 doi:10.20944/preprints202105.0595.v2

Article

Global Surface HCHO Distribution derived from Satellite Observations with Neural Networks Technique

Jian Guan [1,#], Bohan Jin [1,#], Yizhe Ding [2], Wen Wang [1,*] and Guoxiang Li[3], Pubu Ciren[4]

[1] Center for Spatial Information, School of Environment and Natural Resources, Renmin University of China, Beijing 100872, China; guanjiankey@ruc.edu.cn (J, G); 2018200684@ruc.edu.cn (B, J); wenw@ruc.edu.cn (W, W)
[2] School of Statistics and Data Science, Nankai University, Tianjin 300071, China; 1810015@mail.nankai.edu.cn
[3] School of Information, Renmin University of China, Beijing 100872, China; neo@ruc.edu.cn
[4] IMSG Inc. & NOAA/NESDIS/STAR, 5825 University Research Ct., Suite 3250 M Square, College Park, MD 20740, USA; pubu.ciren@noaa.gov
* Correspondence: wenw@ruc.edu.cn; Tel.: (86-10) 88893061
# These authors has contributed to this work equally

Abstract:

Formaldehyde (HCHO) is one of the most important carcinogenic air contaminants. However, the lack of global surface concentration of HCHO monitoring is currently hindering research on outdoor HCHO pollution. Traditional methods are either restricted to small areas or data-demanding for a global scale of research. To alleviate this issue, we adopted neural networks to estimate surface HCHO concentration with confidence intervals in 2019, where HCHO vertical column density data from TROPOMI, in-situ data from HAPs (harmful air pollutants) monitoring network and ATom mission are utilized. Our result shows that the global surface HCHO average concentration is 2.30 μg/m^3. Furthermore, in terms of regions, the concentration in Amazon Basin, Northern China, South-east Asia, Bay of Bengal, Central and Western Africa are among the highest. The results from our study provides a first dataset of the global surface HCHO concentration. In addition, the derived confidence interval of surface HCHO concentration adds an extra layer for the confidence to our results. As a pioneer work in adopting confidence interval estimation into AI-driven atmospheric pollutant research and the first global HCHO surface distribution dataset, our paper will pave the way for the rigorous study on global ambient HCHO health risk and economic loss, thus providing a basis for pollutant controlling policies worldwide.

Keywords: surface formaldehyde; neural network model; interval estimation; TROPOMI; global distribution

1. Introduction

Formaldehyde (HCHO) is a carcinogenic trace gas and toxic pollutant in the atmosphere [1]. It is considered by the U.S. Environmental Protection Agency (EPA) to be one of the most important carcinogens in outdoor air among 187 harmful air pollutants (HAPs) [2], and accounts for more than 50% of the total risk of HAP related cancer in the United States [3]. 13 out of every million people receive nasopharyngeal carcinoma after being exposed to an average concentration of 1 microgram per cubic meter of HCHO for a lifetime [4]. As the most abundant aldehyde compound in the atmosphere, HCHO is one of the major volatile organic compounds (VOCs) and pollutants in the troposphere [5], which has a close relationship with the formation and extinction of O_3 and NO_2 in the atmosphere. HCHO pollution is a global scale issue. Ambient HCHO can be produced naturally and artificially, such as photolysis of isoprene from vegetation [6,7] farmland emissions [8], energy production and automobile exhaust emissions [9,10].

Surface concentration represents the amount of HCHO that people are exposed to, and is the direct data source of health risk estimation. Nevertheless, despite the crucial role of HCHO in human's health and atmosphere, it is difficult to monitor HCHO

 doi:10.20944/preprints202105.0595.v2

systematically and comprehensively by using traditional ground-based methods because of the large error and the expensive cost [11]. As a result, there is still no regular or large-scale monitoring of HCHO over most regions of the world. Most countries and regions with serious pollution fail to measure the surface HCHO concentration. Only in the United States, the HAP sampling network collects HCHO information but is limited to cities and industrial sites [12].

In contrast, remote sensing technology can not only monitor the long-term and large-scale dynamics, but also avoid many interference factors. Currently, there are many satellite missions reporting HCHO vertical column density (VCD) [13], which provides fundamental datasets for many related researches. The main sensors used to measure the concentration of HCHO VCD in the atmosphere include GOME-1 [14], GOME-2 [15], SCIAMACHY [16], OMI [17] and TROPOMI [18]. In terms of precision, TROPOMI is the most advanced atmospheric monitoring spectrometer with the highest resolution, with a swath of 2600km and daily global coverage [19]. However, most satellite-based retrieval can only provide the total column concentration due to their limitation on vertical resolution. Therefore, most studies on ambient HCHO only focus on the total amount in the vertical column in certain regions, such as North America [20], South America [21], Europe [22], Asia [23,24], Africa [7], instead of focusing on its surface concentration.

With the increasing attention towards health risks and photochemical pollution, demand for HCHO surface concentration distribution from the global perspective is growing more urgent. Many efforts have been put to derive surface concentration from total column concentration, such as using the fixed forms of linear models to assess the relationship between VCD and in-situ concentration[1] of NO_2, SO_2, CO, PM [25], and using R^2 to assess the relationship between vertical column density and ground in-situ concentration [26]. However, these methods seem to be less accurate and may only be limited to specific pollutants. In the other few existing studies HCHO surface concentration was derived by applying the vertical distribution profile from, GEOS-Chem model to the satellite-derived total column concentration [27]. However, the atmospheric transportation model itself requires numerous input parameters, which may impede its application to a global scale with a reasonable spatial and telporal resolution. Therefore, our main focus here is to derive the global surface HCHO concentration distribution based from satelllite-derived total column HCHO concentration and a quite limited in-situ HCHO concentration.

Neural network, a powerful machine learning algorithm, has gained its reputation for revealing hidden patterns inside data with a great accuracy in various fields, such as image classification [28], object detection [29], image denoising [30], image synthesis [31], person re-identification [32], etc. However, some algorithms, such as vanilla neural network, do not assign confidence level nor confidence interval to its point estimation results, which is necessary for scientific estimation and public policy decision-making. To quantify uncertainty of results derived from neural networks, a diverse of approaches have been adopted, including Bayesian neural network [33], delta method [34], bootstrap [35], mean variance estimation [35], interpreting dropout as performing variational inference [36]. However, these methods are either computationally demanding or strongly based on assumptions. Quality-driven (QD) method, a method based on LUBE to derive confidence intervals for the neural network, by combining the uncertainty estimating loss and the neural network loss function as a whole [37], is not only compatible with gradient descent algorithms, but also shrink the average confidence interval length up to 10%, compared with previous works [38]. Therefore, to enhance the credibility of our model, this method is leveraged to obtain the interval estimation of surface concentration of HCHO. By combining the point and interval estimation, it is believed to meet a balance between maintaining accuracy and controlling uncertainty in the form of a pre-set confidence level.

[1] In-situ HCHO concentration include surface concentration and high-altitude concentration from ATom flight data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

The potential health impact of HCHO but lack of global surface monitoring data demands an efficient way to get a better understanding of global HCHO surface distribution with limited data. In this paper, as a novel study, we derived the global surface concentration of HCHO in 2019 by feeding TROPOMI VCD data and limited surface HCHO concentration data into a neural network model. In addition, besides the capture of the seasonal changes of key areas, confidence intervals for the derived surface HCHO are also estimated by using QD method. As a novel work on adopting interval estimation in AI-driven atmospheric pollutant research and deriving the first dataset ofglobal HCHO surface distribution, our paper will pave the way for rigorous study on global ambient HCHO health risk and economic loss, thus providing a basis for pollutant controlling policies worldwide.

2. Data and Methods

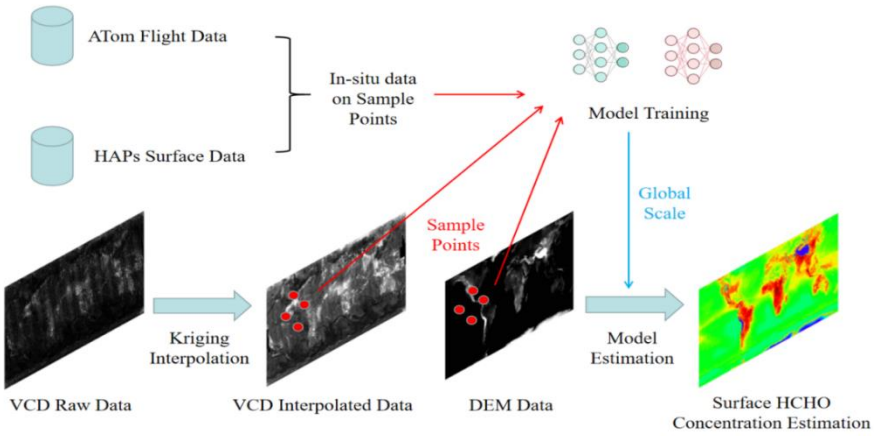


Figure 1. Data processing workflow

To estimate the global distribution of HCHO surface concentration, we used two discrete in-situ data sources and Sentinel-5P TROPOMI VCD data on the corresponding location (as shown by red points in Figure 1) to train our neural network model. Then we apply our model to a global scale and estimate the surface HCHO distribution with confidence intervals.

2.1. Datasets

2.1.1 Sentinel-5P VCD Data

The data of vertical column density (VCD) of HCHO in this study comes from TROPOMI (Tropospheric Monitoring Instrument), which is carried on Sentinel-5P [19]. Sentinel-5P is a global air pollution monitoring satellite launched by ESA on October 13, 2017, as part of the Copernicus project. TROPOMI can effectively observe trace gas components in the atmosphere around the world, including NO_2, O_3, SO_2, HCHO, CH_4, CO and other important indicators closely related to human activities, and can strengthen the observation of aerosols and clouds [39].

In terms of accuracy, TROPOMI is currently the most advanced atmospheric monitoring spectrometer with the highest spatial resolution. The satellite provides global coverage daily with a spatial resolution of 7km×7km and the equator crossing time at about 13:30 local time, which effectively ensures the comparability of data in different regions [19]. Sentinel-5P data are currently available for public access[2].

We use the data of 2019 because a) 2018 is the first year that Sentinel-5P is in operation; the algorithm of the product is not stable then; b) 2020 is within the global COVID-19 pandemic, which might have special impact on anthropogenic sources, making the result less representative in terms of a long-term status.

