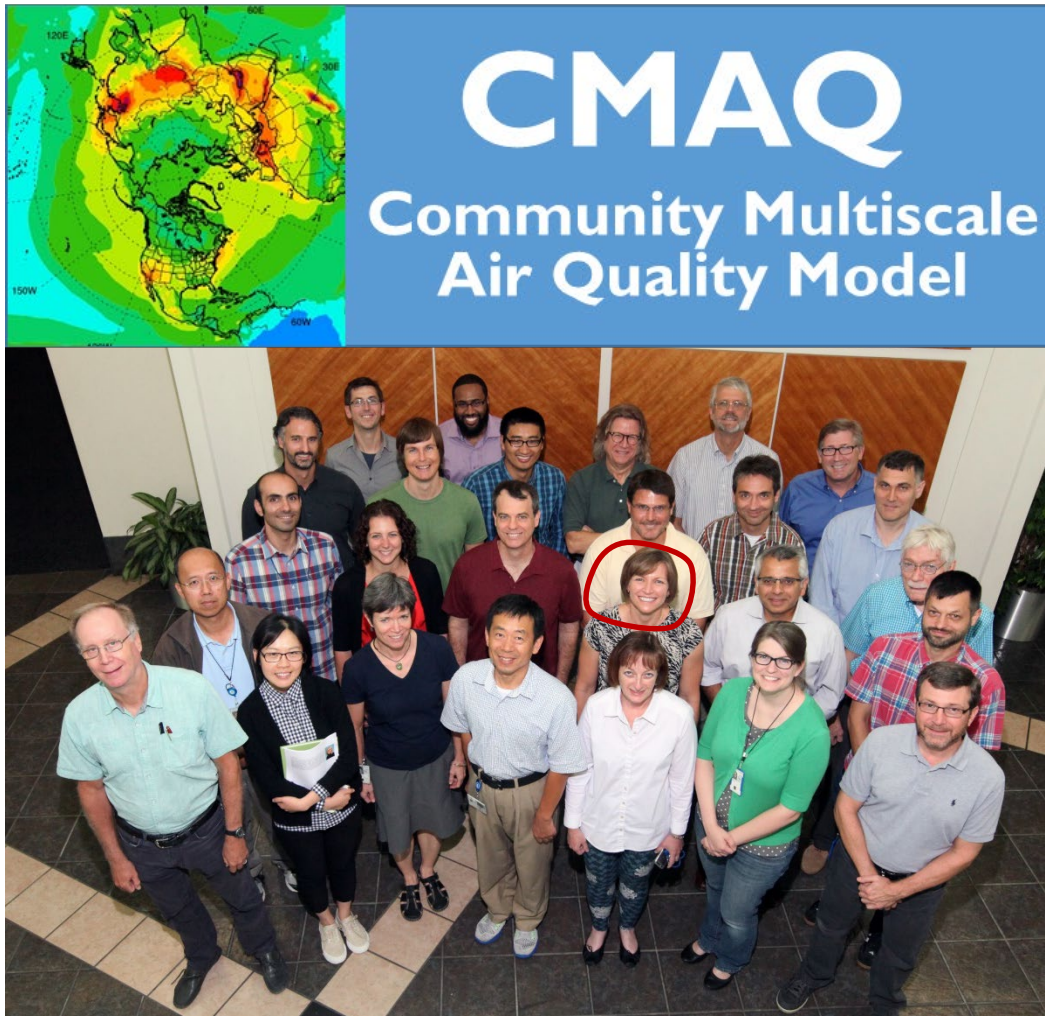
Evaluation of a numerical air quality model

Data wrangling large NetCDF files with R

Kristen Foley
Office of Research and Development
Computational Exposure Division
August 12, 2019

The CMAQ Team:

- Office of Research and Development, ~~NERL/Computational Exposure Division~~ CEMM/Atmospheric & Environmental Systems Modeling Division
- Mix of atmospheric scientists, chemical and environmental engineers, meteorologists, computer scientists, one statistician



https://www.vortech.nl/en/fortran-is-alive/

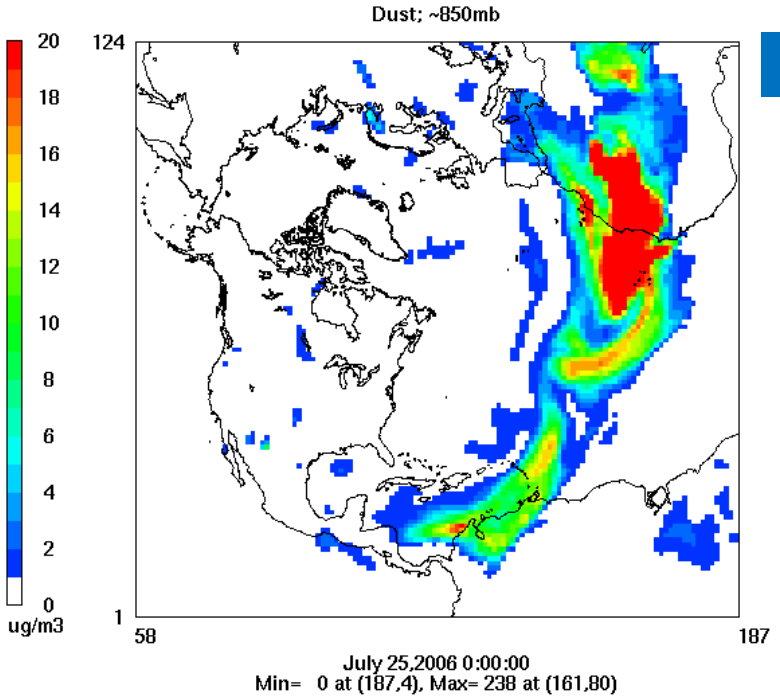
Using R to evaluate a numerical air quality model

