Title

Leveraging EMIT spectroscopy and machine learning to estimate soil texture and particle grain size in dust-producing regions

Journal

Remote Sensing of Environment, 344

ISSN

0034-4257

Authors

Okin, Gregory S
Zhou, Bo
Brodrick, Philip G
et al.

Publication Date

2026-10-01

DOI

10.1016/j.rse.2026.115557

Peer reviewed

Remote Sensing of Environment 344 (2026) 115557

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Leveraging EMIT spectroscopy and machine learning to estimate soil texture and particle grain size in dust-producing regions

Gregory S. Okin [a,*], Bo Zhou [a], Philip G. Brodrick [b], Francisco Ochoa [a], María Gonçalves Ageitos [c,d], Roger N. Clark [e], Michael R. Fischella [f], Carlos Pérez García-Pando [c,g], Longlei Li [h], Natalie Mahowald [h], David R. Thompson [b], Robert O. Green [b]

[a] Department of Geography, University of California Los Angeles, Los Angeles, CA, USA
[b] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
[c] Barcelona Supercomputing Center, Barcelona, Spain
[d] Department of Project and Construction Engineering, Universitat Politècnica de Catalunya – Barcelona Tech, Terrassa, Spain
[e] Planetary Science Institute, Tucson, AZ, USA
[f] U.S. Geological Survey, Southwest Biological Science Center, Flagstaff, AZ, USA
[g] ICREA, Catalan Institution for Research and Advances Studies, Barcelona, Spain
[h] Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA

ARTICLE INFO

Edited by Jing M. Chen

Keywords:
Soil texture
Imaging spectroscopy
EMIT
Random Forest
Grain size
Dust emission

ABSTRACT

The primary mission of NASA's Earth Surface Mineral Dust Source Investigation (EMIT) is to improve our understanding of mineral dust in atmospheric radiative forcing using a novel set of spectroscopically-based mineral observations. In order to estimate mineral mass abundance, information on median surface grainsize is required. Knowledge of clay and silt fractions as well as the sub-components of the sand fractions is required for calculation of median grainsize. Therefore, this study aimed to produce a map of surface soil texture suitable for the intended application. Spectral reflectance data were incorporated into a multi-output Random Forest model to predict mass fraction for seven particle size classes: silt (TSI), clay (Clay), and five sand size classes: very coarse sand (S1, 1–2 mm diameter), coarse sand (S2, 1/2–1 mm), medium sand (S3, 1/4–1/2 mm), fine sand (S4, 1/8–1/4 mm), and very fine sand (S5, 1/16–1/8 mm diameter). To account for the presence of vegetation in EMIT spectral data, spectra of soils were mixed linearly with spectra of green and nonphotosynthetic vegetation in various combinations and amounts and the mixed spectra were used as training data. A five-fold cross-validation approach allowed us to estimate independent errors for each of the texture classes as well as three parameters estimated post hoc: total sand fraction (TSA, calculated as the sum of the five sand classes), mean grain size (GS_{mean}), and median grain size (GS_{median}). Sand fractions have estimated mean absolute errors (MAEs) <6% ($R^2 = 0.54$–0.81), while the error for TSA is <12% ($R^2 = 0.78$). MAE for TSI and Clay is <4% ($R^2 = 0.66$) and < 7% ($R^2 = 0.81$), respectively. Estimated MAE for GS_{median} is 30 μm ($R^2 = 0.78$). For GS_{mean} estimated MAE is 7.3 μm ($R^2 = 0.73$). Above 70% soil cover (below 30% vegetation cover), vegetation did not appear to influence estimates. A final model trained on all available data was used to produce estimates of soil texture class mass fractions for dust-producing areas worldwide at 0.1° resolution suitable for supporting estimates of mineral mass fraction from EMIT data. Despite drastically different estimation approaches, our results are broadly consistent with other global mineral datasets, although differences are apparent. Overall, our approach indicates the value of imaging spectrometer data for estimation of surface grain size for application to EMIT's primary mission of better constraining dust radiative forcing, though additional paired soil spectra-texture measurements would improve results. The products may also be useful for other applications in the Earth sciences, especially those related to global modeling of dust.

* Corresponding author.
E-mail address: okin@ucla.edu (G.S. Okin).

https://doi.org/10.1016/j.rse.2026.115557
Received 1 August 2025; Received in revised form 28 February 2026; Accepted 2 July 2026
Available online 9 July 2026

1. Introduction

Arid regions are major sources of mineral dust, which significantly impacts air quality, human health, and climate systems (Mahowald et al., 2011; Ravi et al., 2011). The Earth Surface Mineral Dust Source Investigation (EMIT) mission's primary goal is to better constrain dust radiative forcing. To this end, the EMIT mission placed an imaging spectrometer (also called 'EMIT') aboard the International Space Station (ISS), measuring surface reflectance in the 380–2500 nm range to estimate mineralogy of the Earth's dust-producing regions (Green et al., 2020).

In the reflected solar spectrum, mineral signatures manifest spectrally through absorption features of mixtures (such as soils), which are estimated for EMIT using the spectral feature matching algorithm Tetracorder (Clark et al., 2024). Many physical attributes of mineral mixtures impact the depth of absorption features. The primary attributes influencing the absorption depth are the abundance of a mineral, the mineral's inherent optical constants (e.g., the complex index of refraction, especially the imaginary part), the size of the absorbing mineral particles, and the size of non-absorbing mineral particles such as quartz and feldspar (Hapke, 1981). However, quantifying a mineral's abundance remains a challenge. In order to estimate mineral (mass) abundance of the top-most layers of a soil (i.e., a few mm, Okin and Painter, 2004) from spectral reflectance, knowledge of the median grainsize of the surface is required (Clark et al., 2024).

The gold standard for soil texture measurements in soil surveys is laboratory measurement of particle size distribution (at minimum, sand, silt, and clay mass fractions, though sometimes additional textural classes are included such as multiple classes of sand from very fine to very coarse), from which soil textural classes are derived (USDA-NRCS, 2007). However, this requires collection of field samples and measurement of soil texture in a laboratory. This is both time- and cost-intensive and efforts to characterize surface soils globally have focused on areas where knowledge of soil texture can support agriculture, as evidenced by compilations of soil pit data (e.g., Batjes et al., 2024). Dust-producing regions largely do not support agriculture due to low rainfall and are not well-sampled by traditional soil surveys. This presents a challenge, as mineral abundance estimation in dust-producing regions requires information about the surface soil texture.

The United Nations Food and Agriculture Organization (FAO) has produced the Harmonized World Soils Database, global soil maps that include surface soil texture (FAO/IIASA, 2012, 2023). These maps are created through rasterization of vector maps produced at 1:5,000,000 or 1:1,000,000 (for China) (FAO/IIASA, 2012). Soil polygons in largely uninhabited arid areas that are the most important dust producers are often coarse and may lack sufficient soil profile data, especially in arid and hyperarid regions, as evidenced by sparse profile coverage (e.g., Batjes, 2009). Estimates of soil texture in areas without a significant number of in situ measurements can be highly uncertain. Furthermore, soil texture in these datasets is reported for the upper layer (e.g., 0–5 cm) which is relatively thick compared to the layer, a few millimeters thick, that is active during dust emission events (Okin, 2022). Therefore, the relevance even of the upper soil layer from these datasets for dust emission modeling isn't clear. In contrast, EMIT directly senses the top few millimeters of soil that are active during dust emission, making it an ideal tool for estimating dust-relevant grainsize.

Digital soil mapping (DSM) (also called predictive soil mapping, PSM) is an approach to estimate soil characteristics in regions where there are no soil samples and beyond the strictures of polygon-based traditional soil mapping (McBratney et al., 2003; Scull et al., 2003). Many large-scale DSM efforts over the last decade-and-a-half have employed machine learning approaches to estimate soil properties (e.g., Chaney et al., 2016; Poggio et al., 2021). These approaches utilize a data-driven implementation of the long-used strategy to use landscape features which covary with soil properties (e.g., Jenny, 1941). Many such approaches now utilize remote-sensing products as inputs, including remote sensing-derived topography (and derived parameters) as well as surface reflectance, primarily from the Landsat and Moderate Resolution Imaging Spectroradiometer (MODIS) instruments (e.g., Poggio et al., 2021).

However, none of these approaches provide global estimates of clay, silt, and the full range of sand fractions (very fine to very coarse) which, because they cover more than one order of magnitude in particle size, are required for estimation of median grainsize. Even recent efforts utilizing imaging spectroscopy (e.g., Shen et al., 2025), in addition to not being global in scale, only estimate sand, silt, and clay fractions. Okin and Painter (2004) have shown that grain size effects in the sand fraction can be observed in imaging spectroscopy data in the top few millimeters, mainly as a decrease in spectral reflectance with increasing grain size, consistent with theory (Hapke, 1981). There is potential that further use of imaging spectroscopy data in estimating soil surface texture may improve estimates of the individual sand fractions. EMIT provides exactly this type of data over sparsely vegetated dust-producing areas of the globe and thus provides an opportunity to simultaneously estimate absorption band depth (i.e., the strength of spectral absorption features indicative of specific minerals as per Clark et al., 2024) and soil surface texture in support of estimating mineral mass abundance.