[2] https://s5phub.copernicus.eu/dhus/#/home

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

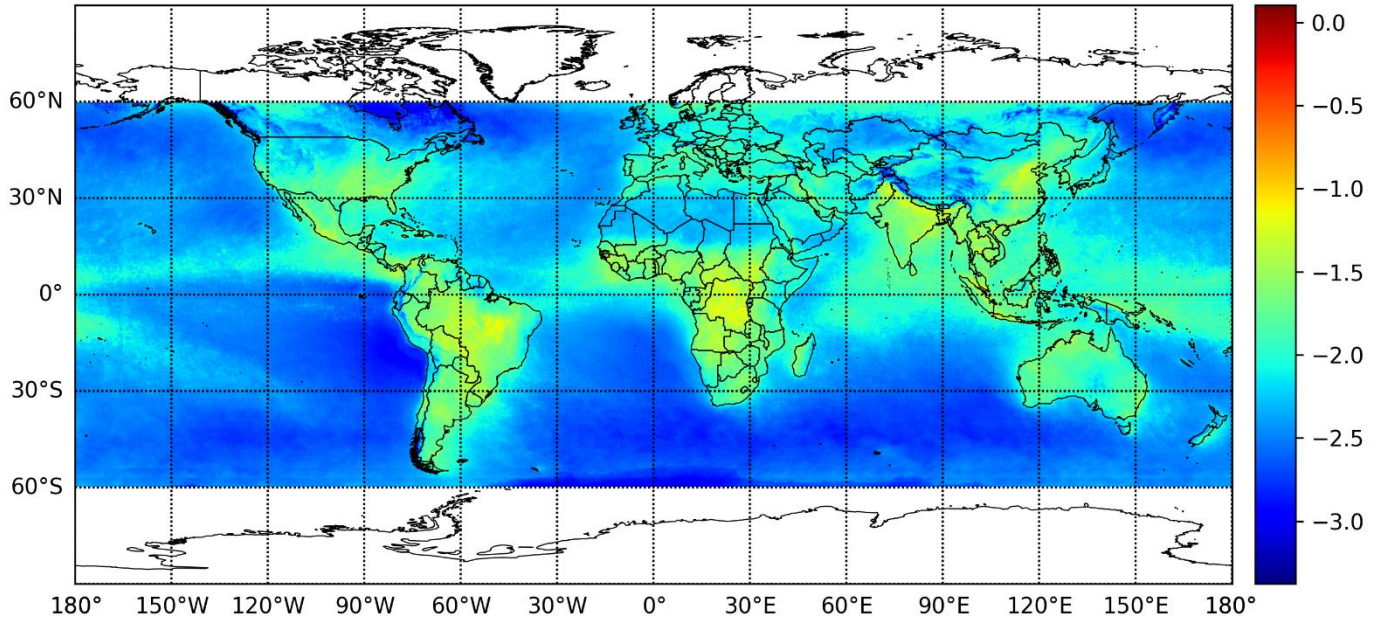


Figure 2. The natural logarithm of global vertical column density (VCD) of HCHO in 2019 after being interpolated and averaged on an annual basis. (unit: mol/m^2)

Offline HCHO data from January 1 to December 31, 2019 are collected. According to the technical documents, data points whose quality index (QA_ value) is less than 0.5 are removed to ensure the best quality. After doing mosaic on the datasets and applying Ordinary Kriging interpolation, we obtained the distribution of global average total column concentration of HCHO with a resolution of 0.05° by 0.05°. The data beyond 60°S and 60°N is discarded due to the sparsity of satellite data and scarceness of human activities , which has little impact on health risk estimation.

2.1.2. In-situ Data

Since our study aims to estimate the surface concentration of HCHO on a global level, we need surface-level concentration data which will cover diverse types of underlying surfaces and also different altitudes to train our model. Therefore, the following two data sources are considered.

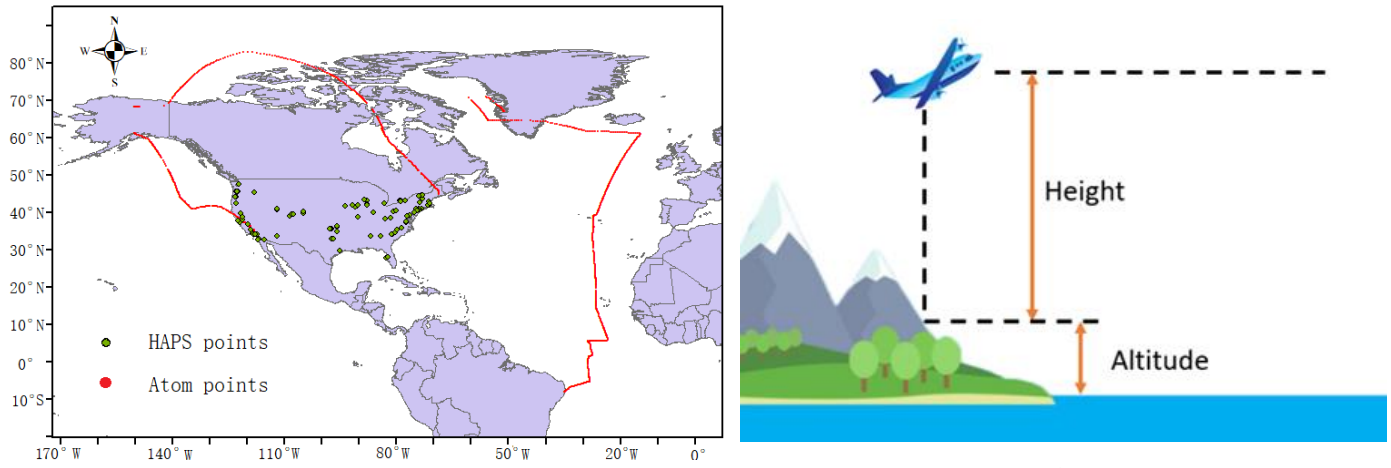
ATom flight data. NASA's atmospheric tomography mission (ATom) is a systematic, global sampling of the atmosphere in the United States from 2016 to 2018, and continuous profile analysis from 0.2km to 12km. The volume mixing ratio of HCHO in air was measured in ATom flight data. A large number of gas and aerosol payloads were deployed on NASA's DC-8 aircraft, and the HCHO on NASA's high-altitude aircraft was measured by ISAF instrument [40,41]. The instrument uses laser-induced fluorescence (LIF) to obtain the high sensitivity needed to detect HCHO in the upper troposphere and lower stratosphere, which has an abundance of 10 parts per trillion. LIF can also achieve quick response to measure the abundance of HCHO in the fine structure outflow of convective storms. These HCHO measurements will be used to elucidate the mechanism of convective transport and to quantify the effects of boundary layer pollutants on ozone photochemistry and cloud microphysics in the upper atmosphere [42].

HAPs ground monitoring data. We obtained ground HCHO observations from EPA SLTS network at https://www.epa.gov/outdoor-air-quality-data, which reports average 24-hour HCHO concentration all around the year. Here, we selected 5965 data points from 109 sites in 2019, covering the whole country, as shown in Figure 2 (a).

These two datasets cover a wide range of latitudes, from -8.1977° S to 82.9404° N, and a diverse variety of landscapes in the U.S. The selection of the HAPs dataset is to ensure that the concentration distribution feature at ground level is represented in our model,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

and the ATom data is to ensure that our model can be generalized and applied to a global extent.



(a) (b)

Figure 3. (a) The geographical distribution of our data, where red represents ATom flight data points and green represents HAPS ground monitoring network. **(b)** The meaning of "Height" and "Altitude" for ATom mission data

Since ATom data are obtained far above the surface, and the vertical distribution of HCHO usually changes largely from ground to 1~2km above [43], we take the "Height" of the aircraft measurements as another input variable in our model to control the impact of vertical distribution along the column. For those HAPS ground monitoring data, we assign 0 as their heights.

2.1.3. Global DEM Data

Since descriptive statistics show a negative relationship between surface altitude and in-situ concentration, with a Pearson's correlation of r=-0.3907 in our in-situ dataset, we use global Digital Elevation Model (DEM) data as one of the input variables—"Altitude", in order to estimate the ground-level concentration. The relationship between variable "Height" and variable "Altitude" is shown in Figure 2 (b).

In our study, we use the Shuttle Radar Topography Mission (SRTM) DEM product and resample it to a resolution of 0.05°. This dataset has an initial resolution of 90m at the equator and is provided in WGS84 projection with a 1 arc resolution[44].

2.2. Data Processing

After collecting and organizing data into formattable structure, we first visualize and preprocess these data. Then, two neural networks are implemented for point and interval estimations by using PyTorch, a well-known deep-learning framework. Our code is available online[3].

The preprocessed data with the ground truth in-situ HCHO concentration are then divided into two groups, training and testing dataset, to train our models. After that, global VCD data are fed into the model to derive global surface level HCHO concentration.

2.2.1 Preprocessing

In theory, a neural network is able to handle input data from a different distribution, however, a significant defect was noticed in the training process without preprocessing, owing to the highly imbalanced, skewed distribution of the HCHO concentration (both column and in-situ). Therefore, we first applied log-transformation to the raw data. The

[3] https://github.com/dingyizhe2000/Interval-HCHO-Concentration-Estimation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

logarithm of the HCHO concentration data shows a bell-shaped distribution, and increments in estimation accuracy have also proven the effectiveness of log-transformation.

2.2.2. Neural Network Architecture

As a universal function approximator, the neural network plays a vital role in helping us derive the point and interval estimations of the HCHO concentration. But instead of training a single network to get these estimations jointly, two separate neural networks are constructed for point and interval estimation respectively, because several experiments which we carried out indicated that a joint model always has to compromise between point estimation and interval estimation, thus greatly reducing the accuracy of point estimation.

Like ordinary multi-layer perceptrons, each neural network in our model contains three input nodes, three BFR blocks (with the ReLUs in the last blocks are disabled). The network for point estimation has one output node, and the other network for interval estimation gets two nodes. The structure of our model is shown in Figure 3.

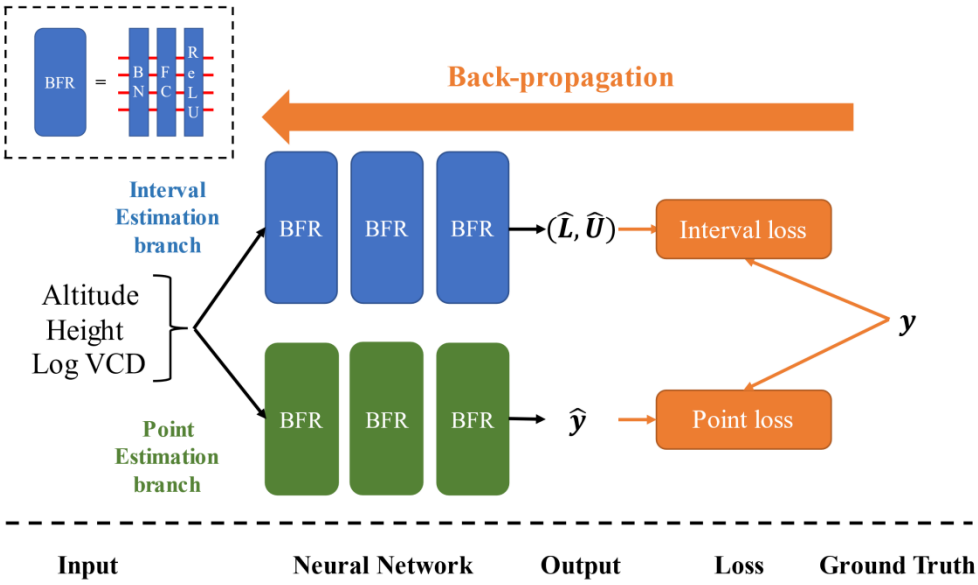


Figure 4. Illustration of two separate neural networks for point and interval estimations respectively. each network has three BFR blocks (with ReLU in the last block disabled).

For the sake of stabilizing the training and prediction procedure, instead of stacking full-connection and non-linear activation layers, we proposed to stack BFR blocks, which are made up of a batch normalization layer, a full connection layer and a ReLU activation layer sequentially.

