1   **OMI Total Column Water Vapor Version 4 Validation and Applications**

2   Huiqun Wang[1], Amir Hossein Souri[1], Gonzalo González Abad[1], Xiong Liu[1] and Kelly Chance[1]

3   [1]. Smithsonian Astrophysical Observatory, 60 Garden Street, Cambridge, Massachusetts 02138,

4   USA

6   *Correspondence to: Huiqun (Helen) Wang (hwang@cfa.harvard.edu)*

8   **Abstract**

9   Total Column Water Vapor (TCWV) is important for the weather and climate. TCWV is

10   derived from the Ozone Monitoring Instrument (OMI) visible spectra using the Version 4.0

11   retrieval algorithm developed at the Smithsonian Astrophysical Observatory. The algorithm uses

12   a retrieval window between 432.0 and 466.5 nm and includes updates to reference spectra and

13   water vapor profiles. The retrieval window optimization results from the trade-offs among

14   competing factors.

15   The OMI product is characterized by comparing against commonly used reference datasets –

16   Global Positioning System (GPS) network data over land and Special Sensor Microwave Imager

17   / Sounder (SSMIS) data over the oceans. We examine how cloud fraction and cloud top pressure

18   affect the comparisons. The results lead us to recommend filtering OMI data with cloud fraction

19   less than f = 0.05 - 0.25 and cloud top pressure > 750 mb (or stricter), in addition to the data

20   quality flag, fitting RMS and TCWV range check. Over land, for f = 0.05, the overall mean of

21   (OMI-GPS) is 0.32 mm with a standard deviation (σ) of 5.2 mm, the smallest bias occurs when

22   TCWV = 10 – 20 mm, and the best regression line corresponds to f = 0.25; Over the oceans, for f

23   = 0.05, the overall mean of (OMI-SSMIS) is 0.4 mm (1.1 mm) with σ = 6.5 mm (6.8 mm) for

24   January (July), the smallest bias occurs when TCWV = 20 – 30 mm, and best regression line

25   corresponds to f = 0.15. For both land and the oceans, the difference between OMI and the

26   reference datasets is relatively large when TCWV is less than 10 mm. The bias for Version 4.0

27   OMI TCWV is much smaller than that for Version 3.0.

28   As test applications of the Version 4.0 OMI TCWV over a range of spatial and temporal

29   scales, we find prominent signals of the patterns associated with El Niño and La Niña, the high

30 humidity associated with a corn sweat event and the strong moisture band of an atmospheric

31 river (AR). A data assimilation experiment demonstrates that the OMI data can help improve the

32 Weather Research and Forecasting model (WRF)'s skill at simulating the structure and intensity

33 of the AR and the precipitation at the AR landfall.

1 Introduction

35   Water vapor is of profound importance for weather and climate. Through condensation, it

36 forms clouds that modify albedo, affect radiation and interact with particulate matter. In addition,

37 latent heat released from water vapor condensation can influence atmospheric energy budget and

38 circulation. Water vapor is the most abundant greenhouse gas, accounting for ~50% of the

39 greenhouse effect (Schmidt et al., 2010). Thus, monitoring the spatial and temporal distributions

40 of water vapor is crucial for understanding water-vapor related processes.

41   Water vapor has been measured using a variety of in situ and remote sensing techniques from

42 the ground, air and space. Satellite data provide global perspective and are indispensable for

43 constraining reanalysis products (Dee et al., 2011; Gelaro et al., 2017). The current satellite

44 water vapor datasets are evaluated through the Global Energy and Water cycle Exchanges

45 (GEWEX) Water Vapor Assessment program (Schröder et al., 2019). These datasets are derived

46 from visible, near infrared (NIR), Infrared (IR), microwave and Global Positioning System

47 (GPS) measurements. Each dataset has its own characteristics and contributes to the

48 understanding of water vapor in its own way. For example, microwave data are useful for both

49 clear-sky and cloudy-sky conditions, but are best suited for non-precipitating ice-free oceans due

50 to the complications associated with land surface emissivity; NIR data are best suited for the

51 land, as the surface albedo is low over the oceans; IR data are available over all surface types,

52 but are strongly influenced by clouds and less sensitive to the planetary boundary layer; visible

53 data are sensitive to the boundary layer over both land and the oceans, but are complicated by

54 uncertainties in clouds and aerosols (Wagner et al., 2013).

55   Total Column Water Vapor (TCWV, also called Integrated Water Vapor - IWV, or

56 Precipitable Water Vapor - PWV) can be retrieved from the 7ν water vapor vibrational polyad

57 band (around 442 nm) despite the weak absorption (Wagner et al., 2013). This made it possible

58 to derive TCWV from instruments measuring in the blue wavelength range. Since water vapor is

59 a weak absorber here, saturation of spectral lines is not of concern (Noël et al., 1999). Moreover,

60   the similarity between the land and ocean surface albedo in the blue wavelength range suggests a

61   roughly uniform sensitivity of the measurement over the globe (Wagner et al., 2013). However,

62   weaker absorption tends to result in larger relative uncertainties, especially for low TCWV

63   amount.

64   Using the visible spectra measured by the Ozone Monitoring Instrument (OMI), Wang et al.

65   (2014) retrieved Version 1.0 TCWV from 430 – 480 nm and publically released the data on the

66   Aura Validation Data Center (AVDC, https://avdc.gsfc.nasa.gov). Wang et al. (2016) found that

67   the Version 1.0 data generally agree with ground-based GPS data over land but are significantly

68   lower than the microwave observations over the oceans. They found that using a narrower

69   retrieval window (427.7 – 465 nm) in Version 2.1 could improve the data over the oceans

70   without adversely affecting the results over land much. However, the Version 2.1 data were only

71   generated for a few test months and not released to the public. An interim Version 3.0 OMI

72   TCWV product was available at AVDC. Compared with Version 2.1, Version 3.0 uses the

73   reference spectrum for water vapor from the latest HITRAN database (Gordon et al., 2016) and

74   that for liquid water from Mason et al. (2016), as well as the newest cloud product (Veefkind et

75   al., 2016). The Version 3.0 retrieval window (427.0 – 467.0 nm) is adjusted from that for

76   Version 2 within 2 nm on each end based on fitting uncertainty for a randomly selected test orbit.

77   This paper focuses on Version 4.0 OMI TCWV which has replaced Version 3.0 on AVDC.

78   We present the Version 4.0 retrieval algorithm which incorporates a more vigorous systematic

79   optimization for the retrieval window and miscellaneous updates. We characterize the

80   performance of the Version 4.0 dataset by comparing with well-established references, such as

81   the GPS network data and the Special Sensor Microwave Imager / Sounder (SSMIS)

82   observations. We also assess the performance of Version 4.0 against that of Version 3.0. To

83   provide practical guide to users of the new data, we investigate the influence of cloud fraction

84   and cloud top pressure on the comparisons. Based on the results, data filtering criteria are

85   recommended. As an additional check on the Version 4.0 product, we show test applications of

86   the data to a range of spatial and temporal scales, including El Niño / La Niña, a corn sweat

87   event and an atmospheric river (AR) event. For the first time, a data assimilation experiment for

88   the AR event demonstrates that the OMI TCWV data can provide useful constraint for weather

89   prediction.

2 Retrieval Algorithm

OMI on board the Aura spacecraft is a UV/Visible imaging spectrometer (Levelt et al., 2006). It has been making daily global observations at a nominal 13×24 km nadir resolution from a 1:30 PM equator crossing time polar orbit since October 2004. The UV-Visible channel of OMI covers 350-500 nm at a spectral resolution of about 0.5 nm.

TCWV is derived from the OMI visible spectrum using a commonly used two-step approach. First, the Slant Column Density (SCD, molecules/cm^2) is retrieved from a spectral fitting algorithm. Then, the Vertical Column Density (VCD, molecules/cm^2) is calculated from the ratio of SCD and Air Mass Factor (AMF) (Palmer et al., 2001). VCD can be converted to TCWV using 10^{23} molecules/cm^2 = 29.89 mm. The details of the two-step procedure can be found in González Abad et al. (2015). The specifics of Version 4.0 are discussed below.

The Version 4.0 spectral fitting parameters are summarized in Table 1. In the nonlinear least square fitting, we consider wavelength shift, under-sampling, closure polynomials (3rd order multiplicative and additive), reference spectroscopic spectra of water vapor, interfering molecules (O_3, NO_2, O_4, liquid water, $C_2H_2O_2$ and IO) and Raman scattering (the Ring effect, vibrational Raman scattering of air and the water Ring effect). In comparison with previous versions, Version 4.0 no longer fits common mode (i.e. the mean of the fitting residual, González Abad et al., 2015). It turns out that the common mode for land is different than that for ocean (Wang et al., 2014). Previous retrievals derive a common mode for each orbit swath using the pixels in the low latitudes which often includes both land and ocean scenes. Thus, the derived common mode depends on the proportion of land versus ocean pixels of the spacecraft orbit and is not universally suitable for all the pixels of the swath. Statistics for Orbit 10423 show that although the mean of SCD differs little between the retrievals with and without common mode in the fitting (0.1 mm), the standard deviation of SCD between them can be significant (1.7 mm). Most of the settings in Table 1 are shared between Version 3.0 and 4.0, except that Version 3.0 uses HITRAN 2016 (Gordon et al., 2016) as the water vapor reference spectrum, includes common mode in the fitting, but does not consider vibrational Raman scattering of air (Lampel et al., 2015a). We revert to the HITRAN 2008 water vapor spectrum (Rothman et al., 2009) in Version 4.0 because validation results show that it leads to better agreements with the GPS and SSMIS TCWV data (Section 3). We did not apply the correction of Lampel et al. (2015b) to the

120   HITRAN 2008 water vapor spectrum. It is recently found that HITRAN 2016 is adversely

121   affected by an issue with line broadening for water vapor in the blue wavelength range and

122   improvements are being made for the next HITRAN release (the HITRAN group, personal

123   communication).

124   **Table 1.** Parameters used in Version 4.0 spectral fitting for OMI total column water vapor.

Wavelength shift	Solar reference spectrum	Dobber et al. (2008)
Target	H_2O	288K, Rothman et al. (2009)
Interference molecules	O_3	228K, Brion et al. (1993)
	NO_2	220K, Vandaele et al. (1998)
	O_4	293K, Thalman and Volkamer (2013)
	Liquid water	Mason et al. (2016)
	$C_2H_2O_2$	296K, Volkamer et al. (2005)
	IO	298K, Spietz et al. (2005)
Raman scattering	Ring effect	Chance and Spurr (1997)
	Water Ring	Chance and Spurr (1997)
	Air Vibrational Raman	Lampel et al. (2015a)
Other	Additive polynomial	3rd order
	Multiplicative polynomial	3rd order
	Under-sampling	Chance et al. (2005)

125

126   To optimize the retrieval window, we randomly selected OMI Orbit number 10426 (on July

127   1, 2006) to examine the effect of varying the starting and ending wavelengths around the 7ν

128   water vapor absorption band. The orbit swath contains 60×1644 ground pixels and covers parts

129   of Australia, the Pacific, China and other areas. We systematically adjust the starting wavelength

130   within 426.0-435.0 nm and the ending wavelength within 460.0-468.5 nm, both at 0.5 nm steps.

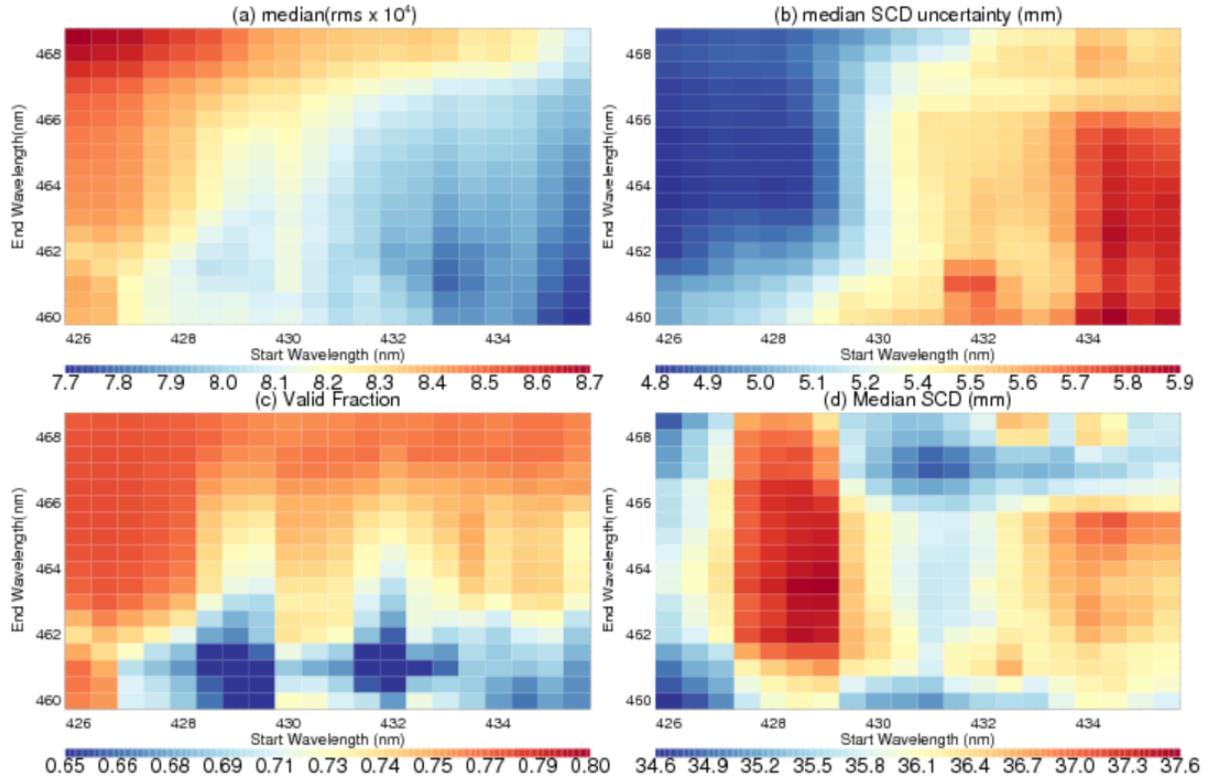



Figure 1. Sensitivity of the retrieval to the start and end wavelengths (nm) of the retrieval window for OMI Orbit number 10426. (a) Median of fitting RMS×10^4; (b) median of water vapor SCD fitting uncertainty in mm; (c) valid fraction for retrievals; (d) median SCD in mm.