Here, we describe an approach to estimate soil surface grain size from EMIT data at 0.1° scale in dust-producing regions relevant to the EMIT primary mission. These estimates account for vegetation cover and include sand size fractions (i.e., very fine to very coarse) in addition to silt and clay. The resulting data have been used in the first estimates of surface mineral mass abundance for the EMIT mission.

Our maps of median grain size and soil texture have potential uses beyond supporting retrieval of mineral abundance from EMIT within the context of dust models. Dust emission is a highly non-linear, threshold-controlled process where the threshold for entrainment (specifically, the wind shear velocity at which particles are transported by wind) is a key parameter (e.g., Bagnold, 1941; Gillette and Passi, 1988). Grain size is an important determinant of the threshold for entrainment (e.g., Darmenova et al., 2009; Gillette et al., 1980) and global estimates of grain size in dust-producing regions could greatly improve estimates of when and where aeolian transport that emits dust occurs. The amount of dust produced by saltating particles is also highly dependent upon soil texture, specifically clay content (e.g., Marticorena and Bergametti, 1995). One of the main approaches in modern global models of dust emission (i.e., based on Kok et al., 2014b) depends on the clay content, and thus global maps of clay content still have potential to contribute directly to our understanding of dust emission from the world's deserts (e.g., Kawai and Matsui, 2025; Klose et al., 2021; Kok et al., 2014a; Kok et al., 2018; Li et al., 2022; Mailler et al., 2017; Tian et al., 2021).

2. Material and methods

Although vegetation cover in dust-producing regions is generally low, it is frequently non-zero (e.g., Okin, 2008). Accordingly, we adopted an approach that would allow us to estimate soil surface texture in the presence of moderate amounts of vegetation. The overall approach for this study was to 1) simulate training data by mixing laboratory spectra of soils with known soil texture with vegetation spectra in various ratios, 2) apply a multi-output machine learning approach to estimate texture based on simulated mixed reflectance, and 3) use the trained model to produce a global-scale map of soil texture in areas mapped as unconsolidated (e.g., not on bedrock, which does not produce globally important dust emission). By employing multi-output machine learning approaches, this study aims to simultaneously predict multiple soil properties—such as soil composition and particle sizes—while ensuring internal consistency among predictions. Although the comparison of sparse in situ point measurements (i.e., profile data) in dust-producing areas with the derived coarse-scale global maps is not possible, efforts were made to compare our results to another global soil texture dataset (Poggio et al., 2021) using a landform approach.

2.1. Training data

Application of machine-learning approaches requires paired observations of soil spectral properties and soil texture. To our knowledge, no dataset currently provides sufficient coverage of surface reflectance and particle-size distributions for the upper few millimeters of soil, the layer sensed by EMIT and most relevant for dust emission. As a practical alternative, we drew on the global ICRAF–ISRIC Soil VNIR Spectral Library (Shepherd and Mantel, 2006), which contains paired laboratory reflectance spectra and soil texture measurements.

The library reports texture as total silt (TSI), total clay, and five sand size classes: very coarse sand (S1, 1–2 mm), coarse sand (S2, 0.5–1 mm), medium sand (S3, 0.25–0.5 mm), fine sand (S4, 0.125–0.25 mm), and very fine sand (S5, 0.0625–0.125 mm). Summing S1–S5 provides an estimate of total sand fraction (TSA). Each texture measurement is paired with a reflectance spectrum, but multiple samples may exist for a single soil profile because measurements were collected at different depths. These depth-specific samples represent distinct physical laboratory specimens with independent texture and spectral properties. Treating all samples as independent observations therefore substantially expands the number of usable texture–spectrum pairs and, because we are not attempting to make depth-dependent predictions, the fact that they are from different depths is irrelevant. Following this approach, our training dataset comprised more than 4000 paired samples.

Spectra in the ICRAF-ISRIC library are available at 10-nm band spacing. Spectra were interpolated to 1-nm resolution then convolved to EMIT bands, with approximately 7.5 nm spectral resolution (Thompson et al., 2024). Spectroscopically active minerals in the SWIR which are commonly found in soils (e.g., hematite, goethite, kaolinite, montmorillonite, calcite, gypsum) have absorption features broader than 10-nm (Clark et al., 2003), meaning that this approach to convolution of library spectra to EMIT wavelengths is acceptable. Both absolute reflectance and first derivative of reflectance were used as predictor variables. Absolute reflectance is directly related to particle size (Hapke, 1981) but can also be influenced by conditions unrelated to soil texture (surface roughness, slope). The first derivative of reflectance, on the other hand, eliminates total brightness information while highlighting how reflectance changes across the spectrum. These changes are caused by the depth and width of absorption features which are related to particle size (Clark et al., 2024). They also help highlight compositional (mineralogical) features that are related to soil texture. However, not all EMIT wavelengths were useful for this task. Specifically removed were wavelengths lower than 454 nm (subject to increased noise and scattering residuals), from 1298–1505 as well as 1775–1980 nm (strong absorption of light by atmospheric water vapor nearly completely obscuring surface observations), and > 2301 nm (reducing residuals due to high noise when atmospheric water vapor levels are high) (Coleman et al., 2024).

Training data were simulated by mixing spectra, convolved to EMIT wavelengths, of green vegetation (GV) and nonphotosynthetic vegetation (NPV), also convolved to EMIT wavelengths, with the soil spectra described above. GV (n = 425) and NPV (n = 303) spectra were taken from available online libraries (Meyer et al., 2022; Ochoa et al., 2024). We used the spectral mixing approach of Ochoa et al. (2025) to simulate mixed spectra. In short, we randomly choose one spectrum from each class (ρ_{NPV}, ρ_{GV}, ρ_{soil}) and fractions (f_{NPV},f_{GV},f_{soil}) generated from a Dirichlet distribution, which requires that the fractions sum to one and that the fractions exist on the interval [0,1]. f_{soil} was limited to ≥65%, a threshold below which we anticipated that we would not be able to accurately estimate soil texture. Vegetation obscures the surface and, indeed, initial models showed that below 65%, soil texture couldn't be retrieved accurately even in simulation. Simulated spectra were then calculated as $\rho_{sim} = f_{gv}\rho_{gv} + f_{npv}\rho_{npv} + f_{soil}\rho_{soil}$. This approach for simulating spectra has been used by many studies (e.g., Dennison et al., 2019; Okin and Gu, 2015; Okin et al., 2001) and resulted in a total of 50,000 simulated mixed spectra. Because each of these simulated spectra is linked with a single soil spectrum from the ICRAF-ISRIC library (Shepherd and Mantel, 2006), each is associated with a soil texture measurement, allowing us to use these simulated spectra as training data. This approach follows that of other authors (e.g., Okujeni et al., 2013; Schug et al., 2020; Schug et al., 2018; Senf et al., 2020) who train machine learning models on simulated mixed data, although here our goal is to uncover information about an associated soil characteristic in a mixed spectrum, rather than trying to use the machine learning model to estimate mixed fractions themselves.

2.2. Model training and cross-validation

Because the soil texture components are constrained to sum to 100% in all cases, we utilized a multi-output modeling approach, employing a single model to predict all variables simultaneously (Borchani et al., 2015). For its simplicity, robustness, and reliability, we selected the Random Forest algorithm (RF) (Breiman, 2001) as the machine learning model. To maximize prediction accuracy, we implemented hyperparameter tuning for best values of n_estimators (final value = 80), min_samples_split (final value = 3), min_samples_leaf (final value = 15), min_impurity_decrease (final value = 0.2), max_leaf_nodes (final value = 90), max_features (final value = 0.3), and max_depth (final value = 110) (Probst et al., 2019). To make the tuning process more efficient, we first used random search to identify approximate optimal parameter ranges for each hyperparameter. This was followed by grid search to exhaustively explore the optimal parameter combinations within those ranges (Dhilsath and Samuel, 2021).

Our simulation approach meant that multiple simulated spectra originated from each soil sample and therefore the simulations are not truly independent of one another. This creates an unusual complexity in splitting training and testing data, because care must be taken to ensure that spectra using any single soil spectrum are not included in both training and testing sets (Kapoor and Narayanan, 2023). Accordingly, the simulated training dataset was divided into five groups such that all simulated spectra associated with any soil spectrum were only placed in one group. The model was then trained on four of the five groups (internally using an 80/20 training/testing split) and predictions were made for the fifth group. This was repeated five times, using each group as the testing dataset once. Error metrics (mean absolute error, MAE; root mean squared error, RMSE; coefficient of determination, R^2) were calculated on the independent predictions of the five groups once assembled post-prediction.

