561 - 0.571 (10)	183.95
7	0.615 - 0.625 (10)	165.02
8	0.665 - 0.675 (10)	150.93
9	0.677 - 0.685 (08)	146.91
10	0.705 - 0.715 (10)	138.76
11	0.775 - 0.785 (10)	115.51
12	0.860 - 0.880 (20)	97.19
13	0.990 - 1.030 (20)	74.93

dataset for validating the retrieved OCM-3 LSA using the proposed ML models.

MODIS aerosol optical depth & water vapor (MCD19A2.061)

The MCD19A2 V6.1 data product acquired from GEE platform (https://developers.google.com/earthengine/datasets/catalog/MODIS_061_MCD19A2_GRANULES) is a combination of observations from MODIS terra and aqua using the Multi-angle Implementation of Atmospheric Correction (MAIAC) algorithm It provides daily level-2 gridded observations of AOD 0.55 μm (Green band) and column-WV over land at 1 km resolution, retrieved from near-IR bands at 0.94 μm. Both of these are essential for understanding its influence on aerosol behaviour, atmospheric moisture content and air quality in real environmental conditions (Lyapustin *et al.,* 2022). Both these datasets of AOD and WV were re-gridded to match the spatial resolution of the OCM-3 and then were subsequently utilized to generate LSA observations.

Eco stress spectra library

The eco-stress spectral library version 1.0 (Meerdink, *et al.,* 2019) is a widespread collection of high-resolution spectral reflectance data for a diverse range of materials such as natural, manmade and vegetation (including non-photosynthetic vegetation). This study incorporates such diverse surface reflectance's for simulations across varied landforms to match the real-world scenarios and to achieve higher precision for LSA retrieval. The methodology is divided into three steps (Fig. 1) for retrieving LSA from OCM-3 data, addressing atmospheric, geometric, and surface variability to optimize accuracy.

Methodology

This study utilizes satellite observations of Oceansat-3 Top-of-atmosphere radiance together with MODIS-derived aerosol optical depth (AOD) and water vapor (WV) products over varied spatial and temporal coverages. Gujarat and Rajasthan were observed on 11th February 2024 and 01th November 2024, while Haryana was covered on 11th February 2024 and 01 November 2023. For Madhya Pradesh, observations were obtained for 20th February 2024 and 06th November 2024.

Initially, the 6S RTM was utilized to simulate at-sensor radiance for OCM-3's 13 spectral bands and broadband range (0.4 - 2.5 μm) across diverse atmospheric, solar, surface, and viewing. Broadband surface reflectance obtained from these simulations served as the target variable whereas key geometric parameters - Solar Zenith Angle (SZA), View Zenith Angle (VZA), Solar Azimuth Angle (SAZ), and View Azimuth Angle (VAZ), Atmospheric variables - Aerosol Optical Depth (AOD), Water Vapor (WV) and TOA-Reflectance were considered as input features (Equation-2) while training all four models. From a total of 1,76,400 simulations per band (i.e. in total 22,93,200 simulations), a merged dataset was created splitting 70% of samples (i.e. 1,23,480) as training dataset and remaining 30% (52,920 samples) as testing dataset to enhance model robustness and adaptability.

$$LSA = f\,(Refl_{TOA}, WV, Aero, SZA, SAZ, VZA. VAZ) \quad (2)$$

Second, the three machine learning models (RF, XGBoost, MLR) developed using these training datasets, were then applied to observed OCM-3 TOA-reflectance and MODIS products (AOD, WV) re-gridded at the same resolution offers a precise alternative to generate high resolution retrieved LSA estimates over four states of India. Lastly, these retrieved LSA were then validated against MODIS (MCD43A3) products to confirm the accuracy of the estimates and hence the model performance.

6S radiative transfer simulations

The 6S (Second Simulation of a Satellite Signal in the Solar Spectrum) Radiative Transfer Model (RTM) (Vermote *et al.,*1997) was employed to simulate diverse atmospheric and surface conditions for the OCM-3 spectral bands. Simulations were carried out under varying geometric and atmospheric conditions, including different aerosol types, surface conditions, water vapor levels, solar and viewing geometries. Further, to enhance the surface recognition beyond the default three classes in 6S mainly water, sand and vegetation, 100 different LULC spectral classes from the ECOSTRESS Spectral Library were used as input surface reflectance. This allowed the generation of a comprehensive lookup table (Table 2), which played a critical role in improving the accuracy of the LSA retrieval by efficiently integrating these training datasets to develop the machine learning models.