In previous versions, the fitting window is selected based on the fitting uncertainty (Wang et al., 2014, 2016). For Version 4.0, we consider the following four factors. (1) Figure 1a shows that the median of the fitting Root Mean Squared error (RMS) is smaller toward the lower right corner of the domain (i.e., longer start wavelength and shorter end wavelength); (2) Figure 1b shows that the medium fitting uncertainty of water vapor SCD decreases toward the upper left corner; (3) Figure 1c shows that the fraction of valid retrievals for the orbit generally increases toward the upper part of the domain. Valid retrievals here refer to those that pass the main data quality check (MDQFL = 0) and have positive SCDs. The main data quality check ensures that the fitting has converged, the SCD is $< 5 \times 10^{23}$ molecules/cm^2 (149.45 mm) and within 2σ of the fitting uncertainty. The SCD threshold here is meant to filter out large outliers. For reference, the largest TCWV of the GPS and SSMIS datasets used in Section 3 is about 75 mm. At low latitudes where TCWV is large, more than 90% of the OMI AMFs are between 0.5 and 2.0; (4) The length of the retrieval window increases with the difference between the end and start

148   wavelengths. The general patterns exhibited by Orbit number 10426 in Figure 1 also hold for

149   Orbit number 10423 which cuts across the Pacific near the dateline.

150   Ideally, we would like to have small fitting RMS to reduce the residual's amplitude and

151   structure, a small fitting uncertainty to reduce error, a large fraction of valid data to increase data

152   volume and a long retrieval window to include more information into the fitting. However, these

153   criteria cannot be met simultaneously. As a compromise, we select the wavelength interval

154   between 432.0 nm and 466.5 nm as the retrieval window for Version 4.0. For Orbit number

155   10426, this leads to a median fitting RMS of 8.1×10^{-4}, a median SCD uncertainty of 5.4 mm, a

156   valid fraction of 0.75 and a window length of 34.5 nm (Figure 1). Figure 1d shows that the

157   median SCD for Orbit number 10426 varies between 34.6 mm and 37.6 mm. This 3 mm

158   difference corresponds to an 8% variation and exhibits a complex pattern within the domain. The

159   Version 4.0 retrieval window leads to a median SCD of 35.5 mm for Orbit number 10426 which

160   is near the beginning of the middle third of the SCD range. The ratio between the median SCD

161   uncertainty and the median SCD (i.e., the relative SCD uncertainty) is about 0.15. Note that this

162   value is for the whole orbit which includes a wide range of SCDs. As shown in Supplementary

163   Figure 1, the relative SCD uncertainty is >1.2 for SCD = 0 – 10 mm, drops to about 0.4 for SCD

164   = 10 – 20 mm, and to about 0.1 for SCD > 40 mm.

165   The AMF is calculated by convolving scattering weights with the shape of water vapor

166   vertical profile (González Abad et al., 2015). The scattering weight is interpolated from the same

167   look-up table as that used in Wang et al. (2016). The scene specific information used in the AMF

168   calculation is listed in Table 2. By propagating typical errors for surface albedo (15%), cloud

169   fraction (10%) and cloud top pressure (15%), we find that the AMF error due to scattering

170   weight for a typical orbit (number 10426) is mostly < 3%, though for cloudy pixels, the error can

171   be 15% or more. Version 4.0 uses the 0.5°×0.667° monthly mean MERRA-2 water vapor profile

172   (Gelaro et al., 2017) for the month and year corresponding to the retrieval, while previous

173   versions used the monthly mean of 2007 for all years. To evaluate the error associated with gas

174   profiles, we compare the TCWV calculated using the daily MERRA-2 profile against that

175   calculated using the monthly MERRA-2 profile for July 2006 (for TCWV within the 0 – 75 mm

176   range). Results show that (TCWV(daily) – TCWV(monthly)) has a mean (median) of 0.3 mm (0

177   mm) with a standard deviation of 5.0 mm. When comparing the TCWV calculated using the

178   daily MERRA-2 profile against that calculated using the daily ERA-Interim profile for July

179 2006, we find that (TCWV(MERRA-2) – TCWV(ERA-Interim)) has a mean (median) of -0.1

180 mm (0 mm) with a standard deviation of 2.8 mm. Thus, gas profiles can introduce substantial

181 scatter to the retrieved TCWV. AMF is highly sensitive to clouds (Wang et al., 2014; Vasilkov et

182 al., 2017). Version 4.0 uses the cloud information from Veefkind et al. (2016). The primary

183 difference with the Acarreta et al. (2004) cloud product used in Version 1.0 and 2.1 is in the

184 cloud top pressure for cloud fraction f < 0.3. In addition to the factors in Table 2, aerosol and

185 surface bi-directional reflectance distribution function (BRDF) influence the AMF (Lorente et

186 al., 2017; Vasilkov et al., 2017), but have not been considered in the retrieval yet.

187 **Table 2.** Parameters used in AMF calculation

Solar Zenith Angle	OMI L1B data
View Zenith Angle	
Relative Azimuth Angle	
Surface Albedo	OMLER (Lambert equivalent reflectance) Kleipool, et al. (2008)
Cloud fraction	OMCLDO2 (derived from O_2-O_2) Veefkind et al. (2016)
Cloud top pressure	
Surface pressure	MERRA-2 monthly data ($0.5° \times 0.667°$), Gelaro et al. (2017)
Water vapor profile	

188

189 **3 Validation**

190 To validate the Version 4.0 OMI TCWV data, we compare them against two commonly used

191 reference datasets – a GPS network dataset for land and a microwave dataset for the oceans.

192 **3.1 OMI and GPS over land**

193 To assess the Version 4.0 OMI TCWV over land, we compare against the GPS network data

194 downloaded from NCAR (rda.ucar.edu/datasets/ds721.1). The GPS data are composed of 2-

195 hourly TCWV at International GNSS Service (IGS), SuomiNet and GEONET stations, and have

196 an estimated error of < 1.5 mm (Wang et al., 2007; Ning et al., 2016). The subset of IGS-

197 SuomiNet data for the whole year of 2006 is used in this paper. The geographical distribution of

198 the stations can be found in Wang et al. (2016). Most of the stations are concentrated in North

199 America and Europe, fewer are scattered on other continents.

200 OMI TCWV data are filtered using the following criteria. The stripes in Level 2 swaths due

201 to systematic instrument error are removed using the SCD scaling procedure described in Wang

202 et al. (2016). The pixels affected by OMI's row anomaly are filtered out

203 (projects.knmi.nl/omi/research/product/rowanomaly-background.php), as well as negative or

204 extremely large (i.e., TCWV > 75 mm) values. For the clear-sky comparison in Figure 3, we

205 require cloud fraction < 5% and cloud top pressure > 750 mb, in addition to MDQFL = 0 and

206 fitting RMS < 0.001. The cloud fraction and cloud top pressure are from the OMCLDO2 cloud

207 product (Veefkind et al., 2016) and are included in the Level 2 OMI product for ease of data

208 filtering. On a typical day (July 1, 2006), among the OMI data that pass the MDQFL and TCWV

209 range test, cloud fraction < 0.05 accounts for 35% of the data, cloud top pressure > 750 mb

210 accounts for 53% of the data and RMS < 0.001 accounts for 72% of the data.

211   To co-locate GPS and OMI data, we select the GPS data observed between 1200 LT and

212 1500LT. This 3-hour local time range covers the OMI overpass time. We average the qualified

213 OMI data within 0.25° longitude × 0.25° latitude of the GPS stations for each day. To minimize

214 the influence of local topography (e.g., mountain peaks, river valleys), if a station's elevation is

215 more than 250 m different than the mean elevation within the corresponding 0.25°×0.25° grid

216 square, then it is excluded from the analysis. The 0.25°×0.25° topography was downloaded from

217 www.temis.nl/data/topo/dem2grid.html. The comparison between OMI and GPS is made for

218 TCWV within the range of 0 – 75 mm as the largest TCWV for the GPS data is about 75 mm.

219 The co-locating procedure leads to about 11,000 co-located data points for the entire year of

220 2006.

221   Figure 2 shows the comparison between the resulting co-located GPS and Version 4 OMI

222 TCWV. The top panel shows the histogram of OMI-GPS (in 0.5 mm bins). The bin from -0.5 to

223 0.0 mm corresponds to the peak of the distribution. The overall mean (median) of OMI-GPS is

224 0.32 mm (0.35 mm), with a standard deviation of 5.2 mm. The mean (median) absolute error is

225 3.9 mm (3.0 mm).

226   The bottom panel of Figure 2 shows the joint distribution of the co-located GPS and Version

227 4.0 OMI data. The count for each 0.5 mm bin is normalized by the maximum of all bins. About

228 34% of the data have TCWV < 10 mm, 72% have TCWV < 20 mm and 90% have TCWV < 30

229 mm. There is a general linear correlation between GPS and OMI data, with a correlation

230 coefficient of r = 0.87 (R^2 = 0.76). The linear regression line (OMI = 2.22 + 0.88 * GPS, where

231 OMI and GPS TCWV are in mm) has a significant positive intercept and a slope that is less than

232   one. This indicates a positive bias of OMI against GPS for small TCWV and a negative bias for

233   large TCWV. Indeed, as indicated at the top of the panel, the mean of OMI-GPS for each 10 mm

234   GPS TCWV bin decreases from 1.7 mm for TCWV = 0 – 10 mm to -2.3 mm for TCWV = 40 –

235   50 mm, though the fraction of data for TCWV > 40 mm is < 3%. The corresponding standard

236   deviation (σ) increases from 3.5 mm to 7.9 mm. The minimum bias of 0.2 mm occurs for TCWV

237   in the 10 – 20 mm bin. The large positive bias of the 0 – 10 mm bin (as compared with the

238   TCWV of the bin) has significant adverse effect on the regression line. For TCWV > 10 mm, the

239   regression line (OMI = 1.51 + 0.91 $\times$ GPS) is better.

240   In comparison, although Version 3.0 OMI is similarly correlated with GPS (correlation

241   coefficient r = 0.86), it has a much larger positive bias of 2.8 mm (with a standard deviation of

242   5.5 mm). The large bias is attributed to the much larger SCD of Version 3.0 (Supplementary

243   Figure 2b), as the AMFs of both versions roughly follow the 1:1 line (Supplementary Figure 2a).

244   Sensitivity tests show that the larger Version 3.0 SCD is primary due to the water vapor

245   reference spectrum. If the water vapor reference spectrum in Version 4.0 is replaced with that of

246   Version 3.0 (Test 1), then the median SCD increases by about 4.5 mm for Orbit 10423

247   (Supplementary Figure 2c). Modifying the retrieval window for Version 3.0 cannot sufficiently

248   reduce the retrieved SCD, therefore cannot make significantly better agreement with the

249   reference TCWV data. As Version 4.0 shows better performance, this paper focuses on

250   characterizing Version 4.0 to provide useful information to potential users. In subsequent

251   discussions, OMI data refer to Version 4.0 unless specified otherwise.