Derived metrics were calculated from the multi-output model output. Total sand (TSA) fraction was calculated as the sum of all sand (S1-S5) fractions. Mean grain size (GS_{mean}) was obtained through a weighted average, weighing the grain size particle size class centers in microns by the predicted mass fraction for each texture component. Particle size class centers in microns ([1, 30, 94, 188, 375, 750, 1500] for [clay, silt, very fine sand, fine sand, medium sand, coarse sand, and very coarse sand]) were calculated as the simple average of the minimum and maximum grain size for the particle size class. To calculate median grain size (GS_{median}), the cumulative distribution was calculated for each prediction (x = particle size class centers in microns and y = the corresponding cumulative percentage of data in the particle size class associated with x and all finer particle size classes). A linear interpolation was used to estimate the median (50th percentile) grainsize in microns. This empirical approach does not make any assumptions about the shape of the overall particle size distribution. Non-uniform and even multimodal distributions can be accommodated.

To derive independent error metrics—ensuring no overlap between training and testing data—we used a cross-validation approach. Cross-validation is a statistical method to estimate the independent prediction accuracy of models by withholding random samples from the dataset for testing while training the model on the remaining data. However, because spectra were convolved, we introduced an additional mechanism to ensure that mixed spectra derived from the same source soil spectra appear only in the training or testing set, not both, to prevent training and testing with the same soil spectrum (Kapoor and Narayanan, 2023).

Cross-validation was necessary for error assessment. However, given the relatively limited number of soil samples, and the fact that increasing the amount of training data for RF models typically improves model performance, a final model was trained using all of the available training data. Reported accuracies are from the cross-validation, meaning they are likely conservative relative to the final model performance when compared against the training dataset (Ramezan et al., 2019).

2.3. EMIT data and global mapping

The EMIT mission has been described in detail elsewhere (Coleman et al., 2024; Green et al., 2020; Thompson et al., 2024). Briefly, EMIT provides data in 285 bands from 400 to 2500 nm with approximately 7.5 nm band width (at half-max) at approximately 60-m resolution. Data limitations inherent in being housed on the ISS mean that spectroscopic data could only be collected over specific areas which, for the EMIT primary mission (Mid-2022 – Early 2024), were globally important dust-producing regions using a target mask produced by the EMIT team. The EMIT sampling strategy means that multiple estimates of surface reflectance are available within (and just outside) the EMIT target mask. Fractional cover of GV, NPV, and soil is also estimated for each EMIT spectrum as part of the creation of the L3 products (Brodrick et al., 2023).

Soil texture fractions (i.e., soil mass fraction in S1-S5 sand fractions, TSI, and Clay) were estimated for all EMIT spectra using the RF model trained on all available data. Only pixels with >65% soil fractional cover were estimated. Data collected between 08/20/2022 (mission start) and 10/20/2024 were used in this study. Scenes with more than 50% cloud cover were excluded entirely.

Calculations were run at native EMIT resolution then mosaicked by first reprojecting to a 0.00055° grid for global consistency using nearest neighbor. Mosaics were generated by selecting pixels with the minimum solar zenith angle, after pixel-level cloud and water masking. Pixels were masked for fractional cover (only pixels with soil >65% were included) using the EndMember Combination Monte Carlo, $E(MC)^2$, approach of Ochoa et al. (2025). The (vector) Global Unconsolidated Materials (GUM) map compiled by Börker et al. (2018) was rasterized to the same 0.00055° grid (using the center point) and pixels that were not mapped as unconsolidated (i.e., bedrock) were omitted when averaging soil texture estimates to 0.1° x 0.1° resolution.

2.4. Comparisons with other global soil texture datasets

Our approach resulted in estimates of soil texture across the world's dust-producing drylands at 0.1° resolution for the purpose of supporting the EMIT mission's goal of estimating surface mineralogy and dust radiative forcing as well as providing a surface grain-size dataset for future modeling efforts. Independent validation of this dataset is complicated by its coarse resolution and the fact that there are a very limited number of in situ soil texture estimates from the drylands that produce most of the world's dust. In short, the very areas where we might most need in situ samples to support the EMIT mission are the areas where such samples are the scarcest. Thus, one approach to determining the potential utility of our soil texture measurements is to compare them with other global datasets such as SoilGrids 2.0 (Poggio et al., 2021). SoilGrids does not include information about sand texture classes (i.e. very fine to very coarse sand), and so comparisons here were limited to clay, silt, and total sand fractions using SoilGrids (from the surface soil layer) data downsampled to the 0.1° resolution of the EMIT-derived fractions. Polygons from the GUM map of Börker et al. (2018) were extracted for all aeolian deposits, dunes (a subset of all aeolian deposits), loess, and lacustrine deposits (Fig. 1). Geomorphic processes that lead to these types of Earth surface deposits result in surface soil textures dominated by specific texture classes. Aeolian deposits and dunes, for example, are typically sandy (e.g., Goudie, 2022). Loess deposits are, by definition, finer-grained and should be dominated by silt-sized particles (e.g., Crouvi et al., 2010). Lacustrine deposits can be complex mixtures of many particle sizes, but are expected to be dominated by silt and clay fractions, though this can be complicated by heterogeneous mixtures of particles (Schnurrenberger et al., 2003). Lacustrine deposits are also often considerably smaller than the 0.1° resolution here (FAO/IIASA, 2023). For this reason, we used both 0.1° resolution results and native resolution (60 m) EMIT estimates of grainsize to examine these deposits. Nonetheless, use of lacustrine deposits is complicated by the fact that screening algorithms used by EMIT often interpret large, bright playa/pan systems as clouds which are omitted from downstream computations, including texture estimates. This means that not all lacustrine deposits which are in the Börker et al. (2018) dataset and also in the EMIT target mask will have estimates of grainsize, even at native resolution.

3. Results

Overall, the multi-output RF model of soil texture performs well, with MAE ranging from 1.8% to 6.2%. Total sand (TSA), calculated as the post hoc sum of S1-S5, generally has the highest error of ~12%, due to summing uncertainties of the soil sand components (Fig. 2). Error is generally the same for samples >70% soil cover, with samples 65% -70% soil cover exhibiting considerably higher error across all grain sizes (Fig. 2 and Table 1).

Variation in the distribution of samples in each particle size class means that both coefficient of determination (R^2) and MAE are required to evaluate the models. There are samples across the entirety of the (0–100%) clay fraction (Fig. 2) and this particle size class exhibits both low MAE (6.2) and high R^2 (0.81). In contrast, the S2 class exhibits low MAE (1.8), due to the generally small fractions in this particle size class (Table 1), with low R^2 (0.54). Thus, despite the low error, the S2 class should be considered the least-well predicted particle size.

Mean and median grain size are generally predicted well (Fig. 3, Table 1). As with the individual fractions, the accuracy worsens (MAE increases) below 70% soil cover (Table 1).

Global maps of EMIT-derived sand, silt, and clay fractions differ from those from SoilGrids (Fig. 4). Global correlations between EMIT- and SoilGrids-derived fractions within the EMIT target mask are poor (sand: 0.08, silt: 0.21, clay: 0.08). Throughout much of EMIT-derived areas in the Northern Hemisphere, SoilGrids predicts lower sand fraction than our approach, though in southern Africa and some parts of Australia this is reversed. SoilGrids predicts considerably higher silt content in Asia and North America than our approach while predicting lower silt content in southern Africa and Australia. Across the Northern Hemisphere, EMIT largely predicts lower clay content than SoilGrids and higher content in southern Africa and Australia. Distributions of sand, silt, and clay fractions in South America are largely similar between SoilGrids and our approach.

Across individual landform types, our approach generally estimates greater sand fractions compared to SoilGrids for aeolian landforms (59% vs 55%), especially sand dunes (62% vs 51%). Our approach generally estimates loess landforms to be sandier (44% vs. 32%) than SoilGrids. However, this appears to be mainly due to estimation of the finer sand fractions because, when we calculate the median grain size from our results, loess landforms exhibit a clear mode in the 50–60 μm range (Fig. 5). Aeolian landforms, in contrast, exhibit a coarser mode (around 150 μm, fine sand), consistent with their origin. Dune deposits, where active aeolian transport tends to winnow away clay and silt fractions exhibit very few cells with median grain sizes finer than sand (Fig. 5).