Implementation of machine learning (ML) models

Retrieving LSA from satellite data involves complex, non-linear interactions of surface reflectance, atmosphere, and land cover properties. ML models like RF, Decision Trees (DT), and Gradient Boosting handle large datasets more effectively than traditional methods, yielding higher-resolution albedo predictions while reducing retrieval uncertainties (Chen *et al.,* 2023). Data-driven approaches including ML, improve land surface classification, thus enhancing LSA estimation across diverse regions. In this study, regression models (RF, XGBoost, MLR) estimate broadband albedo using theoretical 6S-RTM simulations and OCM-3 satellite observations. Modelling and validation were implemented in

Table 2: Variations in input parameters for 6S Radiative Transfer Model (RTM) simulations

Parameters	Variations	Total permutations
Solar zenith angle (SZA°)	0, 10, 20, 30, 40, 50, 60	7
View zenith angle (VZA°)	0, 10, 20, 30, 40, 50	6
Aerosol optical depth (AOT)	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7	7
Water vapor (WV - g/cm²)	0.25, 1, 2, 3, 4, 4.5	6
Land cover types (ECOSTRESS spectral library)	vegetation, sand, water, man-made, minerals, soils, mixed classes	100

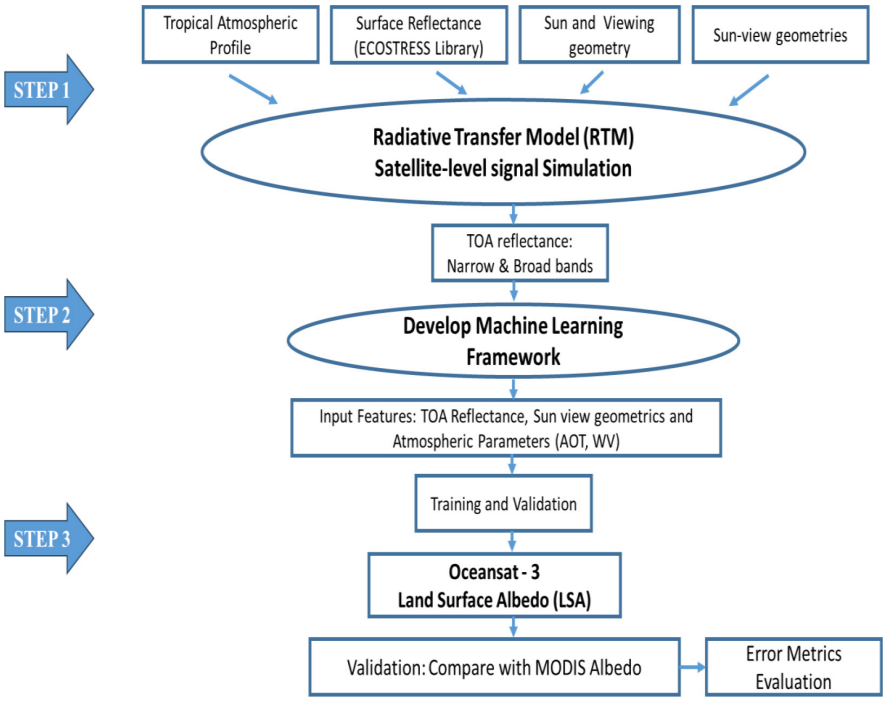

Fig. 1: Flowchart illustrating the methodology for LSA estimation using machine learning.

Python, detailing the principles and parameter settings for each algorithm.

Multiple linear regression (MLR)

MLR models the linear relationship between a dependent variable and multiple independent variables, estimating outcomes with a linear combination of predictors. Coefficients indicate how changes in each variable affect the outcome, and the model uses Ordinary Least Squares (OLS) to minimize prediction errors. The mathematical formulation for MLR (Montgomery *et al.,* 2012) is commonly expressed as:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n+\epsilon \quad (3)$$

Where: Y- is the dependent variable, $X_1,X_2,\ldots, X_n$ - are the independent variables, β_0 - is the intercept term, β_1, β_2,. βn - are the coefficients corresponding to each independent variable, and ϵ - is the error term, representing unexplained variance.

MLR is widely used for estimation and understanding factor influences, but assumptions like linearity and homoscedasticity must be checked for reliability.

Random forest (RF)

RF, as defined by Breiman (2001), is an ensemble learning model that combines multiple decision trees to produce a single result. It handles complex datasets, continuous variables, and reduces overfitting, making it suitable for classification and regression tasks. In this study, the RF regression model was optimized via extensive hyperparameter tuning using a randomized search over 50+ combinations with 5-fold cross-validation. Parameters like number of estimators, tree depth, minimum samples for splits and leaves, feature selection, and bootstrap sampling were tuned to minimize RMSE on test data. The final optimized configuration is detailed in Table 3.

