Formaldehyde (HCHO) as a Hazardous Air Pollutant: Mapping surface air concentrations from satellite and inferring cancer risks in the United States

Lei Zhu†*, Daniel J. Jacob‡, Frank N. Keutsch§, Loretta J. Mickley†, Richard Scheffe‖,
Madeleine Strum‖, Gonzalo González Abad‖, Kelly Chance⊥, Kai Yang⊭, Bernhard Rappenglück○, Dylan B. Millet◆, Munkhbayar Baasandorj◆

†John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. ‡Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA. §Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. ‖U.S. Environmental Protection Agency, Durham, NC, USA. ⊭Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA. ⊭Department of Atmospheric and Oceanic Science, University of Maryland College Park, College Park, MD, USA. ○Department of Earth and Atmospheric Sciences, University of Houston, Houston, TX, USA. ◆Department of Soil, Water, and Climate, University of Minnesota, Minneapolis–Saint Paul, MN, USA.
ABSTRACT Formaldehyde (HCHO) is the most important carcinogen in outdoor air among the 187 hazardous air pollutants (HAPs) identified by the US Environmental Protection Agency (EPA), not including ozone and particulate matter. However, surface observations of HCHO are sparse and the EPA monitoring network could be prone to positive interferences. Here we use 2005–2015 HCHO column data from the OMI satellite instrument, validated with high-quality aircraft data and oversampled on a 5×5 km² grid, to map surface air HCHO concentrations across the contiguous US. Results are in good agreement with high-quality observations from urban sites and a factor of 2 lower than data from the EPA network. We thus estimate that up to 6600–13200 people in the US will develop cancer over their lifetimes by exposure to outdoor HCHO. The main HCHO source in the US is atmospheric oxidation of biogenic isoprene. The HCHO yield from atmospheric oxidation of isoprene decreases as the concentration of nitrogen oxides (NOₓ≡NO+NO₂) decreases, so that NOₓ emission controls to improve ozone air quality have a co-benefit in reducing HCHO-related cancer risks.
INTRODUCTION

Formaldehyde (HCHO) in outdoor air is a known carcinogen. Exposure to a mean HCHO concentration of 1 μg m\(^{-3}\) (about 0.7 ppb at STP) over one’s lifetime will cause up to 13 people in a million to develop lung and nasopharyngeal cancer according to the US Environmental Protection Agency (EPA).\(^1\) HCHO is one of 187 hazardous air pollutants (HAPs) identified by the EPA\(^1\) to cause cancer or other serious health impacts in ambient outside air. It is by far the most important HAP in terms of health risks, accounting for over 50% of the total HAPs-related cancer risks in the US.\(^2,\)\(^3\) The second most important HAP is benzene (≈10%). Unlike most other HAPs, HCHO is not mainly associated with local anthropogenic hotspots but instead is widely present across the US as a product of the oxidation of volatile organic compounds (VOCs) including in particular biogenic isoprene.\(^4\) The HAPs sampling network in the US is limited to urban/industrial sites\(^2\) and thus cannot properly quantify the health risk posed by ambient HCHO. Here we use 11 years of HCHO observations from the OMI satellite instrument\(^5\) with 5×5 km\(^2\) spatial resolution enabled by an oversampling technique\(^6\) to map HCHO surface air concentrations over the contiguous US and infer cancer risks on a national scale.

EPA reports HCHO as an ambient air toxic using data from 300–400 sites operated by states, local agencies, and tribes (SLTs network), including ≈50 national air toxics trends sites (NATTS network). HCHO is collected by 2,4-dinitrophenylhydrazine (DNPH) coated cartridges and then analyzed by high-performance liquid chromatography (HPLC), known as the EPA compendium method TO-11A\(^7\). HCHO measured using this method has potential interferences by ozone\(^8\) and NO\(_2\).\(^9,\)\(^10\) High-quality HCHO measurements in surface air are available only from occasional field campaigns.\(^11-14\) Satellites provide a continuously operating high spatial resolution data set. HCHO satellite data over the US were recently validated using aircraft observations.\(^15\)
Although satellites only measure total HCHO columns (molecules per cm$^2$ of surface), the bulk of that column is in the boundary layer$^{15-17}$ and surface concentrations can therefore be inferred.