Batch normalization (BN) is first introduced to address Internal Covariate Shift, a phenomenon referring to the unfavorable change of data distributions in the hidden layers. Just like the data standardization, BN forces the distribution of each hidden layer to have exactly the same means and variances dimension-wisely, which not only regularizes the network, but also accelerates the training procedure by reducing the dependence of gradients on the scale of the parameters or of their initial values [45].

Full connection (FC) layer is connected immediately after the BN layer in order to provide linear transformation, where we set the number of hidden neurons as 50. The output from the FC layer is non-linearly activated by ReLU function [46,47].

2.2.1.2 Loss function

Objective functions with suitable forms are crucial for applying stochastic gradient descent algorithms to converge while training. Though point estimation only needs to take the precision into consideration, two conflicting factors are involved in evaluating

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

the quality of interval estimation – higher confidential level usually yields an interval with greater length and vice versa.

Point estimation loss. Instead of fancy forms, we found that a l_1 loss is sufficient for training rapidly:

$$L_{point} = E|y - \hat{y}|.$$

Interval estimation loss is relatively complex compared to point estimation loss. The QD-loss takes the confidential level and interval length into consideration simultaneously[38]:

$$L_{interval} = MPIW + \eta \cdot \{0,\ (1-\alpha) - PICP\}^2.$$

On one hand, to control the confidential level of the interval estimator, α is set to indicate at most how many(proportionally) intervals failing to cover the true value can be tolerated. We set multiple $\alpha's$, including 0.05, 0.10, 0.20, in our model to derive interval predictions of various confidential level and average coverage length, and it is verified that higher α yields shorter intervals. $PICP$ indicates the covering rate of intervals:

$$PICP = P\{L < y < U\} \approx \frac{1}{n}\sum_{i=1}^{n} \quad I\left\{\widehat{L_J} < y_i < \widehat{U_j}\right\},$$

where $I\left\{\widehat{L_J} < y_i < \widehat{U_j}\right\} = 1$ if and only if $\widehat{L_J} < y_i < \widehat{U_j}$, else it equals to 0.

On the other hand, the average length of intervals subject to $PICP > 1 - \alpha$ should be minimized. However, intervals that fail to capture their corresponding data point should not be encouraged to shrink further. The average interval length to penalize is, therefore,

$$MPIW = \frac{1}{\sum_{i=1}^{n} \quad I\{\widehat{L_J} < y_i < \widehat{U_j}\}} \sum_{j=1}^{n} \quad (\widehat{U_j} - \widehat{L_J})\ \widetilde{k_J},$$

where $\widetilde{k_j} = \sigma\left(s \cdot \left(y_j - \widehat{L_j}\right)\right) \cdot \sigma(s \cdot (\widehat{U_j} - y_j))$, works as a continuous approximation towards "hard" $I\left\{\widehat{L_J} < y_i < \widehat{U_j}\right\}$. Since the sigmoid function σ is known for providing a differentiable alternative to discrete stepwise functions, and $s = 160$ is a super-parameter for smoothness.

3. Results

3.1. Point Estimation

Point estimation model in this study shows a relatively high accuracy and is generally consistent with previous studies on the vertical distribution of HCHO. Figure 5. shows the point estimation value of in-situ concentration with the change of vertical column density (VCD) and height, when altitude at sea level is fixed. It is seen that in-situ concentration is negatively correlated with the height and positively correlated with VCD.

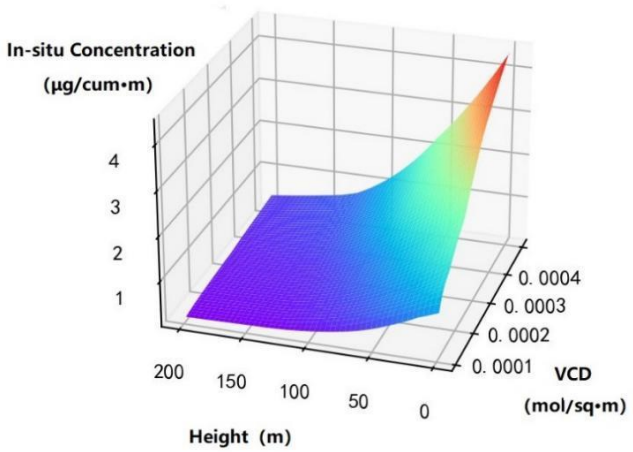


Figure 5. The model's estimation of in-situ concentration on certain height, vertical column density of HCHO when the altitude is fixed at 0 m

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

To evaluate the performance of our model, statistics including MAE[4] and RMSE[5] are calculated based on the training and testing datasets respectively. As shown by Table 1, both MAEs and RMSEs are relatively small, which indicates that the model performs well in the point estimation.

Table 1. MAE and RMSE of point estimation for surface concentration (unit: μg/m3)

Dataset	MAE	RMSE
Training	1.294	1.018
Testing	1.295	1.075

By loading the global DEM, logarithm VCD and the height (0 m at surface) into the model, the annual average of the global surface HCHO distribution map was derived. As shown in Figure 5, there are generally 6 regions where HCHO surface concentration is high, namely the Amazon area, south east U.S., Central and Western Africa, North Eastern India, South East Asia, and North China, with an average concentration of more than 4 μg/m3. The seasonal change of HCHO in these key areas is discussed in section 3.3.

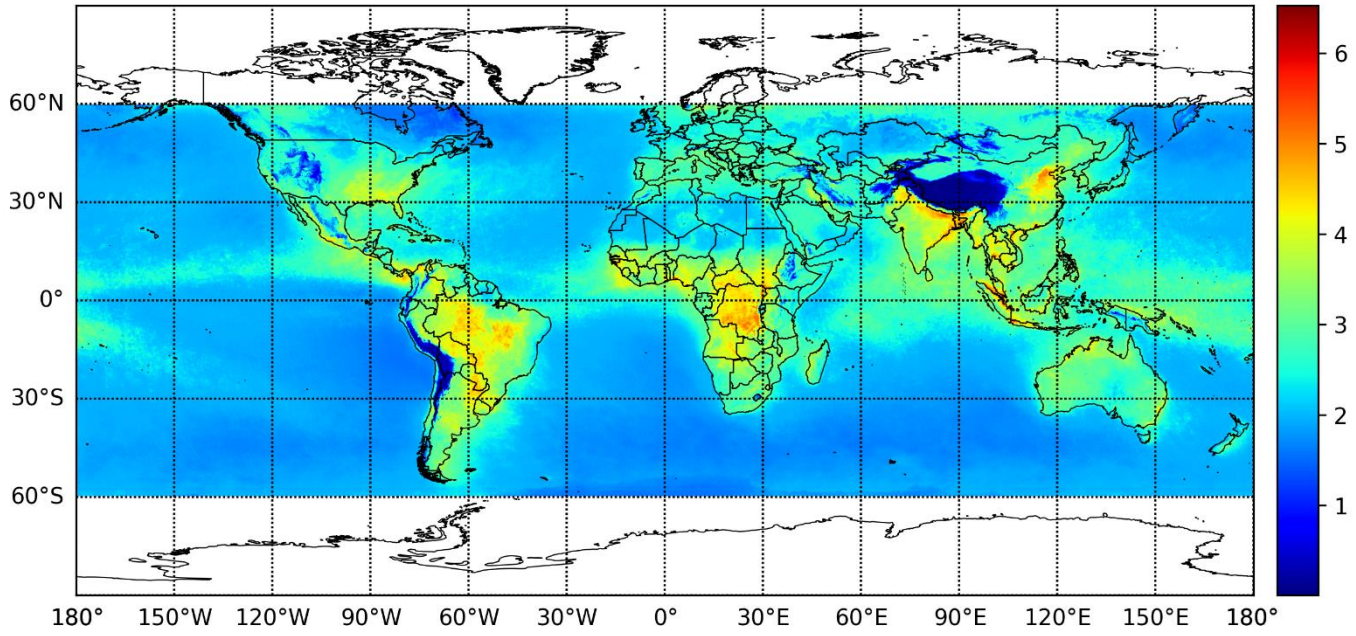


Figure 6. Annual average of global surface concentration in 2019. (unit: μg/m[3])

The uneven distribution of HCHO concentration on the sea and land surface is also noticed in Figure 6, which shows the HCHO concentration is relatively lower and more homogeneous on the sea surface than on the land. Statistics given in Table 2 have also confirms this fact. It is seen that the annual mean of surface HCHO concentration is about 2.21 μg/m3 over ocean and 2.77 μg/m3 over land.

Table 2. Statistics of surface HCHO concentration of sea surface, land surface and combined (unit: μg/m3)

	Standard Dev.	Mean	Minimum	Maximum
Sea	0.414	2.12	1.49	6.22
Land	0.859	2.77	0.006	6.53
Global	0.644	2.30	0.006	6.53

[4] MAE: Mean Average Error

[5] RMSE: Root Square of Mean Square Error

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

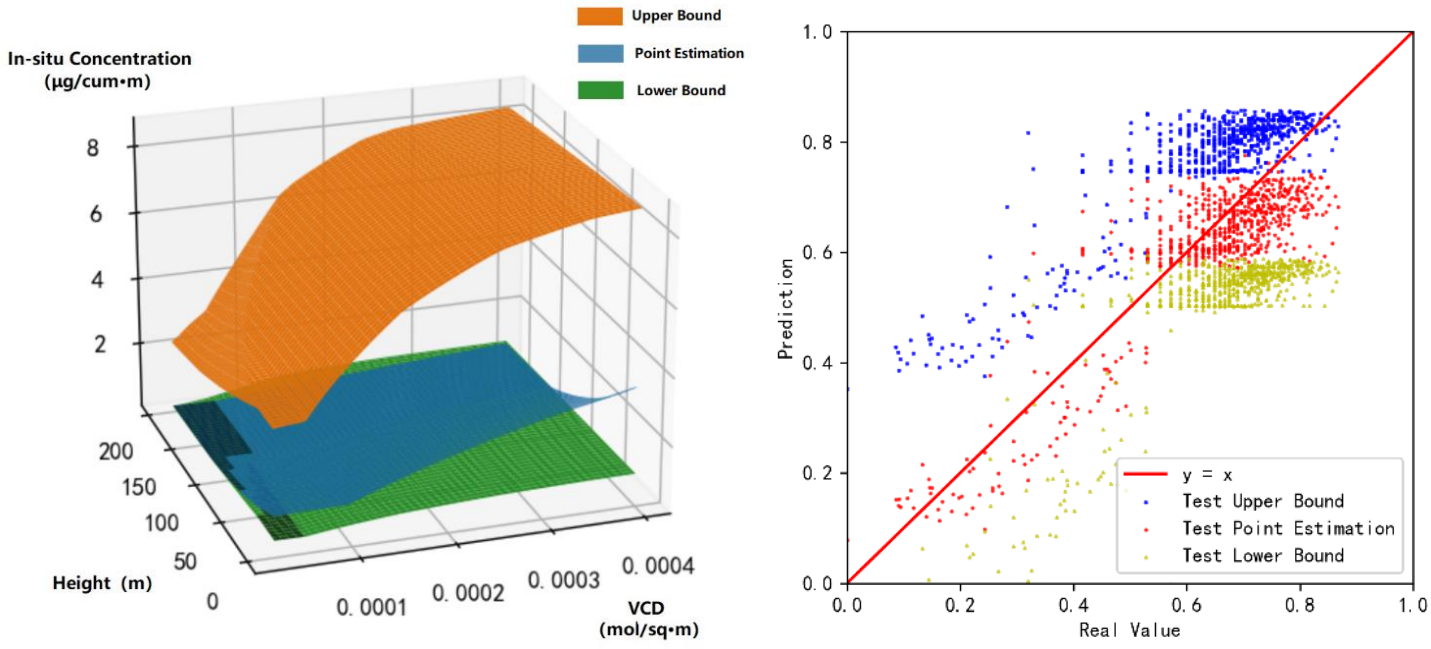
Cities, as the regions with the densest population, deserve specific attention towards their surface HCHO concentration due to its known and potential harm to people living there. Table 3 shows the surface concentration of HCHO of some of the typical cities in these regions, where Jakarta and Singapore, two major cities (country) in South East Asia, rank the highest and the second highest, reaching to 6.18 and 5.83 μg/m3, respectively.

