All-Weather Retrieval of Total Column Water Vapor From Aura OMI Visible Observations

Jiafei Xu and Zhizhao Liu, *Member, IEEE*

***Abstract*—Total column water vapor (TCWV), retrieved from satellite remotely sensed measurements, plays a critically important role in monitoring Earth's weather and climate. The ozone monitoring instrument (OMI) can obtain daily near-global TCWV observations using the visible spectra. The observational accuracy of OMI-estimated TCWV under cloudy-sky conditions is much poorer than OMI-measured clear-sky TCWV. Satellite-based OMI-derived TCWV data, observed with little cloud contamination, are solely used, which, in general, are limited and discontinuous observations. We propose a practical machine learning-based TCWV retrieval algorithm to derive TCWV over land from OMI visible observations under all weather conditions, considering multiple dependable factors linked with OMI TCWV and air mass factor. The global TCWV data, observed from 6000 global navigation satellite system (GNSS)-based training stations in 2017, are utilized as the expected TCWV estimates in the algorithm training process. The retrieval approach is validated in 2018–2020 across the world using ground-based TCWV from additional 4,465 GNSS-based verification stations and 783 radiosonde-based verification stations. The newly retrieved TCWV estimates remarkably outperform operational OMI-retrieved water vapor data, regardless of cloud fraction and TCWV levels. In terms of root-mean-square error, it is overall reduced by 90.44% from 56.38 to 5.39 mm and 90.19% from 53.23 to 5.22 mm compared with GNSS and radiosonde TCWV, respectively. The retrieval algorithm stays stable, both temporally and spatially. This research provides a valuable technique to precisely retrieve OMI-based TCWV data records under all weather conditions, which could be applicable to other satellite-borne visible sensors like GOME-2, SCIAMACHY, and TROPOMI.**

***Index Terms*—Global navigation satellite system (GNSS), machine learning, ozone monitoring instrument (OMI), radiosonde, retrieval, total column water vapor, visible.**

I. Introduction

TOTAL column water vapor (TCWV) is a crucial climatical parameter that is associated with the hydrological cycle, atmospheric circulation, and energy budget [1], [2], [3], [4]. It is a common magnitude to quantify all the atmospheric water vapor content that is enclosed in a vertical column of a cross-section unit, which is also known as precipitable water vapor (PWV) and integrated water vapor (IWV) [5], [6]. Atmospheric water vapor is considered highly varying in the spatial and temporal dimensions [7], [8], highlighting the importance of proper spatiotemporal-resolution TCWV measurements.

TCWV can be measured from ground-based instruments, e.g., radiosonde and global navigation satellite system (GNSS), which are, in general, stationwise observations [9], [10]. Both radiosonde-based and GNSS-based TCWV measurements are utilized to validate other water vapor observations, which are regarded as the truth of TCWV [11], [12], [13]. The ground-based radiosonde only has one or two TCWV data records per day because of the influence of weather conditions [12], while the GNSS can monitor TCWV hourly without the limitation of weather conditions [14]. GNSS-based TCWV observations are also frequently employed as the expected water vapor estimates in algorithm construction and development [15], [16], [17].

In addition, satellite-borne instruments are also utilized to gain TCWV measurements due to the advance of remote sensing [6], [18]. In contrast to ground-based stationwise water vapor observations, satellite-based remotely sensed measurements of TCWV, in general, are at a proper spatiotemporal resolution with a local or even global coverage [19], which are advantageous to be applied in atmospheric water vapor distribution monitoring [20], [21], [22]. Satellite remotely sensed TCWV retrievals can be derived from different spectra regions, such as 432–465 nm visible and 900–940 nm near-infrared. However, satellite-sensed TCWV data have two limitations: 1) the accuracy is low compared to ground-based observation techniques such as radiosonde and GNSS [23], [24], [25], [26]; 2) the accuracy is even degraded when measurements are made under cloudy-sky conditions, particularly for TCWV retrievals from visible or near-infrared channels [5], [11], [12], [27], [28].

The ozone monitoring instrument (OMI) is a spectrometer sensor that operates on the Aura platform launched on 15 July 2004 [29], [30]. It has two ultraviolet channels and one visible channel, with the spectra between 0.27 and 0.50 μm [29]. The OMI sensor is developed based on the previous heritage of the global ozone monitoring experiment (GOME) and the total ozone mapping spectrometer (TOMS) [29]. The OMI/Aura can provide TCWV estimates over land and ocean, which are derived utilizing the visible blue measurements from 432.0 to 465.5 nm [31]. In the operational OMI-based TCWV retrieval approach, the slant column density (SCD) is initially estimated using a non-linear spectral fitting of OMI-sensed radiance observations [31], [32]. Then, the OMI-estimated SCD measurements are converted into the vertical column density (VCD) through air mass factor (AMF) [31], [32], [33].

Received 19 August 2024; revised 9 December 2024; accepted 23 December 2024. Date of publication 26 December 2024; date of current version 10 January 2025. This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project PolyU/RGC 15221620/B-Q80Q, Project 15205821/B-Q84W, and Project 15212622/B-Q94L. *(Corresponding author: Zhizhao Liu.)*

The authors are with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong (e-mail: garfield.xu@connect.polyu.hk; lszzliu@polyu.edu.hk).

Digital Object Identifier 10.1109/JSTARS.2024.3523048

The differential optical absorption spectroscopy (DOAS) approach [34], [35] has also been used to retrieve TCWV from visible observations of multi-satellite sensors, such as the GOME [36], GOME-2 [37], SCanning Imaging Absorption spectroMeter for Atmospheric CartograpHY (SCIAMACHY) [38], and TROPOspheric Monitoring Instrument (TROPOMI) [39]. The primary difference between DOAS and OMI retrieval approaches is that the OMI retrieval is fitted in the intensity space, instead of the optical thickness space [32]. The key step of both DOAS and OMI retrieval models is to determine the conversion factor between SCD and VCD, namely AMF. The calculation of AMF is based on look-up tables derived using a radiative transfer model, which takes cloud fraction, solar zenith angle, sensor zenith angle, relative azimuth angle, surface pressure, cloud top pressure, and surface albedo into consideration [31]. For cloudy-sky conditions, the AMF estimation error can be 15% or more, which can bring notable TCWV retrieval uncertainties in the presence of clouds [31], [39]. It is thus recommended to use OMI-based TCWV data records with little cloud contamination, as clouds can cause significant TCWV errors [31]. Yet this will lead to limited and discontinuous TCWV data records from the OMI sensor. The retrieval performance of OMI/Aura TCWV is required to be further enhanced for obtaining more good-quality water vapor data observations, particularly for TCWV observations with cloud contamination.

Various enhanced retrieval approaches have been proposed to upgrade the quality of satellite-retrieved TCWV measurements [40], [41], [42], [43]. To list some, Preusker et al. [42] proposed a new water vapor retrieval approach to estimate day-time TCWV observations over land from the Sentinel-3 satellites under clear sky conditions. Comparisons between new TCWV with GNSS TCWV had a better agreement, with a decrease in the root-mean-square error (RMSE) from 2.23 to 1.35 mm in North America. Artificial intelligence has improved our understanding of Earth systems, including the atmosphere [44]. A new TCWV retrieval method based on machine learning was proposed to obtain an enhancement in the retrieval performance of TCWV estimates from the moderate resolution imaging spectroradiometer (MODIS), considering several factors associated with MODIS-based water vapor [15], [45]. The retrieval approach reduced the RMSE of operational MODIS-estimated all-sky TCWV data measurements by above 50%, as the cloud-related parameter was utilized in the retrieval model [45]. In addition, Wang et al. [46] developed an improved retrieval algorithm for OMI visible data, which exhibited an overall median of –0.8 mm and a standard deviation of 5.7 mm against GNSS TCWV over land under relatively clear-sky conditions. To our knowledge, little research has been published on the improvement of the retrieval performance of operational satellite-sensed visible TCWV estimates so far, especially for cloudy-sky conditions.

Here, we propose a novel TCWV retrieval algorithm based on light gradient boosting machine (LightGBM) to retrieve TCWV estimates from OMI visible measurements of the Aura spacecraft under all sky conditions, which takes into account both clear- and cloudy-sky conditions. The retrieval approach uses a machine learning model, i.e., LightGBM, to describe the functional relationship between TCWV with OMI-based SCD and several factors associated with OMI TCWV and AMF, which is different from the conventional DOAS and OMI retrieval methods that utilize AMF based on look-up tables and radiative transfer codes [31], [32], [47]. It is expected that this study can bring an indication of improving the retrieval performance of remotely sensed TCWV estimates from satellite-based visible observations.

II. Data and Preprocessing

The OMI instrument of the Aura spacecraft can offer worldwide TCWV measurements based on the visible blue spectra of 432.0 to 465.5 nm, which have a spatial resolution of 13 km x 24 km at nadir [31]. In this study, the recently released OMI TCWV Version 4 data products were used, which are openly available at the Aura Validation Data Center (AVDC) [31].

The operational OMI water vapor product consists of VCD observations, which can be converted into TCWV based on 10^{23} molecules/cm^2 $= 29.89$ mm [31]. The OMI-based SCD measurements, employed in our retrieval algorithm, are recovered based on VCD and AMF using the following equation [31]:

$$\mathrm{SCD} = \mathrm{VCD} \cdot \mathrm{AMF}. \tag{1}$$

In addition, the TCWV measurements, obtained from worldwide GNSS stations from the Nevada Geodetic Laboratory [48], were used in this research, derived from GNSS-sensed observations based on Bevis et al. [14], [49]. The global radiosonde-based observations, collected from the Integrated Radiosonde Archive Version 2 [50], were also employed to generate reference TCWV based on Zhang et al. [51].

The worldwide TCWV observations, obtained from GNSS and radiosonde sites, were utilized in algorithm development and validation, with their distributions presented in Fig. 1. The worldwide TCWV observations at 6000 GNSS training sites in 2017, were used for the training of the retrieval method. In the algorithm assessment process, the global ground-based TCWV datasets during 2018–2020, collected from additional 4465 GNSS testing sites and 783 radiosonde sites, were utilized.

Although no strict criteria were applied in selecting the GNSS stations for algorithm training and validation, the stations chosen are globally distributed as evenly as possible. This strategy aims to reduce potential biases in the training process, ensuring our retrieval method achieves optimal performance. It also allows for independent verification of the TCWV retrievals.

It is critically essential to spatiotemporally match satellite-based OMI measurements with ground-based GNSS/radiosonde observations, aiming to achieve the training and assessment of the retrieval algorithm. In this study, the center of OMI pixels must be closest to the location of GNSS or radiosonde sites, and the distance between OMI with GNSS or radiosonde must not exceed 10 km. The temporal difference between paired OMI–GNSS data measurements must not exceed 30 min. The discrepancy in time between OMI and radiosonde has to be smaller than 1 h.