Presentation outline

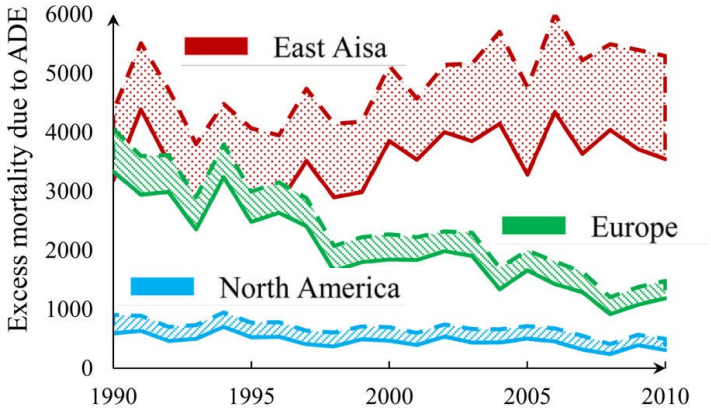
- What is CMAQ?
- What is a NetCDF formatted file?
- How do I visualize and evaluate 10s-100s of GB of model output?

CMAQ: Exploring Air Pollution from Global Scales to Local Impacts

Windblown dust generated in the Sahara Desert and transported to the Southeast US

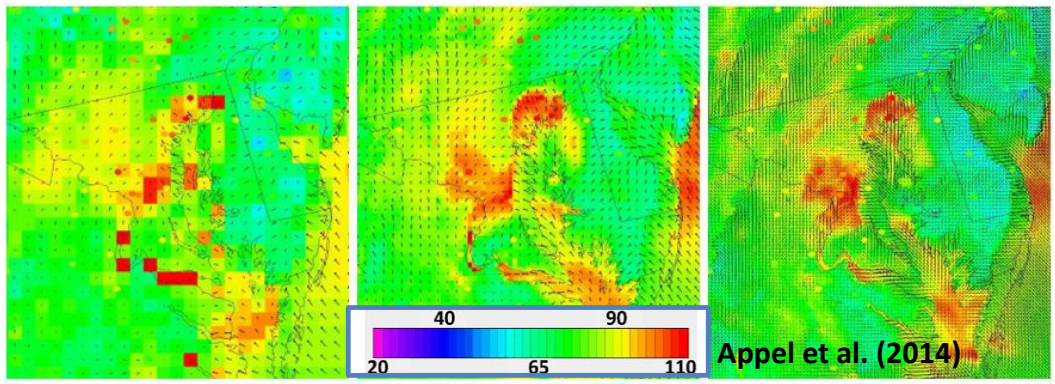


Health impacts from Trends in Aerosol Cooling



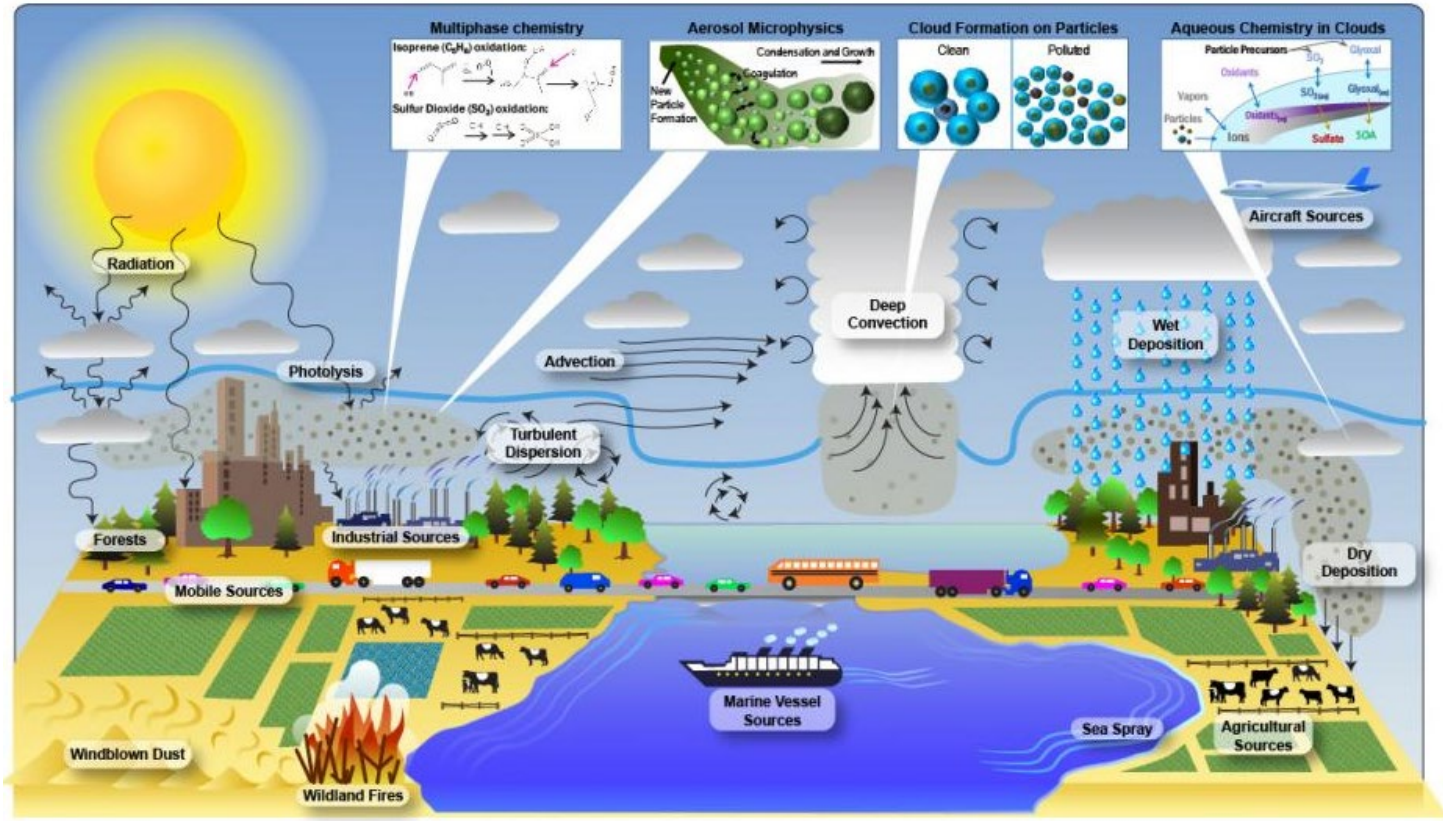
Xing et al. (2016) doi: 10.1021/acs.est.6b00767

Ozone and 10-m wind vectors over Maryland at 12-km, 4-km, 1-km horizontal grid spacing



