 doi:10.20944/preprints202105.0595.v1

*Article*

# Mapping Global Surface HCHO Distribution with Confidential Interval by Satellite Observation

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

[1] Center for Spatial Information, School of Environment and Natural Resources, Renmin University of China, Beijing 100872, China; 2018200684@ruc.edu.cn (B, J); 2018200694@ruc.edu.cn (J, G); 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
* Correspondence: wenw@ruc.edu.cn; Tel.: (86-10) 88893061

**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 researches on outdoor HCHO pollution. Traditional methods are either too naïve or data-demanding for a global scale research. To alleviate this issue, we trained two fully-connected neural networks respectively for deriving point and interval estimation of surface HCHO concentration in 2019, where 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, North China, South-east Asia, Bay of Bengal, Central and Western Africa are among the highest. Our study makes up for the global shortage of surface HCHO monitoring and helps people have a clearer understanding of surface concentration distribution of HCHO. In addition, with the help of quality-driven algorithm, interval estimation of surface HCHO concentration is believed to bring confidence to our results. As an early work adopting interval estimation in AI-driven atmospheric pollutant research and the first to map global HCHO surface distribution, our paper will pave way for rigorous study on global ambient HCHO health risk and economic loss, thus providing 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 U.S. Environmental Protection Agency (EPA) to be one of the most important carcinogens in outdoor air among 187 harmful air pollutants (HAP) [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 compounds in the atmosphere, HCHO is one of the major volatile organic compounds (VOCs) and pollutants in 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 can denote 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 systematically and comprehensively by using traditional ground method 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

 doi:10.20944/preprints202105.0595.v1



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 serial and large-scale dynamics, but also avoid many interference factors. Many satellites have been recording HCHO vertical column density (VCD) [13], which provides data foundation 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 in the world. The imaging width is 2600km, covering nearly all parts of the world every day [19]. However, remote sensing only provides the column concentration due to the limit of satellite vertical data collection. 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 concentration distribution from the global perspective is growing more urgent. Previous researchers have used 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 used $R^2$ to assess the relationship between vertical column density and ground in-situ concentration [26]. However, these methods are either too naive or too limited to specific pollutants. In the few existing study addressing HCHO surface concentration, GEOS-Chem model was adopted to utilize remote sensing monitoring data [27]. Nevertheless, based on atmospheric transportation model, it needs numerous input parameters, which impedes its appliance to a global scale, where surface monitoring data are scarce. Therefore, our concern is to derive the global surface HCHO concentration distribution based on vertical air column of HCHO from satellites, with quite limited in-situ HCHO concentration.

Neural network, a powerful machine learning algorithm, has gained its reputation for revealing hidden patterns beneath data with starling accuracy in various fields, such as image classification [28], object detection [29], image denoising [30], image synthesis [31], person re-identification [32], etc. However, vanilla neural network does not assign confidential level nor confidential interval, which is necessary for scientific estimation and public policy decision, to its point estimation results. 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]. But these methods are either computational demanding or require strong assumptions. Quality-driven (QD) method, a method based on LUBE to derive confidential 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 shrinkages the average confidence interval length up to 10%, compared with previous works [38]. So, 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 confidential level.

The potential health impact of HCHO but lack of global monitoring data calls for an efficient way to get better understanding of global HCHO surface distribution with limited data. In this paper, we, for the first time, derived the global surface concentration of HCHO in 2019 by feeding TROPOMI VCD data and limited surface HCHO concentration data into neural network models. We also captured the seasonal changes of key areas and gave interval estimation of surface HCHO by QD method. As an early work adopting interval estimation in AI-driven atmospheric pollutant research and the first to map global

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

 doi:10.20944/preprints202105.0595.v1



HCHO surface distribution, our paper will pave the way for rigorous study on global ambient HCHO health risk and economic loss, thus providing basis for pollutant controlling policies worldwide.

## 2. Materials and Methods

### *2.1. Materials*

#### 2.1.1 Sentinel-5P VCD Data

The data of vertical column density 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 the most advanced atmospheric monitoring spectrometer with the highest spatial resolution in the world. The satellite covers all parts of the world every day with an imaging resolution of 7km×7km. Each time the satellite passes through the equator, the time is about 13:30 local time, which effectively ensures the comparability of data in different regions [19]. Sentinel-5P data are currently available from public access[2]. We use the data of 2019 because 2018 is the first year that Sentinel-5P is in operation; the algorithm of the product is not stable then. 2020 witnessed the COVID-19 pandemic, which might have special impact on anthropogenic sources, making the result hard to represent a long-term status.

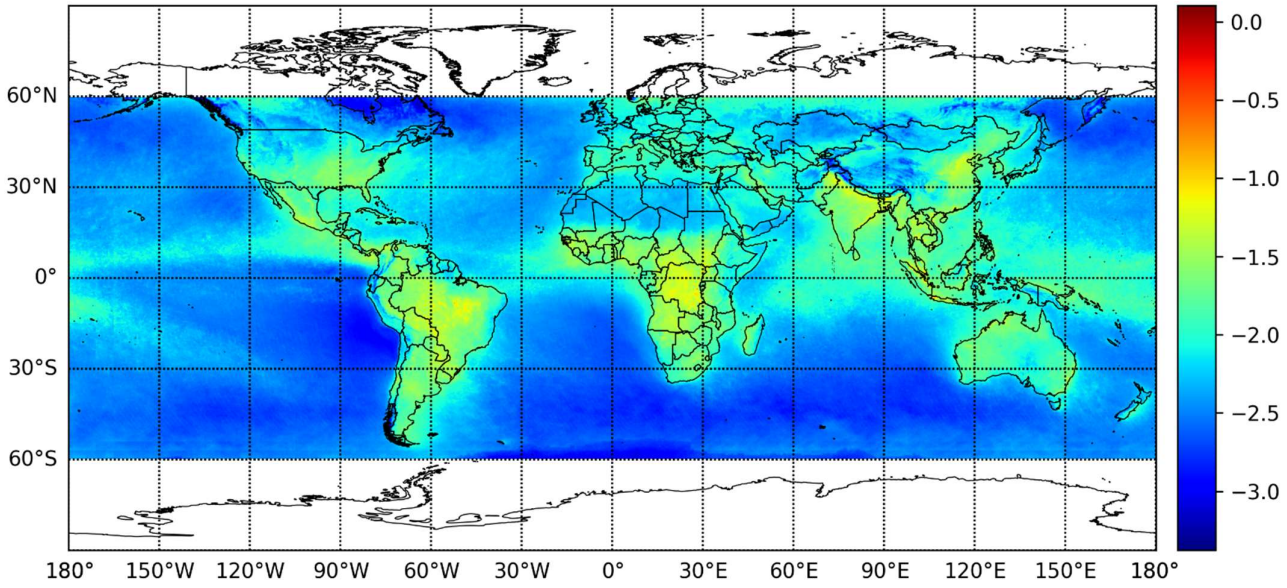



**Figure 1.** The natural logarithm of global vertical column density (VCD) of HCHO in 2019 after being interpolated and averaged on annual basis. (unit: mol/m$^2$)

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. After doing mosaic on the datasets and applying Ordinary Kriging interpolation, we obtained the distribution of global average column concentration of HCHO with 0.05°resolution. Because of the sparsity of satellite data and scarceness of human activities

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1



there, the data beyond 60°S and 60°N is discarded, which has little impact on health risk estimation.