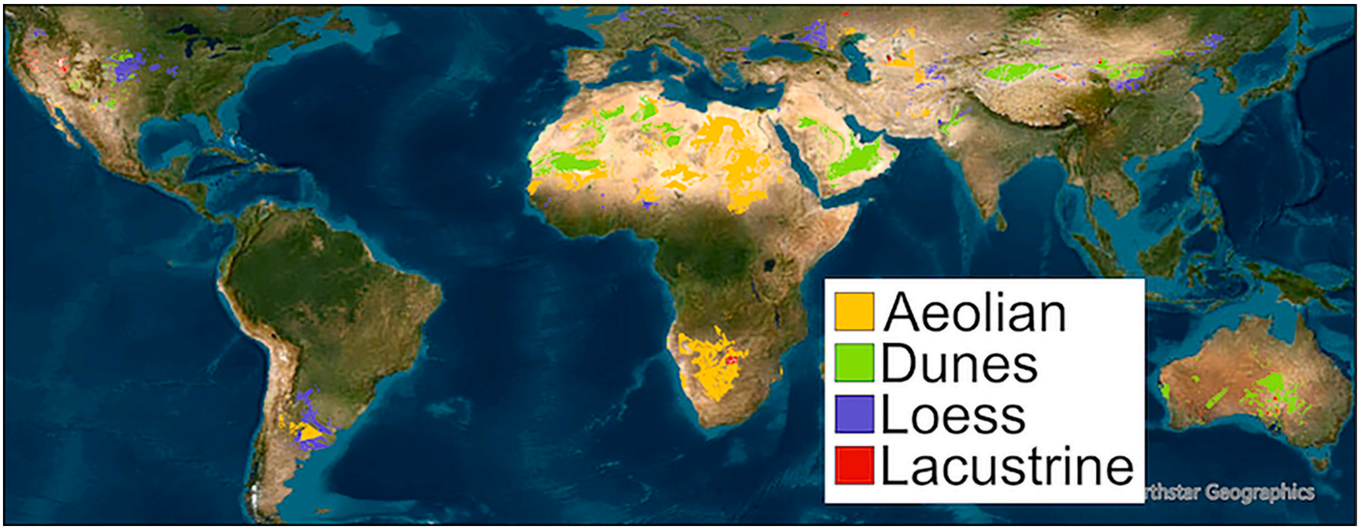

Fig. 1. Landforms from Börker et al. (2018) used for comparisons.

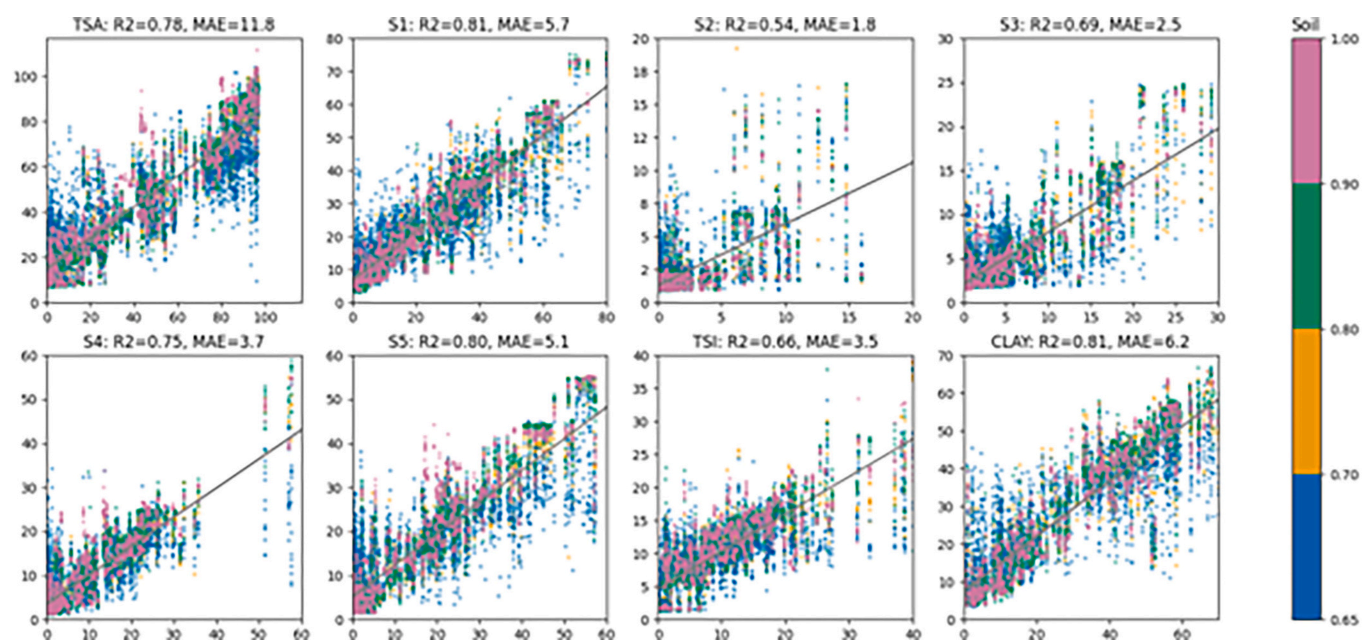

Fig. 2. Five-fold cross validation results for multi-output RF models of soil texture classes (S1: very coarse sand,1–2 mm; S2: coarse sand, 1/2–1 mm; S3: medium sand, 1/4–1/2 mm; S4: fine sand, 1/8–1/4 mm; S5: very fine sand, 1/16–1/8 mm diameter; TSI: total silt, 2 μm- 1/16 mm; Clay, <2 μm). Total sand (TSA) is calculated as the sum of S1-S5. Colors of individual points represent fraction of soil in the simulated spectrum (from 65% - 100% soil). Gray lines are regressions between predicted and actual values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Because SoilGrids does not predict individual sand-sized fractions (e.g., very fine to very coarse sand), meaningful median grain size calculations based on these data are impossible, precluding direct comparison with our approach.

4. Discussion

4.1. Vegetation mixing and training data simulation

One of the challenges inherent in our approach, which was to rely on EMIT-derived spectra of the surface to estimate grain size, is that at the scale of an individual EMIT 60-m GIFOV, the spectral signature of the soil would frequently be contaminated with light reflecting from vegetation. Because the training dataset was based on laboratory spectra of pure soils without vegetation, we needed to artificially add the impact of vegetation by mixing soil spectra with spectra of vegetation (GV and NPV) in various quantities using the approach of Ochoa et al. (2025). We acknowledge that interaction of light with the complex three-dimensional vegetated surface is frequently more complex than the simple linear mixture approach used here (Tanre et al., 1981). However, linear mixing has been shown to be a good first-order approach when unmixing for sparsely vegetated arid regions (Okin et al., 2013) and we do not anticipate that the use of forward linear mixing in generating training data with a vegetation component detracts from the inherently non-linear RF estimation approach used here. Indeed, our approach bears a close similarity to the work of Okujeni et al. (2013) and subsequent approaches (Schug et al., 2020; Schug et al., 2018; Senf et al., 2020). These authors showed that linear mixing to create augmented training data in support of a machine learning prediction serves as a feasible approach.

Table 1
MAE (R^2) from five-fold cross-validation of multi-output RF predictions for all simulated spectra (65–100% soil) and for different ranges of soil cover. TSA, GS_{median}, and GS_{mean} (μm) are calculated post hoc from multi-output texture class fractions.

Texture Class	Percent soil in simulated spectrum				
	65–100%	90–100%	80–90%	70–80%	65–70%
TSA	11.8 (0.78)	10.8 (0.83)	8.3 (0.89)	9.7 (0.86)	16.2 (0.64)
S1	5.7 (0.81)	4.9 (0.88)	3.8 (0.92)	4.8 (0.86)	8.1 (0.66)
S2	1.8 (0.54)	1.7 (0.59)	1.6 (0.63)	1.7 (0.59)	2.0 (0.43)
S3	2.5 (0.69)	2.4 (0.73)	2.0 (0.80)	2.2 (0.75)	3.1 (0.54)
S4	3.7 (0.75)	3.5 (0.83)	2.6 (0.90)	3.0 (0.85)	5.1 (0.58)
S5	5.1 (0.80)	4.6 (0.84)	3.4 (0.90)	4.0 (0.88)	7.2 (0.66)
TSI	3.5 (0.66)	3.4 (0.70)	2.8 (0.77)	3.2 (0.71)	4.3 (0.49)
Clay	6.2 (0.81)	5.6 (0.86)	4.3 (0.90)	5.4 (0.86)	8.8 (0.66)
GS_{median} (μm)	30 (0.78)	26 (0.83)	22 (0.87)	26 (0.83)	40 (0.68)
GS_{mean} (μm)	7.3 (0.73)	6.6 (0.78)	5.7 (0.84)	6.4 (0.79)	9.3 (0.60)

4.2. Model performance and comparison with existing digital soil mapping products

Overall, our cross-validation indicates that, at least for most particle size-classes, as long as soil cover is >70% our approach estimates mass fractions within about 8% of actual values (for multi-output RF model-predicted classes) with <12% for aggregated sand fraction (TSA). In comparison, for a high-resolution predictive soil model that includes soil texture (sand, silt, clay) and is limited to the continental US, Chaney et al. (2019) estimate overall MAE for sand, silt, and clay fractions of 13%, 10%, and 7.8%. Poggio et al. (2021) estimate overall RMSE for sand, silt, and clay fractions for their model to be 18%, 13%, and 13%.