HCHO columns have been observed continuously from space since GOME$^{18}$ (1996–2003) and SCIAMACHY$^{19}$ (2003–2012). Observations are presently available from OMI$^5,20$ (2004–), GOME2A$^{21}$ (2006–), OMPS$^{22,23}$ (2011–) and GOME2B$^{20}$ (2012–). Among those sensors, OMI provides the most suitable data for HCHO mapping due to its daily global coverage, long data record, and fine pixel resolution (13×24 km$^2$ at nadir). Its spatial resolution can be further refined by oversampling, as described below.

METHODS

OMI observations

OMI is a UV/Vis nadir solar backscatter spectrometer launched in 2004 on the Aura satellite in a polar sun-synchronous orbit.$^{24}$ It observes the whole globe daily at 13:30 local time (LT). We use OMI HCHO Version 2.0 (Collection 3) retrievals from the Smithsonian Astrophysical Observatory (OMI-SAO),$^5$ available at http://disc.sci.gsfc.nasa.gov/Aura/data-holdings/OMI/omhcho_v003.shtml. The data archive extends from 2005 to present. We select data for June–August 2005–2015 that (1) pass all the fitting and statistical quality checks (MainDataQualityFlag=0), (2) have cloud fraction less than 0.3 and solar zenith angle less than 60°, and (3) are not affected by the instrumental “row anomaly” (http://projects.knmi.nl/omi/research/product/rowanomaly-background). The single-scene precision is 1×10$^{16}$ molecules cm$^2$,$^5$ which corresponds to about 2 ppb in a 2-km deep well-mixed boundary layer. The precision can be improved by multi-scene averaging.$^{25,26}$ We only use the
summertime data when HCHO columns are highest and detectable from space. HCHO columns in winter are generally below the detection limit.\textsuperscript{27}

Zhu et al.\textsuperscript{15} validated the OMI-SAO product with high-quality HCHO aircraft measurements from the SEAC\textsuperscript{4}RS flight campaign\textsuperscript{28} over the Southeast US in August–September 2013. Aircraft measurements during SEAC\textsuperscript{4}RS were made in situ from 0.3 to 12 km altitude by two independent laser instruments: CAMS\textsuperscript{29} and ISAF.\textsuperscript{30} The two instruments were extremely consistent throughout the campaign ($r=0.99$) with ISAF 10% higher than CAMS. The horizontal patterns from the satellite retrievals were highly correlated with the aircraft and consistent with a dominant source from biogenic isoprene.\textsuperscript{17} However, the retrievals were biased low by 37% relative to the CAMS aircraft data. Here we apply a uniform correction factor of 1.59 to the OMI-SAO retrieval, thus making it unbiased relative to CAMS and 10% lower than ISAF.

**Oversampling method**

Relating concentrations to population exposure requires the highest spatial resolution possible. Temporal resolution is less critical since the HCHO cancer risk is based on a lifetime-averaged exposure. Here we oversample the OMI HCHO data to increase spatial resolution to 5×5 km\textsuperscript{2} through temporal averaging. Oversampling takes advantage of shifting pixel locations and sizes in day-to-day observations\textsuperscript{31, 32} to achieve a spatial resolution finer than pixel size as a temporal average. Oversampling of OMI observations to achieve an effective spatial resolution of a few km has been used previously on urban/regional scales for HCHO,\textsuperscript{6} SO\textsubscript{2}\textsuperscript{32–35} and NO\textsubscript{2}.\textsuperscript{32, 35} The common assumption in all these studies has been to view individual satellite observations as uniformly representative of a circle around the pixel center, with the circle radius optimized to
balance smoothing and noise. This approach is somewhat arbitrary and computationally demanding.