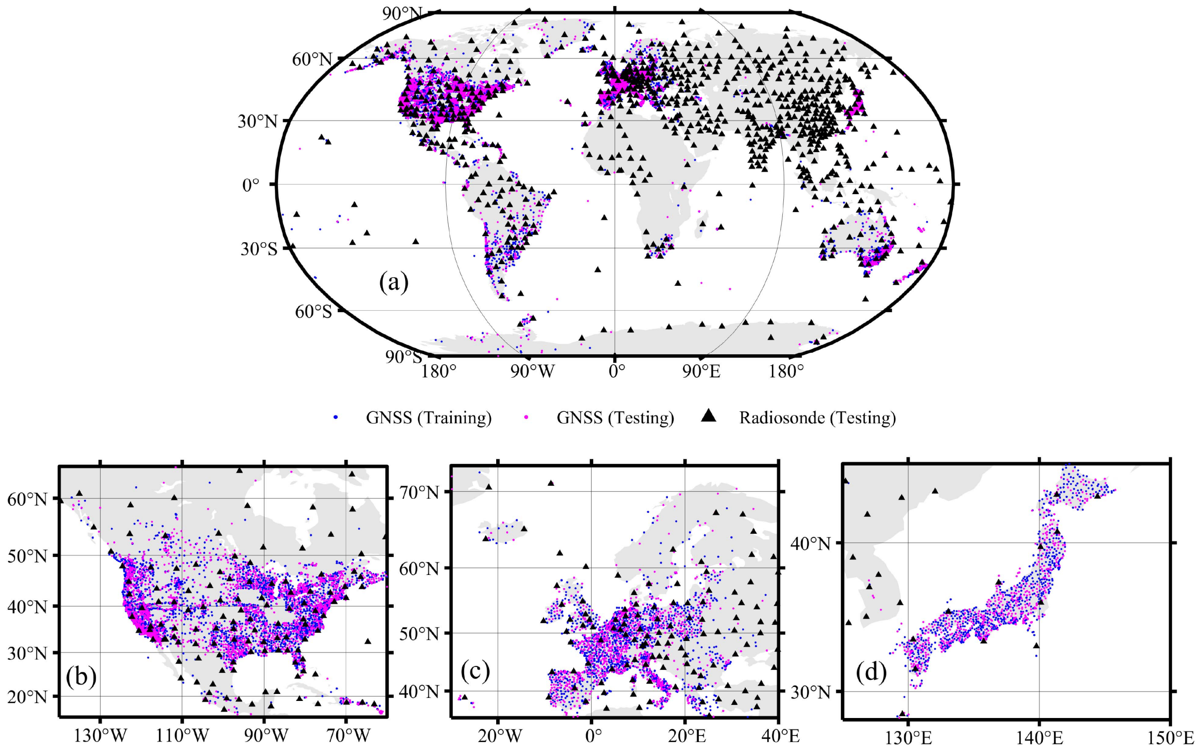

Fig. 1. Worldwide distribution of 6000 GNSS stations used for algorithm training (blue circles), 4465 GNSS stations used for algorithm verification (magenta circles), and 783 radiosonde stations used for algorithm verification (black triangles). (b)–(d) indicate their distributions in North America, Europe, and Japan, respectively.

III. Methodology

A. Theoretical Basis

The observational performance of satellite TCWV data has found in a strong correlation with various factors, including geographical location, TCWV, cloud cover, solar zenith angle, and seasonal variations [10], [26], [31]. As demonstrated in the previous research [15], [16], the all-weather accuracy of satellite-sensed TCWV data can be significantly improved by the joint use of these influence parameters, along with satellite-based measurements.

In addition, AMF plays a critically important role in the estimation of TCWV from satellite-sensed visible measurements, which is influenced by several factors such as solar and viewing angles, cloud properties, surface pressure, and albedo [31], [32].

Given their high correlation with TCWV and AMF, the latitude, longitude, terrain height, month, cloud top pressure, cloud fraction, solar zenith angle, viewing zenith angle, surface pressure, and surface albedo are chose and utilized in the construction of our retrieval approach.

The development of the retrieval algorithm leverages ground-based TCWV data from GNSS observations, which are known for their high reliability and accuracy in all sky conditions [52]. By combining GNSS-derived all-sky TCWV with a multitude of input variables, the retrieval algorithm employs machine learning to capture the atmospheric processes' statistical relationships, enhancing the retrieval of TCWV data under varying sky conditions.

This approach is grounded in the collective insights from previous research [10], [15], [31], ensuring that the retrieval approach effectively models the atmospheric state, thereby improving the accuracy of satellite-derived TCWV estimates.

B. Algorithm Development

1) LightGBM: The LightGBM is a gradient boosting machine learning model, which is an up-to-date implementation of the gradient boosting decision tree (GBDT) [53]. Both GBDT and LightGBM are composed of numerous decision trees that can be employed to solve the classification and regression issues.

It is challenging to achieve our retrieval algorithm using conventional regression methods like linear or exponential functions, as our retrieval method takes many variables into consideration. Instead, a machine learning method, i.e., LightGBM, is used, which is beneficial for addressing the complex multifactor regression problem. Additionally, the LightGBM is an up-to-date implementation method of the GBDT, which has proven a good capability and effectiveness to improve the quality of satellite TCWV estimates [15], [54], [55].

2) All-Sky TCWV Retrievals Using Satellite-Based Visible Data: We develop a practical LightGBM-based TCWV retrieval model to obtain an improvement in the retrieval quality of all-weather TCWV estimates from OMI/Aura visible measurements. The TCWV retrieval from satellite OMI visible measurements is regarded as a regression problem relating TCWV with OMI SCD and several dependence parameters.

TABLE I
PHYSICAL DESCRIPTION OF INPUT ELEMENTS OF THE LIGHTGBM-BASED RETRIEVAL ALGORITHM

Input parameter	Physical description	Reference
SCD	slant column density, directly associated with the determination of atmospheric water vapor	[31] and [32]
LAT	latitude of OMI data, associated with the accuracy of satellite-based water vapor and the spatial variation of water vapor	[10] and [26]
LON	longitude of OMI data, associated with the accuracy of satellite-based water vapor and the spatial variability of water vapor	[10] and [26]
TH	terrain height, associated with the accuracy of satellite-based water vapor	[10] and [12]
MON	month time of OMI data, associated with the accuracy of satellite-based water vapor and the temporal variability of water vapor	[10], [12], and [26]
SZA	solar zenith angle of OMI data, associated with the accuracy of satellite-based water vapor and AMF	[10], [26], and [31]
VZA	viewing zenith angle of OMI data, associated with the accuracy of satellite-based water vapor and AMF	[10], [26], and [31]
CF	cloud fraction of OMI data, associated with the accuracy of satellite-based water vapor, AMF, and sky weather conditions	[12], [26], and [31]
CTP	cloud top pressure of OMI data, associated with the accuracy of satellite-based water vapor and AMF	[31] and [32]
SA	surface albedo of OMI data, associated with the accuracy of satellite-based water vapor and AMF	[31] and [32]
SP	surface pressure of OMI data, associated with the accuracy of satellite-based water vapor and AMF	[31] and [32]

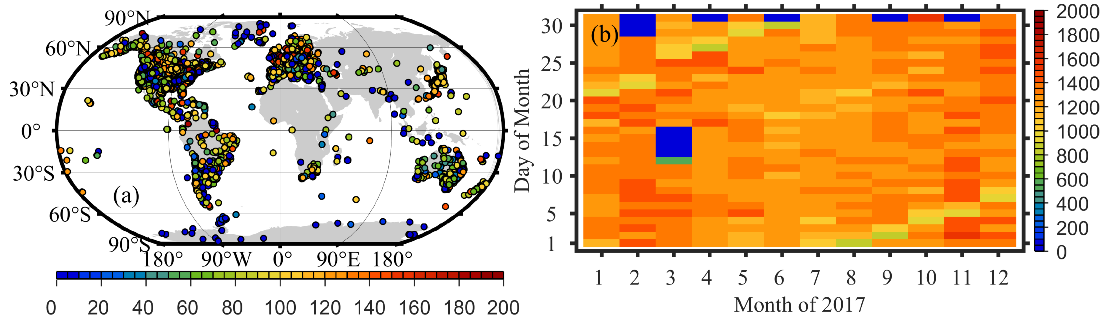

Fig. 2. Spatial-temporal number of collocated OMI and GNSS data in 2017 employed for the training of the newly proposed retrieval algorithm. The color bars denote the number of data points.

The newly retrieved TCWV estimates, determined from OMI visible observations based on LightGBM considering several dependence parameters, can be defined as:

$$\mathrm{TCWV} = \mathrm{LightGBM}(\mathrm{SCD}, \mathrm{LAT}, \mathrm{LON}, \mathrm{TH}, \mathrm{MON}, \\ \times\, \mathrm{SZA}, \mathrm{VZA}, \mathrm{CF}, \mathrm{CTP}, \mathrm{SA}, \mathrm{SP}) \quad (2)$$

where TCWV is the GNSS TCWV in the algorithm training procedure or the new TCWV in the retrieval procedure; SCD is the SCD from OMI visible measurements; LAT, LON, TH, MON, SZA, VZA, CF, CTP, SA, and SP are corresponding to latitude, longitude, terrain height, month, solar zenith angle, sensor zenith angle, cloud fraction, cloud top pressure, surface albedo, and surface pressure, respectively. All factors, used in our retrieval algorithm, can be obtained from the operational AVDC OMI TCWV Version 4 data products. Table I presents the physical description of input elements utilized in our retrieval model.

3) Training of the LightGBM-Based Retrieval Approach: The ground-based TCWV data records, observed from 6000 GNSS stations in 2017, were utilized in the algorithm training process. A total of spatiotemporally paired 461 829 OMI–GNSS data measurements in 2017 were employed, with their spatiotemporal distributions shown in Fig. 2.

Fig. 3 displays the general schematic of training and verification of the newly proposed retrieval model. In this research, the LightGBM-based retrieval model was configured to use GBDT as its boosting type. The hyperparameters were set as follows: the maximum number of leaves per tree ranged from 5 to 300, increasing by 5 at each step; the maximum depth of the trees varied from 1 to 30, with an increment of 1; and the number of trees ranged from 5 to 1000, increasing by 5 at each step. All other parameters of the LightGBM-based retrieval model were set at their default settings. These hyperparameter settings were based on previous studies [56], [57].