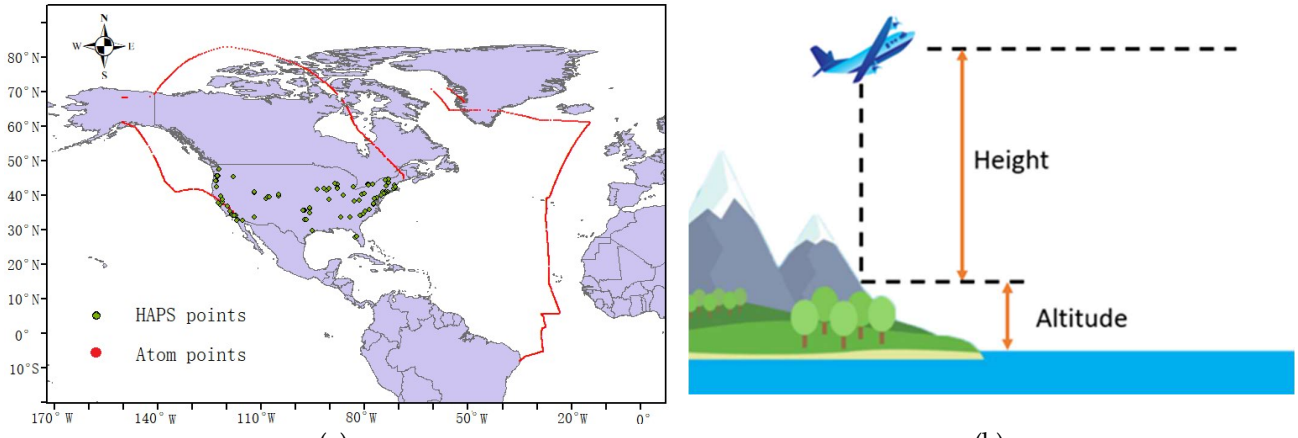
2.1.2. In-situ Data

Since our study aims to estimate the surface concentration of HCHO in a global level, we need data from diverse types of underlying surfaces and different altitudes to train our model. Therefore, the following two data sources are chosen.

**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 of 109 sites in 2019, covering the whole country, as shown in **Figure 2** (a).

These two datasets cover a wide range of altitudes, from -8.1977° S to 82.9404° N, and a diverse variety of landscapes in U.S. The HAPs data ensure that the concentration distribution feature of ground level is emphasized in our model, and the ATom data ensure that our model can be generalized and applied to a global extent. Therefore, our dataset satisfies the requirements of this study.




(a) (b)

**Figure 2. (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 fiercely from ground to 1~2km above [43], we take "Height" 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.

 doi:10.20944/preprints202105.0595.v1



#### 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 to serve as one of the input variables—"Altitude", in estimation of in-situ 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 resolution of 0.05°. This dataset has an initial resolution of 90m at the equator and is provided in WGS84 projection with 1 arc resolution[44].

### *2.2. Data Processing*

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

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

#### 2.2.1 Preprocessing

Though theoretically, a neural network is able to handle input data from different distribution, a significant defect was noticed in the training process without preprocessing, owning to the highly imbalanced, skewed distribution of the HCHO concentration (both column and in-situ). The logarithm of the HCHO concentration data shows a bell-shape 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, neural network plays a vital role in helping us deriving 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 experiments indicate that a joint model always has to compromise between point estimation and interval estimation, thus greatly damaging the accuracy of point estimation.

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1



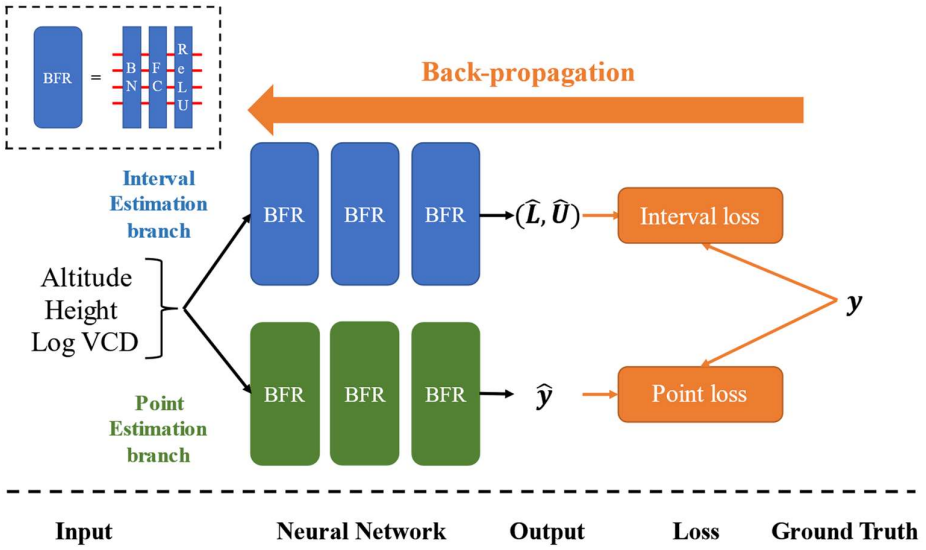



**Figure 3**. We use two separate neural networks for point and interval estimations respectively, each network has three BFR blocks (with ReLU in the last block disabled).

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 firstly 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 algorithm to converge while training. Though point estimation only needs to take the precision into consideration, two conflicting factors are involved in evaluating 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 $\ell_1$ loss is sufficient for training rapidly:

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

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]:

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

 doi:10.20944/preprints202105.0595.v1



On one hand, to control the confidential level of the interval estimator, $\alpha$ 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 $\alpha$ yields shorter intervals. $PICP$ indicates the covering rate of intervals:

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

where $\mathbb{I}\left\{\widehat{L_J} < y_i < \widehat{U_j}\right\} = 1$ 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,

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

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" $\mathbb{I}\left\{\widehat{L_J} < y_i < \widehat{U_j}\right\}$. Since the sigmoid function $\sigma$ 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*

Our point estimation model shows a relatively high accuracy and is generally consistent with previous studies on the vertical distribution of HCHO, where **Figure 4**. shows the point estimation value of in-situ concentration when altitude at sea level is fixed and changing the vertical column density (VCD) and height. In-situ concentration is negatively related with height and positively related with VCD.