252

253

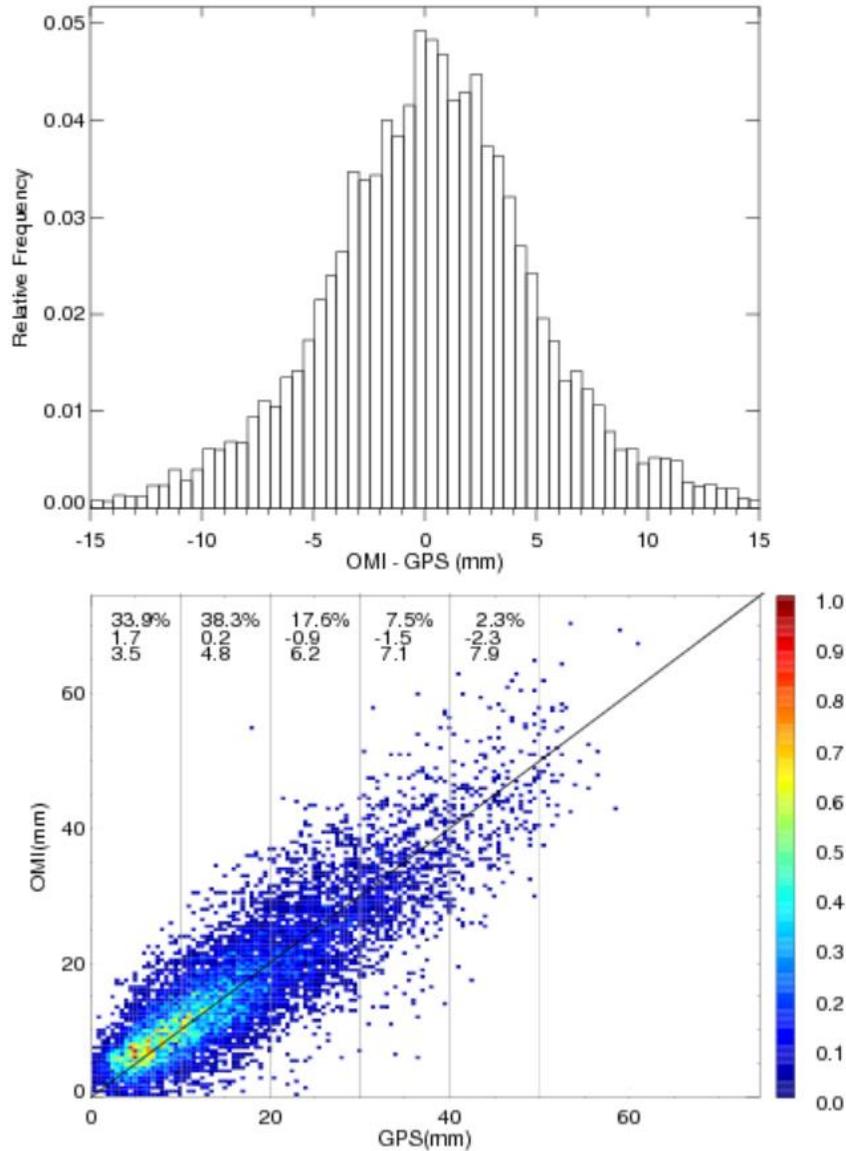



Figure 2. Comparison between co-located GPS and OMI TCWV (mm) for all days in 2006. The
data filtering criteria include cloud fraction < 5%, cloud top pressure > 750 mb, and others
discussed in the text. (Top) Relative frequency of occurrence for OMI-GPS (mm). (Bottom)
Normalized joint distribution of GPS versus OMI TCWV (mm). The three lines of text from top
to bottom indicate the percentage of data points (1st), the mean of OMI-GPS in mm (2nd), and
the standard deviation of OMI-GPS in mm (3rd) for each 10 mm GPS TCWV, respectively. The
1:1 is plotted for reference.

262   OMI TCWV retrieval is highly sensitive to clouds (Wang et al., 2014). In Figure 3, we

263   examine the effect of OMI cloud fraction threshold (f) on the comparison while keeping other

264   data filtering criteria the same as those for Figure 2 (i.e., cloud fraction < f, cloud top pressure <

265   750 mb, MDQFL = 0, fitting RMS < 0.001 and 0 < TCWV<75 mm). From f = 0.05 to f = 0.55,

266   the number of co-located data pairs (N) more than triples, the mean of OMI-GPS increases from

267   0.32 mm to 1.66 mm, the standard deviation of OMI-GPS increases from 5.2 mm to 6.1 mm. The

268   linear correlation coefficient (r) increases from r = 0.87 at f = 0.05 to r ~ 0.90 at f = 0.15, then

269   levels off for larger cloud fraction thresholds. It should be noted that the error in cloud top

270   pressure decreases with cloud fraction in the OMCLDO2 product (Veefkind et al., 2016). As a

271   result, f = 0.05 corresponds to the largest uncertainty in cloud top pressure, and the error will

272   propagate into OMI TCWV through AMF, leading to smaller correlation coefficient than those

273   for larger f values.

274   In addition, as shown by the GPS versus OMI joint distributions for different cloud fraction

275   thresholds in Figure 4, the f ≥ 0.15 cases have larger effective dynamical ranges which tend to

276   favor better correlations. For example, there is a larger fraction of data pairs with TCWV > 30

277   mm for f = 0.15 than for f = 0.05. The regression line for f = 0.15 (OMI = 1.26 + 0.96 * GPS)

278   shows an apparent improvement over that for f = 0.05 (OMI = 2.22 + 0.88*GPS). The best

279   regression line is arguably that for f = 0.25 (OMI = 1.16 + 0.99*GPS) or f = 0.35 (OMI = 1.19 +

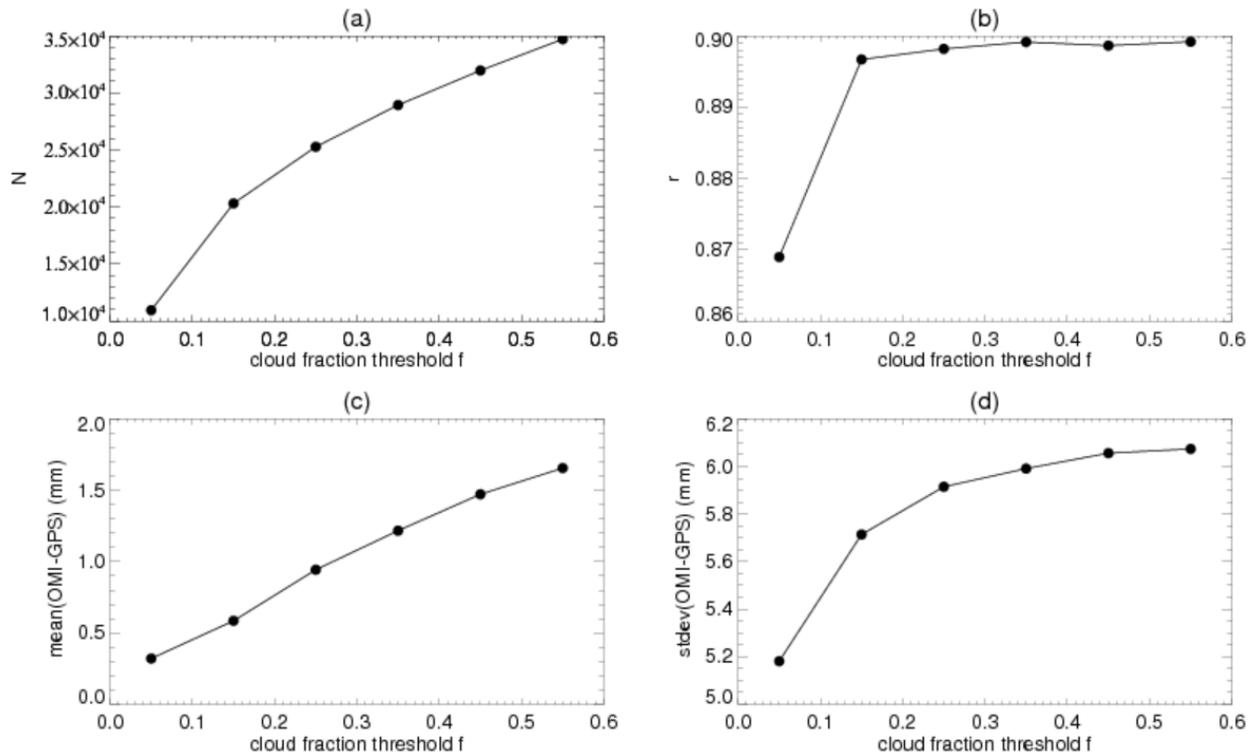
280   1.00*GPS), though the mean bias and scatter are larger than those for f < 0.25 (Figure 4).

281   In brief, f = 0.05 leads to the lowest overall bias and scatter of the co-located data; f = 0.15

282   doubles the number of co-located data pairs and leads to the largest improvement in the

283   correlation coefficient; f = 0.25 (or 0.35) leads to the best linear regression line; the bias and

284   standard deviation increase with cloud fraction threshold. Hence, cloud fraction thresholds in the

285   range of f = 0.05 – 0.25 seems reasonable for filtering OMI TCWV, depending on applications.

286

287   **Figure 3.** Dependence of various parameters on the cloud fraction threshold (f) used for filtering

288   OMI data. Other filtering criteria remain the same as those for Figure 2. The parameters are (a)

289   number of co-located OMI and GPS data pairs; (b) linear correlation coefficient between OMI

290   and GPS TCWV; (c) mean of OMI-GPS in mm; (d) standard deviation of OMI-GPS in mm.

291   Results are derived from the co-located Version 4.0 OMI and GPS data for the whole year of

292   2006.

293

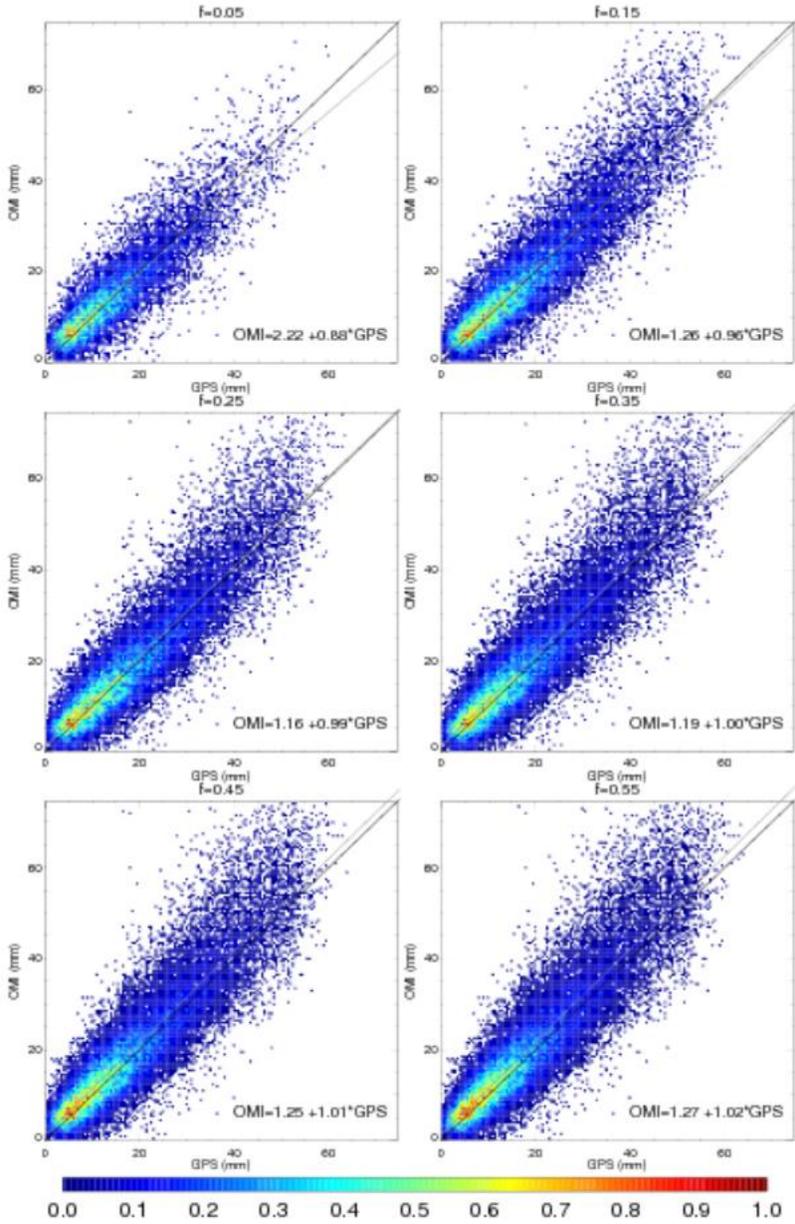



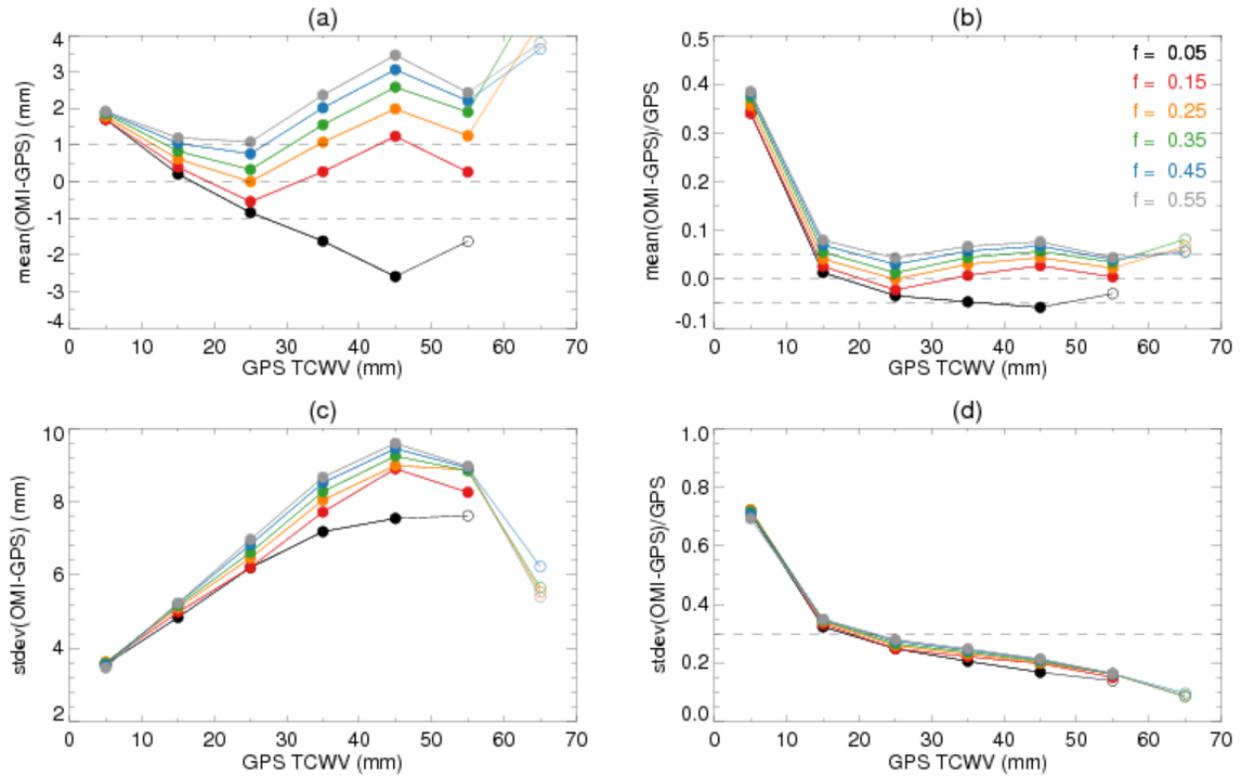
Figure 4. Normalized joint distributions of GPS versus Version 4.0 OMI TCWV for different
cloud fraction thresholds. Results are derived from the co-located data pairs for 2006. The OMI
data filtering criteria are the same as those for Figure 3. In each panel, the 1:1 line is plotted in
black, the linear regression line is plotted in gray and indicated by the formula in the lower right
corner.