The training of the LightGBM was conducted using a 10-fold cross-validation method. About 90% of the data was used for the training of the retrieval method, while 10% of the data was utilized for the testing of the retrieval model. The optimization of the hyperparameters of the LightGBM was conducted based on 2017 training data, evaluated by using the RMSE of the testing data. After the optimization procedure, the maximal tree leaves and depth for base learners were determined as 200 and 10, respectively, with the boosted-tree number of 415.

In the algorithm validation process, we applied the retrieval model to generate new TCWV estimates from worldwide OMI visible observations during the period from 2018 to 2020. The performance of newly derived TCWV data was evaluated using

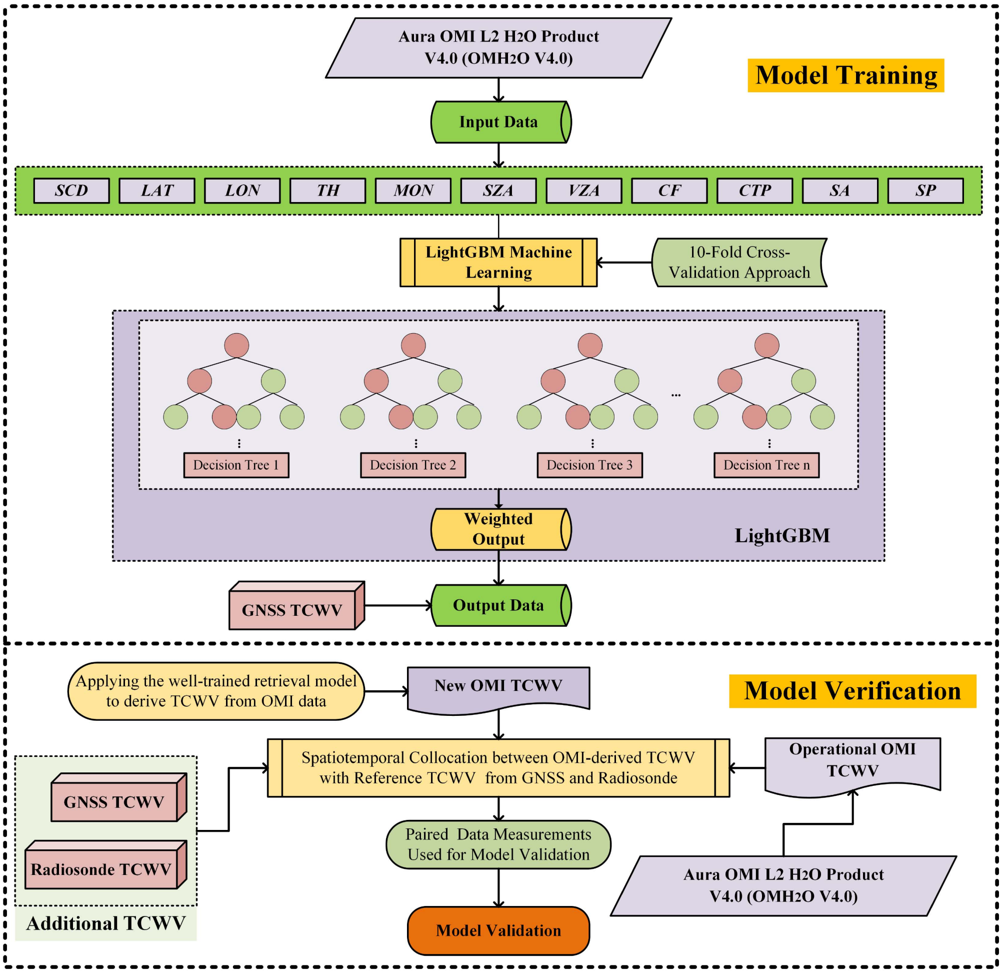

Fig. 3. Overall flowchart of development and verification of the LightGBM-based retrieval algorithm for OMI/Aura visible measurements under all weather conditions.

additional TCWV from ground-based GNSS and radiosonde observations.

The contribution of each input element of the retrieval approach is listed in Fig. 4. The latitude (*LAT*), longitude (*LON*), and solar zenith angle (*SZA*) significantly contribute to the retrieval of OMI TCWV data. In contrast, the terrain height (*TH*), month (*MON*), and surface albedo (*SA*) have the smaller importance to the retrieval of OMI TCWV data, compared to other input parameters. All input parameters, contained in the retrieval method, have contributed to retrieving TCWV estimates from Aura OMI visible observations under all weather conditions. That is, the performance of the retrieval algorithm with a subset of input elements is inferior to that using all input elements. This confirms the effectiveness of our retrieval methodology using these input variables, demonstrating its practicality and reliability.

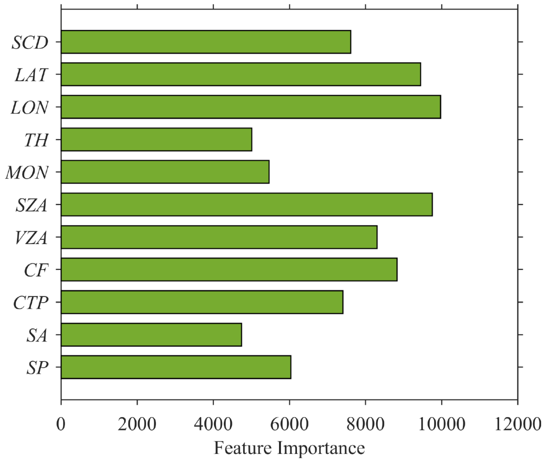

Fig. 4. Importance of input parameters of the LightGBM-based retrieval method for OMI/Aura visible measurements under all weather conditions.

C. Verification of the LightGBM-Based Retrieval Approach

The observational accuracy of newly derived TCWV observations was evaluated using ground-based TCWV measurements from additional 4465 GNSS sites and 783 radiosonde sites during 2018–2020. As presented in Fig. 1, these GNSS-based verification stations differ from the 6000 GNSS-based training stations used in 2017.

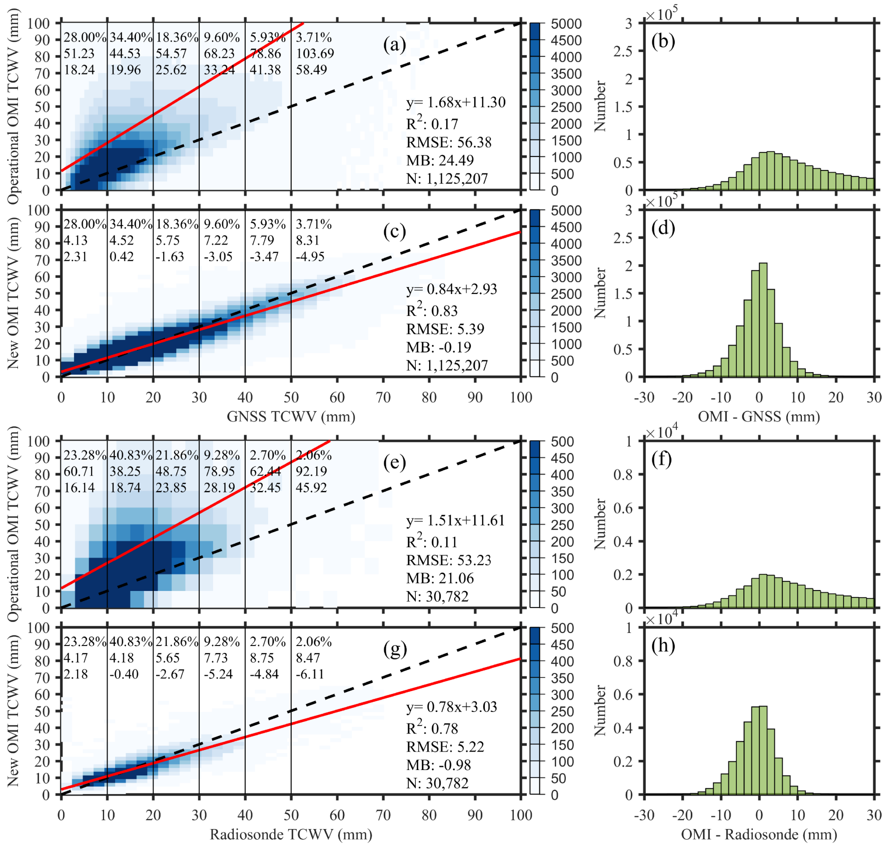

Fig. 5. Worldwide assessment of satellite OMI-based TCWV with ground TCWV from GNSS and radiosonde in 2018 to 2020 under all weather conditions. (a), (c), (e), and (g): scatter plots of OMI TCWV with ground TCWV from GNSS and radiosonde. (b), (d), (f), and (h): histograms of MB of OMI TCWV with ground TCWV from GNSS and radiosonde. The top text, in the left column (a), (c), (e), and (g), indicates the percentage, RMSE, and MB of OMI–GNSS or OMI–radiosonde data measurements for each 10 mm TCWV based on GNSS or radiosonde, respectively. The right-bottom text, in the left column (a), (c), (e), and (g), indicates the general verification metrics for all OMI–GNSS or OMI–radiosonde data measurements. The color bars denote the number of data points.

Three verification factors, that is, R^2, RMSE, and mean bias (MB), were utilized

$$R^2 = \left[\frac{\sum_{i=1}^{N} \left(\mathrm{TCWV}_O - \overline{\mathrm{TCWV}}_O\right)\left(\mathrm{TCWV}_R - \overline{\mathrm{TCWV}}_R\right)}{\sqrt{\sum_{i=1}^{N} \left(\mathrm{TCWV}_O - \overline{\mathrm{TCWV}}_O\right)^2 \left(\mathrm{TCWV}_R - \overline{\mathrm{TCWV}}_R\right)^2}} \right]^2 \quad (3)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\mathrm{TCWV}_\mathrm{O} - \mathrm{TCWV}_R\right)^2} \quad (4)$$

$$MB = \frac{1}{N} \sum_{i\,=\,1}^{N} \left(\mathrm{TCWV}_O - \mathrm{TCWV}_R\right) \quad (5)$$

where TCWV_O is newly derived TCWV estimates from OMI visible observations; $\overline{\mathrm{TCWV}}_O$ is the average of newly derived TCWV estimates from OMI visible observations; TCWV_R is reference TCWV estimates from GNSS and radiosonde observations; $\overline{\mathrm{TCWV}}_R$ is the average of reference TCWV estimates from GNSS and radiosonde observations; N is the number of paired data measurements used in the algorithm verification procedure.