Table 3. Surface HCHO concentration in some typical cities

City Name	Surface HCHO ($\mu g/m^3$)	City Name	Surface HCHO ($\mu g/m^3$)
Jakarta, Indonesia	6.18	Beijing, China	5.23
Singapore	5.83	Patna, India	5.07
Colon, Panama	5.66	Ha Noi, Vietnam	5.06
Kuala Lumpur, Malaysia	5.61	Guangzhou, China	5.00
Dhaka, Bangladesh	5.51	Tianjin, China	4.89
Lagos, Nigeria	5.49	Manaus, Brazil	4.50
Bangkok, Thailand	5.42	Montgomery, U.S.	4.44
Shijiazhuang, China	5.38	Houston, U.S.	4.22
Ho Chi Minh City, Vietnam	5.27	Freetown, Sierra Leone	4.15
Kolkata, India	5.26	Kolwezi, R. D. Congo	3.81

3.2. Interval Estimation

Besides point estimation, the model in this study also provides the estimation of upper and lower bounds of surface concentration of HCHO, so that the uncertainty, or variability of the surface concentration can be evaluated. In Figure 6, the relationship between the estimated upper bound, lower bound and the point estimation are displayed in a 3D space. It is worth emphasizing that the captured uncertainty, or the interval length, delineates the variability of the data itself, not the lower trustworthiness of our model or its estimations.



(a) (b)

Figure 7. (a) The estimated upper bound, point estimation and lower bound (90% CI) as a function of the height and vertical column density of HCHO when the altitude is fixed at 0 m, where CI represents confidence level. The results in this figure is obtained by feeding the equally spaced mock data into the two models. (b) The comparison of the models' estimation and the real value on the testing set. The point estimation lies around the red line, in the middle of upper bound and lower bound (90% CI).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

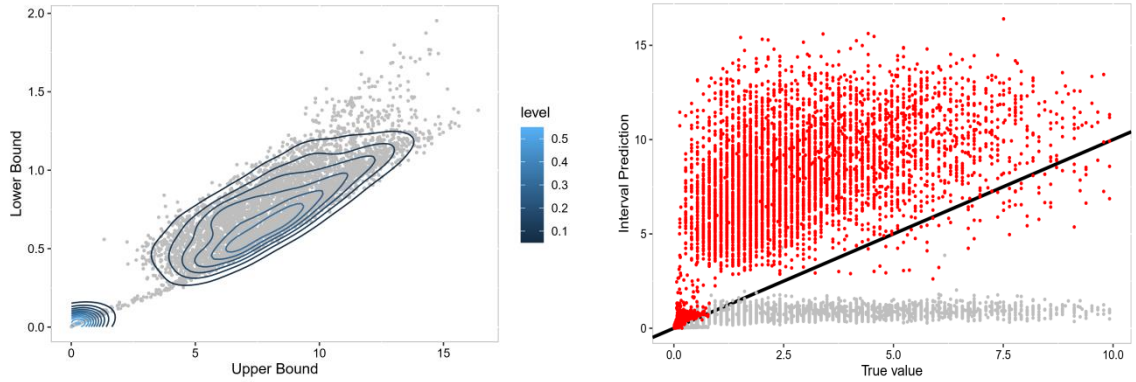
Confidence level, together with the covering length, lay the foundation for the trustworthiness and precision of our interval prediction. As shown in Table 4, interval estimation model obtains the covering rates and the ratio of true values covered by predicted interval, of 94.41% and 88.74%, exceeding the pre-set confidential level $\alpha = 0.9$ and $\alpha = 0.8$, respectively.

In addition, as expected in section 2.2.1.2, a higher confidence level yields a longer average interval length[6], which is 4.530 μg /m³ for $\alpha = 0.9$, 17% more than 3.864 μg /m³ for $\alpha = 0.8$. Such a phenomenon can also be seen in the statistics, shown in Table 4, for minimum, maximum and mean values of upper and lower bounds, respectively for the two confidence levels.

Table 4. Statistics of interval estimation for surface concentration (unit: μg/m³)

α	Covering Rate	Avg Length	Bound	Std	Mean	Min	Max
0.9	94.41%	4.530	U	3.528	7.112	0.00684	16.40
			L	0.354	0.670	0.00193	4.273
0.8	88.74%	3.864	U	3.518	6.446	0.00972	12.35
			L	0.545	0.968	0.00128	1.898

However,the standard deviation of upper bounds seems to be larger than that of the lower bounds under both scenarios in Table 4. From the density scatter plot between these two, shown in Figure 7, It is seen that that the upper bound estimation is not deterministic, though interval estimation successfully covers the true values (and point estimations as shall be discussed below) of surface concentration. Nevertheless, further exploration of seasonal changes of HCHO in some key areas in section 3.3 could basically explain that seasonal variations of surface HCHO may contribute to the majority of the uncertainty of interval estimation.



(a) (b)

Figure 8. (a) The density scatter plot of upper bound (x-axis) against lower-bound (y-axis). (b) scatter plot Relation between point estimation (x-axis) and predicted intervals (y-axis). , Red points for upper bounds and grey points for lower bounds). The black line is the fitted line.

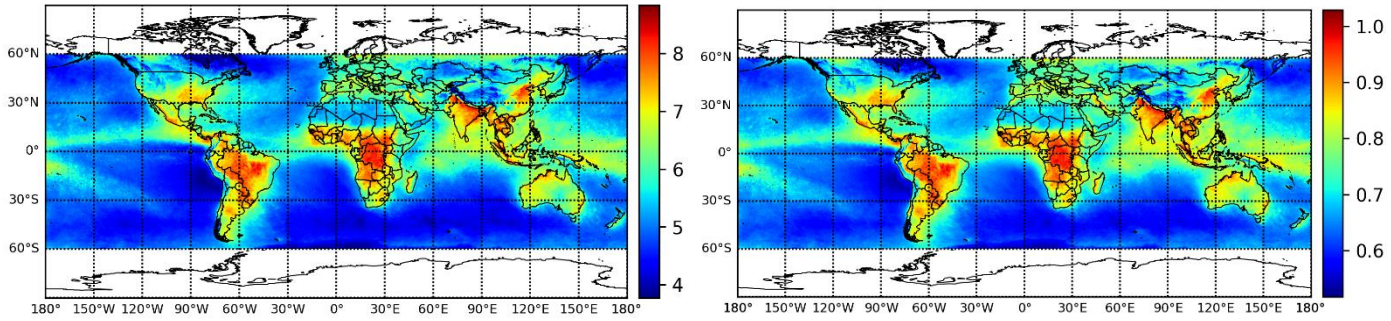
Global distribution of the estimated upper and lower bounds are given in Figure 8(a). It shows that the upper and lower bounds generally share the same global pattern, though with different magnitude, with a range of between 3.77 and 8.83 μg/m³ for upper bounds and from 0.52 to 1.03 μg/m³ for lower bounds. The interval length[6] of 90% confidence interval is 4.77 μg/m³.

As shown in Figure 8(b), the upper and lower bounds share a significantly positive relation, and a majority of predicted interval are in the regions of 0.5~1.0 μg/m³ for lower bound and 5~10 μg/m³ for upper bound. y = x aims for indicating the relative positions of

[6] Interval length = Upper Bound – Lower Bound

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

true values in the predicted intervals. In addition, the predicted intervals can basically cover true values.



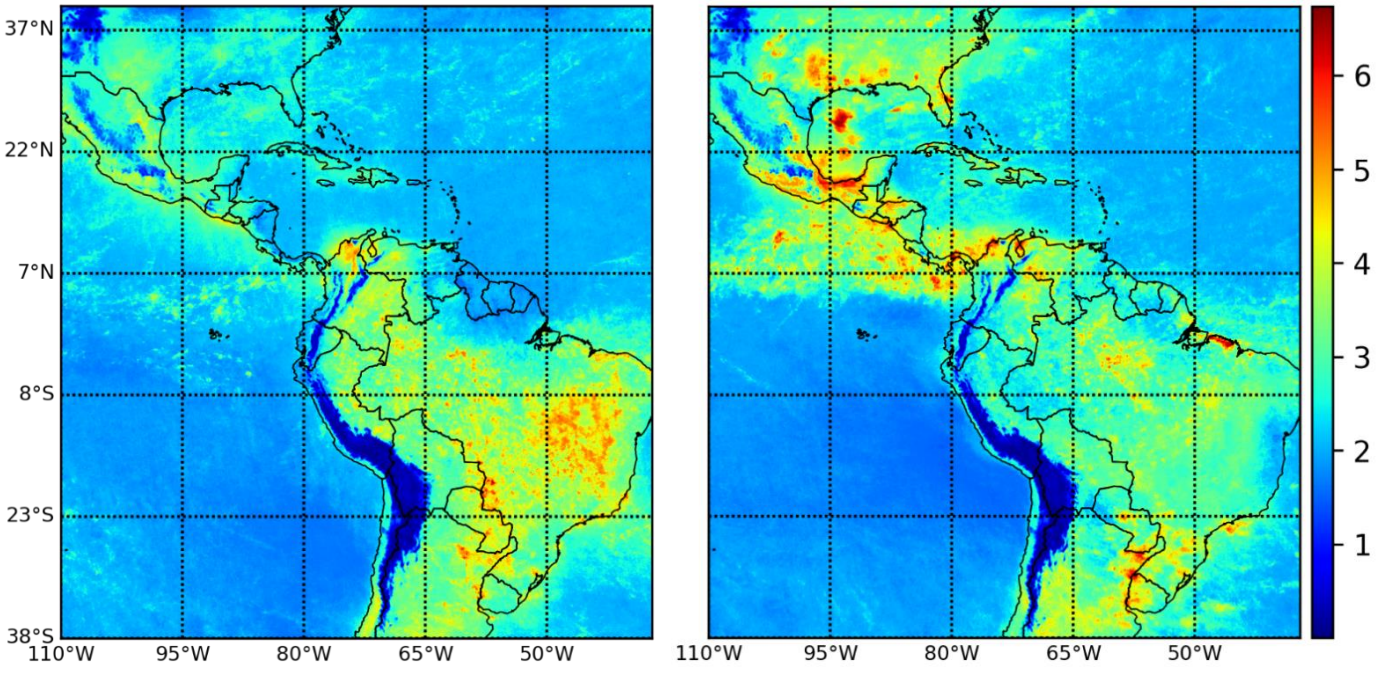
(a) (b)