Extreme gradient boosting (XGBoost)

XGBoost is an optimized boosting algorithm (Chen and Guestrin, 2016) that builds trees sequentially to correct previous errors, improving efficiency and predictive performance over Random Forest, which averages independent trees. It incorporates regularization to prevent overfitting. The XGBoost regression model was extensively tuned using Random Search CV with 100 iterations and 5-fold cross-validation, optimizing parameters like the number of trees, learning rate, tree depth, regularization terms, and pruning

Table 3: Random forest (RF) model specifications

Hyperparameter	Search Space	Best Value Found
n_estimators	[100, 200, 300, 400, 500]	400
max_depth	[5, 10, 20, None]	None
min_samples_split	[2, 5, 10]	3
min_samples_leaf	[1, 2, 4]	1
max_features	['auto', 'sqrt', 'log2']	'sqrt'
bootstrap	[True, False]	False

Table 4: XGBoost configuration and tuning Parameters

Hyperparameter	Search Space Explored	Best Value Found
n_estimators	[100, 200, 300, 500, 700, 1000]	300
learning_rate	[0.001, 0.01, 0.05, 0.1, 0.2, 0.3]	0.05
max_depth	[3, 4, 5, 6, 8, 10]	10
min_child_weight	[1, 3, 5, 7]	1
subsample	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	1
colsample_bytree	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	0.5
gamma	[0, 0.01, 0.1, 0.3, 0.5]	0
reg_alpha	[0, 0.001, 0.01, 0.1, 1, 10]	0
reg_lambda	[0.01, 0.1, 0.5, 1, 1.5, 2, 10]	2

(Table 4), with early stopping to enhance generalization.

Models were trained on 6S-simulated TOA reflectance data including solar geometry, atmospheric variability, and spectral signatures to capture complex surface-atmosphere interactions. This comprehensive training enabled the models to learn non-linear relationships between TOA reflectance and LSA across diverse conditions. The trained models were validated with real OCM-3 satellite data to assess their applicability, and the retrieved LSA values were compared with MODIS products to benchmark accuracy and reliability.

Error analysis

To assess the performance of each of these three ML models, two commonly used metrics such as - Root Mean Square Error (RMSE) and bias were calculated to get insights into the model's accuracy, consistency, and retrieval capability.

The assessment was two stepped:

Test RMSE: Computed using 30% of the theoretical dataset i.e. Test dataset to assess how well the models generalize to unseen simulated data.

***Validation RMSE and Bias*:** Calculated by applying the trained models to real OCM-3 and MODIS (AOD, WV) observations and comparing the estimated LSA values with MODIS LSA products, thereby assessing real-world performance. The formulas used are:

$$\text{Bias}=\text{LSA}_e-\text{LSA}_r \quad (4)$$

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}((\text{Albedo}_e(i)-\text{Albedo}_r(i))^2}{n}} \quad (5)$$

The above equations involve the following terms: LSA_e gives the Estimated values by the models, LSA_r refers to the Reference LSA measurements (whether simulated test samples or MODIS measurements), n is the number of observations.

RESULTS AND DISCUSSION

LSA estimates over varied geographical and environmental conditions

A two-part analysis was conducted to evaluate the accuracy of three ML models (Fig. 2) for estimating LSA. The first step involved developing the three models such as RF, XGBoost and MLR using 70% samples for training, built on inputs such as TOA reflectance, AOD, WV, with solar and viewing geometry. In

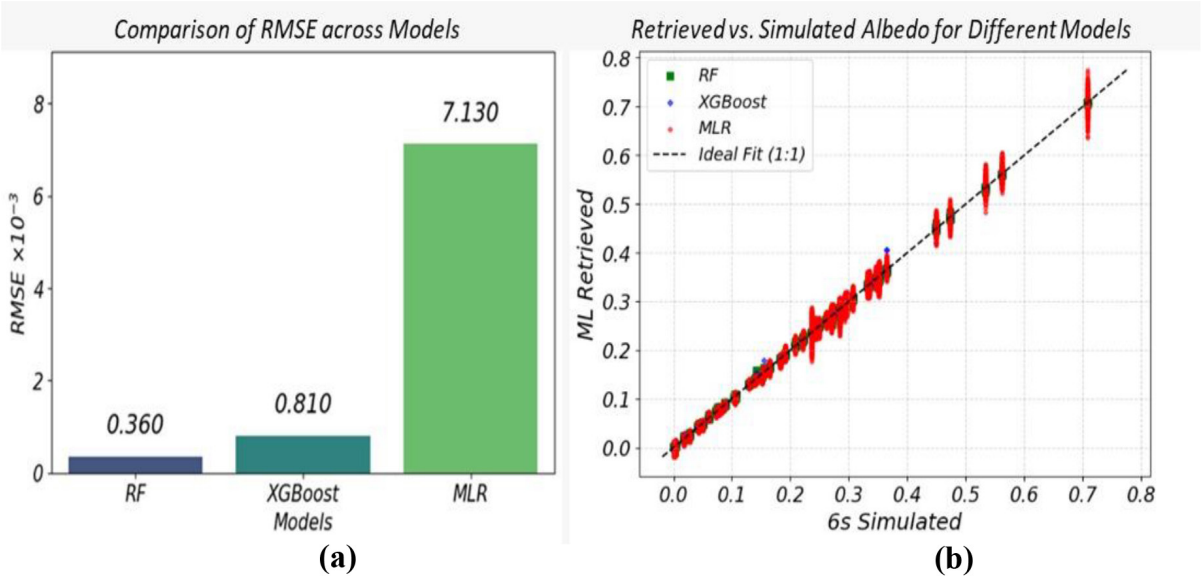

Fig. 2: (a) RMSE ($\times 10^{-3}$) values for three ML models in a bar chart (b) Scatter plot of retrieved versus 6S simulated albedo for three ML models