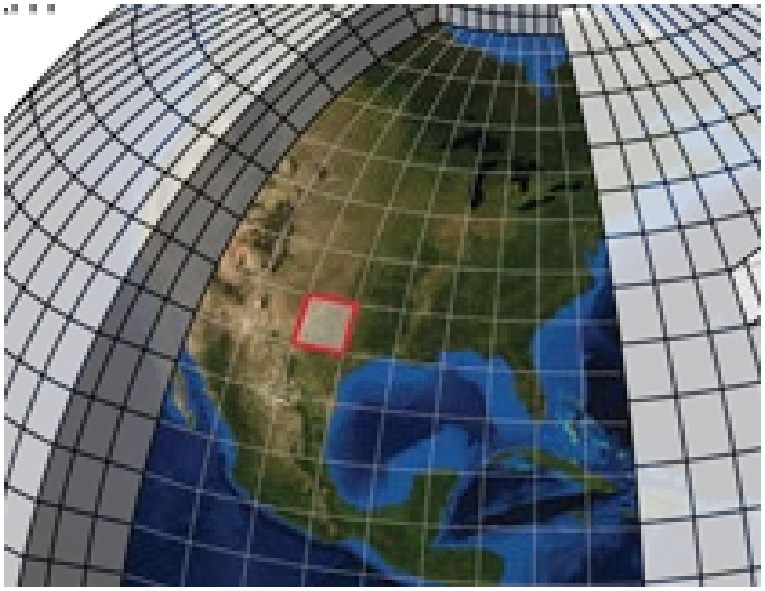
Appel et al. (2014)

Air Quality Modeling: Complex Representation of Complex Atmospheric Processes

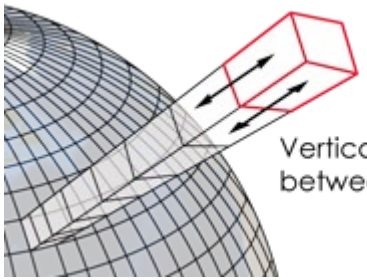


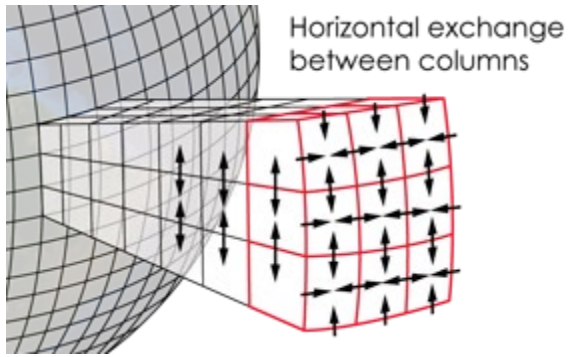
- The model uses numerical methods to solve ordinary and partial differential equations representing chemical transformation, diffusion, advection and removal processes over time
- Main Program: **Fortran**, > 1 million lines of code
- 1 year simulation over the continental US takes ~ 1 week w/ 256 cores