299   To further characterize the effect of cloud fraction threshold on the comparison between GPS
300 and OMI, in Figure 5, we examine the mean and standard deviation (σ) of OMI-GPS for each 10
301 mm GPS TCWV bin. The results are derived from the same sets of co-located GPS and OMI

302   data as those used in Figure 3 and Figure 4. The filled symbols are for the cases where the

303   number of GPS and OMI data pairs within the corresponding TCWV bin is > 1% of the total

304   number of data pairs, and the open symbols are for < 1%. As the filled symbols represent better

305   statistics, we will focus on them below.

306

307   **Figure 5.** Parameters for each 10 mm TCWV bin. Curves with different colors are for different

308   cloud fraction thresholds f as indicated in Panel (b). The OMI filtering criteria remain the same

309   as those for Figure 3 and 4. Symbols are filled if the fraction of data pairs within the TCWV

310   interval is > 1% of all the available data pairs and are open otherwise. The parameters are (a)

311   mean of OMI-GPS in mm, (b) relative bias defined as mean(OMI-GPS)/GPS, (c) standard

312   deviation (σ) of OMI-GPS in mm, and (d) relative scatter defined as σ/GPS. Results are for all

313   days in 2006. Dashed lines are meant to facilitate visualization.

314

315   Figure 5(a) shows that the means of OMI-GPS vary between ±4 mm following "V"-shaped

316   curves whose minima occur in the TCWV = 20 – 30 mm bin except for f = 0.05. The curves shift

317   upward with increasing cloud fraction thresholds, suggesting that OMI cloudy-sky TCWV is

318   generally larger than OMI clear-sky TCWV. Other things being equal, cloud formation indicates

319   water vapor saturation and therefore a larger amount of TCWV than that under clear-sky

320   condition. The smallest absolute bias for 10 < TCWV < 20 mm occurs at f = 0.05, that for 20 <

321   TCWV < 30 mm occurs at f = 0.25, and that for 30 < TCWV < 40 mm occurs at f = 0.15. The f =

322   0.15 and f = 0.25 curves show the best overall performance according to Figure 5(a) as they are

323   within 1 mm of zero for 10<TCWV<40 mm, while other curves come within 1 mm of zero in

324   narrower TCWV ranges. Figure 5(b) shows the relative bias which is defined as the mean of

325   (OMI-GPS)/GPS. The relative biases decrease sharply from ~40% to ~5% as GPS TCWV

326   increases from the TCWV = 0 – 10 mm bin to the TCWV = 10 – 20 mm bin, and generally stay

327   less than ~5 – 10 % for larger TCWV values. Figure 5(c) shows that σ increases from ~3.5 mm

328   for TCWV = 0 - 10 mm to ~9.5 mm for TCWV = 40 - 50 mm (the percentage of data with

329   TCWV > 50 mm is very small). In most cases, larger cloud fraction thresholds correspond to

330   larger σ values. This is consistent with the larger dynamical range (due to a larger fraction of

331   data with high TCWV) for larger cloud fraction threshold (Figure 4). In fact, the relative scatter,

332   defined as the mean of σ/TCWV, shows little difference among the f values (Figure 5d). The

333   relative scatter decreases with TCWV, with the sharpest decrease from ~0.7 to ~0.3 between

334   TCWV = 0 – 10 mm and TCWV = 10 – 20 mm (Figure 5d). The relative scatter continues to

335   decrease for larger TCWV and the overall scatter is about 20%.

336   In short, Version 4.0 OMI agrees with GPS within 1 mm for 10<TCWV<40 mm when f =

337   0.15 and f = 0.25 are used; when f = 0.05 is used, the bias and scatter are the smallest for

338   10<TCWV<20 mm; but, for TCWV < 10 mm, OMI TCWV is too high and has large relative

339   scatter. The latter is expected from the low signal-to-noise ratio when TCWV < 10 mm in the

340   OMI retrieval.

341   **3.2 OMI and SSMIS over ocean**

342   To evaluate Version 4.0 OMI TCWV over the oceans, we compare against the microwave

343   TCWV data from SSMIS on board the Defense Meteorological Satellite Program (DMSP)'s F16

344   satellite. The SSMIS data are derived by Remote Sensing Systems (RSS) using their Version 7

345   algorithm (www.remss.com) and have a retrieval accuracy of better than 1 mm (Wentz, 1997;

346   Mears et al., 2015). For clear-sky comparison, we use the daily 0.25°×0.25° SSMIS data for

347   January and July 2006 and filter out the pixels affected by rain and cloud liquid water. Diedrich

348   et al. (2016) found that the diurnal cycle in TCWV is generally within 1% to 5% of the daily

349 mean, with a minimum between 0600 LT and 1000 LT and a maximum between 1600 LT and

350 2000 LT. To reduce the influence of diurnal cycle, we average the SSMIS data for the ascending

351 and descending orbits of F16 (~2000 LT and 0800 LT in 2006).

352   We generate daily 0.25°×0.25° Level 3 OMI TCWV from the de-striped Level 2 OMI

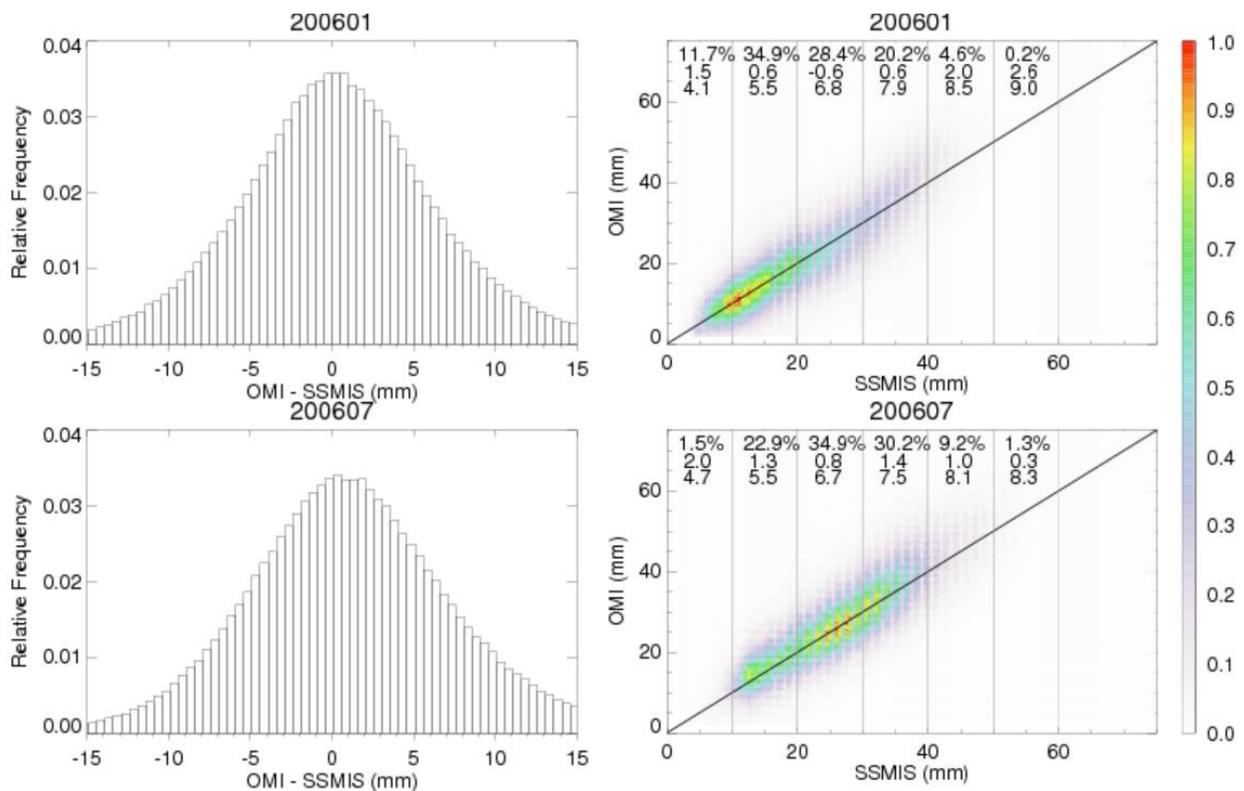
353 swaths, with the requirement that MDQFL = 0, fitting RMS < 0.001, 0<TCWV<75 mm, cloud

354 fraction < 0.05, and cloud top pressure > 750 mb. There are typically 15 Level 2 swaths per day.

355 The gridding program uses a tessellation method that weighs the contribution of a Level 2 data

356 point by its area within the Level 3 grid square and its spectrum fitting uncertainty (Wang et al.,

357 2014, 2016). The filtered daily Level 3 SSMIS and OMI data are compared for each month. We

358 find 548,223 and 847,678 co-located data pairs for January and July 2006, respectively.

359

360 **Figure 6.** Comparisons between Version 4.0 OMI and SSMIS over the oceans for (top) January

361 2006 and (bottom) July 2006. Panels in the left column show the relative occurrence frequency

362 of OMI-SSMIS (mm). Panels in the right column show the normalized joint distribution of

363 SSMIS versus OMI TCWV (mm).

364

365   The left column of Figure 6 shows the distribution of OMI-SSMIS for January and July

366   2006. For July, the mean of OMI-SSMIS is 1.1 mm with a standard deviation of 6.8 mm, the

367   mean absolute error |OMI-SSMIS| is 5.2 mm; for January, the mean error, standard deviation and

368   mean absolute error are 0.4 mm, 6.5 mm and 5.0 mm, respectively. This suggests a slightly better

369   agreement for January than for July. In comparison with the (OMI-GPS) over land (Section 3.1),

370   OMI-SSMIS over the oceans has somewhat larger bias and standard deviation. However, as

371   TCWV over the oceans are generally larger than that over land (compare Figure 6 with Figure

372   2), the relative bias and scatter are actually similar.

373   The right column of Figure 6 shows the normalized joint distribution of SSMIS versus OMI

374   for January and July 2006. The correlation coefficients are r = 0.84 and 0.82 for January and

375   July, respectively. For January, OMI-SSMIS remains within 0.6 mm of zero for TCWV in the 10

376   – 40 mm range, but is 1.5 mm for TCWV in the 0 – 10 mm range (only a small fraction of data

377   pairs have TCWV > 40 mm); for July, OMI-GPS is 0.8 mm for the TCWV = 20 – 30 mm bin,

378   and varies between 0.8 and 1.4 mm for TCWV in the 10 – 50 mm range (only a small fraction of

379   data pairs have TCWV < 10 mm or > 50 mm). For TCWV bins that have > 5% of the data pairs,

380   the standard deviation of OMI-SSMIS vary between 4.1 and 8.1 mm. Overall, Version 4.0 OMI

381   data compare reasonably well with SSMIS data for TCWV in the 10 – 40 mm range, with the

382   smallest bias occurring in the TCWV = 20 – 30 mm bin.

383   The agreement between Version 4.0 OMI with SSMIS is better than that between Version 3.0

384   OMI and SSMIS. For July 2006, using the same data filtering criteria as before, we find that

385   Version 3.0 OMI – SSMIS has a mean of 3.2 mm with a standard deviation of 7.8 mm. The bias

386   is much larger than that for Version 4.0 OMI – SSMIS. Again, this is because of the much larger

387   SCD of Version 3.0 OMI TCWV due to the water vapor reference spectrum (Supplementary

388   Figure 1).

389   Table 3 shows the effect of cloud fraction threshold (f) on the comparison between SSMIS

390   and Version 4.0 OMI TCWV. The comparisons are performed using daily filtered Level 3 data

391   for July 2006. For SSMIS, we only filter out pixels affected by rain. To investigate the influence

392   of clouds, cloud liquid water is not used to filter the SSMIS data here. This is less restrictive than

393   the criteria used for Figure 6 as the SSMIS pixels with cloud liquid water are filtered out in

394   Figure 6 for the "clear-sky" comparison there. For OMI, we require MDQFL = 0, RMS < 0.001,

395   0 < TCWV < 75 mm, cloud top pressure > 750 mb and cloud fraction < f. Results show that OMI

396   is higher than SSMIS by 0.02 – 3.07 mm for f = 0.05 – 0.45. The difference between the f = 0.05

397   case of Table 3 and the f = 0.05 case of Figure 6 is due to the relaxed SSMIS filtering criteria.