Here we developed an improved and faster oversampling method enabling application over the entire contiguous US. Consider a satellite pixel $p$ with HCHO column $\Omega(p)$. The overlap area between the pixel $p$ and oversampling grid cell $i$ is $A(p,i)$. Grid cell $i$ collects $N(i)$ overlapping satellite pixel data points over the oversampling period, from which an average column for that grid cell is calculated. We assume that the averaging weight for each individual satellite observation is proportional to the ratio of the overlap area $A(p,i)$ to the pixel area $S(p)$ and inversely proportional to the error standard deviation $\sigma(p)$ of that observation as reported in the OMI-SAOC product. $S(p)$ varies by a factor of 10 from $\sim$300 km$^2$ at nadir to $\sim$3000 km$^2$ at the outermost swath-angle.$^{24,36}$ $\sigma(p)$ can vary by a factor of 6 from $0.2 \times 10^{16}$ molecules cm$^{-2}$ for background conditions to $1.2 \times 10^{16}$ molecules cm$^{-2}$ in high-concentration regions.$^{5}$ The area- and error-weighted average column for grid cell $i$ is then derived as:

$$\bar{\Omega}(i) = \frac{\sum_{p=1}^{N(i)} \frac{A(p,i)}{S(p)\sigma(p)} \Omega(p)}{\sum_{p=1}^{N(i)} \frac{A(p,i)}{S(p)\sigma(p)}}$$

Besides being computationally fast, this method has the advantage that it fully uses and appropriately weighs the information from all individual satellite observations.

We applied our oversampling method to the OMI observations to produce an 11-year (June–August 2005–2015) summer average map of HCHO columns with $0.05^\circ \times 0.05^\circ$ ($\approx 5 \times 5$ km$^2$) grid resolution. Figure 1 shows the result. Values are highest over the Southeast US and are due to oxidation of biogenic isoprene.$^{4,37}$ The Southeast data were previously validated with the SEAC$^4$RS aircraft observations described above. The HCHO column peaks in the urban areas of Atlanta, Birmingham, and Houston, which could reflect industrial and vehicle sources.$^{6,26,38-40}$
However, these peaks are relatively modest on top of the biogenic enhancement. More detailed inspection of the Atlanta maximum (bottom panel of Figure 1) suggests a source from ring road traffic. HCHO hotspots in the western US are mostly due to fires as biogenic emissions in that part of the country are much lower than in the eastern US. Satellite retrievals of HCHO columns in fire plumes are highly uncertain because of strong sensitivity to plume rise and to light extinction by the smoke particles.41

Deriving annual mean surface HCHO concentrations

We use the summer mean 0.05°×0.05° satellite data for HCHO columns in combination with equation (2) to derive the annual mean surface concentrations required for cancer risk assessments:

\[
\bar{C}(i) = \bar{\Omega}(i)\gamma_1(i)\gamma_2(i)\gamma_3(i)
\]  

(2)

Here \(\bar{C}(i)\) is the annual mean surface air concentration in 0.05°×0.05° grid cell \(i\), \(\bar{\Omega}(i)\) is the summer mean oversampled OMI column in that grid cell (Figure 1), \(\gamma_1(i)\) is the ratio of midday surface air to column concentrations in summer, \(\gamma_2(i)\) is the ratio of 24-h average to midday concentrations in summer, and \(\gamma_3(i)\) is the ratio of annual to summer mean concentrations. We use the GEOS-Chem chemical transport model to infer \(\gamma_1\) and \(\gamma_3\), and surface observations to infer \(\gamma_2\). GEOS-Chem has been used previously to simulate HCHO over the US including comparisons to satellite and in situ observations.4,17,42 Zhu et al.15 and Chan Miller et al.43 find that GEOS-Chem provides an unbiased simulation of SEAC4RS and SENEX aircraft observations over the Southeast US in summer, including boundary layer horizontal patterns and mean vertical profiles.
Here we conducted a GEOS-Chem simulation at 2.0°×2.5° horizontal resolution for the entire year of 2010 using GEOS-5 assimilated meteorological fields\textsuperscript{44} produced at 0.25°×0.3125° resolution by the NASA Global Modeling and Assimilation Office (GMAO).