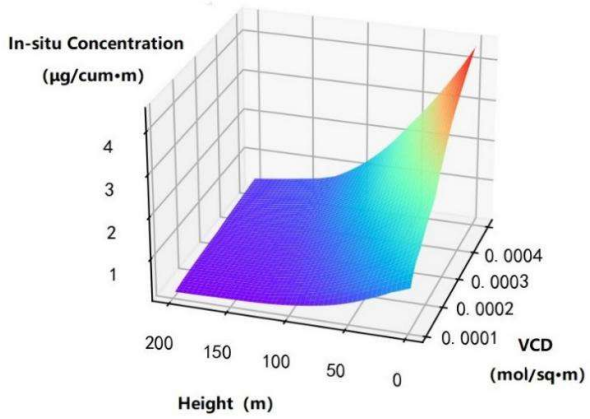



**Figure 4.** The model's estimation of in-situ concentration on certain height, vertical column density of HCHO when the altitude is fixed at 0 m

To evaluate the performance of our model, statistics including MAE and RMSE are calculated based on the training and testing set respectively. As is indicated by **Table 1.** MAE and RMSE of point estimation for in-situ concentration (unit: μg/m3), since MAEs and RMSEs are controlled relatively small, the point estimation from model deserves sufficient confidence.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1



**Table 1.** MAE and RMSE of point estimation for in-situ concentration (unit: μg/m3)

| Dataset | MAE | RMSE |
|---|---|---|
| Training | 1.294 | 1.018 |
| Testing | 1.295 | 1.075 |

By loading global DEM, logarithm VCD and height (0m at surface) into the model, we get the annual average of global in-situ HCHO distribution map. We can see from **Figure 5** that there are generally 6 regions where HCHO in-situ 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/m$^3$. We will get a closer look on the seasonal change of HCHO in these key areas in section 3.3.

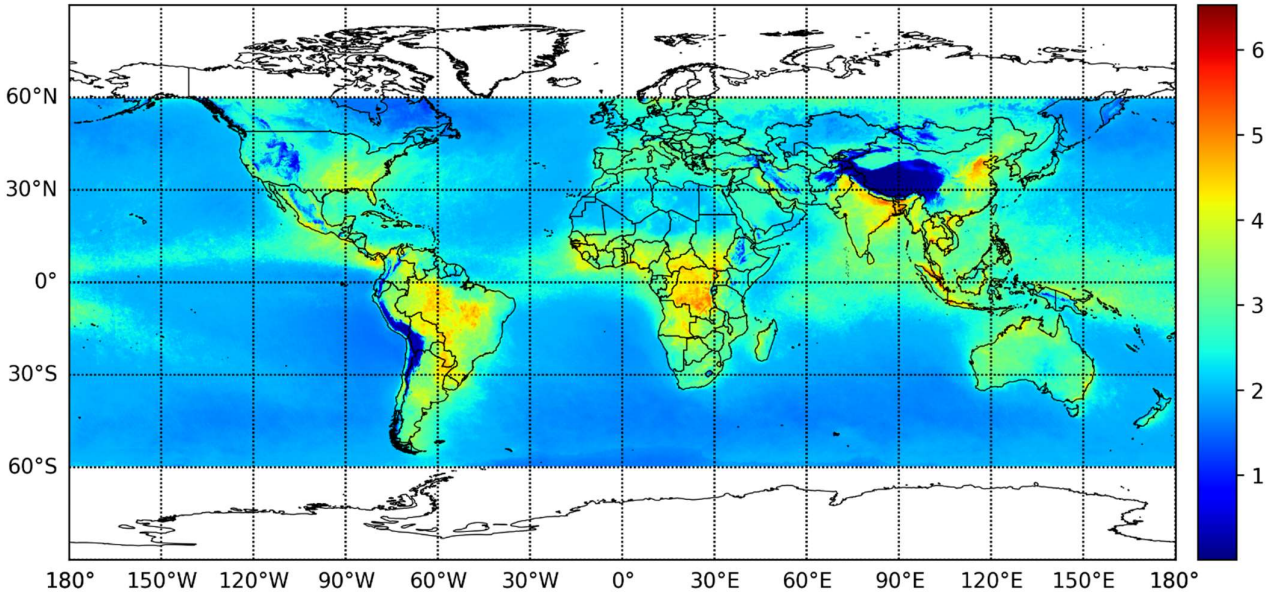



**Figure 5.** Annual average of global in-situ concentration in 2019. (unit: μg/m$^3$)

The uneven distribution of HCHO concentration on the sea and land surface deserves to be mentioned as well. It is obvious that the HCHO is relatively lower and more homogeneous on the sea surface than on the land, statistics in **Table 2** have also confirmed this observation. We hypothesize that sea surfaces with a high concentration are often affected by propagations from nearby continent. This phenomenon is especially obvious in low altitude regions. We call these areas "transmission paths", which will be further discussed in section 4.2.

**Table 2.** Statistics of in-situ 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 |

Cities, as the regions with the densest population, deserve specific attention towards their HCHO concentration due to its known and potential harm to people living there. **Table 3** shows the in-situ concentration of HCHO of some of the typical cities in these

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1



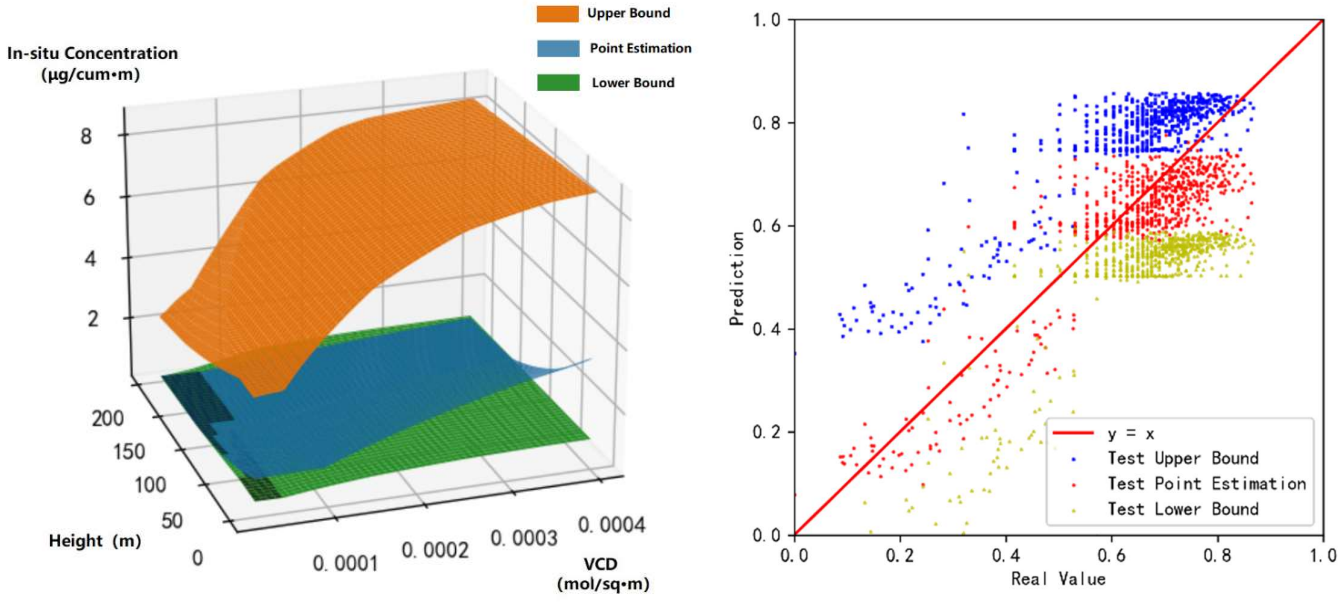
regions, where Jakarta and Singapore, two major cities (country) in South East Asia, rank the first and second, reaching 6.18 and 5,83 μg/m³.