The errors we estimate for our multi-output RF model-predicted classes, as well as the calculated TSA compare favorably with these estimated errors (Table 1). However, we caution against over-interpretating these comparative errors for two reasons: 1) Chaney et al. (2019) and Poggio et al. (2021) report errors for different particle size classes at all depths, whereas we predict only surface texture. It is possible that predictions of Chaney et al. (2019) and Poggio et al. (2021) for surface soils are better than for the entire soil profile, though they aren't reported. 2) Poggio et al. (2021) predicts soil texture and other parameters globally, even where vegetation obscures the surface, whereas our product only predicts soil texture where vegetation cover is low (<35%, or soil cover >65%). It is possible that Poggio et al.'s (2021) errors are lower in the areas within our target mask than globally.

4.3. Spectral and physical sources of uncertainty

As with all machine learning models, the quality of the inference hinges on the quality of the input data. Given that our model was trained with lab and field-based spectra, any deviations between the remote and in situ reflectance will negatively impact the performance of the grain size model. This includes not only discrepancies caused by poor atmospheric retrievals, but also definitional inconsistencies in the forward model. For example, EMIT surface reflectance estimates are directional by definition, and may have spectral variability that is not fully removed by the brightness normalization process used as preprocessing for the random forest. Similarly, the EMIT surface reflectance models the planet as a uniform sphere – deviations from that assumption may introduce additional errors. However, bidirectional effects are generally spectrally broad (Carmon et al., 2023), so the first derivatives of reflectance used as predictors should mitigate them. The use of first derivatives of reflectance should also mitigate the impacts of surface roughness in EMIT GIFOV's because surface slopie or self-shadowing tend to impact overall brightness rather than spectral shape.

4.4. Geomorphic consistency and grain size distributions across landforms

For landforms which should have characteristic soil textures due to their mode of origin (Table 2), our results and the SoilGrids predictions of Poggio et al. (2021) are remarkably similar despite the very different ways they were produced. SoilGrids is produced from a RF model trained from soil property/profile data from the World Soil Information Service (WoSIS, Batjes et al., 2024) using ecological, climatological, hydrological, topographical, and remote sensing data as predictors. Our approach, in contrast, produces predictions from an RF model trained on soil texture data from the ICRAF-ISRIC library (Shepherd and Mantel, 2006), which includes detailed soil textural data and laboratory reflectance spectra. Our method used only reflectance as predictors, ignoring location. Indeed, our product appears to produce more plausible soil texture for certain landforms; aeolian landforms, for example, have higher sand fractions using our approach than using SoilGrids (Table 2). Other landforms have generally similar sand/silt/clay fractions or display median grain size distributions consistent with their formation

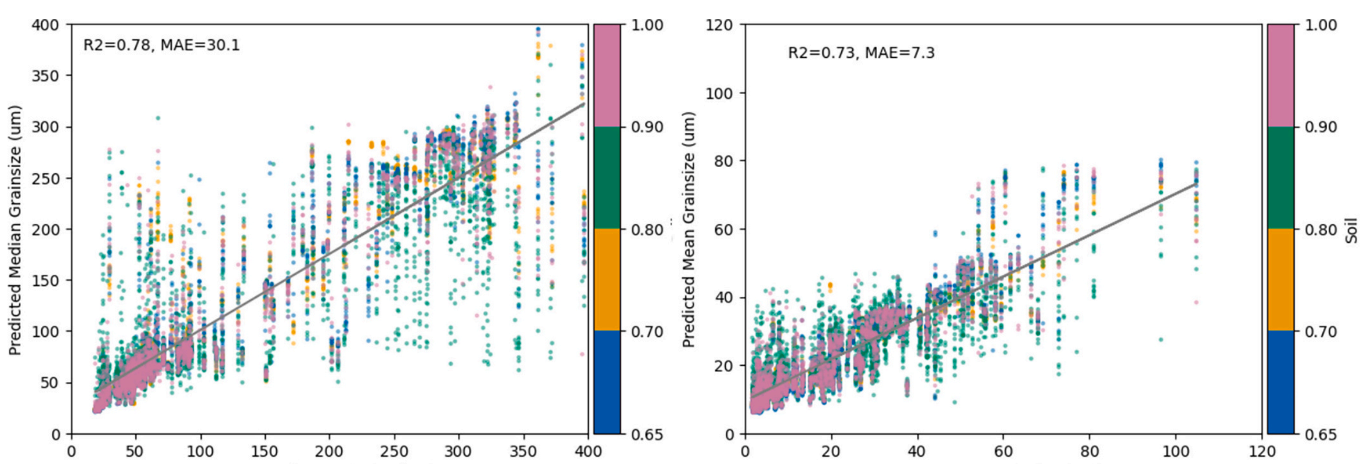

Fig. 3. Median and mean grain size calculated using multi-output RF texture classes from five-fold cross validation. Colors represent the simulated fraction of soil in each of the spectra.

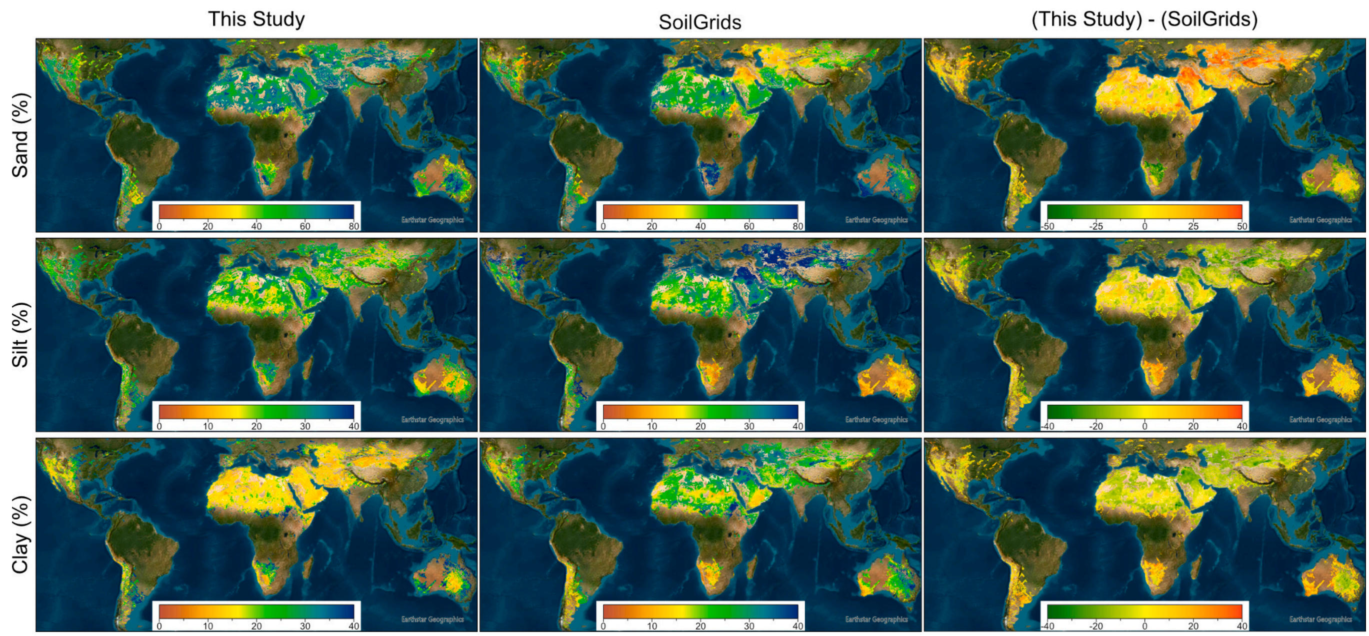

Fig. 4. EMIT-derived (this study), SoilGrids 2.0-derived, and difference in fractions of sand, silt, and clay for areas with EMIT-derived texture.

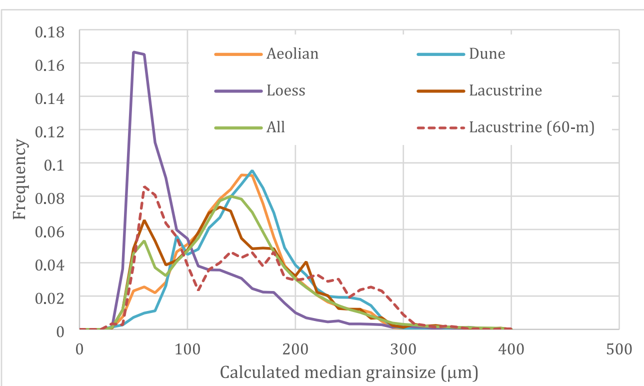

Fig. 5. Calculated median grain size from multi-output RF model for different landforms from Börker et al. (2018).