In order to convert HCHO columns to surface concentrations (scaling factor $\gamma_1$), we sample daily surface HCHO concentrations and total columns from the June–August model output at the OMI overpass time (13:00–14:00 LT; midday here and elsewhere). Figure 2 shows the resulting summer midday mean surface concentrations. The spatial patterns in the OMI data are retained because the GEOS-Chem scaling factors are fairly uniform. Figure 3 compares this product with local measurements from research field campaigns at several urban sites\textsuperscript{12,45} and at Houston (B. Rappenglück, unpublished data). There is good agreement with no significant bias averaged across all sites (-2.8±18%).

To convert midday to 24-h averaged surface air HCHO concentrations (scaling factor $\gamma_2$), we use ground-site HCHO measurements from three field campaigns including (1) CalNex (May–June 2010, Pasadena, California) (2) SOAS (June–July 2013, Brent, Alabama) and (3) SLAQRS (August–September 2013, East St. Louis, Illinois).\textsuperscript{11,13,14} Figure 4 shows the diurnal variations in HCHO concentrations measured at those three sites. HCHO is depleted during the night because of dry deposition.\textsuperscript{46} The 24-h average to midday value ratio is 0.64 at the California site, 0.77 at the Alabama site, and 0.79 at the Illinois site. It is not clear that these differences reflect geographical specificity, therefore we apply a single scaling factor $\gamma_2=0.73$ throughout the US to convert midday to 24-h average summer concentrations.

Finally, we use GEOS-Chem to convert these summertime 24-h averages to annual 24-h averages (scaling factor $\gamma_3$, Figure 5). Surface HCHO concentrations in GEOS-Chem have strong seasonal variations driven mostly by biogenic emissions, with annual to summertime average
ratios of 0.4–0.5 in the Southeast US and 0.6–0.7 in the West. Fires also contribute locally to
seasonality in the West but the resulting inferences are highly uncertain as noted above.
Summertime HCHO values as given in Figure 2 contribute 60% of the annual total in the Southeast
and 40% in the West. Comparison of GEOS-Chem to HCHO aircraft observations from the
WINTER campaign\textsuperscript{47} over the Northeast US in February–March 2015 indicates a 39% low bias in
the model, but the winter season contributes only 5% of the annual mean concentration in the
eastern US and 15% in the west, so the impact of this underestimate would be slight.

{
\vspace{1em}Estimating cancer risks
\vspace{1em}
EPA uses the inhalation unit risk estimate (URE) to quantify the cancer risks of HCHO and
other HAPs.\textsuperscript{1} The URE represents the upper bound for the increased cancer risk from inhalation
exposure to an air concentration of 1 \( \mu g \) m\(^{-3}\) over an individual’s lifetime. Based on the upper
confidence limit of the fitted dose-response curve, the inhalation URE for HCHO is estimated to
be \( 1.3 \times 10^{-5} \ (\mu g \ m^{-3})^{-1} \) by the Agency’s Integrated Risk Information System (IRIS).\textsuperscript{1,2} This means
that individuals exposed to a mean HCHO concentration of 1 \( \mu g \) m\(^{-3}\) (about 0.7 ppb) have a chance
of up to 13 in a million to develop cancer over their lifetime from this HCHO exposure. Risk is
assumed to increase linearly with HCHO concentration.