398   The closest agreement in terms of the mean and standard deviation of OMI-SSMIS occurs when

399   f = 0.05. The number of SSMIS and OMI data pairs more than doubles between f = 0.05 and f =

400   0.15. The linear correlation coefficient varies between 0.82 and 0.85 within the range of f values

401   considered. The best linear regression line (OMI = 0.70 + 1.02 * SSMIS) occurs when f = 0.15.

402   Therefore, for OMI over the oceans, we recommend using cloud fraction threshold f = 0.05 –

403   0.15, in combination with the other usual data filtering criteria, though users are advised to make

404   their own decisions based on their tolerance and applications.

405   **Table 3.** Effect of cloud fraction threshold on the comparison between SSMIS and Version 4.0

406   OMI TCWV for July 2006. f: OMI cloud fraction threshold; N: number of qualifying data pairs;

407   P: Percentage of qualifying data pairs with respect to the total number of qualifying SSMIS data

408   points; Mean: mean of OMI-SSMIS in mm; σ: standard deviation of OMI-SSMIS in mm; MAE:

409   Mean absolute error |OMI-SSMIS| in mm; r: correlation coefficient between SSMIS and OMI;

410   R^2: coefficient of determination for linear regression OMI = b + k * SSMIS, where OMI and

411   SSMIS are in mm; b: Intercept of linear regression; k: slope of linear regression.

f	N	P (%)	Mean	σ	MAE	r	R^2	b	k
0.05	1,048,879	7.4	0.02	7.11	5.39	0.82	0.67	1.43	0.95
0.15	2,837,032	20.0	1.38	7.82	5.84	0.84	0.71	0.70	1.02
0.25	3,932,468	27.8	2.20	8.09	6.09	0.84	0.71	1.11	1.04
0.35	4,819,185	34.0	2.73	8.22	6.24	0.85	0.72	1.45	1.05
0.45	5,537,003	39.1	3.07	8.26	6.32	0.85	0.72	1.62	1.06

412

413   Lowering the value for cloud top pressure threshold also leads to larger bias and scatter. For

414   example, when cloud fraction threshold f = 0.05 and cloud top pressure > 500 mb are used, the

415   mean and standard deviation of OMI-SSMIS become 0.80 mm and 7.9 mm, both are larger than

416   those for f = 0.05 in Table 3, though the linear regression line improves to OMI = 0.63 + 1.01 *

417   RSS due to an increase in the dynamical range of TCWV. It should be noted that the OMCLDO2

418   cloud product shows good agreement with ground-based observations for clouds at altitudes

419   lower than 2.5 km where single cloud layers dominate, but shows significant bias and large

420 scatter for clouds at altitudes higher than 2.5 km where multi-layer clouds dominate (Veefkind et

421 al., 2016). Thus, OMI TCWV data corresponding to low cloud top pressure (high altitude)

422 should be used with caution. Relaxing the filtering criteria for both cloud fraction and cloud top

423 pressure will lead to larger bias and scatter, therefore, it is not recommended. As an example, for

424 cloud fraction < 0.15 and cloud top pressure > 300 mb, the mean (standard deviation) of OMI-

425 SSMIS becomes 2.8 mm (9.0 mm) for July 2006.

426 **4 Applications**

427 **4.1 El Niño / La Niña**

428 In Figure 7, we examine the signals associated with El Niño and La Niña in Version 4.0 OMI

429 TCWV. Panel (a) shows the Multivariate ENSO Index (MEI) from NOAA (Wolter and Timlin,

430 1998) (https://www.esrl.noaa.gov/psd/enso/mei/). Positive (negative) values correspond to El

431 Niño (La Niña) conditions. We examine the anomalies in TCWV for July 2010 (MEI = -1.103,

432 La Niña) and July 2015 (MEI = 1.981, El Niño) in the bottom row. Although these events are

433 strong within the OMI record (from 2005 to the present), they are mild in comparison with the

434 extrema. Between 1950 and 2018, the maximum MEI is 3.008 (in March 1983) and the

435 minimum MEI is -2.247 (in June 1955).

436

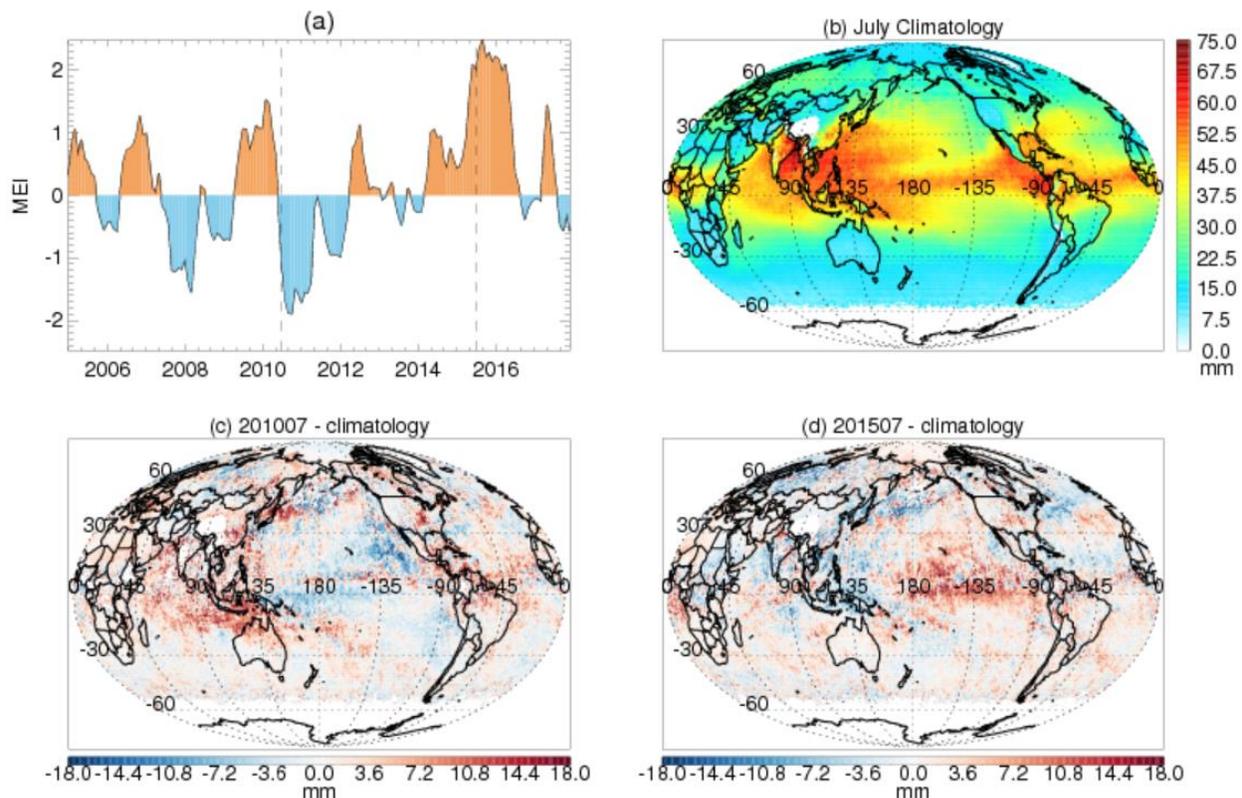



Figure 7. Top row: (a) Multivariate ENSO Index. Dashed vertical lines indicate July 2010 and July 2015; (b) TCWV (mm) climatology for July derived from Version 4.0 OMI data. Bottom row: TCWV anomaly (mm) with respect to the climatology for (c) July 2010 and (d) July 2015.

To examine the changes in OMI TCWV under different conditions, we first generate the monthly Level 3 (0.5°×0.5°) OMI TCWV using the Level 2 data for July 2005 and July 2015 using the method described in Section 3.2 (with a cloud fraction threshold of f = 0.15 and a cloud top pressure threshold of 750 mb). Then, using the same data filtering criteria, we derive the climatology for July using all the Level 2 July data between 2005 and 2015 (Figure 7b). Finally, we plot the deviations from the climatology (mm) for July 2010 and July 2015 in Figure 7(c) and 7(d), respectively.

The TCWV anomalies exhibit large-scale patterns. The pattern for July 2015 largely opposes that for July 2010. Particularly, in July 2015 under El Niño conditions, TCWV are higher in the equatorial central and eastern Pacific and lower in the Indonesia region; while in July 2010 under La Niña conditions, TCWV are lower in the tropical eastern Pacific and equatorial western

453   Pacific and higher in Indonesia and the Indian Ocean. The overall patterns largely conform to the

454   results derived from the Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite Data

455   (HOAPS) data (Shi et al., 2018).

456   **4.2 Corn Sweat**

457   "Corn sweat" refers to a hot and humid condition associated with heat waves which results in

458   large evapotranspiration rate in the Midwestern United States where cropland is often the

459   dominant land usage type. Besides evaporation, transpiration by plants, such as corn, draws

460   water from the soil to the atmosphere, enhancing the humidity and increasing the heat index. A

461   corn sweat event from July 18 to July 24 in 2016 made news in the US. This event is examined

462   in Figure 8 using the Version 4.0 OMI TCWV.

463   Figure 8 (a) and 8(b) show the Level 3 (0.25°×0.25°) OMI TCWV for July 18 - July 24

464   (7-day) and June 1 – August 31 (JJA) in 2016, respectively. The 7-day period corresponds to the

465   corn sweat event. The 0.25°×0.25° Level 3 data are derived using the same filtering criteria as

466   those used for Figure 7. Figure 8(c) indicates the anomaly associated with the corn sweat event

467   relative to the JJA mean. High TCWV is observed for the 7-day period from the Gulf coast to the

468   Midwestern US. Besides the Gulf region, the largest TCWV enhancements (of up to 18+ mm)

469   occur in parts of Iowa (IA), Missouri (MO), Illinois (IL) and Indiana (IN). Elevated TCWV is

470   also observed by several GPS stations in the general area during the same time period, though

471   coincident OMI data are not found at the stations (Supplementary Figure 3). At a few GPS

472   stations, high TCWV persisted a couple more days after July 24 which is most likely related to a

473   change in the weather. As shown by the surface pressure observations at the GPS stations, the

474   Midwest is under the control of a high-pressure system during the corn sweat period and a low-

475   pressure system afterwards (Supplementary Figure 4).

476

477

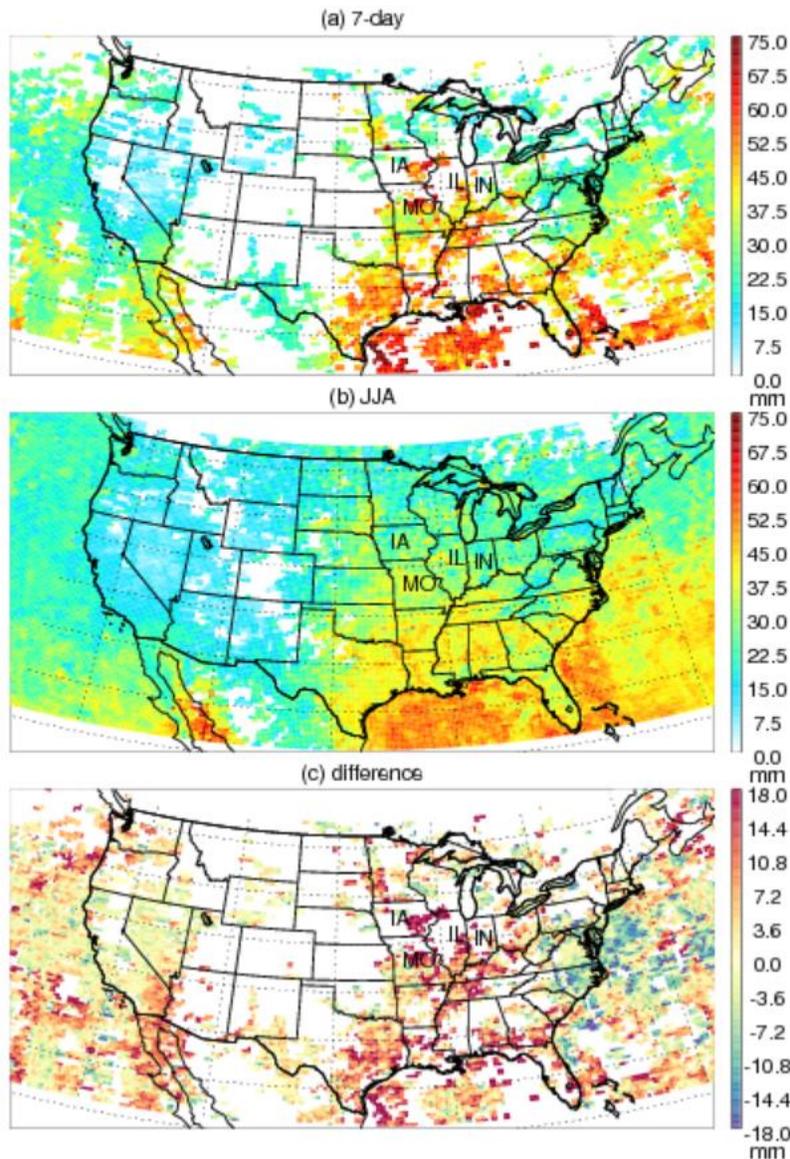



Figure 8. Level 3 (0.25°×0.25°) OMI TCWV (mm) generated using the Level 2 data during (a)

July 18 - July 24, 2016 and (b) June 1 - August 31, 2016. (c) The difference of (a) - (b) in mm.