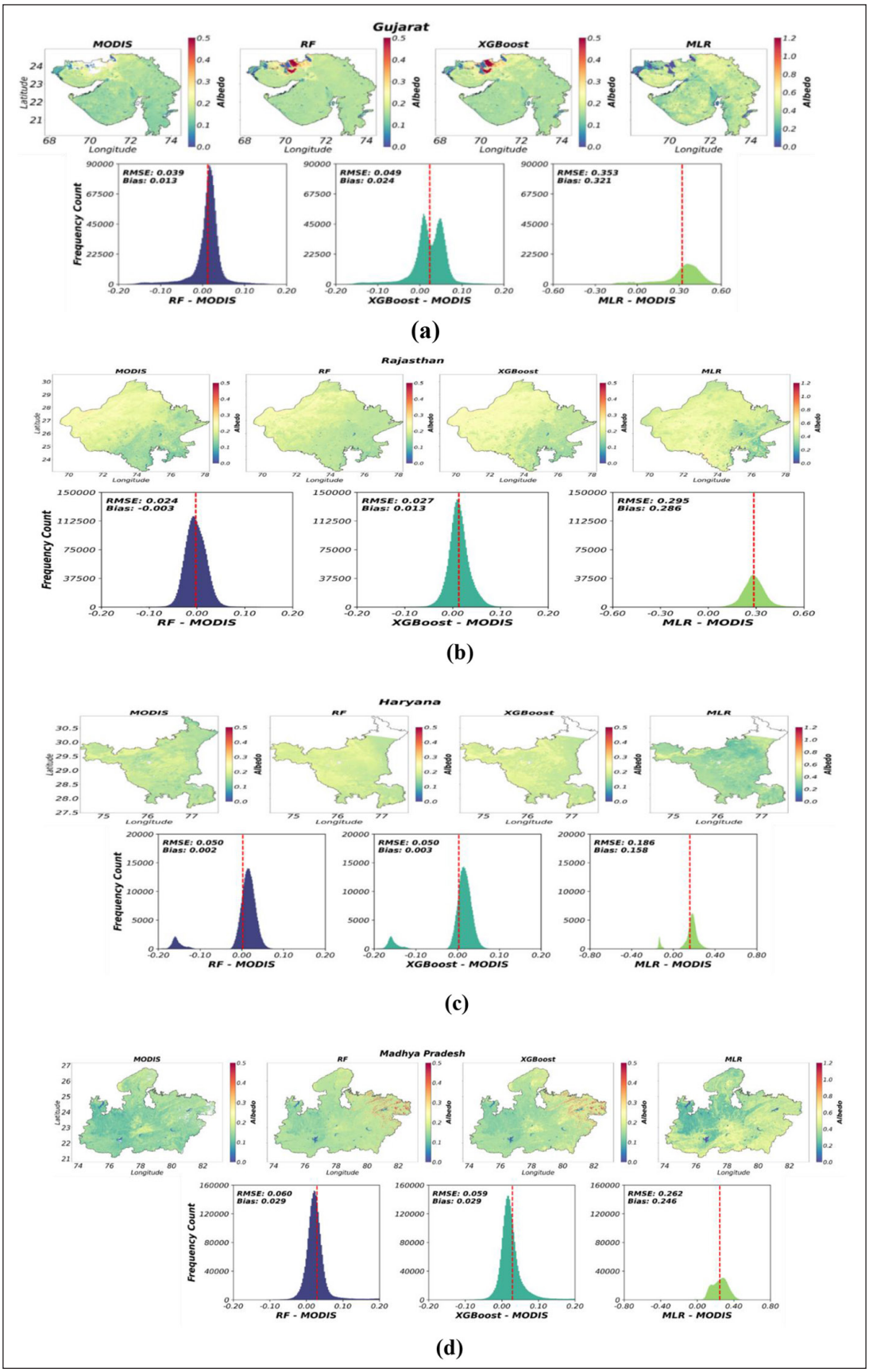

Fig. 3: MODIS and ML model-estimated LSA over (a) Gujarat, (b) Rajasthan, (c) Haryana, and (d) Madhya Pradesh for November, along with corresponding ML model and MODIS albedo difference histogram