Air Quality Modeling: BIG data



Source: https://www.earthmagazine.org



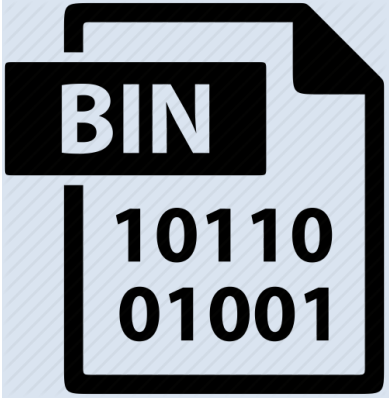


- Requires 100s of GBs of input data, creates TB of output data
- Model output = 4D arrays of hourly output for ~ 200 chemical species

  e.g. 4D array = 460 x 300 horizontal grid (grid spacing = 12km) x 35 vertical layers x 24 hrs
- Output files: **NetCDF** formatted files with CMAQ-specific data structures
- Model simulations, post processing, and evaluation are done on the High-End Scientific Computing (HESC) Linux system at the National Computer Center (NCC)

Network Common Data Form

- Software libraries and data formats developed by Unidata
- Originally designed for sharing weather data
- Used for array-oriented scientific binary data



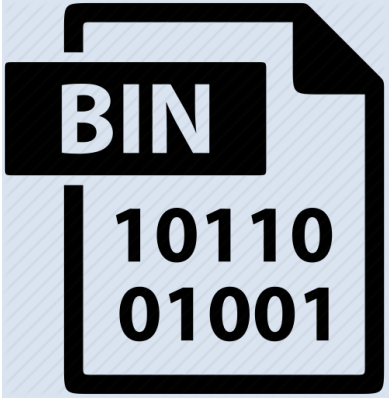
Why binary?

Save space

Save computational effort

Network Common Data Form

- Software libraries and data formats developed by Unidata
- Originally designed for sharing weather data
- Used for array-oriented scientific binary data



	.csv	.nc
Size	14 GB	2.8 GB
Time to write	13 min	0.5 min

netCDF formats

- "Self-describing"
- File header includes dimensions of data arrays, metadata about each variable, global metadata


```
netcdf HR2DAY_LST_ACONC_v521_mpi_intel17.0_4CALIF1_Nodust_201005 {
dimensions:
        TSTEP = UNLIMITED ; // (17 currently)
        DATE-TIME = 2 ;
        LAY = 1 ;
        VAR = 1 ;
        ROW = 325 ;
        COL = 225 ;
variables:
        int TFLAG(TSTEP, VAR, DATE-TIME) ;
                TFLAG:units = "<YYYYDDD,HHMMSS>" ;
                TFLAG:long_name = "TFLAG           " ;
                TFLAG:var_desc = "Timestep-valid flags:  (1) YYYYDD
        float O3_MDA8(TSTEP, LAY, ROW, COL) ;
                O3_MDA8:long_name = "O3_MDA8         " ;
                O3_MDA8:units = "ppbV            " ;
                O3_MDA8:var_desc = "Max-8-hour
// global attributes:
                :IOAPI_VERSION = "$Id: @(#) ioapi library version 3
                :EXEC_ID = "???????????????
                :FTYPE = 1 ;
                :CDATE = 2018243 ;
                :CTIME = 132650 ;
                :WDATE = 2018243 ;
                :WTIME = 132650 ;
                :SDATE = 2010135 ;
                :STIME = 0 ;
                :TSTEP = 240000 ;
                :NTHIK = 1 ;
                :NCOLS = 225 ;
                :NROWS = 325 ;
                :NLAYS = 1 ;
                :NVARS = 9 ;
                :GDTYP = 2 ;
                :P_ALP = 33. ;
                :P_BET = 45. ;
                :P_GAM = -97. ;
                :XCENT = -97. ;
                :YCENT = 40. ;
                :XORIG = -2400000. ;
                :YORIG = -732000. ;
                :XCELL = 4000. ;
                :YCELL = 4000. ;
```

Beginning of header from a CMAQ output file with one variable

When might you encounter netCDF data?

- Commonly used for earth science observations and modeling data:
  - Radar data
  - Satellite data
  - Numerical weather forecast data
  - Global climate modeling data
  - Ocean modeling data
  - **CMAQ data!!**
- Input/output format for many GIS applications

➢ Different datasets will have application-specific data structures and attributes

R Interface to netCDF format data files

- **ncdf4 library**: 17 functions for reading, modifying, writing netCDF data files

  ➢ Good introduction to ncdf4 functions:

  http://geog.uoregon.edu/bartlein/courses/geog490/week04-netCDF.html

- Other packages :
  - RNetCDF – reading and modifying existing netCDF files
  - raster – reading, writing netCDF files, mapping, etc.
  - ncdf4.helpers – tools developed for climate model output
  - M3 – developed by former CMAQ team member, Jenise Swall; specifically designed to handle CMAQ outputs