The abbreviations for the states most affected by the event are indicated in the map.

481

To assess the significance of evapotranspiration for the Midwestern US during the corn sweat

event, we carried out a sensitivity study using the Weather Research and Forecasting (WRF)

model v3.9.1 (Skamarock et al., 2008). The model was run on a 36-km parent domain and a 12-

km nested domain, covering the relevant areas of the US. The physics parameterizations

486   included the WRF Single-Moment (WSM) 6-Class Microphysics (Hong and Lim, 2006), the

487   Kain-Fritsch (KF) subgrid cumulus parameterization (Kain, 2004), the Yonsei University (YSU)

488   planetary boundary layer scheme (Hong et al., 2006), the Noah Land-Surface Model (Ek et al.,

489   2003; Chen and Dudhia, 2001), and the Rapid Radiative Transfer Model (RRTM). Horizontal

490   turbulent diffusion was based on the standard Smagorinsky first-order closure. The initial and

491   lateral boundary conditions were from the 3-hourly NCEP North American Regional Reanalysis

492   (NARR) at 32-km resolution. To reduce the uncertainty associated with lateral boundary

493   condition for the nested domain, we nudged the model in the parent domain toward the

494   reanalysis, but left the nested domain running freely.

495   To diagnose the contribution of evapotranspiration, the model was run from July 19th to July

496   22nd of 2016 with and without evapotranspiration (calculated in the Noah Land-Surface model).

497   The results for July 21st are shown in Figure 9. TCWV is generally lower in the interior of the

498   domain for the run without evapotranspiration (No ET). The higher TCWV in the No ET run

499   near the southern boundary reflects non-linear water vapor transport from the Gulf region.

500   Turning off evapotranspiration not only directly affects the water vapor flux from the surface but

501   also indirectly influences other meteorological variables, such as winds. Thus, there is a

502   difference in the water vapor flux across the domain boundary. The difference between the

503   default and No ET runs in Figure 9 suggests that evapotranspiration contributes about 15 – 25%

504   of the TCWV in the Midwestern US during the corn sweat event. A detailed study incorporating

505   TCWV data with the WRF model will be carried out in future work.

506

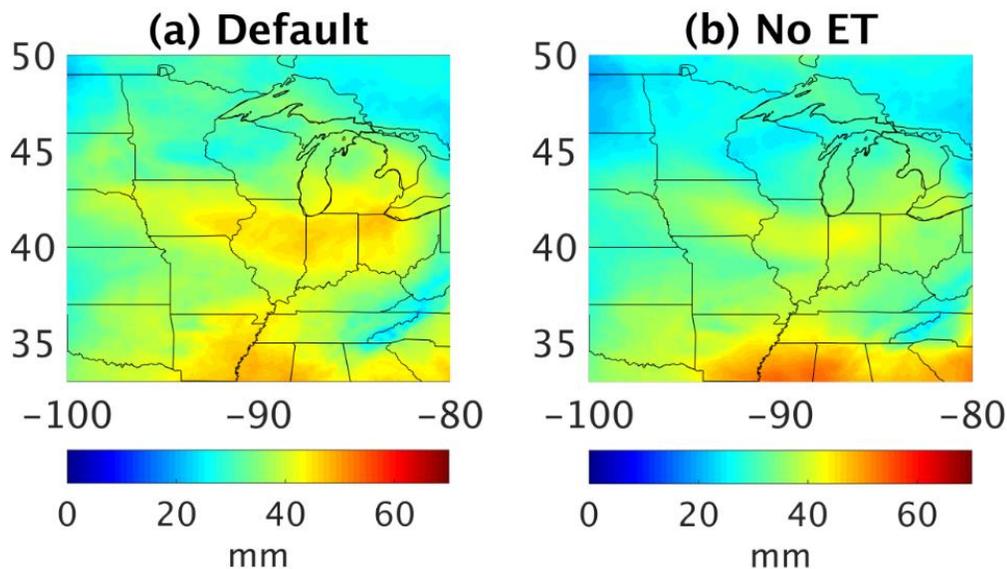



Figure 9. WRF simulations of TCWV (mm) for Midwestern US on 07/21/2016 for the run (a) with and (b) without evapotranspiration.

4.3 Atmospheric River (AR)

4.3.1 An Intense AR in OMI data

ARs are narrow elongated bands with high TCWV in the atmosphere. With flow rates similar to those of large rivers, ARs are highly important in the global hydrological cycle (Zhu and Newell, 1998). Land-falling ARs can lead to heavy orographic precipitation that affects areas such as the west coast of North America and Europe (Gimeno et al., 2014; Neiman et al., 2008b).

The extreme AR of November $6^{th} – 7^{th}$, 2006 brought devastating flood to the Pacific Northwest – the region in western North America bounded by the Pacific to the west and the Cascade mountain range to the east. This AR is described in detail in Neiman et al., 2008a. The signature of this AR is captured in the Version 4.0 OMI TCWV data. The left column of Figure 10 shows the Level 3 OMI TCWV and its anomaly on November 6th, 2006. The Level 3 data are generated following the same procedure as that used for Figure 8. Although many pixels are missing because of the cloud filtering (cloud top pressure > 750 mb, cloud fraction < 0.15) and other criteria, the leading edge of the AR is noticeable as an elongated band of high TCWV (15+ mm above the climatology) extending from Hawaii to Northern California (indicated by arrows in Figure 7(b) and 7(c)). The position of the AR in OMI TCWV agrees well with that in Special Sensor Microwave/Imager (SSM/I) microwave observation (Neiman et al., 2008a).

The right column of Figure 10 shows the Level 3 OMI ozone mixing ratio interpolated to 200 mb and its anomaly. The OMI ozone data are retrieved using the SAO ozone profile algorithm (Liu et al., 2010; Huang et al., 2017, 2018). The climatology is derived by averaging all monthly Level 3 data for November from 2004 to 2017. The global distribution of ozone at 200 mb shows low mixing ratio in the low latitudes and high mixing ratio in the high latitudes, opposite to the global distribution of TCWV. The anomaly shows a curvilinear band of high ozone that is parallel to the AR in the left column, but is located further to the west. This feature indicates intrusion of ozone rich stratospheric air along the polar front, and is associated with the same extra-tropical cyclone as the AR is.

536

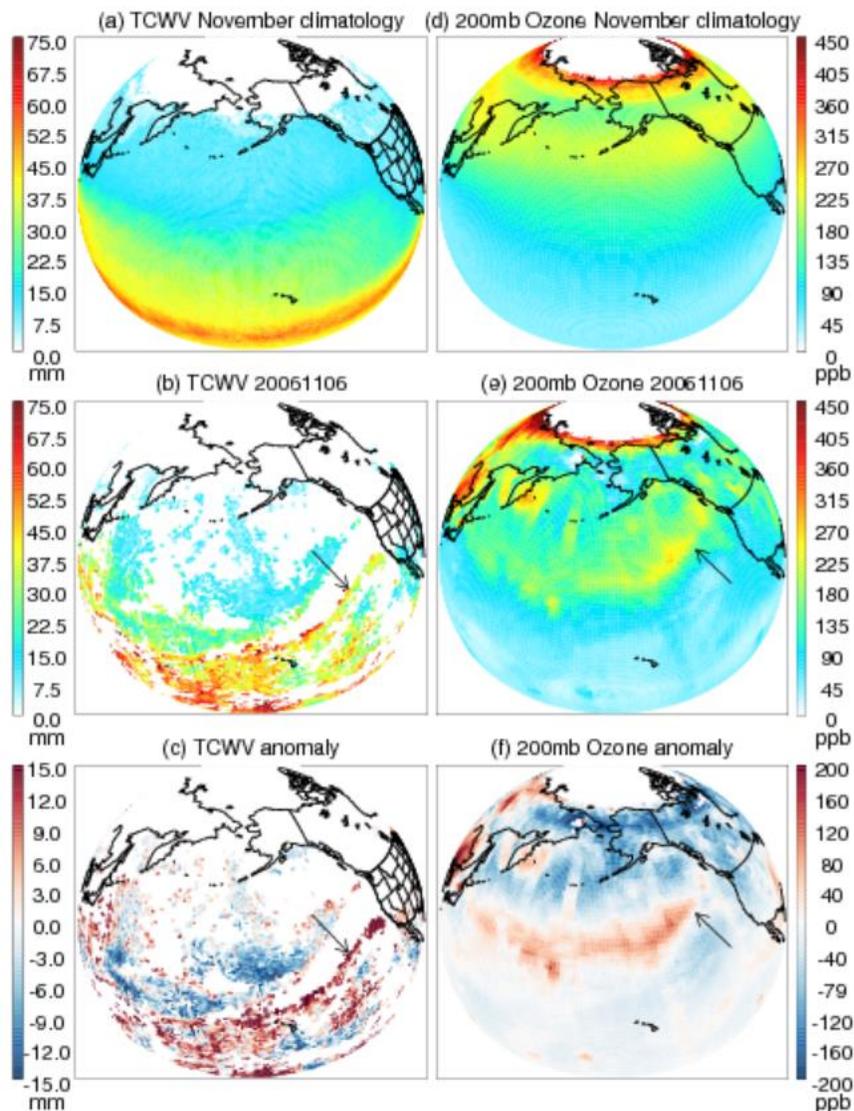



Figure 10. The Level 3 (top row) climatology, (middle row) data on November 6th, 2006 and

(bottom row) anomaly on November 6th, 2006 with respect to the climatology for (left column)

Version 4.0 OMI TCWV (mm, 0.5°×0.5°) and (right column) OMI ozone mixing ratio (ppb,

1°×1°) interpolated to 200 mb.

4.3.2 OMI TCWV Assimilation for the AR

To evaluate the potential of OMI water vapor data to improve numerical weather forecasts,

we conducted a data assimilation experiment from November 2nd to November 8th of 2006 using

WRF v3.9.1 and Version 4.0 OMI TCWV. The model was configured with a 27-km (290×270

546   surface grid points with 51 vertical levels), a 9-km (586×586×51 points) and a 3-km

547   (541×526×51) nested domains in a Lambert projection over the relevant portion of the Pacific

548   and North America (Figure 11 top left). The domains are designed for the November 6 AR event

549   and its associated precipitation at landfall. The model has the same physics parameterizations as

550   those used in Section 4.2 except that a more sophisticated double-moment microphysics scheme

551   is used for quantifying precipitation. The initial and boundary conditions for the 27-km domain

552   were from the 1°×1° NCEP FNL reanalysis. One-way nesting is used for the inner domains. To

553   evaluate the model's skill at simulating the AR and the contribution of OMI TCWV to the

554   quality of the simulation, we did not nudge the run towards the reanalysis, nor assimilate the

555   observed sea surface temperature within the computational domains.

556   The OMI TCWV is assimilated into the model using analytical optimal estimation (Rodgers,

557   2000). This method minimizes the cost function $J(\boldsymbol{x}) = (\mathbf{y} - H\boldsymbol{x})^T \mathbf{E}^{-1}(\mathbf{y} - H\boldsymbol{x}) +$

558    $(\boldsymbol{x} - \boldsymbol{x}^b)^T \mathbf{B}^{-1}(\boldsymbol{x} - \boldsymbol{x}^b)$, where x is the true TCWV, x^b is the a priori TCWV (from the model), y

559   is the observed TCWV, H represents the model Jacobian, $\mathbf{B}$ and $\mathbf{E}$ are the error covariance

560   matrices of the a priori and observation. $\mathbf{B}$ is estimated using the 12-hour and 24-hour forecasts

561   using the National Meteorological Center method (Parrish and Derber, 1992). $\mathbf{E}$ is based on the

562   fitting uncertainties of OMI data.

563   The a posteriori analysis ($\hat{\boldsymbol{x}}$) can be obtained from $\hat{\boldsymbol{x}} = \boldsymbol{x}^b + \mathbf{K}(\mathbf{y} - H\mathbf{x})$, where $\mathbf{K} =$

564    $\mathbf{B}H^T(H\mathbf{B}H^T + W^{-1}\mathbf{E})^{-1}$ is the Kalman gain, $W = \frac{(R^2 - r^2)}{(R^2 + r^2)}$ is the Cressman function to weigh the

565   observations based on their Euclidian distance r to the model grids, and R is the influence radius

566   of the observations. We simply assume R to be 1°, 0.5° and 0.25° for the 27-km, 9-km and 3-km

567   domain to get a quick look at the results in this paper and leave a more vigorous quantification of

568   R to future work. The a posteriori TCWV is solved hourly when OMI data are available and is

569   used to initialize the next simulation window.

570   During the assimilation, we adjust the OMI data using the AMF calculated with the modeled

571   water vapor profile ($OMI_{satellite}^{adjusted} = \frac{OMI_{satellite} \times AMF_{satellite}}{AMF_{model}}$) and the scattering weights provided

572   with the Level 2 OMI data. This can reduce the observational error associated with using the

573   monthly mean water vapor profile in the operational OMI product. The standard deviation of the

574   difference between $AMF_{satellite}$ and AMF_{model} is about 20%.