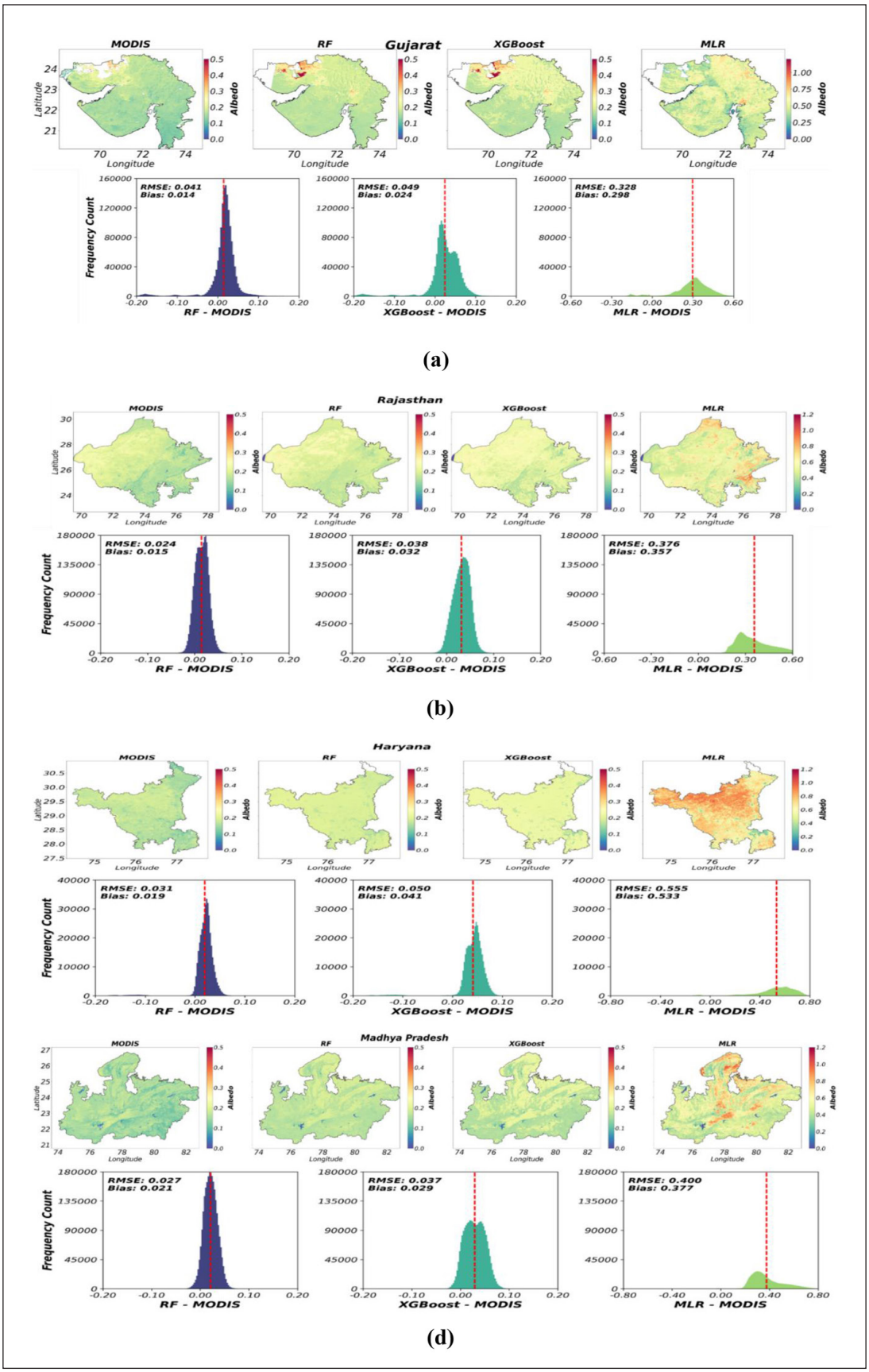

Fig. 4: MODIS and ML model-estimated LSA over (a) Gujarat, (b) Rajasthan, (c) Haryana, and (d) Madhya Pradesh for February, along with corresponding ML model and MODIS albedo difference histogram

Table 5: RMSE and bias values for RF, XGBoost, and MLR across different dates over diverse locations.