**Table 3.** In-situ HCHO concentration in some typical cities

| City Name | In-situ HCHO (μg/m³) | City Name | In-situ HCHO (μg/m³) |
|---|---|---|---|
| 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*

Other than point estimation, our model also provides us with the estimation of upper and lower bounds of in-situ concentration of HCHO, so that we can evaluate the uncertainty, or fluctuation, of the in-situ concentration. In **Figure 6**, the relationship between estimation of upper bound, lower bound and the point estimation which is acquired in section 3.1 are visualized in a 3D space. We would like to emphasize that the captured uncertainty, or the interval length, delineates the fluctuation range of the data itself, not the lower trustworthiness of our model or its estimations.




(a) (b)

**Figure 6.** (a) The estimation of in-situ upper bound, point estimation and lower bound (90% CI) on certain height, vertical column density of HCHO when the altitude is fixed at 0 m, where CI is short for confidential level. This figure is acquired by feeding equally spaced mock data into the two models. (b) The comparison of our models' estimation and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1



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).

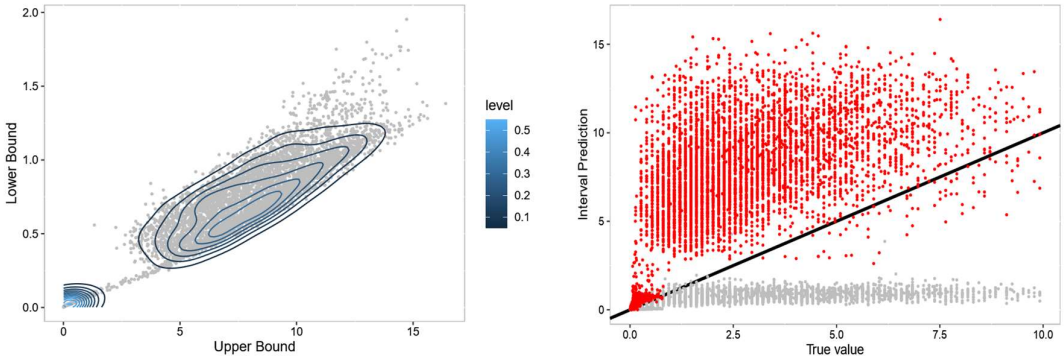
Confidential level, together with the covering length, lays the foundation for the trustworthiness and precision of our interval prediction. As we shall see in **Table 4**, our interval estimation model obtains the covering rates, the ratio of true values covered by predicted interval, of 94.41% and 88.74%, which both exceeds the pre-set confidential level $\alpha = 0.9$ and $\alpha = 0.8$, respectively.

What's more, as what we have expected in section 2.2.1.2, a higher confidential level yields a longer interval length, which is 4.530 for $\alpha = 0.9$, 17% more than 3.864 for $\alpha = 0.8$. Such a phenomenon can also be configured via statistics for minimum, maximum and mean values for upper and lower bounds for the two confidential levels respectively in the table.

**Table 4.** Statistics of interval estimation for in-situ concentration (unit: μg/m3)

| $\alpha$ | 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, we witness a quite greater standard deviation of upper bounds comparing to the lower bounds' one in both scenarios in **Table 4**. By plotting the upper and lower bounds as are in **Figure 7** below, it is self-evident that upper bound estimation is not deterministic, though interval estimation successfully covers the true values (and point estimations as shall discussed below) of in-situ concentration. Nevertheless, further exploration of seasonal changes of HCHO in some key areas in section 3.3 could basically explain that seasonal variations of in-situ HCHO may contribute to the majority of the uncertainty of interval estimation.



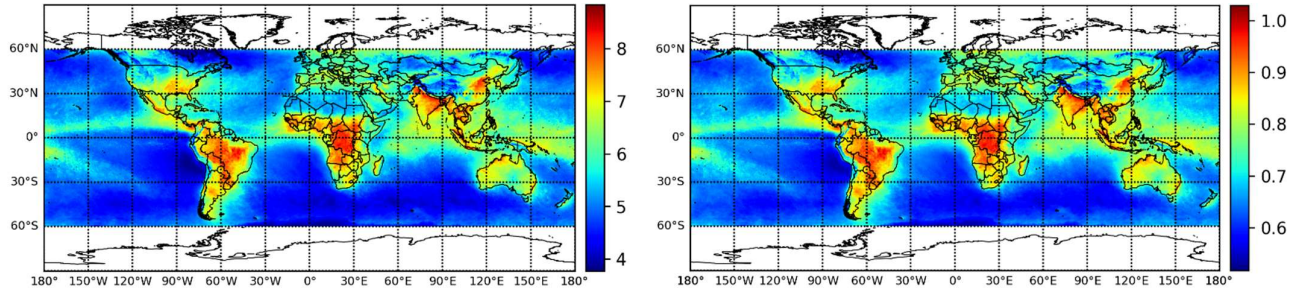

(a) (b)

**Figure 7.** (a) The joint density distribution for upper bound (x-axis) and lower-bound (y-axis). We observe that 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 for lower bound and 5~10 for upper bound. (b) 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, $y = x$, aims for indicating the relative positions of true values in the predicted intervals. We observe that our predicted intervals can basically cover true values.

**Figure 8** is obtained when the model is applied to a global scale. It shows that the upper and lower bounds generally share the same global pattern, though with different scales, with a range of between 3.77 and 8.83 μg/m$^3$ for upper bounds and from 0.52 to 1.03 μg/m$^3$ for lower bounds. The average length of 90% confidential interval is 4.77μg/m$^3$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1





