1. VOC selection

A quantified standard will make it clear to decide which VOC species should be assimilated. Here, we propose a VOC factor to quantify which VOC species should be included in the assimilation system.

A 50-member ensemble of free forecasts without any assimilation is conducted to calculate these factors. The free-forecast period is Jul 10th to Aug 10th in 2022, other model settings have no difference from cycling assimilation. Meteorology initial conditions(MIC), lateral boundary conditions(MLBC), and emissions are perturbed for this ensemble forecast. MIC and MLBC are created by adding Gaussian random noise to NCEP FNL product using WRFDA. Only temperature, water vapor, velocity, geopotential height, and dry surface pressure fields are perturbed. The perturbed emission fields are created by adding Gaussian random noise, with a standard deviation of 10%, to the prescribed anthropogenic emission. Unlike the meteorology fields, every emission variable is perturbed. No chemical IC is provided, and chemical LBC is the idealized profile embedded within the WRF-Chem model. Hourly model output after Jul 13th is used for VOC factor calculation.

The VOC factor considers mainly two parts: its correlation to given observation variables and its potential influence on other chemical variables. The correlation is calculated in model space. For each hourly output file at each grid point, the VOC state variable x and observation state variable y in 50 ensembles are used to calculate a temporary correlation, as shown in (Equ 1), i,m,t stands for ensemble, model grid, and time. After looping over all grid points(M) and time periods(T), a spatial and temporal average is conducted to get the final correlation of a given VOC-observation pair, as shown in (Equ 2).

$$corr_{m,t}(x,y) = \frac{\sum_{i=1}^{N}\left(x_{i,m,t}-\overline{x_{m,t}}\right)\left(y_{i,m,t}-\overline{y_{m,t}}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_{i,m,t}-\overline{x_{m,t}}\right)^2\sum_{i=1}^{N}\left(y_{i,m,t}-\overline{y_{m,t}}\right)^2}} \tag{1}$$

$$corr(x,y) = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} corr_{m,t}(x,y)}{M*T} \tag{2}$$

The potential influence of a specific VOC species on other chemical variables is nearly impossible to accurately quantify, considering the complication in VOC reactions. Here we try to give a coarse estimation, the logic is that if the percentage of a VOC species in all VOCs differs greatly in

prescribed emission and averaged model output, it indicates that this VOC species is actively taking part in reactions, thus may have a larger potential influence. We quantify the potential influence by taking the absolute value of this difference in percentage, as shown in (Equ 3). x is the given VOC species, conc(x) and conc(ALL) is the spatial-temporal averaged model concentration for x and for all VOCs; emiss(x) and emiss(ALL) is the spatial-temporal averaged prescribed emissions for x and for all VOCs.

$$influence(x) = abs(\frac{conc(x)}{conc(ALL)} - \frac{emiss(x)}{emiss(ALL)}) \tag{3}$$

The VOC factor for a given VOC-observation pair can now be calculated simply by multiplying correlation and influence together, as shown in (Equ 4).

$$VOCfactor(x, y) = corr(x, y) * influence(x) \tag{4}$$

It should be noted that a VOC species with a higher factor value can not promise a better assimilation performance. But we think it is acceptable for this study, for we are not trying to select only one best VOC, but a group of VOCs to be assimilated. By selecting VOCs with factor values up to some given threshold, it can be expected that these selected VOCs contain most species with positive assimilation effects(correlated and active in reactions), and avoid negative effects brought by small-value VOCs(not correlated and/or not active in reactions). Thus, by doing this selection, a better performance than assimilating all VOCs could be expected.