Preliminary list of predictors, ancillary variables, and predictands for the Machine Learning model.

Chi Li

1. Predictors:

1) Satellite retrieval:

All the data below are contained in the BEHR output (/volume1/share-sat/SAT/BEHR/BEHR_Files/us/daily/) for each OMI pixel, so convenient to be collected.

'SlantColumnAmountNO2_Trop': The magnitude of this tropospheric SCD is directly relevant to NO_x source; This quantity could be calculated as:

'SlantColumnAmountNO2_Trop' = 'SlantColumnAmountNO2'–'ColumnAmountNO2Strat' * 'AmfStrat'

'MODISAlbedo': Apart from affecting the sensitivity (scattering weights) at surface, brighter surface usually corresponds to bare soil or urban area, so this variable might supplement the WRF-chem emission input for potentially missing NO_x source.

'BEHRScatteringWeightsClear', 'BEHRScatteringWeightsCloudy': Representing weighting of each layer to the AMF. Including these sensitivities would tell the learning model which layers are critical and important. **Potentially, this would relieve some influence of uncertainties in "unimportant layers".**

'CloudFraction', 'CloudPressure', 'CloudRadianceFraction', 'GLOBETerrainHeight', 'MODISCloud': Besides related with the prediction of AMF in clear sky and cloudy scenes, these cloud properties from satellite observation might also contain some information about CCN (aerosol seeds that form cloud droplets), humidity and vertical convection.

2) WRF-Chem parameters:

We need to use the location and time information in the satellite data to find the corresponding WRF-Chem grids (see "Ancillary data" below). Some inputs are from the monthly WRF-Chem outputs (shown in red below, /share-wrf1/BEHR-WRF/MonthlyProfiles/us), and some others are from the instantaneous hourly outputs (in blue, e.g. /share-wrf1/BEHR-WRF/Outputs/us/).

'Clim_NO2': Climatological NO_2 profile for 12 months in 2012, represent an averaged profile shape for the location and month. **The latter parameters are included to modify this climatological shape.**

'E_NO': Anthropogenic NO_x emission (as NO equivalents), representing local enhancement (or reduction) of NO_x source near surface.

'EBIO_NO': Biogenic NO emission.

'PBLH': Boundary layer condition, related with convection and vertical extent of NO_2 plume at surface.

'EBIO_ISO': isoprene emission, largely representing biogenic emission and its influence on VOC_R, which affects NO_x lifetime.

'IC_FLASHCOUNT', 'CG_FLASHCOUNT': Lightning flash rate. Introduce modifications in the upper troposphere.

'Adjacent_Influence': Theoretically, this term is sensitive to wind speed, wind direction and all the above source variables at all the adjacent grid boxes that are close to the location of OMI pixel in the previous τ hours (τ is the lifetime of NO_2). But including all the information is cumbersome.

Practically, we assume that the previous local terms are dominating, and use the climatological NO_2 profile at the adjacent pixels and instantaneous wind information to estimate their contribution. A tentative parameterization could be:

a) For the central grid (represented by *o*), define a characteristic lifetime τ, τ could be a conventional value for each month (e.g. 2 hours), or could be fitted using EMG. Then use τ to define a spatial window *d* (i.e. $d = u*\tau$), where *u* is a typical wind speed or from reanalysis data (e.g. 3 m/s).

b) We propose the total adjacent contribution to be parameterized as

$$Adj(l) = \sum^{k \in d} \frac{C_k(l)}{\overline{u_k}} \exp\left(-\frac{d_k}{\tau \overline{u_k}}\right)$$

Where *l* represents each vertical layer, $C_k(l)$ is the climatological mixing ratio (or number density) for a nearby grid *k* at layer *l*, $\overline{u_k}$ is the averaged wind speed (projected to the direction from *k* to the central grid *o*) of grid *k* at layer *l* within the time window τ, and d_k is the distance from *k* to *o*. This parameterization assumes a 1-D exponential decay on each layer and neglect cross-wind and vertical dilution.

2. Predictands:

'BEHRAMFTrop', 'BEHRAMFTropVisOnly' : Clear and cloudy AMFs calculated using the WRF-Chem profiles (/volume1/share-sat/SAT/BEHR/BEHR_Files/us/daily/). As we discussed before, predicting AMF rather than profile might avoid unnecessary fitting at levels that AMF is not sensitive to.

(Optional) AMF re-calculated using aircraft profiles (We will need to re-calculate these AMFs by using the scattering weights, cloud fraction from the BEHR file).

3. Ancillary data:

These data are either needed for collocation purpose, or for potentially binning (spatially or temporally) in future data training if necessary (/volume1/share-sat/SAT/BEHR/BEHR_Files/us/daily/).

'BEHRPressureLevels', Pressure levels that correspond to the scattering weight. These are reference vertical coordinate used to interpolate the in situ or WRF NO2 profiles (BEHR output files)

'Latitude', 'Longitude', 'FoV75Area', 'FoV75CornerLatitude', 'TiledArea', 'TiledCornerLatitude', 'TiledCornerLongitude': Geolocation and PIXCOR information for finding the collocated WRF grid or in situ record

'RelativeAzimuthAngle', 'ViewingZenithAngle', 'SolarZenithAngle': Observing geometry

4. Notes and discussion:

    1) The spatial distribution (extent) of SCD might contain information about wind direction and NO_x lifetime. Tentatively, we could regrid SCDs at nearby 2-3 rows and columns to lat-long coordinate and include these adjacent SCDs in our predictors to capture this information.

    2) The WRF-Chem PBLH and wind data are online simulation outputs. In realistic implementation, we need to replace these with reanalysis or observation data.

    3) Current WRF-Chem simulations do not include fire emissions, and the newest lightening parameterization from Qindan is only available for 4 months.

    4) It would be interesting to see if the Machine Learning model using the “wrong lightening”, “wrong wind”, etc. would work to better (at least similar to WRF-Chem) reproduce the observation (aircraft data) based AMF after introducing more realistic data inputs.

2-8-19

SCD – slant column data from satellite

AMF corrects slant to vertical

To actual data VCD

Surface and cloud from satellite data

Using the Inputs, we are trying to predict the AMF

AMF is a function of a lot of data, which outputs a coefficient to calculate VCD

VCD = SCD / AMF

SCD = AMF * VCD

WRF-Chem is a chemical transport model used to predict the vertical profile of NO2.

Trying to avoid the vertical profile calculation and just use the inputs to WRF – Chem to calculate AMF.

These predictors are actually estimated, by car emissions

VCD is what BEHR is trying to produce

Satellite data is in the BEHR output

We have calculated AMFs,

Trying to use lat, long, and time to find predictors for WRF-Chem.

Records are the satellite data 1x1

Vertical profiles are 3d maps,

My current job:

Write Python. Or repurpose MatLab code to get our output

Try to find all the WRF-Chem pixels within the satellite Field of View (FOV) and average them for each satellite record.

If deemed unnecessary, use nearest pixel. (WRF Chem resolution is 12km, OMI is 13X24 km FOV box I satellite points downward, 2-4 pixels in each OMI; might be a bigger FOV box if satellite is slanted.)

File with MATLAB code in my notebook: BEHR-core-utils/…../rProfile_WRF.m

Function is called avg_apriori(), to find overlapping WRF chem pixels in the OMI FOV.