Sample R code for reading/mapping CMAQ data

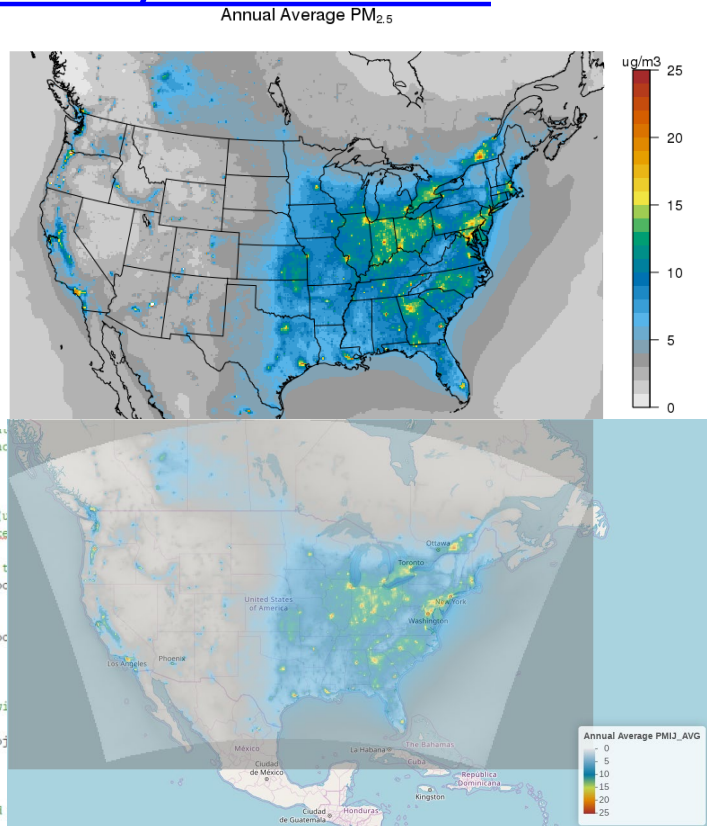
- 2002 -2014 Daily average CMAQ output for 13 species (including SO_2, NO_2, O_3, EC, OC, $PM_{2.5}$, SO_4^{2-}, NO_3^-, NH_4^+) for 2002 -2014 available online:

  https://dataverse.unc.edu/dataverse/cmascenter

Sample R code available

- Open .nc file
- Read in 3D array of daily average $PM_{2.5}$
- Create annual average $PM_{2.5}$
- Map with image.plot in Lambert projected coordinates
- Project to lon/lat raster object and map with Leaflet

```
library(ncdf4)
library(fields)
library(maps)
library(M3)
library(raster)
library(leaflet)

setwd("/work/MOD3EVAL/fib/data_warehouse/")
cctm.file <- "DAILY_ACONC_LST_CMAQv5.0.2_cb05tucl_WRF3.4

#> Open the CCTM file.
cctm.in <- nc_open(cctm.file)

#> Create a list of all model variables in cctm.file. A
so remove the first element of the list.
all.mod.names <- unlist(lapply(cctm.in$var, function(var

#> Create a list units for all the model variables in th
an I/O API file, so remove the first element of the list
#> Use gsub to strip out extra spaces.
all.mod.units <- gsub(" ","",unlist(lapply(cctm.in$var,

#> Pull out the time steps and the grid coordinates asso
#> These functions from the M3 library are wrappers for func
date.seq <- get.datetime.seq(cctm.file)
format.date.seq <- format.Date(date.seq,"%m/%d/%Y")

#> Lambert projected coordinates of the grid cell CENTERS (u
#> These are the unique x, y coordinates of the grid cell ce
regular grid.
#> You can also use the get.coord.for.dimension() function t
x.proj.coord <- get.coord.for.dimension(cctm.file,"col")$coo
length(x.proj.coord)
#[1] 459
y.proj.coord <- get.coord.for.dimension(cctm.file,"row")$coo
length(y.proj.coord)
#[1] 299

#> Also get the grid cell centers of all of the grid cell wi
raster.
xy.proj.coord.meters <- expand.grid(x.proj.coord*1000,y.proj
dim(xy.proj.coord.meters)
#[1] 137241      2

#> Projection information character string that can be used
```



Using R to evaluate the Community Multiscale Air Quality Model (CMAQ)

- New versions of CMAQ are released every 1.5-3 years
- Evaluation is an important part of preparing for a new release
- Need evaluation tools that will:
  - Compare model results across different model configurations or versions
  - Quickly QA model output against a standard set of observational data
  - Evaluate model performance across pollutants, spatial and temporal scales, parts of the distribution

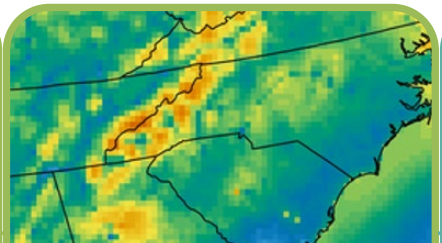
Using R to evaluate CMAQ

Aggregate and Transform

Input: CMAQ output

Output: netcdf

Software: Fortran program



Visualize/QA

Input: CMAQ post-processed output

Output: graphics files, html

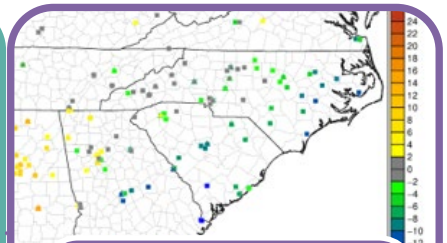
Software: R, VERDI, Ncview

Match to Obs

Input: CMAQ post-processed output + observation data (text files)

