Supplemental materials

Figure

Figure S1. (a) Distribution of vertical column density (mmol/sq·m); (b) Distribution of log (vertical column density); (c) Distribution of log (in-situ concentration); (d) Distribution of in-situ concentration (μg/cum·m)

Method

1. Health Risk

2. Neural Network Architecture

Figure

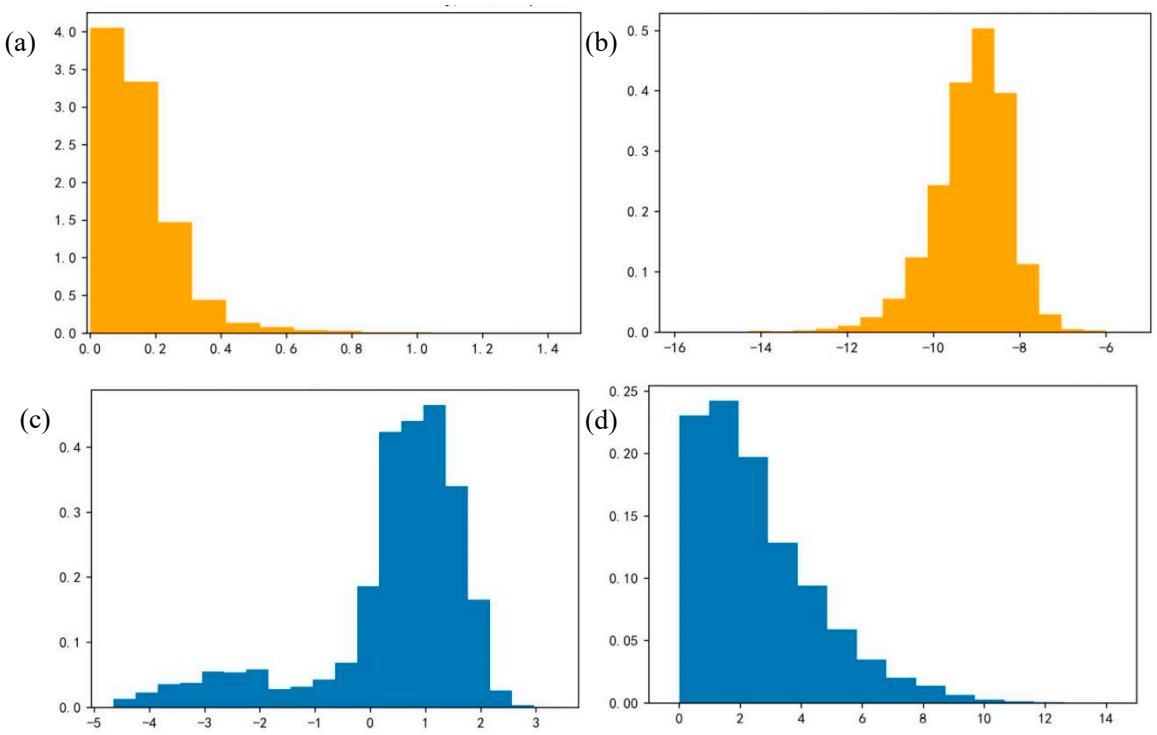

Figure S1. (a) Distribution of vertical column density (mmol/sq·m); (b) Distribution of log (vertical column density); (c) Distribution of log (in-situ concentration); (d) Distribution of in-situ concentration (μg/cum·m)

After log-transformation is applied, the HCHO concentration data shows a bell-shaped distribution, which is more friendly for neural network training. Increments in estimation accuracy have also proven the effectiveness of log-transformation.

Method

1. Health Risk

Inhalation unit risk estimate (URE) was used to quantify the cancer risks of HCHO. According to the integrated risk information system (IRIS) of the EPA, the inhalation URE for HCHO is estimated to be 1.3×10^{-5} $(\mu g\ m^{-3})^{-1}$, based on the upper confidence limit of the fitted dose-response curve [4] (p. 2). The relationship between HCHO concentration and health risk is as follows:

$$R = 1.3 \times 10^{-5} C \quad (5)$$

where R represents the health risk of cancer, and C represents the surface concentration of HCHO.

The relationship between health risk, population and number of cases is as follows:

$$N = P \times R \quad (6)$$

where N represents the number of cases, and P represents the population of the city.

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.

Input and output. Like ordinary multi-layer perceptrons, each neural network in our model contains three input nodes for taking in height, altitude and logarithm VCD at each location respectively (if additional covariates are added, the input nodes number can be adjusted correspondingly). After some transformations (to be discussed later), the point estimation branch containing one output node shall give the point estimated results, whereas interval estimation branch will derive the upper and

lower bounds simultaneously. The structure of our model is shown in Figure 3.

Linear and nonlinear transformation. Neural networks are made up of alternative linear (called fully-connected layer, FC layer) and non-linear transformations (named activation function). Given results from the last layer denoted as z, a FC layer shall transform z affinely: $Az + b$, where A is a linear transformation matrix and b is the bias to be learnt. So, mathematically, two successive FC layers is equivalent to a single FC layer, explaining the necessity of introducing nonlinear activation functions between FC layers. There are a number of activation functions, including Sigmoid, Tanh, and we have selected ReLU [51,52] (p. 8): $\sigma(z) = max\{0, z\}$ dimensionally for its simplicity and effectiveness.

Batch normalization (BN). It is common sense to normalize input data in machine learning. And in fact, issues including gradient exploding and diminishing shall also be alleviated if the distributions of outputs from medium layers are regularized either. That's how batch normalization was introduced. Just like the data standardization, BN layers force the distribution of each hidden layer to have zero means and unit standard deviations dimension-wisely, by storing the means and standard deviations of the output from last layers beforehand. BN can not only regularize the network, but also accelerate the training procedure by reducing the dependence of gradients on the scale of the parameters or of their initial values [50] (p. 7).

As mentioned above that neural networks are made up of alternative linear and non-linear transformations, we propose to combine a BN layer, a FC layer and a ReLU activation in order to form a block named BFR block. Each network consists three BFR blocks to process the data sequentially, where each FC layer has 50 hidden neurons (output dimension). Our BFR constructed network is expected to help stabilizing the training and prediction procedure.