Figure 9. (a) Distribution of upper bound from 90% surface concentration estimation, (b) Distribution of lower bound from 90% surface concentration estimation. (unit: μg/m3)

3.3. Seasonal Changes of HCHO in Some Key regions

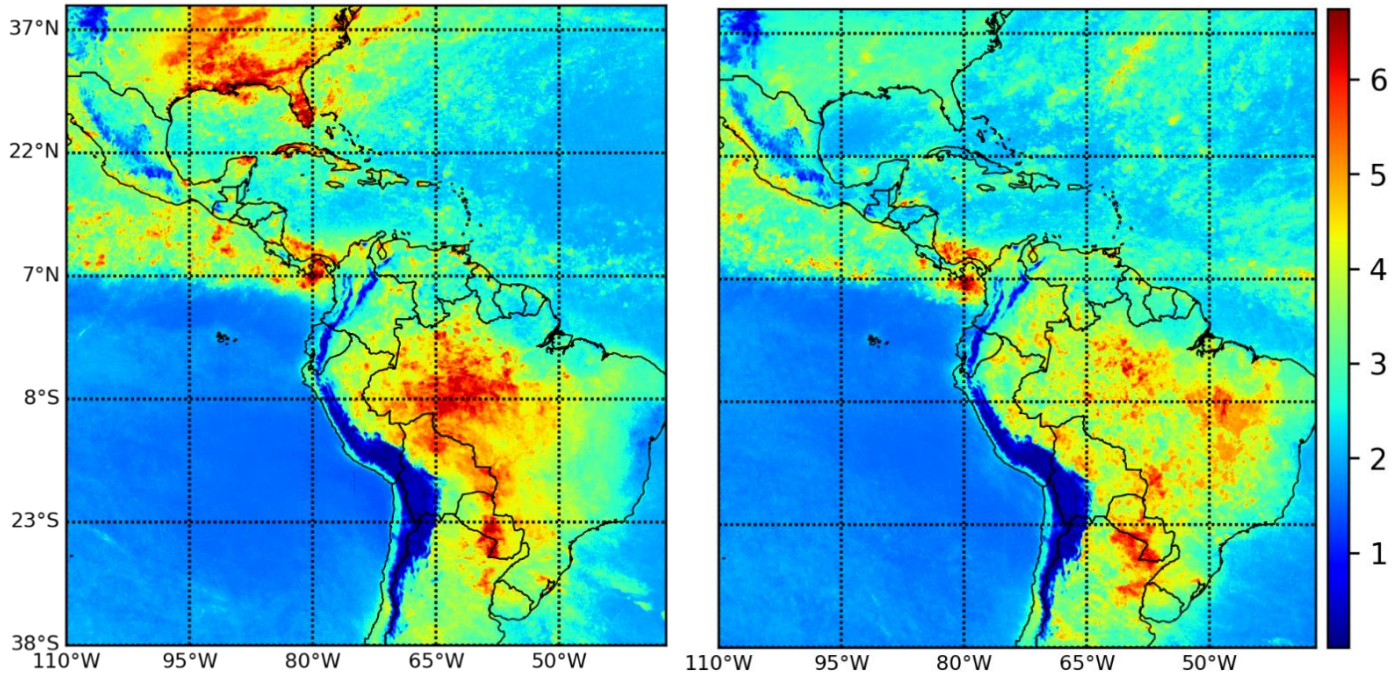
To better understand the seasonal variation of surface HCHO, the distribution pattern of four typical months of some key areas where surface concentration is relatively high are analyzed.

America. Figure 9 shows the surface concentration of February, May, August, and November in South America and around Caribbean Sea. Amazon Basin, Paraguay, and Eastern Central America have a high HCHO surface concentration in November and February, while the south-east coast of U.S. has the highest concentration in November and are almost free from HCHO pollution in February and May. The Andes Mountains has a significantly low concentration, with a value of less than 0.5 μg/m^3.



(a) (b)

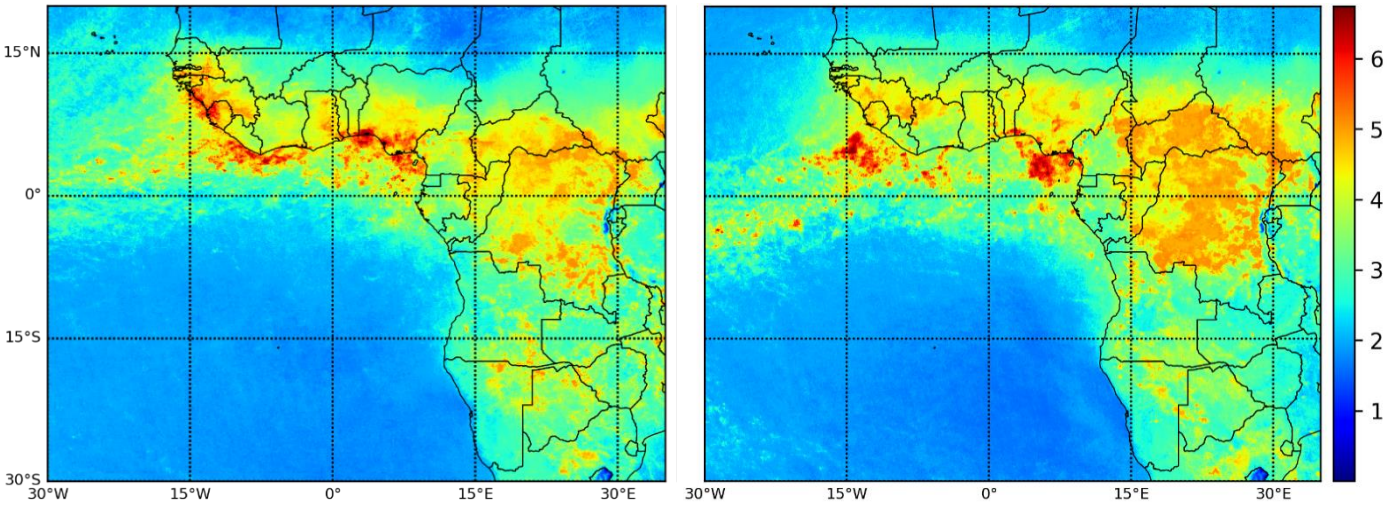
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2



(c) (d)

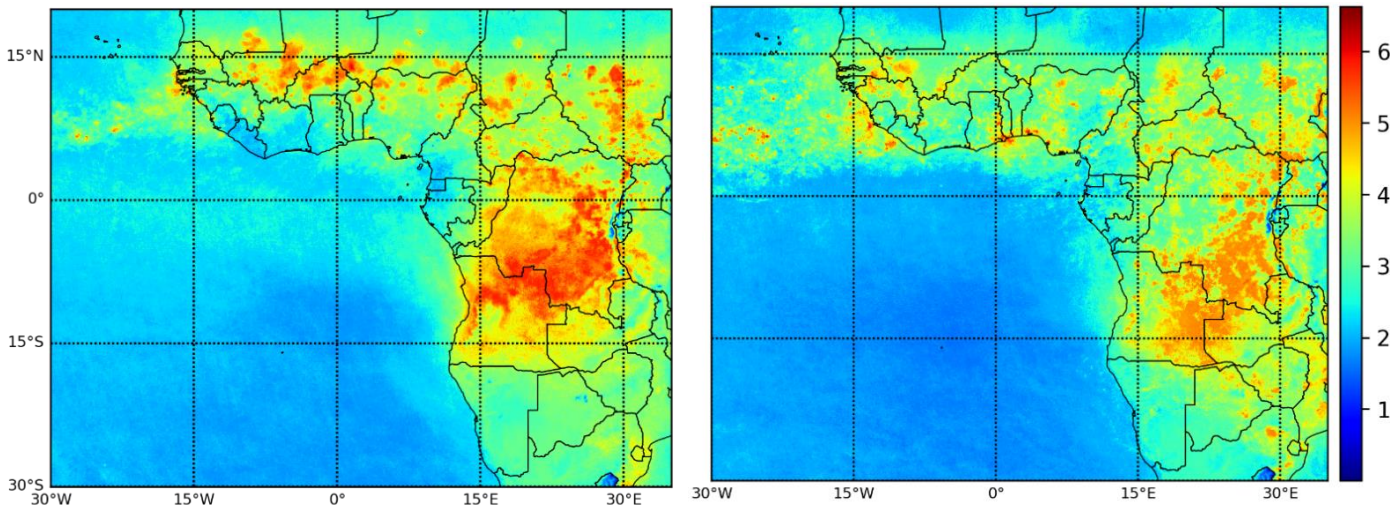
Figure 10. Surface concentration of HCHO in Central and South America in some typical months. (a): February, (b): May, (c): August, (d): November. (unit: μg/m^3)

Africa. As shown in Figure 10, there are two regions in Africa whose HCHO surface concentration is relatively high. One is in the south of R. D. Congo around the city of Kolwezi, a mining center with a humid subtropical climate. The surface concentration of HCHO here reaches its maximum in February. The other pollution belt stretches along the Gulf of Guinea, which is famous for its rainforest climate.



(a) (b)

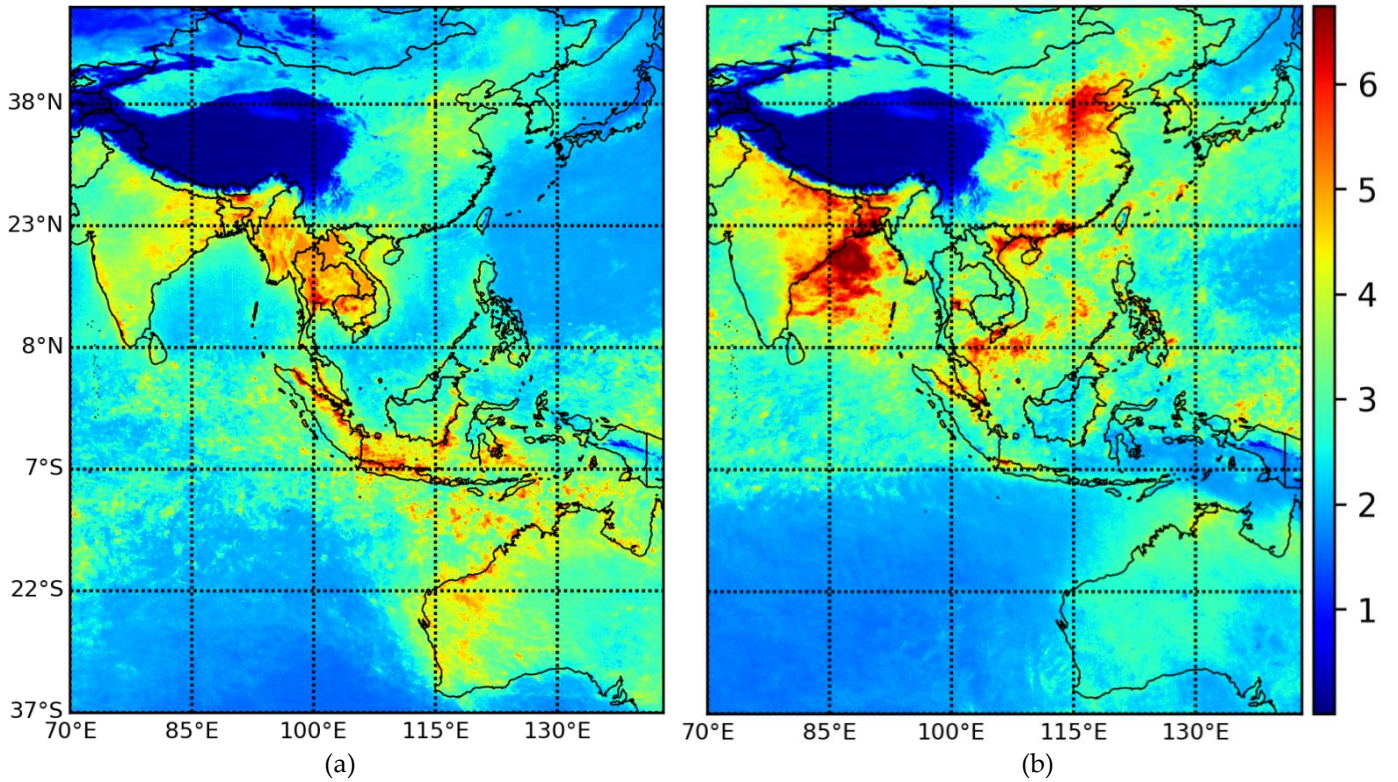
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2