Output: .csv

Software: Fortran program



Evaluate

Input: .csv files

Output: MySQL database, graphics files

Software: R based package AMET

4 TB Model data

90 GB Model data

.5 GB Obs data

1 GB Paired Model/Obs data

Post processing 1 month of model output for the continental US

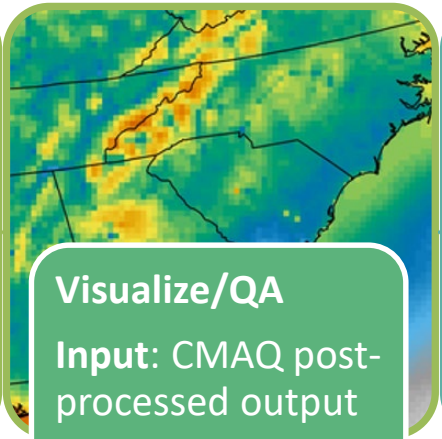
Using R to evaluate CMAQ

Aggregate and Transform

Input: CMAQ output

Output: netcdf

Software: Fortran program



Visualize/QA

Input: CMAQ post-processed output

Output: graphics files, html

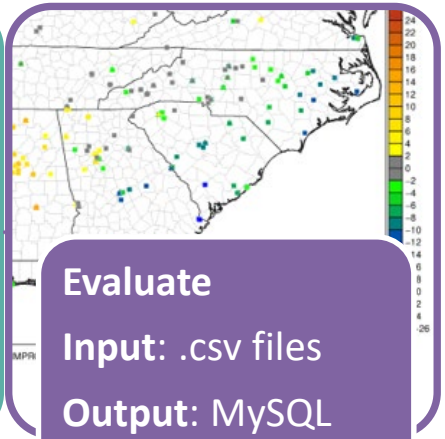
Software: **R**, VERDI, Ncview

Match to Obs

Input: CMAQ post-processed output + observation data (text files)