IV. Results

A. Overall Performance

Fig. 5 presents the global 2018–2020 validation result between OMI TCWV versus GNSS (radiosonde) TCWV under all weather conditions. It is found that the new TCWV retrievals, estimated based on our retrieval approach, significantly outperformed operational OMI-derived TCWV estimates, compared to reference TCWV data from GNSS and radiosonde.

The R^2 between OMI and GNSS was enhanced from 0.17 to 0.83, with a decrease in RMSE of 90.44% from 56.38 to 5.39 mm and a decrease in MB from 24.49 to –0.19 mm. Taking radiosonde TCWV as reference, the new TCWV estimates presented an increase in R^2 from 0.11 to 0.78, a decrease in RMSE of 90.19% from 53.23 to 5.22 mm, and a decrease in MB from 21.06 to –0.98 mm compared to operational OMI-retrieved TCWV measurements. For paired OMI–GNSS and OMI–radiosonde TCWV data measurements, the MB distribution of new OMI-derived TCWV observations was more centered between –10 and 10 mm than operational OMI water vapor estimates.

At different TCWV levels, the newly retrieved TCWV data records also performed much better than operational OMI TCWV observations. In particular, the RMSE values of new TCWV were 4.13 to 8.31 mm with GNSS TCWV and 4.17 to

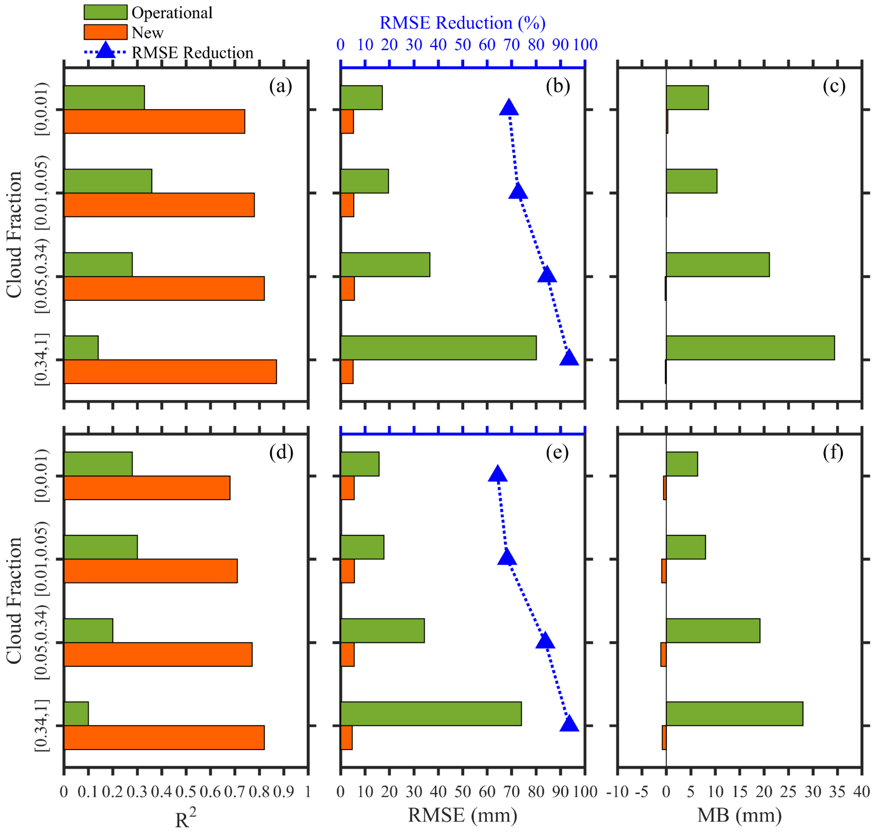

Fig. 6. Worldwide assessment of satellite OMI-based TCWV with ground TCWV from GNSS and radiosonde in 2018 to 2020 under cloud fraction [00.01), [0.01, 0.05), [0.05, 0.34), and [0.34, 1] conditions. (a)–(c) R^2, RMSE, and MB between OMI-based TCWV and GNSS-based TCWV. (d)–(f) R^2, RMSE, and MB between OMI-based TCWV and Radiosonde-based TCWV.

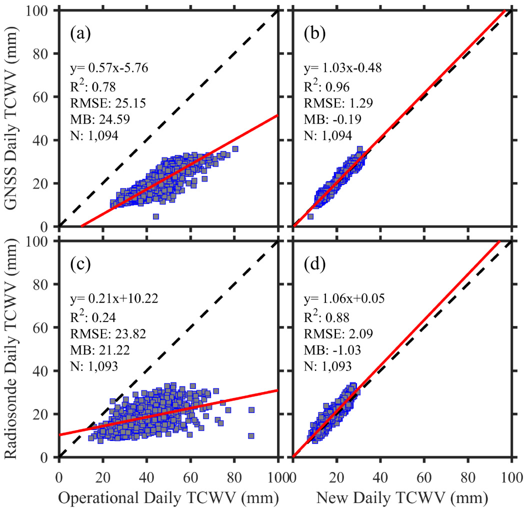

Fig. 7. Daily assessment of satellite OMI-based TCWV with ground TCWV from GNSS and radiosonde in 2018 to 2020 under all weather conditions.

8.75 mm with radiosonde TCWV, which were much smaller than operational TCWV observations (RMSE = 44.53–103.69 mm and 38.25–92.19 mm, respectively). It should be mentioned that both operational and new TCWV measurements exhibited a decreased retrieval performance with the increment of TCWV. This could be due to the amount of TCWV, which had a good consistency with the previous study [26].

B. Cloudy Performance

The collocated OMI–GNSS and OMI–radiosonde data observations were also grouped into different cloud fraction levels based on Platnick et al. [58]. The cloud fraction [00.01), [0.01,0.05), [0.05,0.34), and [0.34,1] conditions indicate that the Aura OMI-based TCWV data observations were retrieved at the cloud fraction of 0 to 0.01 (i.e., confident clear), 0.01 to 0.05 (i.e., probably clear), 0.05 to 0.34 (i.e., probably cloudy), and 0.34 to 1 (i.e., confident cloudy), respectively. When the cloud fraction was between 0 and 1, namely [01], the OMI/Aura TCWV data records are derived under both clear and cloudy sky conditions, i.e., all weather.

It is measured in Fig. 6 that the newly derived TCWV data records had higher correlation coefficient (R^2), lower RMSE, and smaller MB than operational OMI water vapor observations at cloud fraction [00.01), [0.01,0.05), [0.05,0.34), and [0.34,1] levels, when compared with GNSS and radiosonde TCWV measurements. The LightGBM-retrieved TCWV estimates presented R^2, RMSE, and MB of 0.74–0.87, 5.12–5.64 mm, and –0.26–0.25 mm with GNSS-based reference TCWV, significantly outperforming operational OMI TCWV retrievals (correlation coefficient: 0.14–0.37; RMSE: 17.07–80.13 mm, and

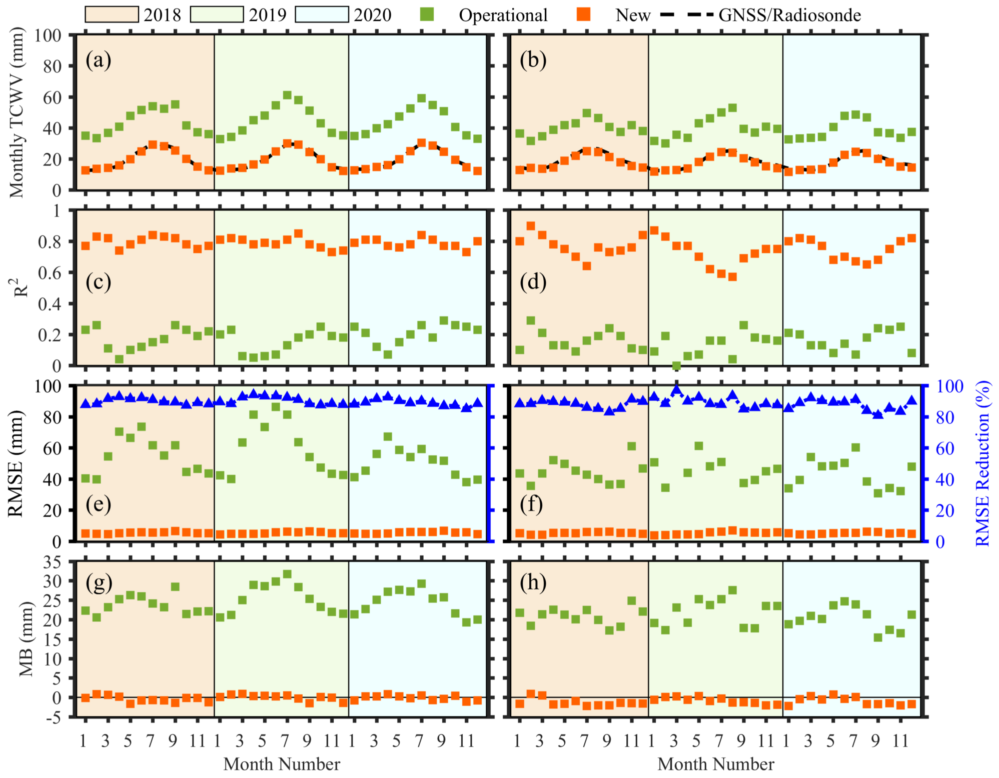

Fig. 8. Monthly assessment of satellite OMI-based TCWV with ground TCWV from GNSS and radiosonde in 2018 to 2020 under all weather conditions. The left column (a), (c), (e), and (g) shows the monthly assessment between OMI-based TCWV and GNSS-based TCWV, while the right column (b), (d), (f), and (h) shows the monthly assessment between OMI-based TCWV and radiosonde-based TCWV. The blue triangles illustrate the diminution in RMSE between new and operational TCWV retrievals.

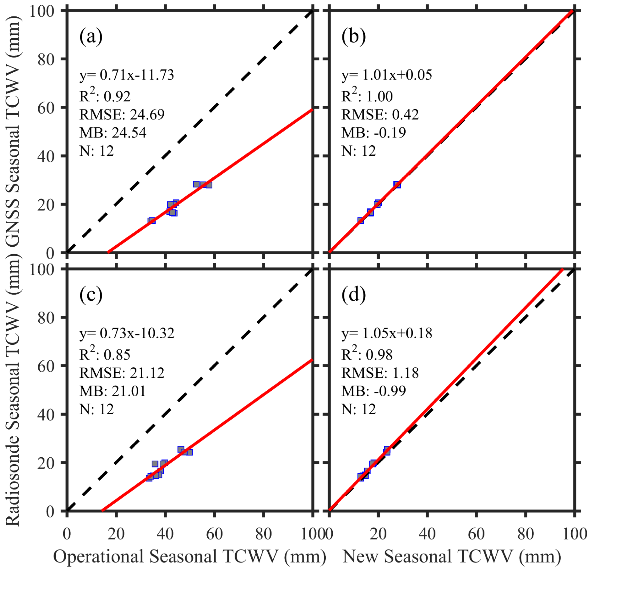

Fig. 9. Seasonal assessment of satellite OMI-based TCWV with ground TCWV from GNSS and radiosonde in 2018 to 2020 under all weather conditions.