(e.g., loess, Crouvi et al., 2010). Although at 0.1° resolution, lacustrine deposits are considerably coarser in the EMIT-derived product than in SoilGrids, when compared at finer resolution (60-m native to EMIT), the EMIT and SoilGrids sand, silt, and clay fraction estimates are nearly the same. The apparent coarseness of lacustrine deposits at 0.1° x 0.1° is due to their small size such that averaging values from the deposits with neighboring deposits results in mixing results from fine lacustrine deposits with coarser neighboring geomorphic surfaces.

Very few landforms exhibit median grain sizes below 40 μm in our global predictions (Fig. 5). This could be due to at least two non-exclusive reasons. The first reason relates to the distribution of silt- and clay-dominated soils on the surface. Very few silt- and clay-dominant landforms in the EMIT target mask with >65% soil exposure exist at scales larger than 0.1°. For example, the distribution of vertisols (soils which are nearly entirely clay at the surface and therefore have median grain size <40 μm) globally is extremely limited, with little overlap between those areas modeled here (e.g., the world's largest area of vertisols is in an area of India not included in the EMIT target mask, Sanchez et al., 2009). Although some Quaternary lacustrine deposits do exist that would have clay- and silt-dominant grain sizes, these tend to be in depressions smaller than our 0.1° grid resolution. The second reason for so few areas exhibiting median grain sizes below 40 μm in our global predictions concerns the predictions themselves. Although our training dataset included samples with very fine median grain sizes (down to 20 μm), this is the extreme low end of the distribution. It is well known that RF and other approaches tend to "regress to the mean", overestimating the low end of the distribution and underestimating the upper end of the distribution (Belitz and Stackelberg, 2021). Although we have not directly predicted median grain size, this effect is nonetheless apparent in our results. Indeed, nearly all predictions for samples with true median grain size <40 μm are greater than the actual value (Fig. 3). Thus, it is not surprising that our approach has difficulty correctly predicting grain size in very fine-texture soils. That said, median grain size for landforms expected to have the finest soils, loess and lacustrine deposits (at 60 m), do show a distinct mode of 50–60 μm. A direct comparison cannot be made with SoilGrids, as this calculation requires prediction of the various sand fractions. However, this does suggest that, even though our approach predicts lower silt content than

Table 2
Comparison of predicted median grainsize (μm) between this study and SoilGrids for different major landform types (from Börker et al. 2018, Fig. 1). Because of their small size, grain size retrievals at native 60-m EMIT resolution were compiled for lacustrine deposits in addition to the coarse resolution (0.1° x 0.1°) product presented here.

Landform Type	Sand		Silt		Clay		Number of cells
	SoilGrids	This Study	SoilGrids	This Study	SoilGrids	This Study	
Aeolian	55.3 ± 14.0	59.4 ± 11.0	23.4 ± 6.8	17.3 ± 6.9	21.3 ± 6.8	17.3 ± 6.9	60,825
Dune	51.4 ± 12.1	61.9 ± 9.4	26.0 ± 6.3	15.8 ± 7.3	22.6 ± 6.3	15.8 ± 6.1	26,702
Loess	32.0 ± 15.4	44.0 ± 12.5	42.6 ± 6.9	27.8 ± 5.4	25.4 ± 6.9	27.8 ± 9.4	10,257
Lacustrine	48.3 ± 16.9	54.0 ± 12.1	28.1 ± 7.2	21.5 ± 6.2	23.6 ± 7.2	21.5 ± 9.0	2367
Lacustrine (60 m)	53.3 ± 16.3	55.3 ± 15.6	22.1 ± 14.1	22.1 ± 8.4	24.6 ± 6.6	22.6 ± 10.6	2952

SoilGrids for loess deposits, it nonetheless captures the fine-grained nature of these landforms. For lacustrine deposits at 60-m resolution, our estimates and SoilGrids produce nearly identical sand, silt, and clay fractions.

5. Conclusions

A central motivation for this work is that estimation of mineral mass abundance from imaging spectroscopy requires knowledge of the median grain size of the soil surface. Grain size exerts a first-order control on absorption band depth by influencing photon path length and scattering behavior, such that mineral abundance cannot be inferred from surface reflectance without knowledge of median grainsize (Clark et al., 2024). Despite the importance of grain size, no existing global soil dataset provides spatially explicit estimates that can be used to estimate median grainsize. Indeed, available global-scale products report only aggregate sand, silt, and clay fractions and are commonly derived from soil profile measurements representing centimeter- to meter-scale depths rather than the millimeter-scale surface sensed by imaging spectroscopy that is active during dust emission.

The approach presented here addresses this key gap by retrieving individual sand–size fractions, together with silt and clay, directly from EMIT data and using these fractions to derive median grain size across dust-producing regions globally. By combining laboratory spectral libraries, simulated vegetation–soil mixtures, and a multi-output Random Forest framework, we produced a dust-relevant dataset of particle–size fractions and median grain size consistent with EMIT surface reflectance observations. This capability enables physically consistent interpretation of spectroscopic absorption features and represents a necessary step toward robust estimation of mineral abundance in dust-producing regions. More broadly, the resulting grain-size product introduces a new observational constraint on aeolian processes and dust emission modeling that has not previously been available from global soil information systems.

Cross-validation results indicate that the modeling framework retrieves individual texture fractions with generally low error under sparsely vegetated conditions, and accuracy decreasing only when soil exposure falls below ~70%. Comparisons with SoilGrids demonstrate regional differences in predicted sand, silt, and clay fractions, reflecting the contrasting observational bases of the two datasets. Landform-based analyses suggest that the EMIT-derived textures are physically plausible and, in some cases, more consistent with geomorphic expectations than SoilGrids, particularly for aeolian deposits and dunes. These findings support the value of surface-sensitive spectroscopic observations for characterizing soil physical properties in drylands where traditional soil surveys remain sparse.

Several limitations remain. Model performance depends on the representativeness of available paired spectral–texture measurements and may be affected by uncertainties in atmospheric correction, surface anisotropy, and the simplified treatment of vegetation–soil mixing. The coarse spatial resolution of the final product and scarcity of in situ validation data in major dust source regions, especially at scales relevant here further constrain quantitative assessment. Future work should prioritize expanded field campaigns that collect co-located soil texture and spectral measurements, improved representation of surface heterogeneity and crusting, and exploration of alternative modeling approaches that explicitly enforce compositional and physical constraints.

Nonetheless, the dataset developed here provides an enabling component for the next generation of mineral abundance retrievals from imaging spectroscopy and contributes new information relevant to dust research. Because dust emission thresholds and dust production efficiency are strongly controlled by particle size distribution and clay content, global maps of median grain size and texture fractions offer valuable inputs for process-based dust emission models and climate simulations. As imaging spectroscopy missions continue to expand, methods that jointly retrieve mineralogical and physical soil properties, such as the framework demonstrated here, will play an increasingly important role in improving quantitative understanding of mineral dust sources and their climatic impacts.

CRediT authorship contribution statement

Gregory S. Okin: Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Bo Zhou:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Philip G. Brodrick:** Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization. **Francisco Ochoa:** Methodology, Investigation, Formal analysis. **María Gonçalves Ageitos:** Writing – review & editing, Validation. **Roger N. Clark:** Writing – review & editing, Conceptualization. **Michael R. Fischella:** Writing – review & editing, Methodology, Conceptualization. **Carlos Pérez García-Pando:** Writing – review & editing, Validation, Methodology, Conceptualization. **Longlei Li:** Writing – review & editing, Validation, Conceptualization. **Natalie Mahowald:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **David R. Thompson:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Robert O. Green:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Gregory S. Okin reports financial support, administrative support, and article publishing charges were provided by NASA. Longlei Li reports financial support was provided by NASA. Natalie Mahowald reports financial support was provided by NASA. Roger N Clark reports financial support was provided by NASA. Carlos Perez Garcia-Pando reports financial support was provided by Spain Ministry of Economy and Competitiveness. Carlos Perez Garcia-Pando reports financial support was provided by European Research Council. Maria Goncalves Ageitos reports financial support was provided by Spain Ministry of Economy and Competitiveness. Maria Goncalves Ageitos reports financial support was provided by European Research Council. Maria Goncalves Ageitos reports financial support was provided by European Regional Development Fund. Carlos Perez Garcia-Pando reports was provided by European Regional Development Fund. Longlei Li reports financial support was provided by US Department of Energy. Carlos Perez Garcia-Pando reports financial support was provided by AXA Research Fund. Maria Goncalves Ageitos reports financial support was provided by AXA Research Fund. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

EMIT is supported by the National Aeronautics and Space Administration Earth Venture Instrument Program, under the Earth Science Division of the Science Mission. A portion of this research was performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. We acknowledge the support and assistance of NASA's International Space Station Program. The USGS authors' contribution to this published work was prepared by U.S. federal government employees as part of their official duties and constitutes a “work of the United States government,” and is considered to be in the public domain and therefore domestic copyright does not apply. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. L.L. acknowledges assistance from EMIT, as well as from the Department of Energy (DOE) under award DE-SC0021302. US Government Support Acknowledged. CPG-P and MGA acknowledge support from the ERC under the Horizon 2020 research and innovation programme through the FRAGMENT project (grant agreement no. 620 773051), the Spanish Ministerio de Economía y Competitividad through the HEAVY project (grant no. PID2022-140365OB-I00 618 funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU), the Horizon Europe Programme under Grant Agreement No 101137680 via project CERTAINTY, and the AXA Research Fund through the AXA Chair on Sand and Dust Storms at BSC.