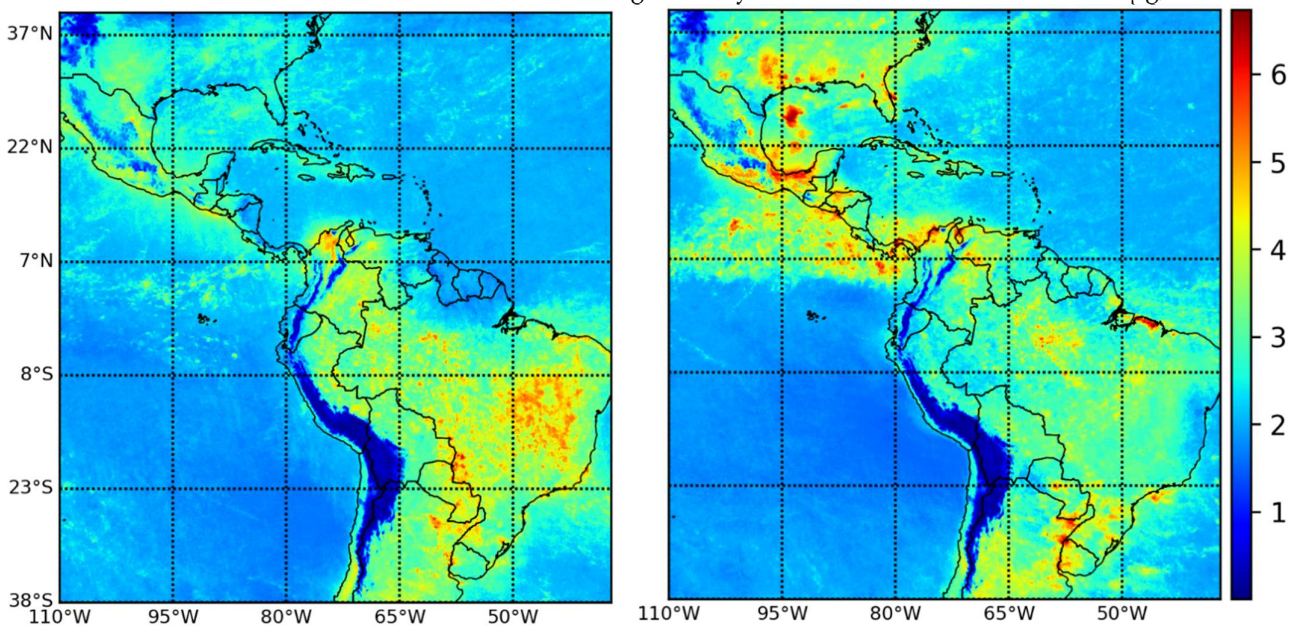

(a) (b)

**Figure 8.** (a) Distribution of upper bound from 90% in-situ concentration estimation, (b) Distribution of lower bound from 90% in-situ concentration estimation. (unit: μg/m3)

*3.3. Seasonal Changes of HCHO in Some Key Areas*

To better understand the seasonal variation of surface HCHO, four typical months of some key areas where in-situ concentration is relatively high are visualized. The reason why typical months, instead of season average, are visualized is that they can provide a stronger contrast and represent the trend better.

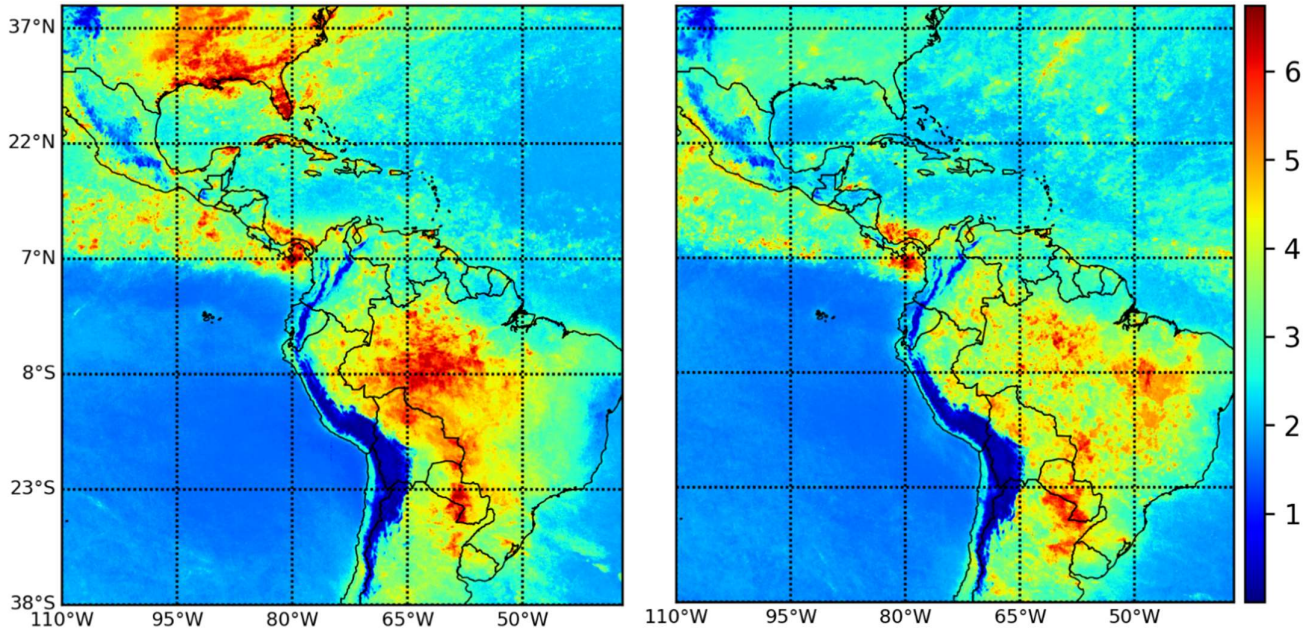
**America. Figure 9** shows the in-situ concentration of February, May, August, and November in South America and around Caribbean Sea. We can see that the Amazon Basin, Paraguay, and Eastern Central America have a high HCHO in-situ concentration in November and February, while the south-east coast of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




(a) (b)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1

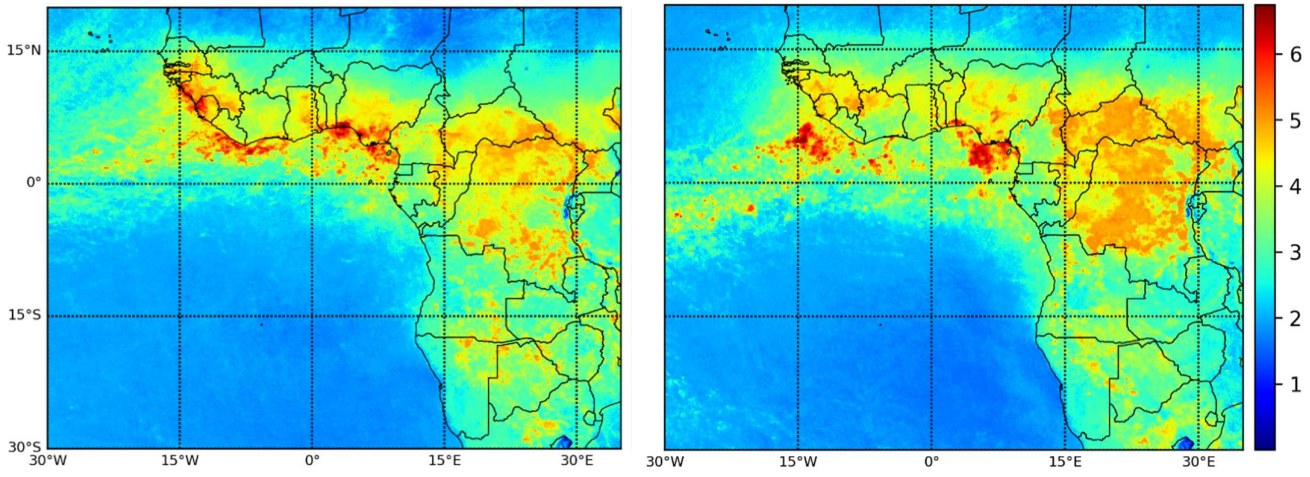





(c) (d)