Location	Date	RF		XGBoost		MLR	
		RMSE	Bias	RMSE	Bias	RMSE	Bias
Gujarat	11th Feb, 2024	0.041	0.014	0.049	0.024	0.328	0.299
	01st Nov, 2024	0.039	0.013	0.049	0.024	0.353	0.321
Haryana	11th Feb, 2024	0.031	0.019	0.050	0.041	0.555	0.533
	01st Nov, 2024	0.050	0.004	0.050	0.005	0.186	0.163
Rajasthan	11th Feb, 2024	0.023	0.015	0.038	0.032	0.376	0.357
	01st Nov, 2024	0.024	-0.003	0.027	0.013	0.295	0.286
Madhya Pradesh	20th Feb, 2024	0.027	0.021	0.037	0.029	0.400	0.377
	06th Nov, 2024	0.060	0.029	0.059	0.029	0.262	0.246

the second step, the accuracy of each of these models was assessed using the remaining 30% test samples of the theoretical dataset. As shown in Fig. 2a, the RF model achieved lowest RMSE (0.360 x 10-3), depicting higher precision in learning the nonlinear relations between the parameters of the theoretical dataset. XGBoost, another tree-based model was followed with a slightly higher RMSE (0.810 x 10-3) while the basic MLR model performed very poorly with an RMSE of 7.130 x 10-3, highlighting its limitations due to the linearly assumption. Similarly, the scatter plot (Fig. 2b) also shows that the estimation made by the tree-based models (RF & XGBoost) closely aligns with the 1:1 reference line, whereas the MLR models displays noticeable deviations. These findings derived from the theoretical dataset confirms that the ensemble-based models like RF and XGBoost are more suitable for capturing the complex dynamics simulated by the 6S-RTM, laying a reliable foundation for further testing the model's performance for actual observations collected from OCM-3 sensor. Therefore, we now proceed to evaluate the performance of these three models using the data collected over the diverse landforms of India, while also considering temporal variability in the measurements.

Spatial and temporal comparisons (Fig. 3 & Fig. 4) were carried out using the MODIS-derived LSA across multiple regions such as Gujarat (11th February, 01st November), Haryana (01st November, 11th February), Rajasthan (11th February, 01st November), Madhya Pradesh (20th February, 06st November). In Gujarat, the RF model consistently outperformed the remaining two models, achieving RMSE values ranging from 0.021 to 0.041 and Bias observed to be very close to zero. XGBoost followed closely, particularly for December with an RMSE of 0.021 signifying that the tree-based models were well-and more effective as compared to the simple MLR model which gave higher errors for these conditions.

For Madhya Pradesh, as for 20th February the RF models had attained lower RSME values 0.020 with almost negligible bias, thereby highlighting the model's strong performance for regions with mixed terrain and land use conditions. Whereas on 6th November, both RF and XGBoost models maintained more accuracy (RMSE: 0.06) indicating that for varying seasonal and surface characteristics, the tree-based ensemble models can deliver more efficient LSA estimates.

In Rajasthan, the variances between the models were minimal. For instance, for 1st November the two models, RF (0.024), and XGBoost (0.027) generated similar results depicting that given sufficient feature information, both models can give better accuracy for LSA estimation. Similar, the RF model performed better in Haryana on 11th February with RMSE of 0.031, although XGBoost and MLR models demonstrated comparatively lower performances.

The regional analysis was conducted to evaluate the accuracy of different ML models in estimating LSA over diverse landforms, with an aim of assessing the robustness and generalizability of these models based on the performance indicators such as RMSE to quantify average prediction errors and Bias values to reflect the tendency to systematically overestimate or underestimate LSA (Table 5).

However, the MLR model consistently showed the lowest performance across all regions and time period, with RMSE values exceeding 0.3 and reaching as high as 0.555 in Haryana and notably higher bias valued as well. These results highlight the limitations of linear models in capturing the complex and nonlinear patterns present in the observations especially when applied to diverse land surfaces.

Overall, the analyses suggest that while RF model consistently demonstrate higher accuracy and was found to be well-suited for capturing complex spatiotemporal patterns, other tree-based ensemble model like XGBoost also offers reliable estimates. Their ability to deliver comparable results, coupled with lower computational requirements with the ease of implementation, makes them suitable for applications where computational efficiency and rapid deployment are essential.

CONCLUSION

This study was undertaken to systematically evaluate the ability of multiple machine learning (ML) models to reliably estimate clear sky land surface albedo (LSA) across diverse landforms over India. Three widely used ML models Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multiple Linear Regression (MLR) were trained using synthetic data generated by the 6S Radiative Transfer Model (RTM) and validated both on simulated test samples and actual satellite observations, including MODIS LSA products. The findings indicate that certain ML models, particularly RF, are more effective at capturing complex patterns in the data, resulting in higher accuracy and lower prediction errors. RF not only exhibited the best overall performance across varied terrains but also closely matched the MODIS LSA product,

thereby demonstrating its strong generalization capabilities. Its consistent accuracy when trained on physically based simulations underlines its adaptability and strength for geophysical parameter retrieval. Although XGBoost gave comparable performance, but it consistently trailed Random Forest in estimative accuracy. The MLR model, constrained by its assumption of linearity, showed comparatively higher errors and limited capability in modelling the nonlinear nature of LSA variations across India. In conclusion, this study demonstrates the effectiveness of integrating ML algorithms with radiative transfer simulations for satellite based LSA estimation. The proposed approach offers a promising framework for large scale remote sensing applications, providing more accurate and consistent LSA retrieval using data from EO sensors like OCM-3.