Data availability

Model predictions at 0.1° x 0.1° resolution are available at: https://zenodo.org/records/16700296. The code for producing texture class fractions from EMIT data at native EMIT resolution (grainsize.py) is available from the EMIT Science Data System Level 3 repository (https://github.com/emit-sds/emit-sds-l3).

References

Bagnold, R.A., 1941. The Physics of Blown Sand and Desert Dunes. Methuen, New York.
Batjes, N.H., 2009. Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. Soil Use Manag. 5, 124–127.
Batjes, N.H., Calisto, L., de Sousa, L.M., 2024. Providing quality-assessed and standardised soil data to support global mapping and modelling (WoSIS snapshot 2023). Earth Syst. Sci. Data 16, 4735–4765.
Belitz, K., Stackelberg, P.E., 2021. Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. Environ. Model Softw. 139, 105006.
Borchani, H., Varando, G., Bielza, C., Larrañaga, P., 2015. A survey on multi-output regression. Wiley Interdiscipl. Rev. Data Min. Knowledge Discov. 5, 216–233.
Börker, J., Hartmann, J., Amann, T., Romero-Mujalli, G., 2018. Terrestrial sediments of the earth: development of a global unconsolidated sediments map database (GUM). Geochem. Geophys. Geosyst. 19, 997–1024.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brodrick, P., Okin, G., Ochoa, F., Thompson, D., Clark, R., Ehlmann, B., Keebler, A., Miller, R., Mahowald, N., Ginoux, P., Garcia-Pando, C., Goncalves, M., Green, R., 2023. EMIT L3 Aggregated Mineral Spectral Abundance and Uncertainty 0.5 Deg V001. accessed 2025-06-15 in. NASA EOSDIS Land Processes Distributed Active Archive Center.
Carmon, N., Berk, A., Bohn, N., Brodrick, P.G., Dozier, J., Johnson, M., Miller, C.E., Thompson, D.R., Turmon, M., Bachmann, C.M., Green, R.O., Eckert, R., Liggett, E., Nguyen, H., Ochoa, F., Okin, G.S., Samuels, R., Schimel, D., Song, J.J., Susiluoto, J., 2023. Shape from spectra. Remote Sens. Environ. 288, 113497.
Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C. W., Odgers, N.P., 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67.
Chaney, N.W., Minasny, B., Herman, J.D., Nauman, T.W., Brungard, C.W., Morgan, C.L. S., McBratney, A.B., Wood, E.F., Yimam, Y., 2019. POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. Water Resour. Res. 55, 2916–2938.
Clark, R.N., Swayze, G.A., Wise, R., Livo, K.E., Hoefen, T.M., Kokaly, R.F.K., Sutley, S.J., 2003. USGS digital spectral library splib05a. USGS, Denver, CO.
Clark, R.N., Swayze, G.A., Livo, K.E., Brodrick, P.G., Noe Dobrea, E., Vijayrangan, S., Green, R.O., Wettergreen, D.S., Candela, A., Hendrix, A., Garcia-Pando, C.P., Pearson, N., Land, M.D., Gonzalez-Romero, A., Querol, X., EMIT Team, TREX Team, 2024. Imaging spectroscopy: earth and planetary remote sensing with the PSI Tetracorder and expert systems from rovers to EMIT and beyond. Plan. Sci. J. 5, 276.
Coleman, R.W., Thompson, D.R., Brodrick, P.G., Ben Dor, E., Cox, E., Pérez García-Pando, C., Hoefen, T., Kokaly, R.F., Meyer, J.M., Ochoa, F., Okin, G.S., Pearlshtien, D.H., Swayze, G., Green, R.O., 2024. An accuracy assessment of the surface reflectance product from the EMIT imaging spectrometer. Remote Sens. Environ. 315, 114450.
Crouvi, O., Amit, R., Enzel, Y., Gillespie, A.R., 2010. Active sand seas and the formation of desert loess. Quat. Sci. Rev. 29, 2087–2098.
Darmenova, K., Sokolik, I.N., Shao, Y.P., Marticorena, B., Bergametti, G., 2009. Development of a physically based dust emission module within the weather research and forecasting (WRF) model: assessment of dust emission parameterizations and input parameters for source regions in central and East Asia. J. Geophys. Res.-Atmos. 114, D14201.
Dennison, P.E., Qi, Y., Meerdink, S.K., Kokaly, R.F., Thompson, D.R., Daughtry, C.S.T., Quemada, M., Roberts, D.A., Gader, P.D., Wetherley, E.B., Numata, I., Roth, K.L., 2019. Comparison of methods for modeling fractional cover using simulated satellite hyperspectral imager spectra. Remote Sens. 11, 2072.
Dhilsath, F.M., Samuel, S.J., 2021. Hyperparameter tuning of ensemble classifiers using grid search and random search for prediction of heart disease. In: Computational Intelligence and Healthcare Informatics, pp. 139–158.
FAO/IIASA, 2012. Harmonized World Soil Database (Version 1.2). FAO, Rome, Italy. IIASA, Laxenburg, Austria.
FAO/IIASA, 2023. Harmonized World Soil Database (Version 2.0). FAO, Rome, Italy. IIASA, Laxenburg, Austria.
Gillette, D.A., Passi, R., 1988. Modeling of dust emission caused by wind erosion. J. Geophys. Res. 93, 14223–14242.
Gillette, D.A., Adams, J., Endo, A., Smith, D., Kihl, R., 1980. Threshold velocities for input of soil particles into the air by desert soils. J. Geophys. Res., 85, 5621–5630.
Goudie, A.S., 2022. Aeolian processes and landforms. In: Burt, T.P., Goudie, A.S., Viles, H.A. (Eds.), The History of the Study of Landforms or the Development of Geomorphology: Volume 5: Geomorphology in the Second Half of the Twentieth Century. Geological Society of London.
Green, R.O., Mahowald, N., Ung, C., Thompson, D.R., Bator, L., Bennet, M., Bernas, M., Blackway, N., Bradley, C., Cha, J., Clark, P., Clark, R., Cloud, D., Diaz, E., Ben Dor, E., Duren, R., Eastwood, M., Ehlmann, B.L., Fuentes, L., Ginoux, P., Gross, J., He, Y., Kalashnikova, O., Kert, W., Keymeulen, D., Klimesh, M., Ku, D., Kwong-Fu, H., Liggett, E., Li, L., Lundeen, S., Makowski, A., Miller, R., Mouroulis, P., Oaida, B., Okin, G.S., Ortega, A., Oyake, A., Nguyen, H., Pace, T., Painter, T.H., Pempejian, J., Garcia-Pando, C.P., Pham, T., Phillips, B., Pollock, R., Purcell, R., Realmuto, V., Schoolcraft, J., Sen, A., Shin, S., Shaw, L., Soriano, M., Swayze, G., Thingvold, E., Vaid, A., Zan, J., 2020. The Earth surface MIneral dust source invesTigation: an Earth science imaging spectroscopy mission. In: 2020 IEEE Aerospace Conference, pp. 1–15.
Hapke, B., 1981. Bidirectional reflectance spectroscopy. 1. Theory. J. Geophys. Res. 86, 3039–3054.
Jenny, H., 1941. Factors of Soil Formation, a System of Quantitative Pedology. McGraw-Hill, New York.
Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4, 100804.
Kawai, K., Matsui, H., 2025. Contributions of dust source regions to ice nucleating particles in mixed-phase clouds simulated with a global climate-aerosol model. J. Clim., 38, 4925–4939.
Klose, M., Jorba, O., Gonçalves Ageitos, M., Escribano, J., Dawson, M.L., Obiso, V., Di Tomaso, E., Basart, S., Montané Pinto, G., Macchia, F., Ginoux, P., Guerschman, J., Prigent, C., Huang, Y., Kok, J.F., Miller, R.L., Pérez García-Pando, C., 2021. Mineral dust cycle in the Multiscale Online Nonhydrostatic AtmospheRe CHemistry model (MONARCH) Version 2.0. Geosci. Model Dev. 14, 6403–6444.
Kok, J., Albani, S., Mahowald, N., Ward, D., 2014a. An improved dust emission model–Part 2: evaluation in the community earth system model, with implications for the use of dust source functions. Atmos. Chem. Phys. 14, 13043–13061.
Kok, J., Mahowald, N., Fratini, G., Gillies, J., Ishizuka, M., Leys, J., Mikami, M., Park, M.-S., Park, S.-U., Van Pelt, R., 2014b. An improved dust emission model-Part 1: model description and comparison against measurements. Atmos. Chem. Phys. 14, 13023–13041.
Kok, J.F., Ward, D.S., Mahowald, N.M., Evan, A.T., 2018. Global and regional importance of the direct dust-climate feedback. Nat. Commun. 9, 241.
Li, L., Mahowald, N.M., Kok, J.F., Liu, X., Wu, M., Leung, D.M., Hamilton, D.S., Emmons, L.K., Huang, Y., Sexton, N., Meng, J., Wan, J., 2022. Importance of different parameterization changes for the updated dust cycle modeling in the community atmosphere model (version 6.1). Geosci. Model Dev. 15, 8181–8219.
Mahowald, N., Ward, D.S., Kloster, S., Flanner, M.G., Heald, C.L., Heavens, N.G., Hess, P. G., Lamarque, J.-F., Chuang, P.Y., 2011. Aerosol impacts on climate and biogeochemistry. Annu. Rev. Environ. Resour. 36, 45–74.
Mailler, S., Menut, L., Khvorostyanov, D., Valari, M., Couvidat, F., Siour, G., Turquety, S., Briant, R., Tuccella, P., Bessagnet, B., Colette, A., Létinois, L., Markakis, K., Meleux, F., 2017. CHIMERE-2017: from urban to hemispheric chemistry-transport modeling. Geosci. Model Dev. 10, 2397–2423.
Marticorena, B., Bergametti, G., 1995. Modeling the atmospheric dust cycle: 1. Design of a soil-derived dust emission scheme. J. Geophys. Res. 100, 16415–16430.
McBratney, A.B., Mendonca Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Meyer, T., Okin, G.S., Ochoa, F., Brodrick, P.G., 2022. Kalahari Ecosystem Endmember Set. Data set. Available on-line [http://ecosis.org] from the Ecological Spectral Information System (EcoSIS). https://doi.org/10.21232/fEsyLrWo.
Ochoa, F., Brodrick, P.G., Okin, G.S., Chadwick, K.D., Thompson, D.R., Green, R.O., LeGrand, S.L., Ochoa Gonzalez, J.A., Angel, Y., Raiho, A., Braghiere, R., Eckert, R., Schneider, F.D., Kreisberg, A., Thompson, K., Dao, P., Davis, F., 2024. Drylands Spectral Libraries in Support of EMIT. Data Set. Available on-line [http://ecosis.org] from the Ecological Spectral Information System (EcoSIS). https://doi.org/10.21232/QijXnpFt.
Ochoa, F., Brodrick, P.G., Okin, G.S., Ben-Dor, E., Meyer, T., Thompson, D.R., Green, R. O., 2025. Soil and vegetation cover estimation for global imaging spectroscopy using spectral mixture analysis. Remote Sens. Environ. 324, 114746.
Okin, G.S., 2008. A new model for wind erosion in the presence of vegetation. J. Geophys. Res. Earth Surf. 113, F02S10.
Okin, G.S., 2022. Where and how often does rain prevent dust emission? Geophys. Res. Lett. 49, e2021GL095501.
Okin, G.S., Gu, J., 2015. The impact of atmospheric conditions and instrument noise on atmospheric correction and spectral mixture analysis of multispectral imagery. Remote Sens. Environ. 164, 130–141.
Okin, G.S., Painter, T.H., 2004. Effect of grain size on remotely sensed spectral reflectance of sandy desert surfaces. Remote Sens. Environ. 89, 272–280.
Okin, G.S., Okin, W.J., Murray, B., Roberts, D.A., 2001. Practical limits on hyperspectral vegetation discrimination in arid and semiarid environments. Remote Sens. Environ. 77, 212–225.
Okin, G.S., Clarke, K.D., Lewis, M.M., 2013. Comparison of methods for estimation of absolute vegetation and soil fractional cover using MODIS normalized BRDF-adjusted reflectance data. Remote Sens. Environ. 130, 266–279.