**Figure 9.** In-situ concentration of HCHO in Central and South America in some typical months. (a): February, (b): May, (c): August, (d): November. (unit: μg/m³)

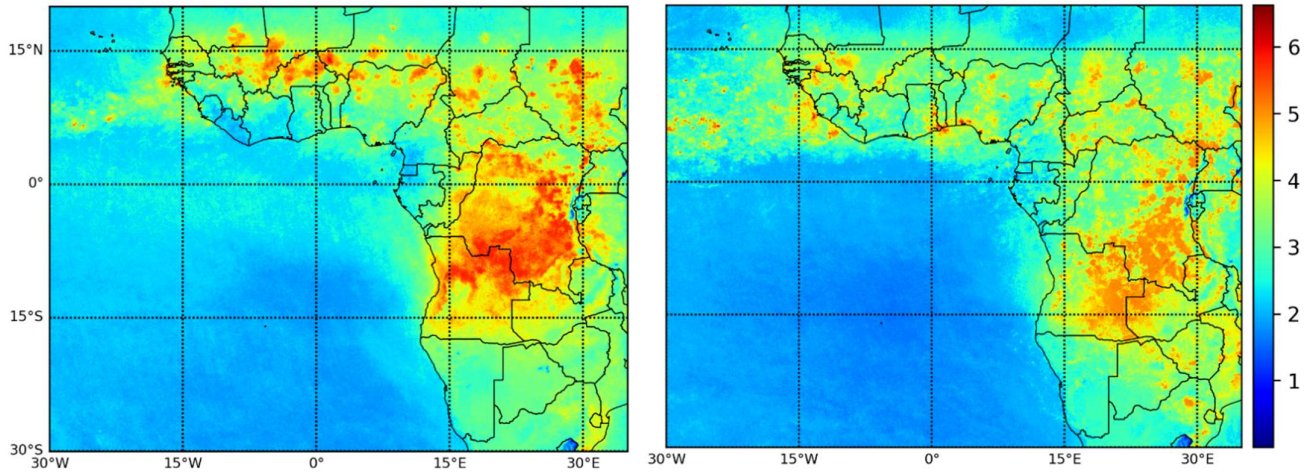
**Africa.** As is shown in **Figure 10**, there are two regions in Africa whose HCHO in-situ concentration is relatively high. One is in the south of R. D. Congo around the city of Kolwezi, a mining center with humid subtropical climate. The in-situ concentration of HCHO here reaches its climax in February. 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: 25 May 2021 doi:10.20944/preprints202105.0595.v1

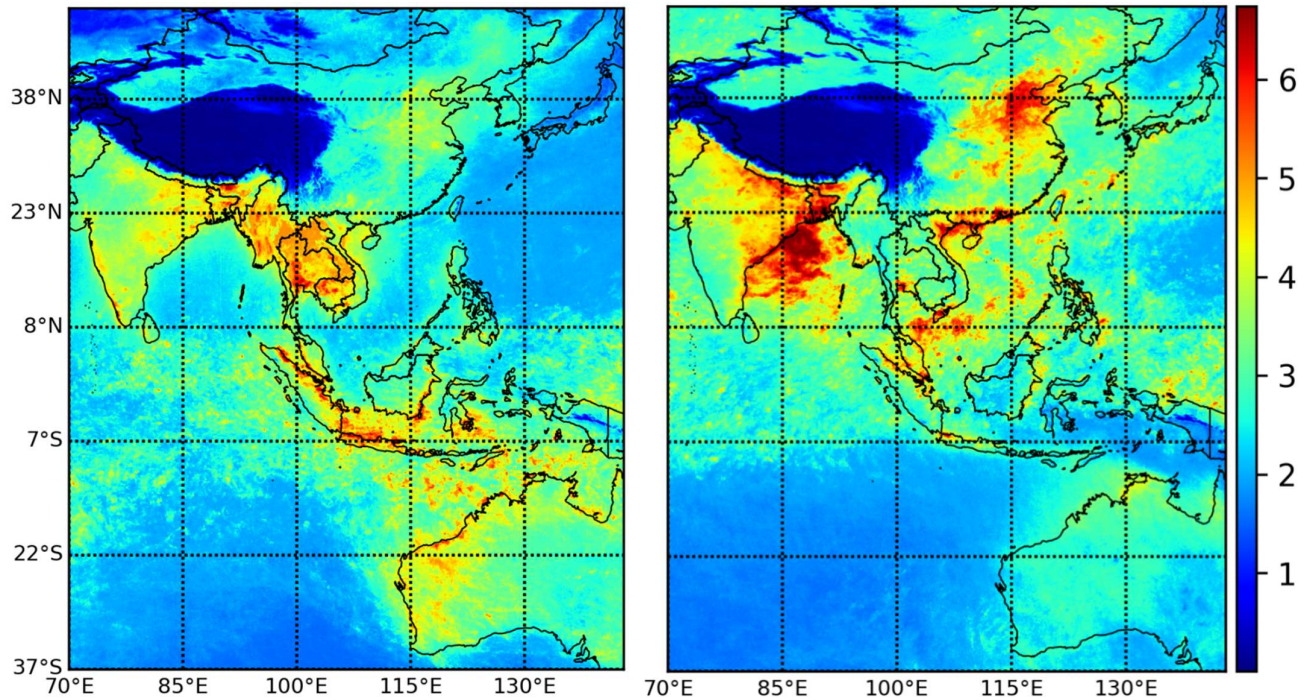





(c) (d)

**Figure 10.** In-situ concentration of HCHO in Central and South Africa in some typical months. (a): January, (b): April, (c): July, (d): October. (unit: μg/m3)