\vspace{1em}RESULTS AND DISCUSSION
\vspace{1em}
\textit{Annual mean surface HCHO concentrations}
\vspace{1em}
Figure 6 shows the annual mean HCHO concentrations in surface air across the US as
derived from equation (2). The patterns largely reflect those of the original OMI column data in
Figure 1.
Figure 7 compares these values with 2005–2015 annual averages from the EPA SLTs network. SLTs sites report 24-h average concentrations every 6 days. Annual averages are computed here for sites with full yearly coverage (at least 12 samples per quarter) for at least 9 years of the 2005–2015 period. 50 SLTs sites meet the above criteria and their locations are shown in the top panel of Figure 7.

We find from Figure 7 that OMI-derived surface HCHO is a factor of 2 lower than the EPA SLTs data, despite being unbiased relative to high-quality urban summer measurements (Figure 3). The SLTs HCHO data could have positive interferences, as noted in the Introduction. On the other hand, GEOS-Chem may underestimate HCHO during winter. There is moderate correlation between the OMI-derived and SLTs data ($r=0.56$). Here we will view the factor of 2 difference as representing a range of uncertainty for annual mean surface HCHO concentrations and hence cancer risks.

National cancer risks from outdoor HCHO exposure

Figure 6 shows the distribution of cancer risks in the US inferred from the OMI-derived mean surface HCHO concentrations, based on a conversion factor of 1.23 $\mu$g m$^{-3}$ ppb$^{-1}$ at 298 K and 1 atm. We estimate the total national cancer risk from HCHO exposure by convolving the OMI-derived distribution of cancer risks in Figure 6 with gridded 0.05°×0.05° population data for 2015.$^{48}$ We infer in this manner that 6600 people in the US will develop cancer at some point in their lives due to exposure to outdoor HCHO. As pointed out above, OMI-derived HCHO concentrations are a factor of 2 lower than measured at the EPA SLTs sites. Thus we increase the OMI-derived cancer risk by a factor of 2 to get an upper estimate. The resulting upper bound number of lifelong cancer occurrences due to exposure to outdoor HCHO over the US is in the
range 6600–13200, \( i.e., \) one person in \( 2\text{–}4\times10^5 \) based on a US population of 319 million. This cancer risk is well above the \( 1\text{ in }10^6 \) level generally considered as the threshold of tolerable risk.\textsuperscript{49}, \textsuperscript{50} Based on a US life expectancy of 78.7 years (2012), it translates into 84–168 cancer cases per year in the US caused by exposure to outdoor HCHO.