MB = 8.60–34.41 mm). Similarly, the R^2 between TCWV from OMI and radiosonde was enhanced to 0.68–0.82, with RMSE and MB improved to 4.68–5.64 mm and –1.13—0.59 mm, after the employment of the LightGBM-based retrieval approach.

Taking GNSS TCWV as reference, the RMSE of operational TCWV retrievals dropped 68.95% from 17.07 to 5.30 mm, 72.75% from 19.63 to 5.35 mm, 84.58% from 36.58 to 5.64 mm, and 93.61% from 80.13 to 5.12 mm under cloud fraction [00.01), [0.01,0.05), [0.05,0.34), and [0.34,1] conditions, respectively. At the same time, the new TCWV estimates presented a minimum decrease in RMSE of 64.35% from 15.71 to 5.60 mm at the cloud fraction [00.01) level and a maximum decrease in RMSE of 93.68% from 74.04 to 4.68 mm at the cloud fraction [0.34,1] level, compared to radiosonde-derived reference TCWV observations.

The RMSE of our TCWV retrievals consistently ranged from 5 to 6 mm across different cloud fraction conditions, demonstrating the robustness of our algorithm. In contrast, the RMSE values for operational OMI TCWV retrievals varied widely from 15 mm (cloud fraction [0, 0.01)) to 80 mm (cloud fraction [0.34, 1]). The operational TCWV data showed significantly larger RMSE values in higher cloud fraction conditions, while our TCWV retrievals maintained relatively stable RMSE values between 5 and 6 mm. As a result, the reduction in RMSE between the operational data and our TCWV data was more pronounced at higher cloud fractions, demonstrating that our model achieves larger RMSE reductions as cloud fractions increase.

It should be mentioned that our retrieval approach showed minimal dependence on cloud fraction. This may be because our method was developed based on machine learning using high-accuracy all-weather TCWV data from GNSS observations. This model is capable of capturing the statistical relationships of atmospheric processes in various sky conditions, improving the retrieval of TCWV data in all sky conditions.

C. *Temporal Performance*

In Fig. 7, the daily averaged TCWV data, calculated using our LightGBM-based retrieval algorithm, performed much better than operational OMI TCWV data, when compared with ground-based GNSS and radiosonde TCWV measurements. In terms of RMSE, it was reduced by 94.87% from 25.15 to 1.29 mm against GNSS TCWV and by 91.23% from 23.82 to 2.09 mm against radiosonde TCWV.

We also listed in Fig. 8 the temporal monthly series assessment of OMI TCWV versus GNSS (radiosonde) TCWV from 2018 to 2020 under cloud fraction [0, 1] conditions. The time-series variation trend of new OMI-based monthly average TCWV had a better consistency with that of monthly average TCWV from GNSS and radiosonde instruments, compared to operational OMI-retrieved monthly mean water vapor measurements.

The monthly correlation coefficient between new TCWV and GNSS TCWV was enhanced to 0.73–0.85, which considerably outperformed operational TCWV estimates ($R^2 = 0.04$–0.29). When compared with radiosonde-measured reference TCWV data, the R^2 of operational TCWV observations was improved from 0.02–0.29 to 0.57–0.90, with the employment of the LightGBM-based retrieval approach.

The newly retrieved TCWV estimates exhibited much smaller monthly RMSE values than operational OMI/Aura TCWV retrievals, when compared to GNSS and radiosonde TCWV measurements. In terms of the diminution in monthly RMSE, the retrieval accuracy of OMI-derived TCWV estimates was improved above 80% in almost all months of 2018–2020, denoting the reliable and effective performance of our retrieval approach in the temporal dimension.

The operational OMI-derived monthly TCWV data, in general, overestimated the monthly mean TCWV values from GNSS and radiosonde measurements, with monthly MB values between 19.26 and 33.15 mm. After the use of the retrieval approach, the monthly MB values between OMI and GNSS (radiosonde) were –2.2–0.88 mm, namely the low magnitude of the underestimation/overestimation of monthly TCWV values.

Additionally, our newly retrieved TCWV data had excellent agreement with ground-based GNSS and radiosonde TCWV data on a seasonal basis, with the results presented in Fig. 9. The RMSE between OMI TCWV and GNSS TCWV decreased by 98.30% from 24.69 to 0.42 mm, after the use of the LightGBM-based retrieval approach. When compared to radiosonde TCWV data, there was a decrease in RMSE of 94.41% from 21.12 to 1.18 mm.

D. *Spatial Performance*

Fig. 10 shows the spatial distribution of TCWV from GNSS, operational OMI water vapor data, and our newly developed retrieval algorithm. Our TCWV data exhibited superior consistency with GNSS-derived reference TCWV data, compared to operational TCWV retrievals. This suggests that our algorithm is robust in retrieving TCWV from OMI measurements, although it was developed using GNSS-sensed stationwise TCWV data.

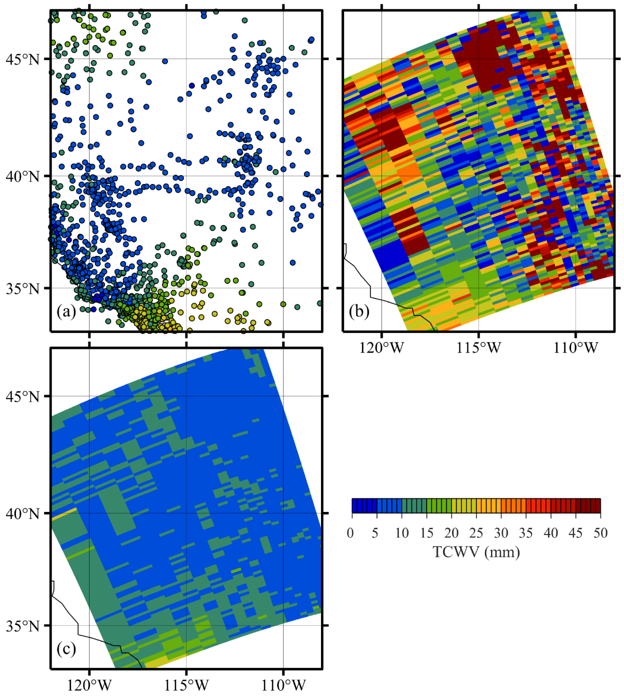

Fig. 10. Spatial distribution of TCWV on 10 May 2020 in the west of the United States, under all weather conditions. (a) GNSS. (b) Operational OMI TCWV data. (c) Our new TCWV retrievals.

The spatial stationwise assessment between OMI TCWV with GNSS (radiosonde) TCWV during 2018–2020 across the globe is displayed in Figs. 11 and 12. The newly retrieved TCWV data records had smaller stationwise RMSE values at almost all GNSS and radiosonde stations, compared with operational Aura OMI-retrieved water vapor estimates. Particularly, the new LightGBM-derived TCWV observations presented RMSE below 10 mm or even smaller than 5 mm at almost all GNSS and radiosonde sites, while stationwise RMSE values of operational TCWV measurements were usually above 20 mm.

In addition, stationwise MB values of newly derived TCWV data in most GNSS and radiosonde stations were found to be –5 to 5 mm, which were superior to operational OMI/Aura water vapor data records that had overall stationwise MB values above 10 mm. In terms of RMSE and MB, the retrieval algorithm, developed in this research, also performed better at almost all GNSS-based and radiosonde-based testing locations, compared with the operational OMI TCWV retrieval approach. This implies that the newly proposed retrieval model is reliable and effective in improving the retrieval performance of operational OMI/Aura TCWV measurements in the spatial dimension.

V. Discussion

A. *Comparison Between This Work and Previous Studies*

In our research, the LightGBM-retrieved all-weather TCWV data records a better consistency with GNSS-measured reference TCWV observations (correlation coefficient: 0.83; RMSE:

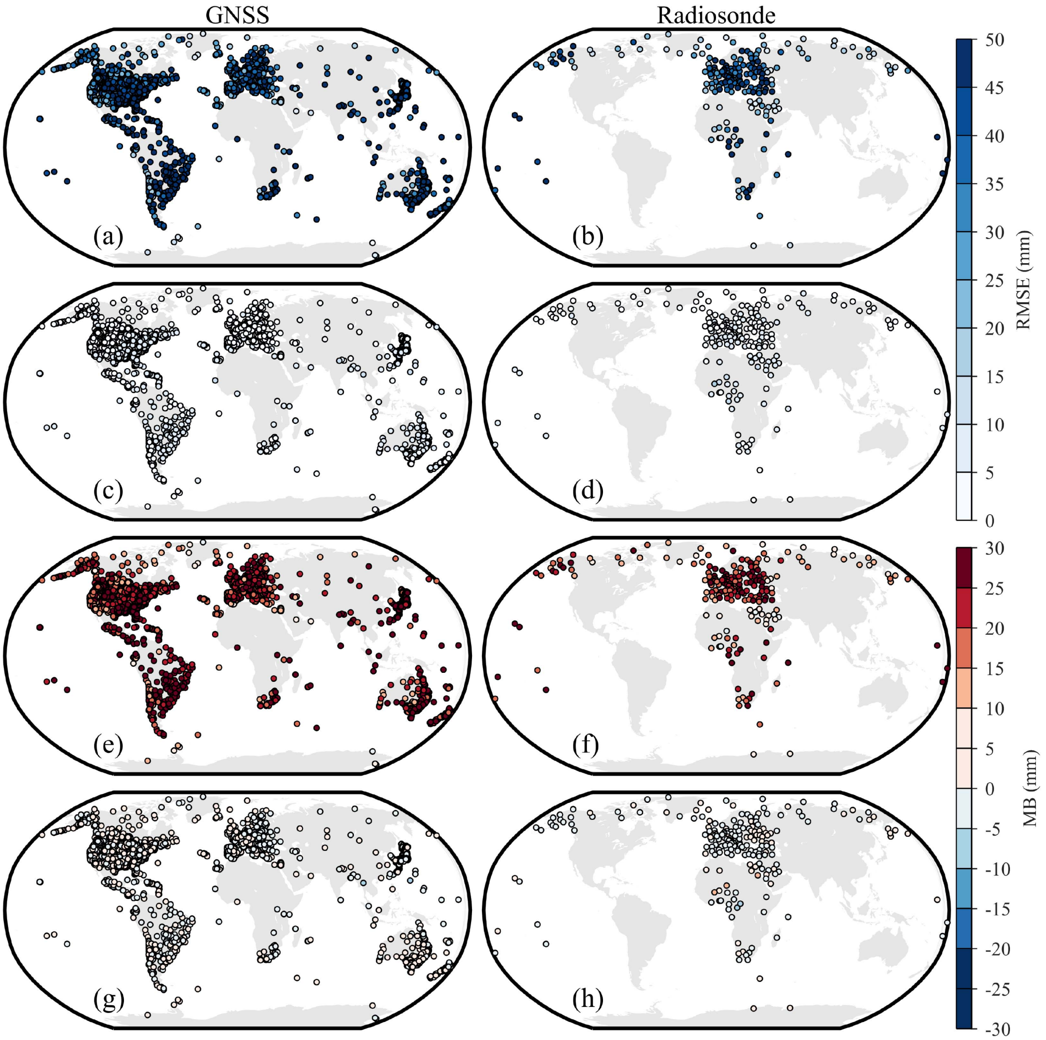

Fig. 11. Spatial assessment of satellite OMI-based TCWV with ground TCWV from GNSS and radiosonde in 2018 to 2020 under all weather conditions. The left column (a), (c), (e), and (g) show the stationwise assessment between OMI-based TCWV and GNSS-based TCWV, while the right column (b), (d), (f), and (h) show the stationwise assessment between OMI-based TCWV and radiosonde-based TCWV. The first, second, third, and fourth rows show the RMSE of operational OMI TCWV estimates, RMSE of new OMI TCWV estimates, MB of operational OMI TCWV estimates, and MB of new OMI TCWV estimates, respectively.

5.07 mm; MB: –0.07 mm), compared with operational OMI-retrieved TCWV data. The newly derived TCWV estimates also agreed better with radiosonde-observed reference TCWV data, with R^2, RMSE, and MB of 0.78, 5.22, and –0.98 mm, respectively.

In terms of RMSE, the observational accuracy of our newly derived all-weather TCWV estimates is comparable to that of common MODIS-derived operational clear-sky TCWV retrievals (RMSE = 4–6 mm) [12], [24], [27], illustrating the capability and effectiveness of the LightGBM-based retrieval model. In addition, the performance of our LightGBM-derived all-weather TCWV data is comparable to that of newly calibrated MODIS-retrieved all-weather TCWV estimates listed in Xu and Liu [54], which were much better than operational MODIS-based all-sky TCWV retrievals with an RMSE above 10 mm [12], [27].

The recent research in Wang et al. [46] showed that the improved land-region OMI TCWV retrievals, observed under clear sky conditions, had an overall MB of −0.7 mm and a standard deviation of 5.7 mm compared with GNSS TCWV data. The observational performance of our newly derived all-weather TCWV estimates is also comparable to that of OMI-derived clear-sky TCWV retrievals in Wang et al. [46]. To our knowledge, this study is the first to refine the retrieval accuracy of TCWV from satellite remotely sensed visible observations under all weather conditions, based on machine learning.

B. Spatiotemporal Stability of the Retrieval Algorithm

The verification datasets in 2018–2020 at GNSS-based and radiosonde-based testing stations, are different from the training datasets in 2017 at GNSS-based training stations, both spatially and temporally. As a result, the performance of the retrieval algorithm is verified independently in the spatial–temporal domain.

In the spatial dimension, the stationwise RMSE and MB discrepancies between TCWV from OMI and GNSS (radiosonde)

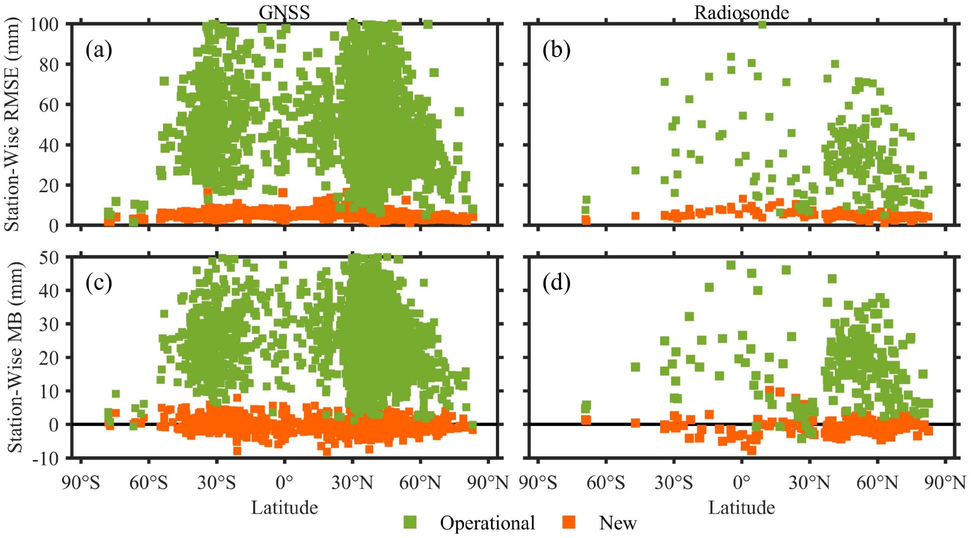

Fig. 12. Latitude-based RMSE and MB distributions of satellite OMI-based TCWV with ground TCWV from GNSS and radiosonde in 2018 to 2020 under all weather conditions. The left column (a) and (c) shows the stationwise RMSE and MB between OMI-based TCWV and GNSS-based TCWV, while the right column (b) and (d) shows the stationwise RMSE and MB between OMI-based TCWV and radiosonde-based TCWV.

were considerably reduced at almost all GNSS and radiosonde locations after the employment of the newly proposed retrieval approach, as the results presented in Figs. 11 and 12. That is, the newly retrieved TCWV data records had smaller stationwise RMSE and MB values than operational Aura OMI-retrieved TCWV estimates. In particular, the new LightGBM-derived TCWV observations presented the RMSE below 10 mm or even smaller than 5 mm in most GNSS and radiosonde stations, while operational OMI TCWV retrievals had an overall stationwise RMSE above 20 mm.

At the same time, the newly retrieved TCWV estimates performed better than operational OMI TCWV data in the temporal dimension, with the results displayed in Fig. 8 and Table II. Among almost all months of 2018–2020, the LightGBM-retrieved TCWV estimates exhibited much lower monthly RMSE and MB values than operational OMI/Aura TCWV retrievals, when compared to GNSS and radiosonde TCWV measurements. Annually, the new TCWV retrievals presented a better consistency with ground-based TCWV from GNSS and radiosonde data (i.e., higher R^2, and lower RMSE and MB), compared with operational TCWV estimates. In particular, the annual RMSE values of operational TCWV retrievals, in general, were reduced by ~90% under cloud fraction [0, 1] conditions.

In terms of RMSE and MB, the retrieval algorithm, proposed in this study, is capable and effective in improving the observational performance of operational OMI TCWV retrievals in the spatial–temporal domain.

C. Limitation and Future Research

Although our newly proposed retrieval method has utilized GNSS-based training stations located on oceanic islands, these ocean-island GNSS stations are technically situated on "land" regions. Therefore, our retrieval method is specifically applicable to land areas, which also encompass areas on oceanic islands. In future work, we will further explore extending our retrieval model to oceanic areas, by training our retrieval method using satellite-sensed microwave TCWV estimates over ocean. Such an extension, utilizing microwave instrument measurements, would be a valuable addition to our current model's capabilities.

Limited GNSS and radiosonde data are available in regions like Africa and Asia for algorithm training and verification. Our future work will focus on refining the retrieval approach's performance in these areas with sparse GNSS and radiosonde data. We will also validate its effectiveness in these regions to ensure robust and reliable results. In addition, we plan to evaluate the performance of the retrieval algorithm under different environmental conditions, even though the retrieval model has already demonstrated a general capability for retrieving TCWV from OMI visible observations across various environmental conditions worldwide. The "row anomaly" of OMI measurements could also have an influence on OMI TCWV retrievals. Therefore, considering the OMI rows has the potential to further improve the retrieval performance of OMI-based TCWV estimates.

VI. Conclusion

We propose a novel LightGBM-based TCWV retrieval algorithm to derive TCWV over land from OMI/Aura visible measurements under all weather conditions, which considers several factors in correlation with AMF and OMI-based TCWV. It is the first time to establish the functional relationship between TCWV with OMI SCD and multiple factors based on a machine learning approach (i.e., LightGBM), which differs from the previous DOAS and OMI retrieval approaches based on look-up tables using a radiative transfer model.

TABLE II
Annual Assessment of Satellite OMI-Based TCWV With Ground TCWV From GNSS and Radiosonde in 2018 to 2020 Under Cloud Fraction [00.01), [0.01, 0.05), [0.05, 0.34), [0.34, 1], and [01] Conditions