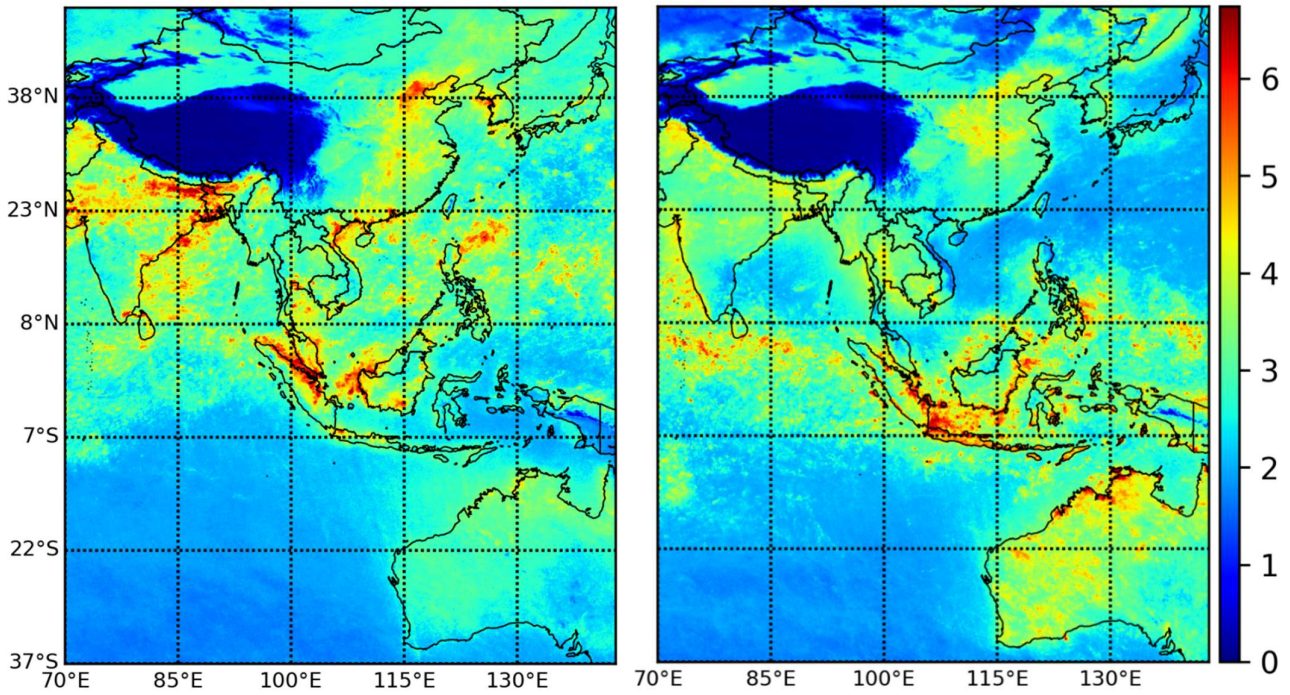
**Indo-Pacific.** As is conveyed in **Figure 11**, there are several regions in In Indo-Pacific whose time of occurrence of high HCHO concentrations differ from each other and may result from different reasons. First, Malaysia and Indonesia islands, both abundant of rainforests, have a relatively high 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 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 the climax around September in 2019.




(a) (b)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1





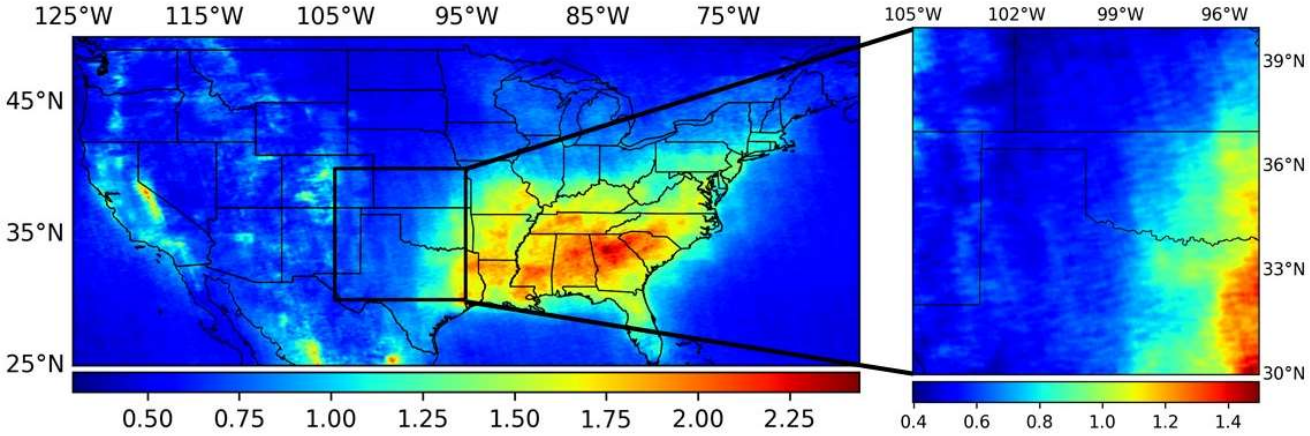

(c) (d)

**Figure 11.** In-situ 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*

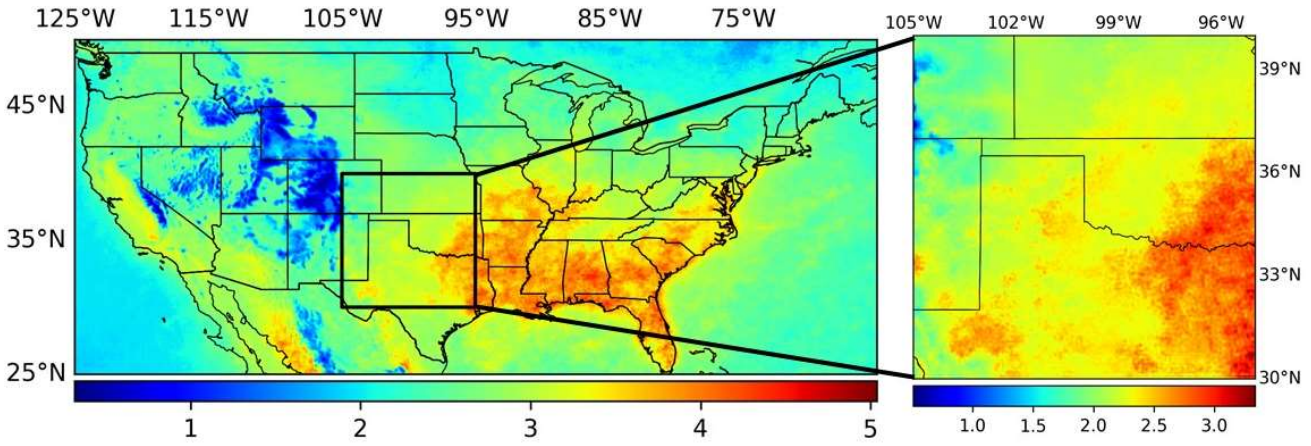
Through the works above, we, for the first time, successfully obtain the global surface distribution of HCHO in 2019 with point and interval estimation. As is seen in **Figure 12**, our result is generally consistent with previous mapping of surface HCHO which is obtained by OMI data and GEOS-Chem model from 2005 to 2016, but shows less noise along the trace of satellite. We can get a clearer view on this phenomenon from the smaller figure on the right side. Our estimation result shows some reversal trend in the Cordillera mountains area. Future research may do some validation on this case. However, since this difference occurs in places where population is sparse, it is not likely to have perceivable influence on the estimation of cancer risks.




(a)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1






(b)

**Figure 12.** (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 us a closer look at the global distribution pattern of HCHO. We notice that HCHO tend to prevail on the plain of continent, instead of on the ocean or on high altitude areas. According to previous study, this can be attribute to the scarceness of VOC sources like chemical industry, combustion and rainforests, which are common precursor of the free radical reaction of HCHO production [46-48]. By mapping the distribution of HCHO, we can also preliminarily distinguish two kinds of sources around the world. 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 works are needed to accurately identify the source of these HCHO-polluted area.

In addition, we introduce the interval estimation of neural network in the conversion of in-situ concentration 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 neural network model [51], thus being useful for the application of neural network into the field of atmospheric pollutant or health risk estimation in the future.