(c) (d)

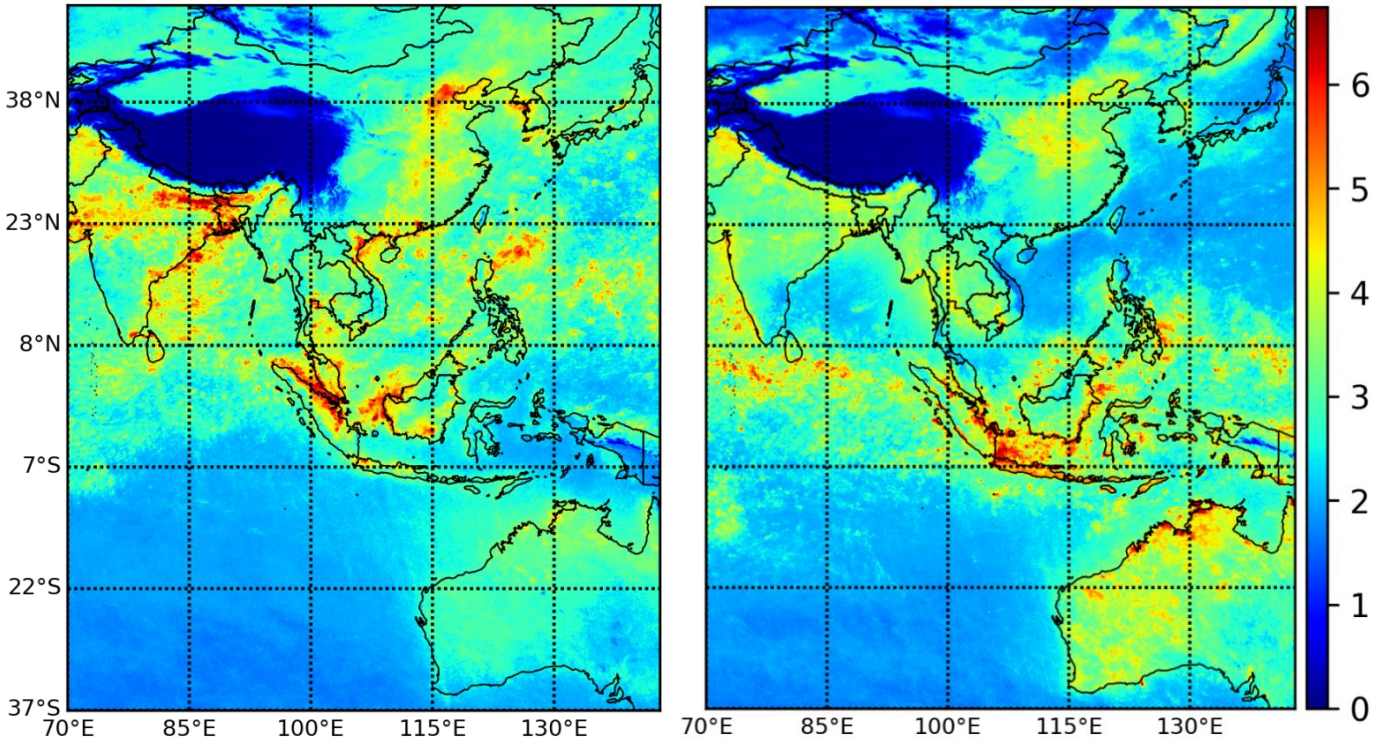
Figure 11. Surface concentration of HCHO in Central and South Africa in some typical months. (a): January, (b): April, (c): July, (d): October. (unit: μg/m3)

Indo-Pacific. As shown in Figure 11, there are several regions in Indo-Pacific whose time of occurrence of high HCHO concentrations differ from each other and may result from several different reasons. First, Malaysia and Indonesia islands, both are abundant in rainforest, have a relatively high concentration all the year round, and reach their maximum in December. Second, the surface HCHO concentration of China-Indochina Peninsula reaches the maximum in June, while the high concentration center moves to the Gulf of Tonkin and Pearl River Delta in the latter half of the year. Third, the Bay of Bengal and the coasts nearby witness a high concentration in September. Forth, the Beijing-Tianjin-Hebei Urban Agglomeration (BTH), which has no rainforest distribution but mass population and economic activities, also has a high HCHO concentration through most of the year. The concentration there reaches its maximum around September in 2019.



(a) (b)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2



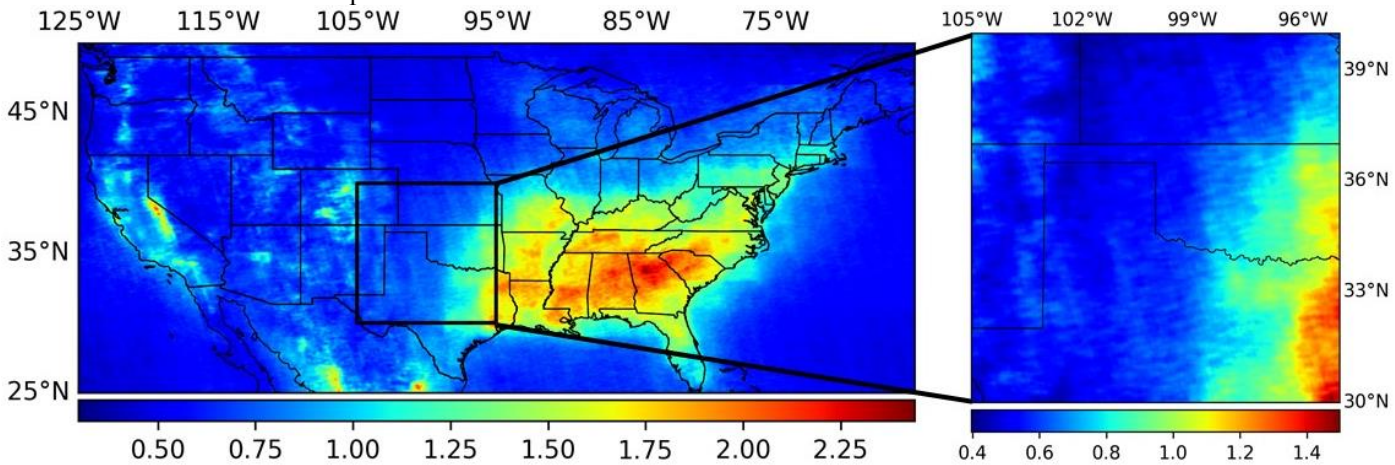
(c) (d)

Figure 12. Surface concentration of HCHO in Indo-Pacific region in some typical months. (a): March, (b): June, (c): September, (d): December. (unit: μg/m3)

4. Discussion

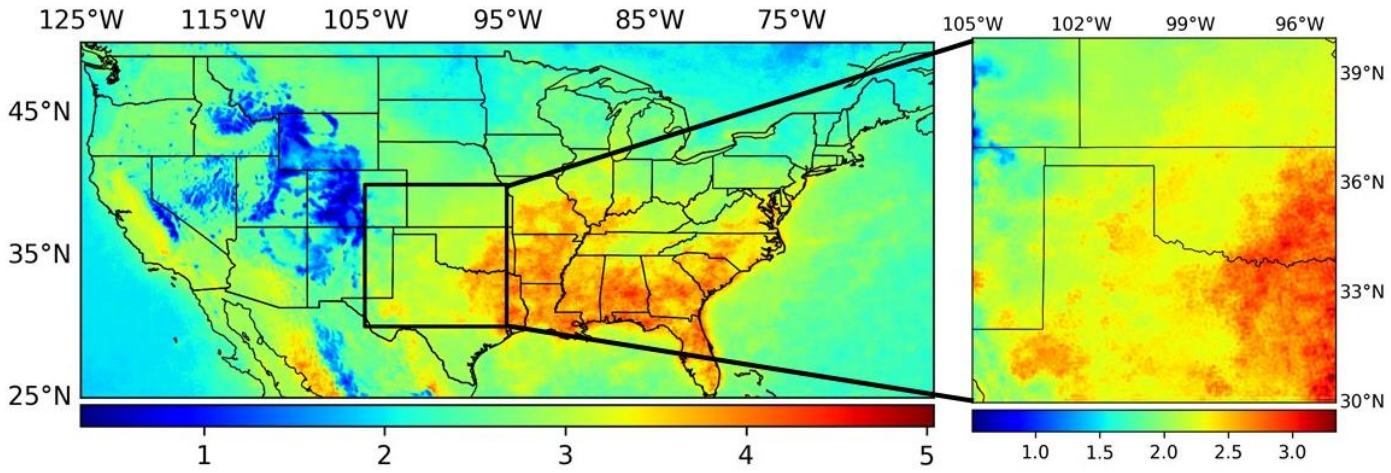
4.1. Consistency and innovativeness

It is clear that the global surface distribution of HCHO with point and interval estimation is able to be obtained successfully by using neural network models described above. As shown in Figure 13, the results obtained through machine learning technique are generally consistent with results from the previous works which is obtained by combining OMI total column HCHO concentration with GEOS-Chem model from 2005 to 2016, but with less noise across the satellite track. It is seen from the blowup box, which is shown on the right of the figure, corresponding to each result respectively, results from previous studies bear a strip across the satellite track, but the new results from this study does not. In addition, the estimation results in this study show some reversal trend in the Cordillera mountains area. Future validation may be needed for this case. However, since this difference occurs in places where population is sparse, it is not likely to have a perceivable influence on the estimation of cancer risks.