Reference Source	Year	Cloud Fraction		Slope	Offset	R^2	RMSE (mm)	MB (mm)	N	Diminution in RMSE
GNSS	2018	[0,0.1)	Operational	0.37	7.29	0.34	15.64	7.87	14 108	66.94%
			New	1.02	−0.31	0.76	5.17	−0.04		
		[0.1,0.05)	Operational	0.37	7.49	0.39	17.66	9.42	46 954	70.33%
			New	1.02	−0.15	0.79	5.24	−0.20		
		[0.05,0.34)	Operational	0.23	10.8	0.30	33.76	19.74	159 764	83.29%
			New	1.02	0.14	0.83	5.64	−0.56		
		[0.34,1]	Operational	0.07	16.69	0.14	80.33	34.21	146 141	93.59%
			New	1.02	0.08	0.87	5.15	−0.44		
		[0,1]	Operational	0.10	15.28	0.16	55.81	23.72	366 967	90.36%
			New	1.02	0.05	0.84	5.38	−0.45		
	2019	[0,0.1)	Operational	0.32	7.84	0.32	17.82	9.30	14 078	70.15%
			New	0.98	−0.12	0.74	5.32	0.48		
		[0.1,0.05)	Operational	0.33	8.09	0.37	20.05	10.90	45 937	73.42%
			New	0.99	−0.04	0.78	5.33	0.16		
		[0.05,0.34)	Operational	0.19	11.20	0.27	37.38	21.74	164 462	85.13%
			New	0.99	0.33	0.82	5.56	−0.07		
		[0.34,1]	Operational	0.05	17.02	0.11	88.60	35.46	149 847	94.33%
			New	0.98	0.46	0.87	5.02	−0.02		
		[0,1]	Operational	0.08	15.50	0.14	61.79	25.43	374 324	91.41%
			New	0.98	0.31	0.84	5.31	0.00		
	2020	[0,0.1)	Operational	0.32	8.07	0.33	17.62	8.62	15 066	69.30%
			New	0.96	0.30	0.74	5.41	0.29		
		[0.1,0.05)	Operational	0.30	8.57	0.35	20.98	10.74	49 395	73.88%
			New	0.98	0.24	0.77	5.48	0.12		
		[0.05,0.34)	Operational	0.19	11.31	0.28	38.27	21.69	172 426	85.08%
			New	0.98	0.49	0.81	5.71	−0.15		
		[0.34,1]	Operational	0.09	15.79	0.18	70.24	33.55	147 029	92.60%
			New	0.99	0.54	0.86	5.20	−0.24		
		[0,1]	Operational	0.12	13.92	0.21	51.15	24.31	383 916	89.29%
			New	0.98	0.44	0.83	5.48	−0.13		
Radiosonde	2018	[0,0.1)	Operational	0.32	10.96	0.26	15.38	6.03	363	60.08%
			New	1.02	0.38	0.64	6.14	−0.81		
		[0.1,0.05)	Operational	0.33	11.22	0.29	15.93	6.88	1 243	64.09%
			New	1.01	1.27	0.72	5.72	−1.48		
		[0.05,0.34)	Operational	0.19	13.54	0.22	32.30	18.03	4 525	82.32%
			New	1.02	1.18	0.80	5.71	−1.64		
		[0.34,1]	Operational	0.09	13.33	0.18	59.89	28.50	4 585	92.37%
			New	1.04	0.29	0.84	4.57	−0.87		
		[0,1]	Operational	0.11	14.87	0.15	44.86	20.81	10 716	88.25%
			New	1.03	0.69	0.81	5.27	−1.27		
	2019	[0,0.1)	Operational	0.30	9.93	0.30	16.16	7.68	370	67.33%
			New	0.94	1.38	0.68	5.28	−0.33		
		[0.1,0.05)	Operational	0.29	10.56	0.31	18.64	8.98	1 167	69.21%
			New	0.96	1.41	0.69	5.74	−0.71		
		[0.05,0.34)	Operational	0.16	12.35	0.22	34.20	19.27	4 407	83.68%
			New	0.97	1.24	0.74	5.58	−0.77		
		[0.34,1]	Operational	0.03	16.29	0.05	97.85	29.67	4 240	95.20%
			New	1.01	0.64	0.81	4.70	−0.86		
		[0,1]	Operational	0.04	16.35	0.07	67.39	22.00	10 184	92.22%
			New	0.99	1.00	0.76	5.24	−0.78		
	2020	[0,0.1)	Operational	0.30	10.93	0.28	15.57	5.29	327	65.77%
			New	1.03	0.17	0.72	5.33	−0.63		
		[0.1,0.05)	Operational	0.29	10.71	0.33	18.63	8.21	1169	70.69%
			New	0.94	1.73	0.72	5.46	−0.66		
		[0.05,0.34)	Operational	0.13	13.48	0.16	36.38	20.12	4422	85.29%
			New	0.97	1.43	0.77	5.35	−0.96		
		[0.34,1]	Operational	0.08	14.21	0.16	57.41	25.45	3964	91.67%
			New	0.99	0.98	0.81	4.78	−0.88		
		[0,1]	Operational	0.10	14.38	0.15	44.31	20.36	9882	88.40%
			New	0.98	1.23	0.78	5.14	−0.88		

The newly retrieved all-weather TCWV data records have $R^2 = 0.83$, RMSE = 5.39 mm, and MB = –0.19 mm and $R^2 = 0.78$, RMSE = 5.22 mm, and MB = –0.98 mm compared to GNSS and radiosonde TCWV, respectively, which considerably outperform operational OMI-retrieved TCWV estimates that show R^2 below 0.2, RMSE above 50 mm, MB above 20 mm. At different TCWV and cloud fraction levels, the OMI-based TCWV estimates, determined using our retrieval model, also perform better with reference TCWV from GNSS and radiosonde, indicating the effective and reliable performance of the retrieval algorithm. The temporal monthly-series and spatial stationwise TCWV discrepancies between paired OMI–GNSS and OMI–radiosonde data observations, in general, are reduced after the employment of the retrieval approach.

Overall, the retrieval algorithm offers an effective and reliable means to derive all-weather TCWV estimates from Aura OMI visible measurements, which performs stably in the spatiotemporal dimension. It can help obtain more good-quality OMI-observed TCWV data records, without the use of filtering criteria that are frequently employed in the previous research. The newly proposed retrieval method has significant potential to be applicable to other similar satellite-borne visible instruments, such as GOME, GOME-2, SCIAMACHY, and TROPOMI.

Acknowledgment

The authors gratefully acknowledge the Aura Validation Data Center, Nevada Geodetic Laboratory, and Integrated Global Radiosonde Archive Version 2 for offering OMI, GNSS, and radiosonde measurements, respectively. The authors would also like to thank the editorial team and the five anonymous reviewers for their invaluable comments and suggestions that help improve the quality of the manuscript.

References

[1] R. D. Cess, "Water vapor feedback in climate models," *Science*, vol. 310, no. 5749, pp. 795–796, Nov. 2005.

[2] I. M. Held and B. J. Soden, "Water vapor feedback and global warming," *Annu. Rev. Energy Environ.*, vol. 25, no. 1, pp. 441–475, Nov. 2000.

[3] S. C. Sherwood, R. Roca, T. M. Weckwerth, and N. G. Andronova, "Tropospheric water vapor, convection, and climate," *Rev. Geophys.*, vol. 48, no. 2, Apr. 2010.

[4] K. E. Trenberth, J. T. Fasullo, and J. Kiehl, "Earth's global energy budget," *Bull. Amer. Meteorol. Soc.*, vol. 90, no. 3, pp. 311–323, Mar. 2009.

[5] J. Vaquero-Martínez et al., "Water vapor satellite products in the European Arctic: An inter-comparison against GNSS data," *Sci. Total Environ.*, vol. 741, Nov. 2020, Art. no. 140335.

[6] Y. J. Kaufman and B.-C. Gao, "Remote sensing of water vapor in the near IR from EOS/MODIS," *IEEE Trans. Geosci. Remote Sens.*, vol. 30, no. 5, pp. 871–884, Sep. 1992.

[7] L. Huang et al.., "A new model for vertical adjustment of precipitable water vapor with consideration of the time-varying lapse rate," *GPS Solut.*, vol. 27, no. 4, Jul. 2023, Art. no. 170.

[8] K. E. Trenberth, J. Fasullo, and L. Smith, "Trends and variability in column-integrated atmospheric water vapor," *Climate Dyn.*, vol. 24, no. 7/8, pp. 741–758, Mar. 2005.

[9] L. Huang et al., "High-precision GNSS PWV retrieval using dense GNSS sites and in-situ meteorological observations for the evaluation of MERRA-2 and ERA5 reanalysis products over China," *Atmospheric Res.*, vol. 276, Oct. 2022, Art. no. 106247.

[10] J. Xu and Z. Liu, "Water vapour products from ERA5, MERSI-II/FY-3D, OLCI/Sentinel-3A, OLCI/Sentinel-3B, MODIS/Aqua and MODIS/Terra in Australia: A comparison against in situ GPS water vapour data," *Q. J. R. Meteorol. Soc.*, vol. 149, no. 753, pp. 1435–1458, Apr. 2023.

[11] M. Antón, D. Loyola, R. Román, and H. Vömel, "Validation of GOME-2/MetOp-A total water vapour column using reference radiosonde data from the GRUAN network," *Atmospheric Meas. Tech.*, vol. 8, no. 3, pp. 1135–1145, Mar. 2015.

[12] J. Xu and Z. Liu, "Evaluation of precipitable water vapor product from MODIS and MERSI-II NIR channels using ground- based GPS measurements over Australia," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 8744–8758, 2022.

[13] B. Radhakrishna, T. N. Rao, and G. S. V. Chandrakanth, "Total column water vapor from INSAT-3D: Assessments with ground-based GNSS receivers and GMI datasets at different temporal scales," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5003408.

[14] M. Bevis, S. Businger, T. A. Herring, C. Rocken, R. A. Anthes, and R. H. Ware, "GPS meteorology: Remote sensing of atmospheric water vapor using the global positioning system," *J. Geophys. Res. Atmospheres*, vol. 97, no. D14, pp. 15787–15801, Oct. 1992.

[15] J. Xu and Z. Liu, "Enhanced all-weather precipitable water vapor retrieval from MODIS near-infrared bands using machine learning," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 114, Nov. 2022, Art. no. 103050.

[16] J. Xu and Z. Liu, "Long-term calibration of satellite-based all-weather precipitable water vapor product from FengYun-3A MERSI near-infrared bands from 2010 to 2017 in China," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4104114.

[17] J. He and Z. Liu, "Applying the new MODIS-based precipitable water vapor retrieval algorithm developed in the North Hemisphere to the South Hemisphere," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4100812.

[18] J. C. Alishouse, S. A. Snyder, J. Vongsathorn, and R. R. Ferraro, "Determination of oceanic total precipitable water from the SSM/I," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 5, pp. 811–816, Sep. 1990.

[19] J. Vaquero-Martínez et al., "Validation of MODIS integrated water vapor product against reference GPS data at the Iberian Peninsula," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 63, pp. 214–221, Dec. 2017.

[20] X. Calbet, C. Carbajal Henken, S. DeSouza-Machado, B. Sun, and T. Reale, "Horizontal small-scale variability of water vapor in the atmosphere: Implications for intercomparison of data from different measuring systems," *Atmospheric Meas. Tech.*, vol. 15, no. 23, pp. 7105–7118, Dec. 2022.

[21] B. Chen and Z. Liu, "Global water vapor variability and trend from the latest 36 year (1979 to 2014) data of ECMWF and NCEP reanalyses, radiosonde, GPS, and microwave satellite," *J. Geophys. Res. Atmospheres*, vol. 121, no. 19, pp. 11442–11462, Oct. 2016.

[22] R. Wang and Y. Liu, "Recent declines in global water vapor from MODIS products: Artifact or real trend?," *Remote Sens. Environ.*, vol. 247, Sep. 2020, Art. no. 111896.

[23] S. Gong, D. Fiifi Hagan, J. Lu, and G. Wang, "Validation on MERSI/FY-3A precipitable water vapor product," *Adv. Space Res.*, vol. 61, no. 1, pp. 413–425, Jan. 2018.