ACKNOWLEDGEMENT

This research was conducted as part of the project titled as "Retrieval of Land Surface Albedo (LSA) using Oceansat-3 OCM data", funded by the Space Applications Centre (SAC), Indian Space Research Organisation (ISRO). The authors are thankful to the Director of SAC, the Principals of PTSCS and NVPAS for their support, NRSC (ISRO) for access to OCM-3 data via Bhoonidhi, and NASA for MODIS products. Gratitude is also extended to all who contributed directly or indirectly to this work.

Conflict of interest Statement: The authors declare that there is no conflict of interest related to this article.

Sources of Funding: This research is funded by SAC-ISRO titled "Retrieval of Land Surface Albedo (LSA) using Oceansat-3 OCM data".

Data availability Statement: Data will be made available on request.

Authors Contribution: **A. M. Kureshi:** Writing original draft, At sensor radiance simulations, Data curation, Algorithm Development, Analysis and Visualization; **V. N. Pathak:** Important suggestions, Conceptualization, Software, formal analysis; **D. B. Kardani:** Writing: original draft, Simulations, Analysis; **J. A. Dave:** Review, editing, Important suggestions; **D. B. Shah:** Review & editing, Conceptualization, Research Supervision, Methodology; **T. P. Turakhia:** Writing and editing, Research Supervision, Algorithm development, Data Curation and Analysis; **A. Gujrati:** Validation, Conceptualization, Resources, Scientific inputs; **M. R. Pandya:** Validation, Conceptualization, Resources, Scientific inputs; **H. J. Trivedi:** Validation, Supervision, Review.

Disclaimer: The contents, opinions, and views expressed in the research article published in the Journal of Agrometeorology are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

Publisher's Note: The periodical remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

Belgiu, M., and Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photog. Remote Sens.,* 114: 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bijlwan, A., Pokhriyal, S., Ranjan, R., Singh, R. K., and Jha, A. (2024). Machine learning methods for estimating reference evapotranspiration. *J. Agrometeorol.,* 26(1): 63-68. https://doi.org/10.54386/jam.v26i1.2462

Boussetta, S., Balsamo, G., Dutra, E., Beljaars, A., and Albergel, C. (2015). Assimilation of surface albedo and vegetation states from satellite observations and their impact on numerical weather prediction. *Remote Sens. Environ.,* 163: 111-126. https://doi.org/10.1016/j.rse.2015.03.009

Breiman, L. (2001). Random forests. *Mach. Learn.,* 45 (1): 5-32. https://doi.org/10.1023/A:1010933404324

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785

Chen, H., Lin, X., Sun, Y., Wen, J., Wu, X., You, D., Cheng, J., Zhang, Z., Zhang, Z., Wu, C., Zhang, F., Yin, K., Jian, H., and Guan, X. (2023). Performance assessment of four data-driven machine learning models: A case to generate Sentinel-2 albedo at 10 meters. *Remote Sens.,* 15 (10): 2684. https://doi.org/10.3390/rs15102684

Dave, J. A., Pandya, M. R., Shah, D. B., Varchand, H. K., Parmar, P. N., Trivedi, H. J., Pathak, V. N., Singh, M., and Kardani, D. B. (2023). Comparative analysis of two parameter-dependent split window algorithms for the land surface temperature retrieval using MODIS TIR observations. *J. Agrometeorol.,* 25(4): 510-516. https://doi.org/10.54386/jam.v25i4.2286

He, T., Wang, D., and Qu, Y. (2018). Land surface albedo. In S. Liang (Ed.), Comprehensive remote sensing (Vol. 5, Earth's energy budget, pp. 140-162). Oxford, UK: Elsevier. https://doi.org/10.1016/B978-0-12-409548-9.10370-7

ICAR. (2025, August). Indian Council of Agricultural Research, Gujarat. https://icar.org.in/en/node/17264

Kumar, P., Bhattacharya, B. K., Nigam, R., Kishtawal, C. M., and Pal, P. K. (2014). Impact of Kalpana-1 derived land surface albedo on short-range weather forecasting over the Indian subcontinent. *Remote Sens. Environ.,* 152: 227-239. https://ui.adsabs.harvard.edu/link_gateway/2014JGRD..119.2764K/doi:10.1002/2013JD020534

Lohare, J., Nair, R., Sharma, S. K., Pandey, S. K., and Ramakrishnan, S. (2023). Geospatial based land use/land cover change detection in Jabalpur district, Madhya Pradesh. *Biol. Forum Int. J*., 15 (10), 585-592.