(a)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2



(b)

Figure 13. (a) Average surface distribution of HCHO over the U.S. from 2005 to 2016, Zhu et al [27] (b) Average surface distribution of HCHO in 2019 over the U.S. predicted by our model. (unit: ppb)

The result of global surface concentration estimation of 2019 gives a closer look at the global distribution pattern of HCHO. Obviously, HCHO tends to prevail on the plain of the continent, instead of on the ocean or on high altitude areas. According to previous study, this can be attributed to the scarceness of VOC sources like chemical industry, combustion and rainforest, which are common precursors of the free radical reaction of HCHO production [46-48]. By mapping the distribution of HCHO, two kinds of sources around the world can be distinguished preliminarily. One is plant-related, including Amazon, South East Asia and Gulf of Guinea, the other is human-related, including North China Plain and Pearl River Delta [49,50]. More work is needed to accurately identify the source of these HCHO-polluted areas.

In addition, we introduce the interval estimation of neural network model in the conversion from VCD to global surface concentration of HCHO for the first time, increasing the credibility of the model by providing uncertainty information. This new idea can make up for the deficiency of inexplicability of the neural network model [51], thus being useful for the application of neural network models into the field of estimating atmospheric pollutant or health risk in the future.

4.2. Limitations and potential improvements

Despite the consistency and innovativeness mentioned before, the shortage of in-situ data is also hindering the further improvement of the model accuracy. On one hand, the existing HCHO in-situ concentration data is seriously insufficient in both spatial and temporal dimensions. Only the United States monitors HCHO in-situ concentration routinely. Even if ATom data are also adopted, in-situ concentration data in low latitude regions is still sparse, which may lead to estimation bias in the low latitude areas such as Asia and Africa. On another hand, it is also difficult to reach a better result by adding more covariates into our model. Experiments with additional covariates input, such as latitude and months, have failed with degenerated or overfitting outputs. In addition, the large gap between true values and the upper bounds from our interval estimation model may suggest a heterogeneous in-situ concentration of HCHO distribution in different months or seasons, Since the model is required to give the interval estimations on the scale of a whole year, rather than on a fine time scale. The seasonal changes of HCHO in some key areas as discussed in section 3.3 has also shown this phenomenon directly.

Therefore, as more HCHO in-situ monitoring network develop, a larger amount of data from more diverse sites could enable scientists to adopt a careful designation of temporal data input and could help give a better estimation towards in-situ concentration of HCHO. Meanwhile, with more Sentinel-5P data accumulating over time, the model in this study can take more factors, including latitude and seasons, into consideration, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2021 doi:10.20944/preprints202105.0595.v2

could provide more precise estimation of a global scale health risk and economic loss based on specific regions and seasons. Besides the significance of the health risk, the results from this study can also help research on the generation of photochemical pollution, the concentration of VOC, NO_2 and other photochemical reaction related pollutants.

4.3. Health risk of HCHO in major cities

Table 5. Potential number of cancer cases in typical cities if HCHO surface concentration remains the amount of 2019

City Name	Patients per million	Population	Number of cases
Jakarta, Indonesia	80.34	32,275,000	2,593
Singapore	75.79	5,930,000	449
Kuala Lumpur, Malaysia	72.93	7,820,000	570
Dhaka, Bangladesh	71.63	17,425,000	1,248
Lagos, Nigeria	71.37	13,910,000	993
Bangkok, Thailand	70.46	15,975,000	1,126
Shijiazhuang, China	69.94	3,765,000	263
Ho Chi Minh City, Vietnam	68.51	10,690,000	732
Kolkata, India	68.38	15,095,000	1,032
Beijing, China	67.99	21,250,000	1,445
Patna, India	65.91	2,320,000	153
Ha Noi, Vietnam	65.78	8,140,000	535
Guangzhou, China	65.00	19,965,000	1,298
Tianjin, China	63.57	13,655,000	868
Manaus, Brazil	58.50	2,020,000	118
Houston, U.S.	54.86	6,285,000	345
Freetown, Sierra Leone	53.95	1,755,000	95
Kolwezi, R. D. Congo	49.53	515,000	26

HCHO, as one of the most important carcinogens in the outdoor environment [2], draws little attention due to the lack of ground measurements of HCHO in most countries and regions for a long time, leading to the shortage of knowledge about health and economic loss. Even if the vertical column density of HCHO is currently available and does settle parts of concern about these issues, it is the ground level HCHO concentration that reflects the actual amount of concentration people are exposed to.

Taking 2019 as an example, it is assumed that the HCHO concentration is always the same as this year. According to the inhalation unit risk estimate from EPA and population data[4,54]. Health risks in main high-risk cities are calculated and given in Table 5. It is indicated that more than a thousand people have the potential to get cancer due to exposure to HCHO in Jakarta, Dhaka, Bangkok, Kolkata, Beijing and Guangzhou. Jakarta has the highest potential patients due to exposure, with a number of up to 2593. Jakarta, Singapore, Kuala Lumpur, Dhaka and Lagos are the most prevalent cities, with 80.34, 75.79, 72.93, 71.63, 71.37 potential patients per million. The main cities with high health risk concentration in Southeast Asia, which was previously neglected by the academia, may become the next hotspots for research in HCHO pollution and health risk.

5. Conclusions

With the benefit from a quality-driven interval estimation algorithm designed for neural network, we are able to derive the confidence interval and a precise point

estimation of 2019 global surface HCHO on different confidence levels with a limited amount of data. By mapping the HCHO surface concentration distribution, we found that Southeast Asia, North China, Central and Western Africa, and the rainforest area of Latin America have a relatively more serious HCHO pollution than the rest regions. Major cities in these regions, such as Bangkok, Beijing, Guangzhou, Singapore, have an annual concentration over 5.00 μg/m^3. The health effects from such high levels of HCHO pollution deserve more attention from the academia and governments.

Our work paves a way for research on formaldehyde-related cancers, and provides guidance for policy making and insurance pricing. To the best of our knowledge, we are the first to map the global distribution of HCHO and provide insights on its potential health risks. With more HCHO VCD data from Sentinel-5P accumulated, the surface concentration of HCHO dataset covering a longer period of time will be generated, which will help for better assessment of the global risk distribution of formaldehyde-related cancers.

Author Contributions: Conceptualization, W.W.; Methodology, J.G. and Y.D.; Software, J.G., G.L., B.J. and Y.D.; Validation, B.J.; Formal Analysis, J.G. and B.J.; Investigation, B.J. and J.G.; Resources, B.J. and J.G.; Data Curation, B.J. and Y.D.; Writing, B.J., J.G. Y.D. and P.C.; Visualization, B.J. and Y.D.; Supervision, W.W.; Project Administration, W.W.; Funding Acquisition, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (17XNLG09).

Data Availability Statement: The data presented in this study are openly available in https://s5phub.copernicus.eu/dhus/#/home for Sentinel-5P VCD Data; https://www.epa.gov/outdoor-air-quality-data for HAPs ground monitoring data; https://drive.google.com/drive/folders/0B_J08t5spvd8VWJPbTB3anNHamc for Global DEM Data. ATom flight data available in a publicly accessible repository [40,41]. The input data of our model is available in [https://drive.google.com/file/d/1tovF73HogGNEXC1i_jBbnVRHlm1n-ZT_/view?usp=sharing]. And the data presented in this study are available in [https://drive.google.com/file/d/10A2VIEHm22DF_gyCufV-pbgUdYYhNJKf/view?usp=sharing].

Acknowledgments: We would like to appreciate Lei Zhu for providing us with technical support about searching available data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tesfaye, S.; Hamba, N.; Gerbi, A.; Neger, Z. Oxidative Stress and Carcinogenic Effect of Formaldehyde Exposure: Systematic Review & Analysis. *Endocrinol Metab Syndr* **2020**, 9, 319.
2. Scheffe, R.D.; Strum, M.; Phillips, S.B.; Thurman, J.; Eyth, A.; Fudge, S.; Morris, M.; Palma, T.; Cook, R. Hybrid Modeling Approach to Estimate Exposures of Hazardous Air Pollutants (HAPs) for the National Air Toxics Assessment (NATA). *Environ Sci Technol* **2016**, 50, 12356-12364, doi:10.1021/acs.est.6b04752.
3. Blair, A.; Saracci, R.; Stewart, P.A.; Hayes, R.B.; Shy, C. Epidemiologic evidence on the relationship between formaldehyde exposure and cancer. *Scand J Work Environ Health* **1990**, 16, 381-393, doi:10.5271/sjweh.1767.
4. Agency, E.P. Formaldehyde. https://www.epa.gov/sites/production/files/2016-09/documents/formaldehyde.pdf (accessed on 5.21 2021).
5. Jin, X.; Fiore, A.; Boersma, K.F.; Smedt, I.D.; Valin, L. Inferring Changes in Summertime Surface Ozone–NO x–VOC Chemistry over US Urban Areas from Two Decades of Satellite and Ground-Based Observations. *Environ Sci Technol* **2020**, 54, 6518-6529.
6. Javed, Z.; Liu, C.; Khokhar, M.F.; Tan, W.; Liu, H.; Xing, C.; Ji, X.; Tanvir, A.; Hong, Q.; Sandhu, O. Ground-based MAX-DOAS observations of CHOCHO and HCHO in Beijing and Baoding, China. *Remote Sens-Basel* **2019**, 11, 1524.
7. Liu, Y.; Tang, Z.; Abera, T.; Zhang, X.; Hakola, H.; Pellikka, P.; Maeda, E. Spatio-temporal distribution and source partitioning of formaldehyde over Ethiopia and Kenya. *Atmos Environ* **2020**, 237, 117706.
8. Kaiser, J.; Wolfe, G.M.; Bohn, B.; Broch, S.; Fuchs, H.; Ganzeveld, L.N.; Gomm, S.; Häseler, R.; Hofzumahaus, A.; Holland, F., et al. Evidence for an unidentified non-photochemical ground-level source of formaldehyde in the Po Valley with potential implications for ozone production. *Atmos Chem Phys* **2015**, 15, 1289-1298, doi:10.5194/acp-15-1289-2015.
9. Green, J.R.; Fiddler, M.N.; Fibiger, D.L.; McDuffie, E.E.; Aquino, J.; Campos, T.; Shah, V.; Jaeglé, L.; Thornton, J.A.; DiGangi, J.P. Wintertime Formaldehyde: Airborne Observations and Source Apportionment Over the Eastern United States. *Journal of Geophysical Research: Atmospheres* **2021**, 126, e2020J-e33518J.
10. Geddes, J. in *Impacts of Interannual Variability in Biogenic VOC Emissions near Transitional Ozone Production Regimes*, AGU Fall Meeting Abstracts, 2017; **2017**; pp A54B-A56B.