### *4.2. HCHO Transmission path along the equator*

As far as we know, the phenomenon of HCHO transmission path along the equator has not been discussed in previous studies. **Figure 13** shows four transmission paths of HCHO, namely the Central America-Pacific path, the Africa-Atlantic path, the India-Indian Ocean path and the Southeast Asia-Pacific path. These paths are all around the equator, and indicate the possibility that HCHO pollution has significant cross region transmission feature. The path on the west side of America, Africa and Indonesia can be attributed to the constant west wind along the equator [52]. While the path on the east side of Papua New Guinea is hard to explain, since this region should be dominated by southeast trade wind [53]. Future researchers may collect observations data in longer period to see if this is a normalcy and if it has any significant impact on global atmospheric process.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0595.v1



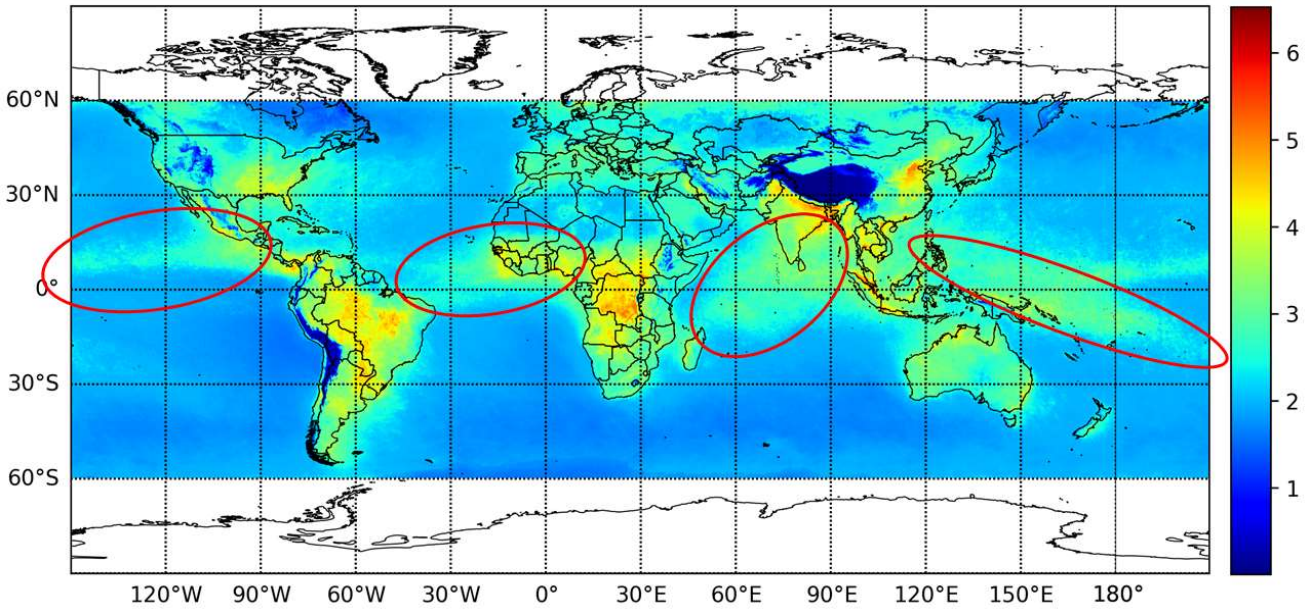



**Figure 13.** Global distribution of surface HCHO and the four transmission paths along the equator. (unit: μg/m$^3$)

*4.3. Health risk of HCHO in major cities*

HCHO, as one of the most important carcinogens in outdoor environment [2], draws little attention due to the lack of ground detection of HCHO in most countries and regions for a long time, leading to the shortage of knowledge about health and economic losses caused by it. Even if the vertical column density of HCHO is currently available and do settle parts of our concerns about these issues, the in-situ HCHO concentration shall bring more benefits, as it can reflect the actual HCHO concentration people exposed to better.

Take 2019 as an example, we assume that the HCHO concentration is always the same as that in 2019. According to inhalation unit risk estimate from EPA and population data[4,54], Health risks in main high-risk cities are calculated in table 5, exhibiting more than a thousand people get cancer due to be exposed in the HCHO in Jakarta, Dhaka, Bangkok, Kolkata, Beijing and Guangzhou. Jakarta has the highest patients due to be exposed, which up to 2593. Meanwhile, Jakarta, Singapore, Kuala Lumpur, Dhaka, Lagos are the highest prevalence cities, and have 80.34, 75.79, 72.93, 71.63, 71.37 people per million. Interestingly, the main cities of high health risk concentrate in Southeast Asia, which is neglected in HCHO pollution and health risk research and Southeast Asia may become the next research focus in HCHO pollution

**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 |

 doi:10.20944/preprints202105.0595.v1



| | | | |
|---|---|---|---|
| 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 |

*4.4. Deficiencies and Prospects*

For one thing, the data of HCHO in-situ concentration is seriously insufficient in spatial and temporal dimensions. Since only 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 low latitude areas like Asia and Africa. It has also been a major obstacle to reaching 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 unfortunately. What's more, the large gap between true values and upper bounds from our interval estimation model may suggest a heterogeneous in-situ concentration of HCHO distribution in different months or seasons, as our model is required to give the interval estimations in the scale of a whole year, rather than finer-grained time scales. Seasonal changes of HCHO in some key areas in visualized section 3.3 has also shown this phenomenon directly.

Therefore, we expect that as HCHO in-situ monitoring network develops, larger amount of data from a more diverse sites could enable us to adopt a careful designation of temporal data input and could help give a better estimation towards in-situ concentration of HCHO. Meanwhile, as Sentinel-5P is accumulating more data, we expect that our model can take more factors, including latitude and seasons, into consideration, which 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, our study is also conducive to researches on the generation of photochemical pollution, the concentration of VOC, $NO_2$ and other photochemical reaction related pollutants.

## 5. Conclusions

With the facilitation of quality-driven interval estimation algorithm designed for neural network, we manage to give the confidential interval and a precise point estimation of 2019 global surface HCHO on different confidential levels with limited amount of data. By mapping the HCHO concentration distribution, we find that Southeast Asia, North China, Central and Western Africa, and the rainforest area of Latin America have relatively more serious HCHO pollution. Major cities in these regions, such as Bangkok, Beijing, Guangzhou, Singapore, have an annual concentration over 5.00μg/$m^3$, the health effects of which is worthy of more attentions from the academia and governments.

Our work paves the way for researches on formaldehyde-related cancers, and provide 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. As HCHO VCD data from Sentinel-5P accumulate, we can map the surface

concentration of HCHO for longer period of time and give more precise estimation of the global risk distribution of formaldehyde-related cancers, which would be more statistically reliable with our confidential intervals.

**Author Contributions:** Conceptualization, W.W.; Methodology, J.G. and Y.D.; Software, G.L., B.J. and Y.D.; Validation, B.J.; Formal Analysis, 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. and Y.D.; 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).