**Co-benefit of NO\textsubscript{x} emission controls for reducing HCHO-related cancer risks**

HCHO in the US originates mainly from the oxidation of biogenic isoprene emitted by vegetation, and thus would seem largely uncontrollable. However, the HCHO yield from isoprene oxidation is higher in the presence of nitrogen oxide radicals (NO\textsubscript{x})\textsuperscript{51,52} and NO\textsubscript{x} in the US is mainly anthropogenic. We conducted a sensitivity GEOS-Chem simulation with no anthropogenic NO\textsubscript{x} emissions and found that HCHO annual mean surface concentrations in the US decrease by 10–30\%. Convolving this decrease with the US population map we find that cancer risks from outdoor HCHO decrease by 20\%. Thus the increment in HCHO cancer risks due to anthropogenic NO\textsubscript{x} is one person in \( 4\text{–}8\times10^6 \), which by itself is above the tolerable threshold of \( 1\text{ in }10^6 \). Anthropogenic NO\textsubscript{x} emissions in the US have decreased by 2.1\% \textsubscript{a\textsuperscript{−}1} over 1991–2013 in response to regulations to improve ozone air quality.\textsuperscript{53} Such efforts to mitigate NO\textsubscript{x} emissions thus have a significant co-benefit in also reducing HCHO-related cancer risks.
Figure 1. Mean OMI HCHO columns over the contiguous US for June–August 2005–2015 with oversampling on a $0.05^\circ \times 0.05^\circ$ ($\approx 5 \times 5 \text{ km}^2$) grid. The bottom panel zooms in on the $\sim 100 \times 100 \text{ km}^2$ Atlanta area. The white circle indicates the Atlanta city center.
**Figure 2.** OMI-derived summer midday HCHO concentrations in surface air. Values are 2005–2015 averages for June–August at 13:00–14:00 local time.
Figure 3. Summer mean midday HCHO concentrations at urban sites. OMI-derived values for 2005–2015 are compared to local measurements in different years. Measurements for Houston are from the Moody Tower in August 2006 and 2010 (B. Rappenglück, unpublished data). Measurements for New York City are from Lin et al. Measurements for Atlanta, Philadelphia, and Nashville are from Dasgupta et al. Error bars represent ±1 standard deviation in the measurements.
Figure 4. Diurnal variation of summertime HCHO concentrations in surface air. Observations are from three field campaigns including CalNex (May–June 2010) in Pasadena, California; SOAS (June–July 2013) in Brent, Alabama; and SLAQRS (August–September 2013) in East St. Louis, Illinois. Error bars are standard deviations in the hourly averaged data. Red lines are 24-h averages.
Figure 5. Ratios of annual to summer (JJA) mean HCHO concentrations in surface air. Values are from the GEOS-Chem model.
Figure 6. OMI-derived annual mean HCHO concentrations in surface air and associated cancer risks. Cancer risk is related to surface HCHO concentration using the EPA inhalation unit risk estimate (URE) of $1.3 \times 10^{-5} \, (\mu g \, m^3)^1$. 
Figure 7. Annual mean 2005–2015 HCHO concentrations from EPA surface network (SLTs) and comparison to OMI-derived surface air concentrations. Top panel shows the network site locations and the annual mean data. Only sites with at least 9 years of complete data for the 2005–2015 period are shown (see text for details). Bottom panel compares the OMI-derived and EPA data for individual sites. Slopes (95% confidence interval) of reduced major axis (RMA) regressions are shown along with the normalized mean bias (NMB), sample size (N) and correlation coefficient (r).
AUTHOR INFORMATION

Corresponding Author: *Email: leizhu@fas.harvard.edu.

*Present address: Utah Department of Environmental Quality, Salt Lake City, UT, USA.

Notes: The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the NASA Earth Science Division. Funding for the SLAQRS measurements was provided by NSF (Grant #1148951). We thank Noelle Selin for valuable discussions on HCHO cancer risks.

REFERENCES


Volkamer, R.; Beirle, S.; Platt, U.; Wagner, T. Simultaneous global observations of glyoxal and

20. De Smedt, I.; Stavrakou, T.; Hendrick, F.; Danckaert, T.; Vlemmix, T.; Pinardi, G.; Theys,
N.; Lerot, C.; Gielen, C.; Vigouroux, C.; Hermans, C.; Fayt, C.; Veefkind, P.; Müller, J.-F.; Van
Roozendael, M. Diurnal, seasonal and long term variations of global formaldehyde columns
inferred from combined OMI and GOME-2 observations. Atmos. Chem. Phys. 2015, 15, 12519–
12545; DOI:10.5194/acp-15-12519-2015.

21. De Smedt, I.; Van Roozendael, M.; Stavrakou, T.; Müller, J.-F.; Lerot, C.; Theys, N.; Valks,
P.; Hao, N.; van der A, R. Improved retrieval of global tropospheric formaldehyde columns from

total columns from the Suomi National Polar-orbiting Partnership Ozone Mapping and Profiler

Observatory Ozone Mapping and Profiler Suite (SAO OMPS) formaldehyde retrieval. Atmos.


43. Chan Miller, C. Glyoxal yield from isoprene oxidation and relation to formaldehyde: chemical mechanism, constraints from SENEX aircraft observations, and interpretation of OMI satellite data. Submitted for publication, 2016.