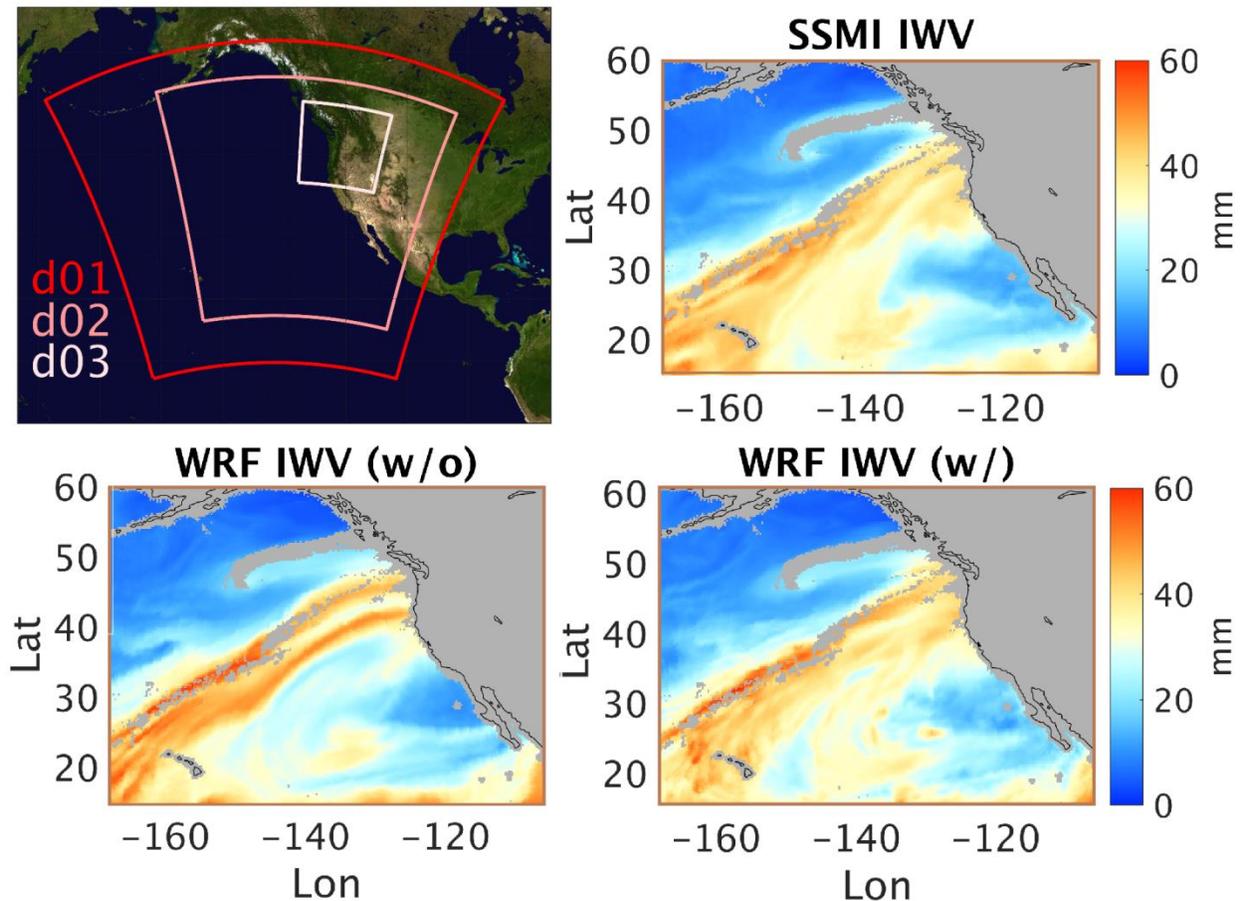



Figure 11. Top left: WRF model domain configuration for the November 2006 AR event. Top right: TCWV observed by SSM/I on November 6th, 2006. Bottom row: TCWV simulated by WRF on the same day (left) without and (right) with OMI TCWV data assimilation. Gray color indicates area with no SSM/I data.

Figure 11 shows the zoomed-in views of the AR on November 6th, 2006. The TCWV independently observed by SSM/I is shown in the upper right panel. The lower left and lower right panels show the model results without and with OMI TCWV assimilation. The model without assimilation shows an AR that is split into two parallel filaments making landfall at separate locations on the west coast of North America, where the TCWV is too high compared to the SSM/I observation, especially for the southern filament. As discussed later, this has a significant impact on precipitation (Figure 12). After assimilating OMI TCWV, the modeled

588   TCWV agrees much better with the SSM/I observation. The spurious southern filament

589   disappeared, the overall shape and amplitude of the AR are significantly improved.

590   The location and intensity of precipitation over land are crucial for local flood control and

591   water resource management, and are closely related to the shape and strength of AR at landfall.

592   The 24-hour accumulated precipitation on November 6 in the 3-km domain is examined in

593   Figure 12. The model output is coarsened to 0.25°×0.25° to match the resolution of the Tropical

594   Rainfall Measuring Mission (TRMM) observation product. The model without OMI data

595   assimilation produces spurious rainfall over the Oregon - California border (box A) as a result of

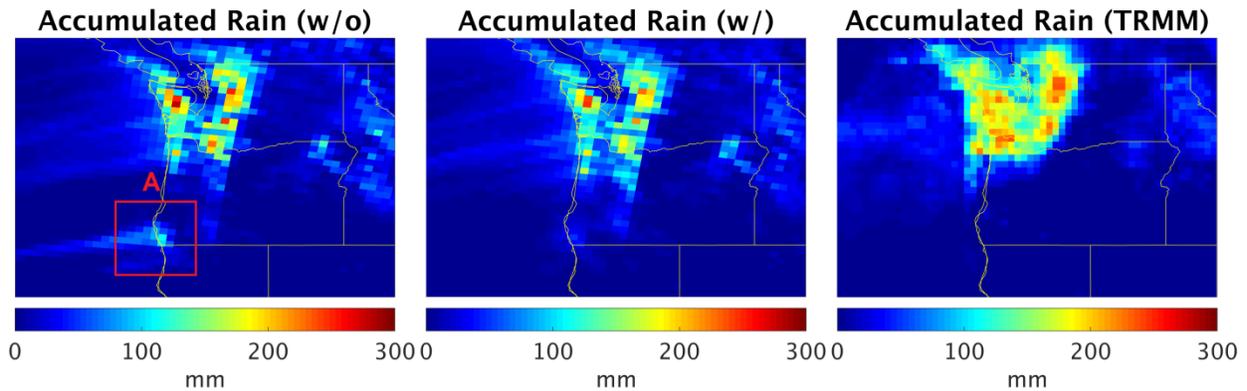
596   the erroneously strong southern filament of the simulated AR (Figure 11, lower left panel). This

597   artifact was removed after OMI data assimilation, showing better agreement with the

598   corresponding TRMM rainfall observation. The difference in rainfall between the assimilation

599   and observation in the Oregon / Washington area is probably related to both the model error and

600   the data error, as well as the data density and distribution. A detailed error attribution for

601   precipitation is beyond the scope of this paper.

602

603   **Figure 12.** The simulated rainfall accumulated from 0000 UTC to 2300 UTC (in mm) on

604   November 6, 2006 for the model (left) without and (middle) with OMI TCWV assimilation. The

605   rightmost panel show the accumulated rainfall observed by TRMM for the same time period.

606   Note that the 3-km model result is coarsened to match the resolution of the TRMM product.

607   Box A highlights the erroneously simulated precipitation in the run without OMI TCWV data

608   assimilation.

609

610   **5 Summary and Conclusion**

611   The Version 4.0 retrieval algorithm for OMI Total Column Water Vapor (TCWV) is presented

612   in this paper. The algorithm follows the usual two-step approach where Slant Column Density

613   (SCD) is derived from spectral fitting and Vertical Column Density (VCD) is obtained through

614   the ratio of SCD and Air Mass Factor (AMF). In Version 4.0, the spectral fitting no longer

615   considers common mode. The retrieval window (432.0 - 466.5 nm) results from a systematic

616   optimization that reflects trade-offs among several factors, including small fitting RMS, small

617   fitting uncertainty, large fraction of successful retrieval and long retrieval window length. The

618   AMF calculation uses the latest OMI O_2-O_2 cloud product (Veefkind et al., 2016) and monthly

619   variable vertical profiles from the MERRA-2 reanalysis (Gelaro et al., 2017).

620   The Version 4.0 OMI TCWV product is compared against the GPS network data over land

621   and the SSMIS microwave observations over the oceans for 2006. Version 4.0 OMI TCWV has

622   much smaller bias than Version 3.0 and has replaced previous versions on the Aura Validation

623   Data Center website. Version 4.0 OMI TCWV is characterized under different cloud conditions.

624   Under "clear-sky" condition (cloud fraction < 5% and cloud top pressure > 750 mb), the overall

625   mean of OMI-GPS over land is 0.32 mm with a standard deviation of 5.2 mm, and the smallest

626   bias occurs when TCWV is between 10 mm and 20 mm; the overall mean of OMI-SSMIS over

627   the oceans is 0.4 – 1.1 mm with a standard deviation of 6.5 - 6.8 mm, and the smallest bias

628   occurs for TCWV between 20 mm and 30 mm. The correlation coefficient between OMI TCWV

629   and the reference datasets realizes the largest gain when the cloud fraction threshold is increased

630   from 5% to 15%. The regression line appears the best when f = 0.25 is used over land and when f

631   = 0.15 is used over the oceans. But, larger cloud fraction leads to larger bias and scatter. Thus,

632   for most applications, we recommend to consider only OMI data with cloud fraction < 5% to

633   25% and cloud top pressure > 750 mb, in addition to main data quality flag = 0, no row anomaly,

634   fitting RMS < 0.001 and 0<TCWV<75 mm. Relaxing the cloud top pressure threshold has a

635   similar effect as relaxing the cloud fraction threshold. TCWV corresponding to low cloud top

636   pressure (high altitude) should be used with caution due to the degraded accuracy for these

637   clouds in the OMCLDO2 product.

638   As example applications of the Version 4.0 OMI TCWV data across a variety of temporal

639   and spatial scales, this paper examines the climate pattern associated with El Niño / La Niña, the

640   enhanced humidity during a week-long corn sweat event in the Midwest US, and the linear band

641   of high TCWV associated with an intense atmospheric river which made landfall on the west

642 coast of North America. Strong signals are found in OMI TCWV for all three examples. A data

643 assimilation experiment shows that the OMI TCWV data can help improve WRF's skill of

644 simulating the shape and intensity of the AR, as well as the accumulated rainfall near the coast.

645 Further improvement of the product can proceed from both spectral fitting and AMF

646 calculation, such as, water vapor reference spectrum, instrument slit-function and solar irradiance

647 for spectral fitting, aerosol correction and surface bi-directional reflectance for AMF calculation.

648

649 **Data availability**

650 The GPS network data are downloaded from NCAR (rda.ucar.edu/datasets/ds721.1). The SSMIS

651 data used in this paper are downloaded from the Remote Sensing Systems

652 (http://www.remss.com/support/data-shortcut/). The Multivariate ENSO Indices are downloaded

653 from NOAA (https://www.esrl.noaa.gov/psd/enso/mei/table.html). OMI TCWV and ozone

654 profile data are released through the Aura Validation Data Center (https://avdc.gsfc.nasa.gov/).

655

656 **Author contribution**

657 Huiqun Wang optimized the OMI TCWV retrieval window, performed the data validation

658 and tested most of the data application described in this paper. Amir Souri performed the WRF

659 simulations and data assimilation experiment presented in this paper. Gonzalo González Abad

660 improved and maintained the SAO retrieval code and implemented OMI TCWV data production

661 for the Aura Validation Data Center. Xiong Liu developed the OMI ozone profile retrieval and

662 provided the relevant data used in the AR application. Kelly Chance is the PI of the NASA grant,

663 and is responsible for the overall direction and execution of the project. Huiqun Wang prepared

664 and revised the manuscript with contributions from all co-authors. All authors contributed to

665 technical and scientific discussions during this project.

666

667 **Competing interests**

668 The authors declare that they have no conflict of interest.

669

670 **Acknowledgement**

671   We thank NASA's ACMAP program (Grant NNX17AH47G) for support.

672 **References**

673 Brion, J., Chakir, A., Daumont, D., Malicet, J. and Parisse, C.: High-resolution laboratory
674   absorption cross section of O3 - temperature effect, Chem. Phys. Lett., 213 (5-6), 610-612,
675   doi:10.1016/0009-2614(93)89169-1, 1993.

676 Chance, K. and Spurr, R. J. D.:, Ring effect studies: Rayleigh scattering, including molecular
677   parameters for rotational Raman scattering, and the Fraunhofer spectrum, Applied optics,
678   36, No. 21, 5224-5230, 1997.

679 Chance, K., Kurosu, T. P., Sioris, C. E.:, Undersampling correction for array detector-based
680   satellite spectrometers. Appl. Opt., 44, 1296-1304, doi:10.1364/AO.44.001296, 2005.

681 Chen, F. and J. Dudhia: Coupling an Advanced Land Surface-Hydrology Model with the Penn
682   State-NCAR MM5 Modeling System. Part I: Model Implementation and Sensitivity. Mon.
683   Wea. Rev., 129, 569-585, 2001.

684 Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., et al.: The ERA-Intrim
685   reanalysis: configuration and performance of the data assimilation system, Q. J. R.
686   Meteorol. Soc., 137, 553-597, 2011.

687 Diedrich. H., Wittchen, F., Preusker, R., and Fischer, J.: Representativeness of total column
688   water vapour retrievals from instruments on polar orbiting satellites, Atmos. Chem. Phys.,
689   16, 8331-8339, doi:10.5194/acp-16-8331-2016, 2016.