Liang, S. (2001). Narrowband to broadband conversions of land

surface albedo I: Algorithms. *Remote Sens. Environ.,* 76 (2), 213-238. https://doi.org/10.1016/S0034-4257(00)00205-4

Liang, S. (2003). A direct algorithm for estimating land surface broadband albedos from MODIS imagery. *IEEE Trans. Geosci. Remote Sens.,* 41 (1): 136-145. http://dx.doi.org/10.1109/TGRS.2002.807751

Lin, X., Liu, Y., Zhang, L., Li, X., Wu, W., and Chen, Y. (2022). Estimating 10-m land surface albedo from Sentinel-2 satellite observations using a direct estimation approach with Google Earth Engine. *ISPRS J. Photog. Remote Sens*., 194: 1-20. https://doi.org/10.1016/j.isprsjprs.2022.09.016

Lyapustin, A., and Wang, Y. (2022). MCD19A2 MODIS/Terra+Aqua land aerosol optical depth daily L2G global 1km SIN grid V061. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MCD19A2.061

Martonchik, J. V., Diner, D. J., Pinty, B., Verstraete, M. M., Myneni, R. B., Knyazikhin, Y., and Gordon, H. R. (1998). Determination of land and ocean reflective, radiative, and biophysical properties using multiangle imaging. *IEEE Trans. Geosci. Remote Sens.,* 36 (4): 1266-1281. https://doi.org/10.1109/36.701077

Meerdink, S.K., Hook, S.J., Roberts, D.A. and Abbott, E.A. (2019). The ECOSTRESS spectral library version 1.0. *Remote Sens. Environ.*, 230: 111196.

Mehta, A., Sinha, V. K., and Ayachit, G. (2012). Land-use/land-cover study using remote sensing and GIS in an arid environment. *Bull. Environ. Sci. Res.,* 1 (3-4), 4-8.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). Introduction to linear regression analysis (5th ed.). Wiley.

NASA. (2015). MODIS/Terra+Aqua BRDF/Albedo daily L3 global 500 m SIN grid V006 (MCD43A3). USGS/Earth Resources Observation and Science Center. https://doi.org/10.5067/MODIS/MCD43A3.006

Pandya, M. R., and Pathak, V. N. (2021). Algorithm theoretical basis document (ATBD) of vegetation indices (NDVI, EVI) and vegetation fraction retrieval from OCM-3 payload onboard Oceansat-3 (Version 1.0). Space Applications Centre, Indian Space Research Organisation. https://bhoonidhi.nrsc.gov.in/bhoonidhi_resources/help/docs/EOS06-ATBD-Vegetation_Indices_and_Fraction-V1.pdf

Qu, Y., Liu, Q., Liang, S., Wang, L., Liu, N., and Liu, S. (2014). Direct-estimation algorithm for mapping daily land-surface broadband albedo from MODIS data. *IEEE Trans. Geosci. Remote Sens.,* 52 (2): 907-919. https://doi.org/10.1109/TGRS.2013.2245670

Schaaf, C. B., Gao, F., Strahler, A. H., Lucht, W., Li, X., Tsang, T., and Roy, D. (2002). First operational BRDF, albedo nadir reflectance products from MODIS. *Remote Sens. Environ.,* 83 (1-2): 135-148. . https://doi.org/10.1016/S0034-4257(02)00091-3

Sharma, H., Burark, S. S., and Meena, G. L. (2015). Land degradation and sustainable agriculture in Rajasthan, India. *J. Indus. Poll. Control,* 31 (1): 7-11.

Shuai, Y., Masek, J. G., Gao, F., Schaaf, C. B., and He, T. (2014). An approach for the long-term 30-m land surface snow-free albedo retrieval from historic Landsat surface reflectance and MODIS-based a priori anisotropy knowledge. *Remote Sens. Environ.,* 152: 467-479. https://doi.org/10.1016/j.rse.2014.07.009

Singh, J., Mansi, Baweja, P., Neha, Arya, I., Chopra, H., Gupta, S., Gandhi, P. B., Singh, P., and Rena, V. (2023). An insight into application of land use land cover analysis towards sustainable agriculture within Jhajjar District, Haryana. *Int. J. Exp. Biol. Agric. Sci.,* 11 (4): 756-766. https://doi.org/10.18006/2023.11(4).756.766

Vermote, E. F., Tanré, D., Deuzé, J. L., Herman, M., and Morcrette, J. J. (1997). Second Simulation of the Satellite Signal in the Solar Spectrum, 6S: An overview. *IEEE Trans. Geosci. Remote Sens.,* 35 (3): 675-686. https://doi.org/10.1109/36.581987

Wang, D., Liang, S. L., He, T., and Yu, Y. (2013). Direct estimation of land surface albedo from VIIRS data: Algorithm improvement and preliminary validation. *J. Geophys. Res. Atmos.,* 118: 12,577-12,586. https://doi.org/10.1002/2013JD020417