Okujeni, A., van der Linden, S., Tits, L., Somers, B., Hostert, P., 2013. Support vector regression and synthetically mixed training data for quantifying urban land cover. Remote Sens. Environ. 137, 184–197.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7, 217–240.

Probst, P., Wright, M.N., Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9, e1301.

Ramezan, C.A., Warner, T.A., Maxwell, A.E., 2019. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. Remote Sens. 11, 185.

Ravi, S., D'Odorico, P., Breshears, D.D., Field, J.P., Goudie, A.S., Huxman, T.E., Li, J., Okin, G.S., Swap, R.J., Thomas, A.D., Van Pelt, R.S., Whicker, J.J., Zobeck, T.M., 2011. Aeolian processes and the biosphere. Rev. Geophys. 49, RG3001.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.d.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital soil map of the world. Science 325, 680–681.

Schnurrenberger, D., Russell, J., Kelts, K., 2003. Classification of lacustrine sediments based on sedimentary components. J. Paleolimnol. 29, 141–154.

Schug, F., Okujeni, A., Hauer, J., Hostert, P., Nielsen, J.Ø., van der Linden, S., 2018. Mapping patterns of urban development in Ouagadougou, Burkina Faso, using machine learning regression modeling with bi-seasonal Landsat time series. Remote Sens. Environ. 210, 217–228.

Schug, F., Frantz, D., Okujeni, A., van der Linden, S., Hostert, P., 2020. Mapping urban-rural gradients of settlements and vegetation at national scale using Sentinel-2 spectral-temporal metrics and regression-based unmixing with synthetic training data. Remote Sens. Environ. 246, 111810.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. 27, 171–197.

Senf, C., Laštovička, J., Okujeni, A., Heurich, M., van der Linden, S., 2020. A generalized regression-based unmixing model for mapping forest cover fractions throughout three decades of Landsat data. Remote Sens. Environ. 240, 111691.

Shen, Q., Shang, K., Xiao, C., Tang, H., Wu, T., Wang, C., 2025. A novel hyperspectral remote sensing estimation model for surface soil texture using AHSI/ZY1-02D satellite image. Int. J. Appl. Earth Obs. Geoinf. 138, 104453.

Shepherd, K., Mantel, S., 2006. A Globally Distributed Soil Spectral Library Visible near Infrared Diffuse Reflectance Spectra. ICRAF (World Agroforestry Centre)/ISRIC (World Soil Information) Spectral Library, Nairobi, Kenya.

Tanre, D., Herman, M., Deschamps, P.Y., 1981. Influence of the background contribution upon space measurements of ground reflectance. Appl. Opt. 20, 3676–3684.

Thompson, D.R., Green, R.O., Bradley, C., Brodrick, P.G., Mahowald, N., Ben Dor, B., Bennett, M., Bernas, M., Carmon, N., Chadwick, K.D., Clark, R.N., Coleman, R.W., Cox, E., Diaz, E., Eastwood, M.L., Eckert, R., Ehlmann, B.L., Ginoux, P., Ageitos, M. G., Grant, K., Guanter, L., Pearlshtien, D.H., Helmlinger, M., Herzog, H., Hoefen, T., Huang, Y., Keebler, A., Kalashnikova, O., Keymeulen, D., Kokaly, R., Klose, M., Li, L., Lundeen, S.R., Meyer, J., Middleton, E., Miller, R.L., Mouroulis, P., Oaida, B., Obiso, V., Ochoa, F., Olson-Duvall, W., Okin, G.S., Painter, T.H., Pérez García-Pando, C., Pollock, R., Realmuto, V., Shaw, L., Sullivan, P., Swayze, G., Thingvold, E., Thorpe, A.K., Vannan, S., Villarreal, C., Ung, C., Wilson, D.W., Zandbergen, S., 2024. On-orbit calibration and performance of the EMIT imaging spectrometer. Remote Sens. Environ. 303, 113986.

Tian, R., Ma, X., Zhao, J., 2021. A revised mineral dust emission scheme in GEOS-Chem: improvements in dust simulations over China. Atmos. Chem. Phys. 21, 4319–4337.

USDA-NRCS, 2007. National Soil Survey Handbook, Title 430-VI. Available at http://soils.usda.gov/technical/handbook.