11. Gratsea, M.; Vrekoussis, M.; Richter, A.; Wittrock, F.; Schönhardt, A.; Burrows, J.; Kazadzis, S.; Mihalopoulos, N.; Gerasopoulos, E. Slant column MAX-DOAS measurements of nitrogen dioxide, formaldehyde, glyoxal and oxygen dimer in the urban environment of Athens. *Atmos Environ* **2016**, 135, 118-131, doi:10.1016/j.atmosenv.2016.03.048.
12. EPA. Outdoor air quality data. https://www.epa.gov/outdoor-air-quality-data (accessed on 3.21 2021).
13. Xin, T.; Jin, X.; Pin-hua, X.; Ang, L.; Zhao-kun, H.; Xiao-mei, L.; Bo, R.; Zi-yang, W. Retrieving Tropospheric Vertical Distribution in HCHO by Multi-Axis Differential Optical Absorption Spectroscopy. *Spectrosc Spect Anal* **2019**, 39, 2325-2331.
14. Chance, K.; Palmer, P.I.; Spurr, R.J.; Martin, R.V.; Kurosu, T.P.; Jacob, D.J. Satellite observations of formaldehyde over North America from GOME. *Geophys Res Lett* **2000**, 27, 3461-3464.
15. Wang, Y.; Beirle, S.; Lampel, J.; Koukouli, M.; Smedt, I.D.; Theys, N.; Li, A.; Wu, D.; Xie, P.; Liu, C. Validation of OMI, GOME-2A and GOME-2B tropospheric NO 2, SO 2 and HCHO products using MAX-DOAS observations from 2011 to 2014 in Wuxi, China: investigation of the effects of priori profiles and aerosols on the satellite products. *Atmos Chem Phys* **2017**, 17, 5007-5033.
16. Jin, X.; Fiore, A.; Boersma, K.F.; Smedt, I.D.; Valin, L. Inferring Changes in Summertime Surface Ozone–NO x–VOC Chemistry over US Urban Areas from Two Decades of Satellite and Ground-Based Observations. *Environ Sci Technol* **2020**, 54, 6518-6529.
17. Zhu, L.; Mickley, L.J.; Jacob, D.J.; Marais, E.A.; Sheng, J.; Hu, L.; Abad, G.G.; Chance, K. Long-term (2005–2014) trends in formaldehyde (HCHO) columns across North America as seen by the OMI satellite instrument: Evidence of changing emissions of volatile organic compounds. *Geophys Res Lett* **2017**, 44, 7079-7086.
18. Vigouroux, C.; Langerock, B.; Bauer Aquino, C.A.; Blumenstock, T.; Cheng, Z.; De Mazière, M.; De Smedt, I.; Grutter, M.; Hannigan, J.W.; Jones, N. TROPOMI–Sentinel-5 Precursor formaldehyde validation using an extensive network of ground-based Fourier-transform infrared stations. *Atmos Meas Tech* **2020**, 13, 3751-3767.
19. Veefkind, J.P.; Aben, I.; McMullan, K.; Förster, H.; de Vries, J.; Otter, G.; Claas, J.; Eskes, H.J.; de Haan, J.F.; Kleipool, Q., et al. TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sens Environ* **2012**, 120, 70-83, doi:10.1016/j.rse.2011.09.027.
20. Millet, D.B.; Jacob, D.J.; Boersma, K.F.; Fu, T.; Kurosu, T.P.; Chance, K.; Heald, C.L.; Guenther, A. Spatial distribution of isoprene emissions from North America derived from formaldehyde column measurements by the OMI satellite sensor. *Journal of Geophysical Research* **2008**, 113, doi:10.1029/2007JD008950.
21. Zhang, Y.; Li, R.; Min, Q.; Bo, H.; Fu, Y.; Wang, Y.; Gao, Z. The controlling factors of atmospheric formaldehyde (HCHO) in Amazon as seen from satellite. *Earth Space Sci* **2019**, 6, 959-971.
22. Curci, G.; Palmer, P.I.; Kurosu, T.P.; Chance, K.; Visconti, G. Estimating European volatile organic compound emissions using satellite observations of formaldehyde from the Ozone Monitoring Instrument. *Atmos Chem Phys* **2010**, 10, 11501-11517.
23. Biswas, M.S.; Choudhury, A.D. Impact of COVID-19 Control Measures on Trace Gases (NO2, HCHO and SO2) and Aerosols over India during Pre-monsoon of 2020. *Aerosol Air Qual Res* **2021**, 20.
24. Sun, W.; Zhu, L.; De Smedt, I.; Bai, B.; Pu, D.; Chen, Y.; Shu, L.; Wang, D.; Fu, T.M.; Wang, X. Global significant changes in formaldehyde (HCHO) columns observed from space at the early stage of the COVID-19 pandemic. *Geophys Res Lett* **2021**, 48, 2e-20e.
25. Yu, T.; Wang, W.; Ciren, P.; Zhu, Y. Assessment of human health impact from exposure to multiple air pollutants in China based on satellite observations. *Int J Appl Earth Obs* **2016**, 52, 542-553.
26. Schroeder, J.R.; Crawford, J.H.; Fried, A.; Walega, J.; Weinheimer, A.; Wisthaler, A.; Müller, M.; Mikoviny, T.; Chen, G.; Shook, M. Formaldehyde column density measurements as a suitable pathway to estimate near-surface ozone tendencies from space. *Journal of Geophysical Research: Atmospheres* **2016**, 121, 13, 13-88, 112.
27. Zhu, L.; Jacob, D.J.; Keutsch, F.N.; Mickley, L.J.; Scheffe, R.; Strum, M.; González Abad, G.; Chance, K.; Yang, K.; Rappenglück, B., et al. Formaldehyde (HCHO) As a Hazardous Air Pollutant: Mapping Surface Air Concentrations from Satellite and Inferring Cancer Risks in the United States. *Environ Sci Technol* **2017**, 51, 5650-5657, doi:10.1021/acs.est.7b01356.
28. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In ACM: New York, **2017**; Vol. 60, pp 84-90.
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans Pattern Anal Mach Intell* **2017**, 39, 1137-1149, doi:10.1109/TPAMI.2016.2577031.
30. Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; Zhang, L. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *Ieee T Image Process* **2017**, 26, 3142-3155, doi:10.1109/TIP.2017.2662206.
31. Ian J. Goodfellow, J.P.; Mehdi, M.B.X.D.; Ozair, S.; Aaron, C.Y.B. Generative Adversarial Nets. In 2014.
32. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Trans Pattern Anal Mach Intell* **2021**, PP, doi:10.1109/TPAMI.2021.3054775.
33. MacKay, D.J. A Practical Bayesian Framework for Backpropagation Networks. *Neural Comput* **1992**, 4, 448-472.
34. Tibshirani, R. A Comparison of Some Error Estimates for Neural Network Models. *Neural Comput* **1996**, 8, 152-163.
35. Heskes, T. Practical confidence and prediction intervals. In 1997.
36. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. **2015**.
37. Khosravi, A.; Nahavandi, S.; Creighton, D.; Atiya, A.F. Lower Upper Bound Estimation Method for Construction of Neural Network-Based Prediction Intervals. *IEEE Transactions on Neural Networks* **2011**, 22, 337-346, doi:10.1109/TNN.2010.2096824.
38. Pearce, T.; Zaki, M.; Brintrup, A.; Neely, A. High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach. In **2018**.
39. Apituley, A.; Pedergnana, M.; Sneep, M.; Pepijn, J.; Loyola, D.; Landgraf, J.; Borsdorff, T. *Sentinel-5 precursor/TROPOMI Level 2 Product User Manual Carbon Monoxide document number*; **2018**; p.
40. Williamson, C.; Kupc, A.; Wilson, J.; Gesler, D.W.; Reeves, J.M.; Erdesz, F.; McLaughlin, R.; Brock, C.A. Fast time response measurements of particle size distributions in the 3–60 nm size range with the nucleation mode aerosol size spectrometer. *Atmos Meas Tech* **2018**, 11, 3491-3509, doi:10.5194/amt-11-3491-2018.
41. Brock, C.A.; Williamson, C.; Kupc, A.; Froyd, K.D.; Erdesz, F.; Wagner, N.; Richardson, M.; Schwarz, J.P.; Gao, R.; Katich, J.M., et al. Aerosol size distributions during the Atmospheric Tomography Mission (ATom): methods, uncertainties, and data products. *Atmos Meas Tech* **2019**, 12, 3081-3099, doi:10.5194/amt-12-3081-2019.
42. Hanisco, T.F.; Bian, H.; Nicely, J.M.; Pan, X.; Hannun, R.A.; St. Clair, J.M.; Wolfe, G.M. ATom: L2 Measurements of In Situ Airborne Formaldehyde (ISAF). In ORNL Distributed Active Archive Center: **2019**.
43. Https Orcidorg, Y.W.; Https Orcidorg, S.D.; Https Orcidorg X, S.D.; Böhnke, S.; Isabelle, D.S.H.O.; Dickerson, R.R.; Dong, Z.; He, H.; Https Orcidorg X, Z.L.; Li, Z., et al. Vertical profiles of NO2, SO2, HONO, HCHO, CHOCHO and aerosols derived from MAX-DOAS measurements at a rural site in the central western North China Plain and their relation to emission sources and effects of regional transport. *Atmos Chem Phys* **2019**, 19, 5417-5449, doi:10.5194/acp-19-5417-2019.

44. Farr, T.G.; Edward, P.A.R.; Kobrick, M.; Rodriguez, M.P.E.; Shaffer, S.; Umland, J.S.J.; Burbank, D.; Alsdorf, A.D. THE SHUTTLE RADAR TOPOGRAPHY MISSION. *Rev Geophys* **2007**, 45.
45. Ioffe, S.; Szegedy, C. in *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, International conference on machine learning, 2015; PMLR: **2015**; pp 448-456.
46. Starn, T.K.; Shepson, P.B.; Bertman, S.B.; Riemer, D.D.; Zika, R.G.; Olszyna, K. Nighttime isoprene chemistry at an urban-impacted forest site. *Journal of Geophysical Research: Atmospheres* **1998**, 103, 22437-22447.
47. Guo, S.; Wen, S.; Wang, X.; Sheng, G.; Fu, J.; Hu, P.; Yu, Y. Carbon isotope analysis for source identification of atmospheric formaldehyde and acetaldehyde in Dinghushan Biosphere Reserve in South China. *Atmos Environ* **2009**, 43, 3489-3495, doi:10.1016/j.atmosenv.2009.04.041.
48. Kean, A.J.; Grosjean, E.; Grosjean, D.; Harley, R.A. On-Road Measurement of Carbonyls in California Light-Duty Vehicle Emissions. *Environ Sci Technol* **2001**, 35, 4198-4204, doi:10.1021/es010814v.
49. Luecken, D.J.; Napelenok, S.L.; Strum, M.; Scheffe, R.; Phillips, S. Sensitivity of Ambient Atmospheric Formaldehyde and Ozone to Precursor Species and Source Types Across the United States. *Environ Sci Technol* **2018**, 52, 4668-4675, doi:10.1021/acs.est.7b05509.
50. Zhu, S.; Li, X.; Cheng, T.; Yu, C.; Wang, X.; Miao, J.; Hou, C. Comparative analysis of long-term (2005-2016) spatiotemporal variations in high-level tropospheric formaldehyde (HCHO) in Guangdong and Jiangsu Provinces in China. *Journal of remote sensing* **2019**, 01, 137-154.
51. Nourani, V.; Paknezhad, N.J.; Tanaka, H. Prediction Interval Estimation Methods for Artificial Neural Network (ANN)-Based Modeling of the Hydro-Climatic Processes, a Review. *Sustainability-Basel* **2021**, 13, 1633.
52. Stimac, J.P. Atmospheric Circulation. https://www.ux1.eiu.edu/~jpstimac/1400/circulation.html (accessed on 5.21 2021).
53. Key, J. Climate variability, extreme and change in the western tropical pacific 2019. In **2020**; Vol. 11, p 220.
54. Demographia. World Urban Areas. https://web.archive.org/web/20180503021711/http://www.demographia.com/db-worldua.pdf (accessed on 5.21 2021).