Output: .csv

Software: Fortran program



Evaluate

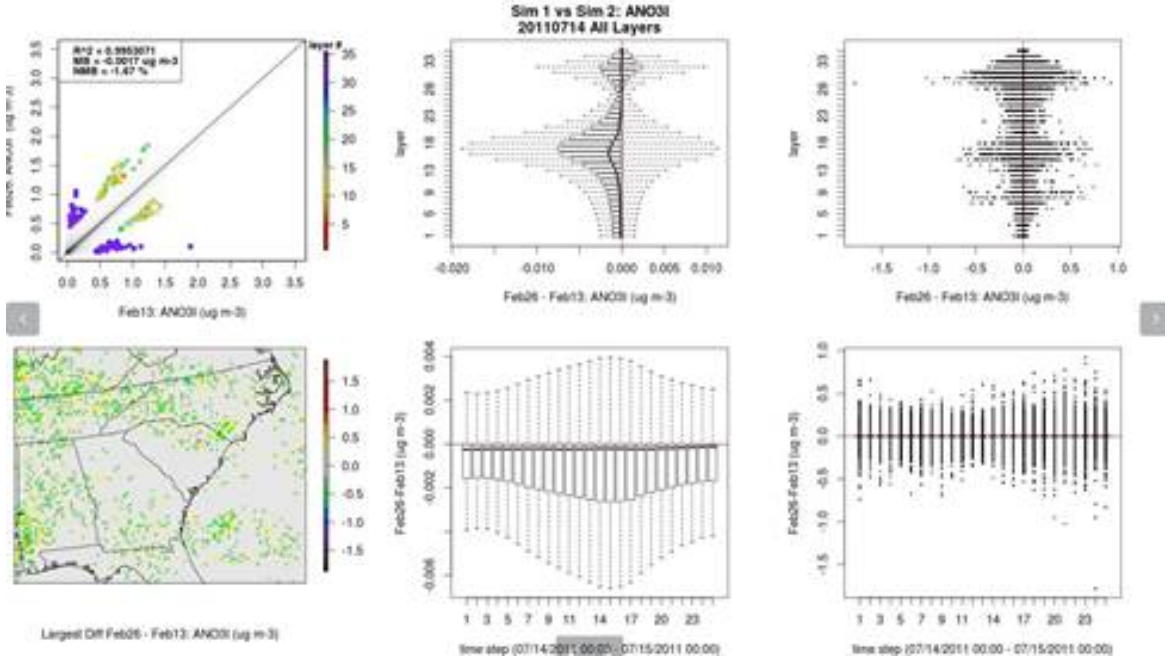
Input: .csv files

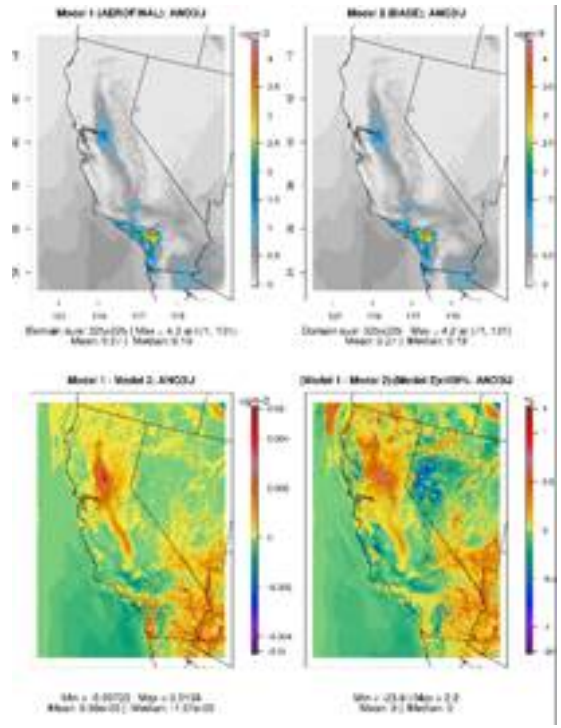
Output: MySQL database, graphics files

Software: **R** based package **AMET**

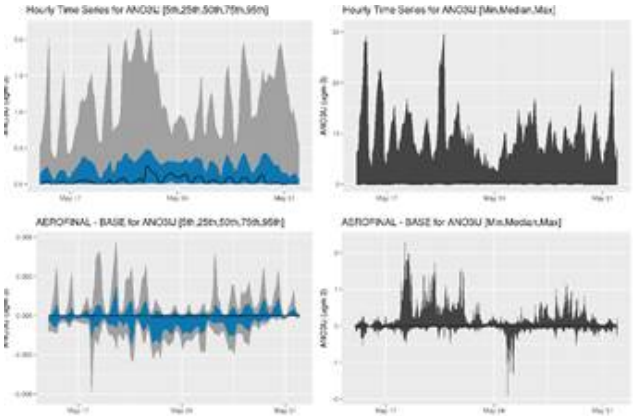
Automate all steps with a combination of Linux shell scripts and R code

QA Code Changes with Batch R Plots

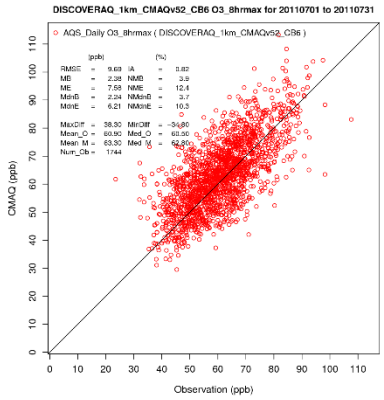


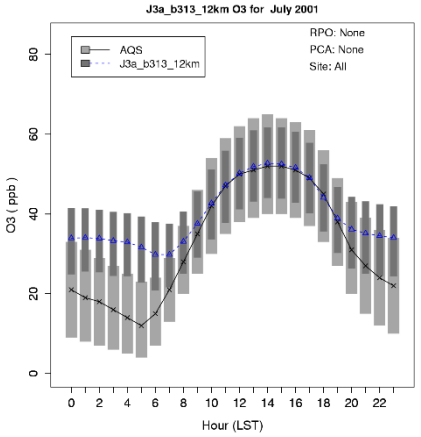


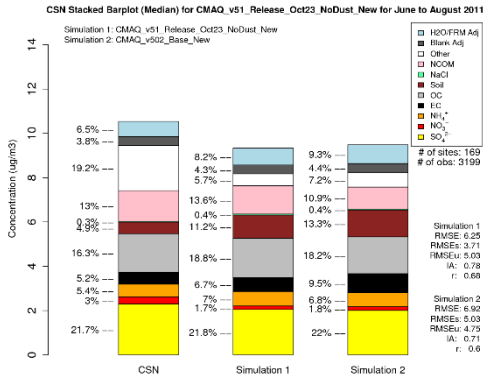
- Model-to-model comparisons (no observations)
- Loop over all species, look for largest differences by hour, vertical layer, spatial location