[24] H. Liu, S. Tang, S. Zhang, and J. Hu, "Evaluation of MODIS water vapour products over China using radiosonde data," *Int. J. Remote Sens.*, vol. 36, no. 2, pp. 680–690, Jan. 2015.

[25] Z. Liu, M. S. Wong, J. Nichol, and P. W. Chan, "A multi-sensor study of water vapour from radiosonde, MODIS and AERONET: A case study of Hong Kong," *Int. J. Climatol.*, vol. 33, no. 1, pp. 109–120, Jan. 2013.

[26] J. Vaquero-Martínez et al.., "Inter-comparison of integrated water vapor from satellite instruments using reference GPS data at the Iberian Peninsula," *Remote Sens. Environ.*, vol. 204, pp. 729–740, Jan. 2018.

[27] J. He and Z. Liu, "Comparison of satellite-derived precipitable water vapor through near-infrared remote sensing channels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10252–10262, Aug. 2019.

[28] J. Xu, Z. Liu, G. Hong, and Y. Cao, "A new machine-learning-based calibration scheme for MODIS thermal infrared water vapor product using BPNN, GBDT, GRNN, KNN, MLPNN, RF, and XGBoost," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5001412.

[29] P. F. Levelt et al., "The ozone monitoring instrument," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 5, pp. 1093–1101, May 2006.

[30] P. F. Levelt et al., "The ozone monitoring instrument: Overview of 14 years in space," *Atmospheric Chem. Phys.*, vol. 18, no. 8, pp. 5699–5745, Apr. 2018.

[31] H. Wang, A. H. Souri, G. González Abad, X. Liu, and K. Chance, "Ozone monitoring instrument (OMI) total column water vapor version 4 validation and applications," *Atmospheric Meas. Tech.*, vol. 12, no. 9, pp. 5183–5199, Sep. 2019.

[32] H. Wang, X. Liu, K. Chance, G. G. Abad, and C. C. Miller, "Water vapor retrieval from OMI visible spectra," *Atmospheric Meas. Tech.*, vol. 7, no. 6, pp. 1901–1913, Jun. 2014.

[33] P. I. Palmer et al., "Air mass factor formulation for spectroscopic measurements from satellites: Application to formaldehyde retrievals from the Global Ozone Monitoring Experiment," *J. Geophys. Res. Atmospheres*, vol. 106, no. D13, pp. 14539–14550, Jun. 2001.

[34] U. Platt, "Differential optical absorption spectroscopy, air monitoring by," in *Encyclopedia of Analytical Chemistry*. Wiley, 2006, doi: 10.1002/9780470027318.a0706.

[35] U. Platt and J. Stutz, "Differential absorption spectroscopy," in *Differential Optical Absorption Spectroscopy* (Physics of Earth and Space Environments). Berlin, Germany: Springer, 2008, pp. 135–174.

[36] T. Wagner, J. Heland, M. Zöger, and U. Platt, "A fast H_2O total column density product from GOME – Validation with in-situ aircraft measurements," *Atmospheric Chem. Phys.*, vol. 3, no. 3, pp. 651–663, Jun. 2003.

[37] K. L. Chan, P. Valks, S. Slijkhuis, C. Köhler, and D. Loyola, "Total column water vapor retrieval for global ozone monitoring experience-2 (GOME-2) visible blue observations," *Atmospheric Meas. Tech.*, vol. 13, no. 8, pp. 4169–4193, Aug. 2020.

[38] S. Noël, M. Buchwitz, and J. P. Burrows, "First retrieval of global water vapour column amounts from SCIAMACHY measurements," *Atmospheric Chem. Phys.*, vol. 4, no. 1, pp. 111–125, Jan. 2004.

[39] C. Borger, S. Beirle, S. Dörner, H. Sihler, and T. Wagner, "Total column water vapour retrieval from S-5P/TROPOMI in the visible blue spectral range," *Atmospheric Meas. Tech.*, vol. 13, no. 5, pp. 2751–2783, May 2020.

[40] J. He and Z. Liu, "Water vapor retrieval from MODIS NIR channels using ground-based GPS data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3726–3737, May 2020.

[41] X. Ma, Y. Yao, B. Zhang, Y. Qin, Q. Zhang, and H. Zhu, "An improved MODIS NIR PWV retrieval algorithm based on an artificial neural network considering the land-cover types," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622412.

[42] R. Preusker, C. Carbajal Henken, and J. Fischer, "Retrieval of daytime total column water vapour from OLCI measurements over land surfaces," *Remote Sens.*, vol. 13, no. 5, Jan. 2021, Art. no. 5.

[43] J. Xu and Z. Liu, "Radiance-based retrieval of total water vapor content from sentinel-3A OLCI NIR channels using ground-based GPS measurements," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 104, Dec. 2021, Art. no. 102586.

[44] T. Zhao et al., "Artificial intelligence for geoscience: Progress, challenges, and perspectives," *Innovation*, vol. 5, no. 5, Sep. 2024, Art. no. 100691.

[45] J. Xu and Z. Liu, "A back propagation neural network-based algorithm for retrieving all-weather precipitable water vapor from MODIS NIR measurements," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5633614.

[46] H. Wang et al., "Development of the MEaSUREs blue band water vapor algorithm - towards a long-term data record," *Atmospheric Meas. Tech. Discuss.*, pp. 1–32, Jun. 2023.

[47] H. Wang, G. Gonzalez Abad, X. Liu, and K. Chance, "Validation and update of OMI Total column water vapor product," *Atmospheric Chem. Phys.*, vol. 16, no. 17, pp. 11379–11393, Sep. 2016.

[48] G. Blewitt, W. C. Hammond, and C. Kreemer, "Harnessing the GPS data explosion for interdisciplinary science," *EOS*, vol. 99, no. 10, Sep. 2018, Art. no. 485.

[49] M. Bevis et al., "GPS meteorology - mapping zenith wet delays onto precipitable water," *J. Appl. Meteorol.*, vol. 33, no. 3, pp. 379–386, Mar. 1994.

[50] I. Durre, X. Yin, R. S. Vose, S. Applequist, and J. Arnfield, "Enhancing the data coverage in the integrated global radiosonde archive," *J. Atmospheric Ocean. Technol.*, vol. 35, no. 9, pp. 1753–1770, Sep. 2018.

[51] Y. Zhang, C. Cai, B. Chen, and W. Dai, "Consistency evaluation of precipitable water vapor derived from ERA5, ERA-Interim, GNSS, and radiosondes over China," *Radio Sci.*, vol. 54, no. 7, pp. 561–571, Jul. 2019.

[52] J. Vaquero-Martínez and M. Antón, "Review on the role of GNSS meteorology in monitoring water vapor for atmospheric physics," *Remote Sens.*, vol. 13, no. 12, Jan. 2021, Art. no. 12.

[53] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30*, I. Guyon et al. Eds., 2017, Accessed: May 14, 2022. [Online]. Available: https://www.webofscience.com/wos/woscc/full-record/WOS:000452649403021

[54] J. Xu and Z. Liu, "Improving the accuracy of MODIS near-infrared water vapor product under all weather conditions based on machine learning considering multiple dependence parameters," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4101115.

[55] J. Xu and Z. Liu, "A gradient boosting decision tree based correction model for AIRS Infrared water vapor product," *Geophys. Res. Lett.*, vol. 50, no. 14, Jul. 2023, Art. no. e2023GL104072.

[56] J. Xu and Z. Liu, "STCFCM: A spatial and temporal cloud fraction-based calibration method for satellite-derived near-infrared water vapor product," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4103611.

[57] J. Xu and Z. Liu, "Machine learning-based retrieval of total column water vapor over land using GMI-sensed passive microwave measurements," *GIScience Remote Sens*, vol. 61, no. 1, Dec. 2024, Art. no. 2385180.

[58] S. Platnick et al., "The MODIS cloud products: Algorithms and examples from Terra," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 2, pp. 459–473, Feb. 2003.

Jiafei Xu received the B.Sc. degree in remote sensing science and technology from Shandong Agricultural University, Tai'an, China, in 2015, the M.Sc. degree in geomatics engineering from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2019, and the Ph.D. degree in geomatics from the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, in 2024.

He was awarded the prestigious "PolyU Distinguished Postdoctoral Fellowship Scheme" in 2024. His research interests include the algorithm development for enhancing precipitable water vapor retrievals from multi-satellite visible, near-infrared, infrared, and microwave data observations as well as the algorithm validation for various ground-based, satellite-based, and reanalysis-based precipitable water vapor data products.

Dr. Xu has published several peer-reviewed journal articles on satellite remote sensing water vapor, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *International Journal of Applied Earth Observation and Geoinformation*, *Geophysical Research Letters*, and other esteemed journals.

Zhizhao Liu (Member, IEEE) received the B.Sc. degree in surveying engineering from the Jiangxi University of Science and Technology, Ganzhou, China, in 1994, the M.Sc. degree in geodesy from Wuhan University, Wuhan, China, in 1997, and the Ph.D. degree in geomatics engineering from the University of Calgary, Calgary, AB, Canada, in 2004.

He is currently a Professor with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong. His research interests include new algorithm development for precise Global Positioning System (GPS) and Global Navigation Satellite System (GNSS), GPS/GNSS precise point positioning (PPP), ionosphere modeling and scintillation monitoring, tropospheric remote sensing and modeling, and GPS/GNSS meteorology. He has more than 20 years of experience in GPS/GNSS research. His group has developed a highly efficient and effective algorithm of cycle slip detection and repair for dual- and multifrequency GNSS carrier phase data. The algorithm his group developed can improve the accuracy of water vapor retrieval from remote sensing satellite data by up to 50%. His group developed China's first GPS PPP-based Precipitable Water Vapor Real-time Monitoring System in the Pearl-River-Delta region in 2012. In 2012, his group established Hong Kong's first GPS/GNSS-based ionosphere scintillation monitoring system (two stations deployed in South and North Hong Kong) with his collaborators. His research group established Hong Kong's first GPS/GNSS-radiosonde water vapor sounding collocation system in 2013 in collaboration with Hong Kong Observatory.

Prof. Liu was the recipient of the inaugural Early Career Award of the Hong Kong Research Grants Council (RGC), Hong Kong, in 2012, and the inaugural Best Conference Paper of the China Satellite Navigation Conference (CSNC), China, in 2013. In 2014, he was nominated by the Hong Kong Observatory for the World Meteorological Organization (WMO) "Norbert Gerbier-MUMM International Award for 2015" for his paper that has developed a method to evaluate the absolute accuracy of water vapor measurements.