690 Dobber, M., Voors, R., Dirksen, R., Kleipool, Q. and Levelt, P.: The high-resolution solar
691   reference spectrum between 250 and 550 nm and its application to measurements with the
692   Ozone Monitoring Instrument, Solar Physics, 249, 2, 281-291, doi:10.1007/s11207-008-
693   9187-7, 2008.

694 Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., et al.: Implementation
695   of Noah land surface model advances in the National Centers for Environmental Prediction
696   operational mesoscale Eta model. *Journal of Geophysical Research:*
697   *Atmospheres*, *108*(D22), 2003.

698 Gelaro, R., McCarty, W., Suarez, M. J., Todling, R., Molod, A., et al.: The Modern-Era
699   retrospective analysis for research and applications, Version 2 (MERRA-2), J. Climate, 30,
700   5419-5454, doi:10.1175/JCI-D-16-0758.1, 2017

701   Gimeno, L., Nieto, R., Vazquez, M. and Lavers, D. A.: Atmospheric rivers: a mini-review,

702   Frontiers in Earth Sci., 2, Art. 2, doi:10.3389/feart.2014.00002, 2014.

703   González Abad, G., Liu, X., Chance, K., Wang, H., Kurosu, T. P. and Suleiman, R.: Updated

704   Smithsonian Astrophysical Observatory Ozone Monitoring Instrument (SAO OMI)

705   formaldehyde retrieval, Atmos. Meas. Tech., 8, 19-32, doi:10.5194/amt-8-19-2015, 2015.

706   Hong, S. Y. and Lim, J. O. J.: The WRF single-moment 6-class microphysics scheme

707   (WSM6). *J. Korean Meteor. Soc*, *42*(2), 129-151, 2006.

708   Hong, S. Y., Noh, Y. and Dudhia, J.: A new vertical diffusion package with an explicit treatment

709   of entrainment processes. *Monthly weather review*, *134*(9), 2318-2341, 2006.

710   Huang, G., Liu, X., Chance, K., Yang, K., Bhartia, P. K., Cai, Z., Allaart, M., Ancellet, G.,

711   Calpini, B., Coetzee, G. J. R., et al.: Validation of 10-year SAO OMI ozone profile

712   (PROFOZ) product using ozonesonde observations, Atmos. Meas. Tech., 10, 7,

713   doi:10.5194/amt-10-255-2017, 2017.

714   Huang, G., Liu, X., Chance, K., Yang, K., Cai, Z.: Validation of 10-year SAO OMI ozone profile

715   (PROFOZ) product using Aura MLS measurements, Atmos. Meas. Tech., 11, 1, 17-32,

716   doi:10.5194/amt-11-17-2018, 2018.

717   Kain, J.S.: The Kain–Fritsch convective parameterization: an update. *Journal of applied*

718   *meteorology*, *43*(1), 170-181, 2004.

719   Kleipool, Q. L., Dobber, M. R., de Hann, J. F., and Levelt, P. F.: Earth surface reflectance

720   climatology from 3 years of OMI data, J. Geophys. Res., 113, D18308,

721   doi:10.1029/2008JD010290, 2008.

722   Lampel, J., Frieß, and Platt, U.: The impact of vibrational Raman scattering of air on DOAS

723   measurements of atmospheric trace gases, Atmos. Meas. Tech., 8, 3767-3787,

724   doi:10.5194/amt-8-3767-2015, 2015a.

725   Lampel, J., Pohler, D., Tschritter, J., Friess, U., Platt, U.: On the relative absorption strengths of

726   water vapour in the blue wavelength range, Atmos. Meas. Tech., 8, 4329-4346,

727   doi:10.5194/amt-8-4329-2015, 2015b.

728 Liu, X., Bhartia, P. K., Chance, K., Spurr, R. J. D., Kurosu, T. P.: Ozone profile retrievals from
729   the Ozone Monitoring Instrument, Atmos. Chem. Phys., 10, 2521-2537, doi:10.5194/acp-
730   10-2521-2010, 2010.

731 Lorente, A., Boersma, K. F., Yu, H., Dorner, S., Hillboll, A., Richter, A., Liu, M., Lamsal, L. N.,
732   Barkley, M., De Smedt, I., Van Roozendael, M., Wang, Y., Wagner, T., Beirle, S., Lin, J.,
733   Krotkov, N., Stammes, P., Wang, P., Eskes, H. J. and Krol, M.: Structural uncertainty in air
734   mass factor calculation for NO2 and HCHO satellite retrievals, Atmos. Meas. Tech., 10,
735   759-782, doi:10.5194/amt-10-759-2017, 2017.

736 Levelt, P. F., van den Oord, G. H., Dobber, M. R., Malkki, A., Visser, H., de Vries, J., Stammes,
737   P., Lundell, J. O. and Saari, H.: The ozone monitoring instrument, T. Geosci. Remote, 44,
738   1093-1101, 2006.

739 Mason, J. D., Cone, M. T. and Fry, E. S.: Ultraviolet (250-550 nm) absorption spectrum of pure
740   water, Applied optics, 55, No. 25, 7163-7172, doi:10.1364/AO.55.007163, 2016

741 Mears, C. A., Wang, J., Smith, D. and Wentz, F. J.: Intercomparison of total precipitable water
742   measurements made by satellite-borne microwave radiometers and ground-based GPS
743   instruments, J. Geophys. Res. Atmos., 120, 2492-2504, doi:10.1002/2014JD022694, 2015.

744 Neiman, P. J., Ralph, F. M., Wick, G. A., Kuo, Y., Wee, T., Ma, Z., Taylor, G. H., Dettinger, M.
745   D.: Diagnosis of an intense atmospheric river impacting the Pacific northwest: storm
746   summary and offshore vertical structure observed with COSMIC satellite retrievals,
747   Monthly Wea. Rev., 136, 4398 – 4420, doi:10.1175/2008MWR2550.1, 2008a.

748 Neiman, P. J., Ralph, F. M., Wick, G. A., Lundquist, J. D., Dettinger, M. D.: Meteorological
749   characteristics and overland precipitation impacts of atmospheric rivers affecting the West
750   Coast of North America based on eight years of SSM/I satellite observations, J.
751   Hydrometeorology, 9, 22-47, doi:10.1175/2007JHM855.1, 2008b.

752 Ning, T., Wang, J., Elgered, G., Dick, G., Wickert, J., Bradke, M., Sommer, M., Querel, R. and
753   Smale, D.: The uncertainty of the atmospheric integrated water vapour estimated from
754   GNSS observations, Atmos. Meas. Tech., 9, 79-92, doi:10.5194/amt-9-79-2016, 2016.

755   Noël, S., Buchwitz, M., Bovensmann, H., Hoogen, R., and Burrows, J. P.: Atmospheric Water
756   Vapor Amounts Retrievd from GOME Satellite data, Geophys. Res. Lett., 26, 1841–1844,
757   1999.

758   Rodgers, C. D.: Inverse methods for atmospheric sounding, theory and practice, Series on
759   Atmospheric, Ocean and Planetary Physics – Vol. 2, Editor: Taylor, F. W., Published by
760   World Scientific Publishing Co. Pte. Ltd., Singapore, 2000.

761   Schröder, M., Lockhoff, M., Shi, L., August, T., Bennartz, R., et al.: The GEWEX Water Vapor
762   Assessment: Overview and introduction to results and recommendations, Remote Sensing,
763   11, 3, doi:10.3390/rs11030251, 2019.

764   Skamarock W. C. and Klemp, J. B.: A time-split nonhydrostatic atmospheric model for weather
765   research and forecasting applications, J. Computational Physics, 227, 2, 3465-3485,
766   doi:10.1016/j.jcp.2007.01.037, 2008.

767   Palmer, P. I., Jacob, D. J., Chance, K., Martin, R. V., Spurr, R. J. D., Kurosu., T. P., Bey, I.,
768   Rantosca, R., Fiore, A. and Li, Q.: Air mass factor formulation for spectroscopic
769   measurements from satellites: Application to formaldehyde retrievals from the Global
770   Ozone Monitoring Experiment, J. Geophys. Res., 106, D13, 14539-14550, 2001.

771   Ralph, F. M., Neiman, P. J. and Wick, G. A.: Satellite and CALJET aircraft observations of
772   atmospheric rivers over the eastern north Pacific ocean during the winter of 1997/98,
773   Monthly Wea. Rev., 132, 1721-1745, doi:10.1175/1520-
774   0493(2004)132<1721:SACAOO>2.0.CO;2, 2004.

775   Rothman, L. S., Gordon, I. E., Barbe, A., Benner, D. C., Bernath, P. F., et al.: The HITAN 2008
776   molecular spectroscopic database, J. Quant. Spectr. Radiat. Tran., 110, 533-572, 2009.

777   Schmidt, G. A., Ruedy, R. A., Miller, R. L. and Lacis, A. A.: Attribution of the present-day total
778   greenhouse effect, J. Geophys. Res., 115, D20106, doi:10.1029/2010JD014287, 2010.

779   Schroder, M., Lockhoff, M., Fell, F., Forsythe, J., Trent, T., Bennartz, R., Borbas, E., Bosilovich,
780   M. G., Castelli, E., Hersbach, H., Kachi, M., Kobayashi, S., Kursinski, E. R., Loyola, D.,
781   Mears, C., Preusker, R., Rossow, W. B. and Saha, S.: The GEWEX water vapor assessment
782   archive of water vapour products from satellite observations and reanalyses, Earth Syst.
783   Sci. Data, 10, 1093-1117, doi:10.5194/essd-10-1093-2018, 2018.

784 Shi, L., Schreck, C. J. III, and Schroder, M.: Assessing the pattern differences between satellite-
785    observed upper tropospheric humidity and total column water vapor during major El Niño
786    events, Remote Sens., 10, 1188, doi:10.3390/rs10081188, 2018.

787 Spietz, P., Martin, J. C. G. and Burrows, J. P.: Spectroscopic studies of the I-2/O-3
788    photochemistry – Part 2. Improved spectra of iodine oxides and analysis of the IO
789    absorption spectrum, J. Photochemistry and Photobiology, 176 (1-3), 50-67,
790    doi:10.1016/j.photochem.2005.08.023, 2005.

791 Thalman, R. and Volkamer, R.: Temperature dependent absorption cross-sections of O2-O2
792    collision pairs between 340 and 630 nm and at atmospherically relevant pressure, Phys.
793    Chem. Chem. Phys., 15, 15371-15381, doi:10.1039/c3cp50968k, 2013.

794 Vandaele, A. C., Hermas, C., Simon, P. C., Carleer, M., Colin, R., Fally, S., Merienne, M. F.,
795    Jenouvier, A. and Coquart, B.: Measurements of the NO2 absorption cross-section from
796    42000 cm(-1) to 10000 cm(-1) (238-1000 nm) at 200 K and 294 K, J. Quant. Spectr.
797    Radiat. Trans., 59, 171-184, doi:10.1016/S0022-4073(97)00168-4, 1998.

798 Vasilkov, A., Qin, W., Krotkov, N., Lamsal, L., Spurr, R., Haffner, D., Joiner, J., Yang, E. and
799    Marchenko, S.: Accounting for the effects of surface BRDF on satellite cloud and trace-gas
800    retrievals: a new approacj based on geometry-dependent Lambertian equivalent reflectivity
801    applied to OMI algorithms, Atmos. Meas. Tech., 10, 333-349, doi:10.5194/amt-10-333-
802    2017, 2017.

803 Veefkind, J. P., de Hann, J. F., Sneep, M. and Levelt, P. F.: Improvements to the OMI O2-O2
804    operational cloud algorithm and comparisons with ground-based radar-lidar observations,
805    Atmos. Meas. Tech., 9, 6035-6049, doi:10.5194/amt-9-6035-2016, 2016.

806 Volkamer, R., Spietz, P., Burrows, J. and Platt, U.: High-resolution absorption cross section of
807    glyoxal in the UV/Vis and IR spectral ranges, J. Photochem. Photobio., 172, 35-46,
808    doi:10.1016/j.jphotochem.2004.11.011, 2005.

809 Wagner, T., Beirle, S., Sihler, H. and Mies, K.: A feasibility study for the retrieval of the total
810    column precipitable water vapour from satellite observations in the blue spectral range,
811    Atmos. Meas. Tech., 6, 2593-2605, doi:10.5194/amt-6-2593-2013, 2013.

812 Wang, H., Liu, X., Chance, K., González Abad, G. and Chan Miller, C.: Water vapor retrieval

813   from OMI visible spectra, Atmos. Meas. Tech., 7, 1901-1913, doi:10.5194/amt-7-1901-

814   2014, 2014.

815 Wang, H., González Abad, G., Liu, X. and Chance, K.: Validation and update of OMI total

816   column water vapor product, Atmos. Chem. Phys., 16, 11379-11393, doi:10.5194/acp-16-

817   11379-2016, 2016.

818 Wentz, F. J.: A well-calibrated ocean algorithm for special sensor microwave/imager, J.

819   Geophys. Res., 102(C4), 8703-8718, doi:10.1029/96JC01751, 1997.

820 Wolter, K. and Timlin, M.S.: Measuring the strength of ENSO events - how does 1997/98 rank?

821   Weather, 53, 315-324, 1998.

822 Zhu, Y. and Newell, R. E.: A proposed algorithm for moisture fluxes from atmospheric rivers,

823   Monthly Weather Rev., 126, 3, 725-735, doi:10.1175/1520-

824   0493(1998)126<0725:APAFMF>2.0.CO;2, 1998.