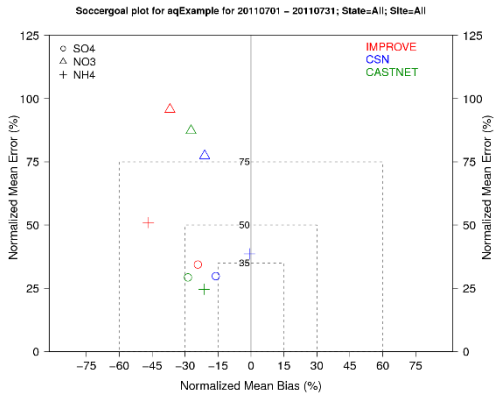


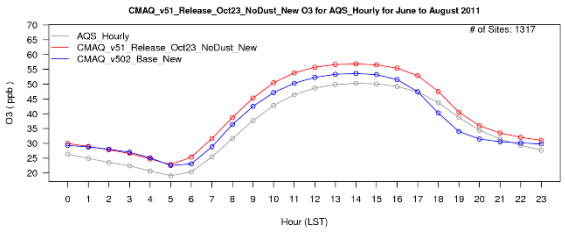
Atmospheric Model Evaluation Tool

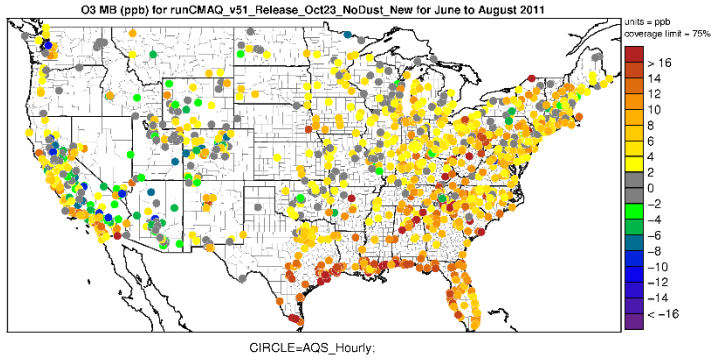


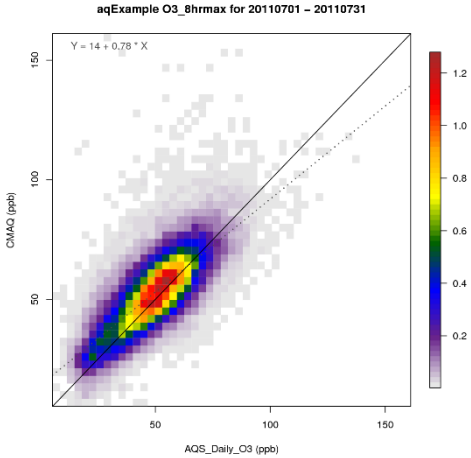












All R plots!

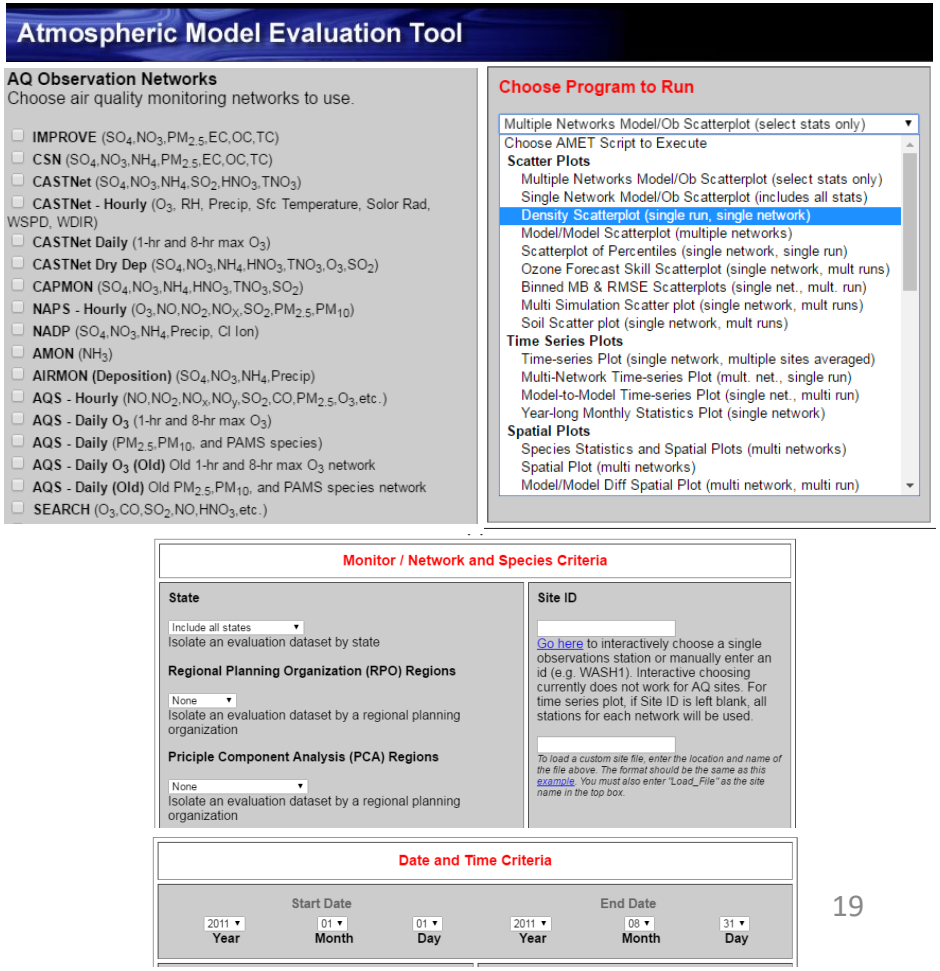
- Open source software publicly available on GitHub and developed by **Wyat Appel** (air quality eval.), **Rob Gilliam** (meteorological eval.)
- Evaluation of air quality model output against routine networks, e.g. AQS, IMPROVE, CSN, CASTNET, NADP, SEARCH, AMON, FLUXNET

Atmospheric Model Evaluation Tool

- Select model simulation, air quality monitor network data, plotting options through drop-down menus and selection boxes
- Underlying MySQL database allows for easy subsetting by location/time
- Clicking “Run Program” will run a single R plotting script

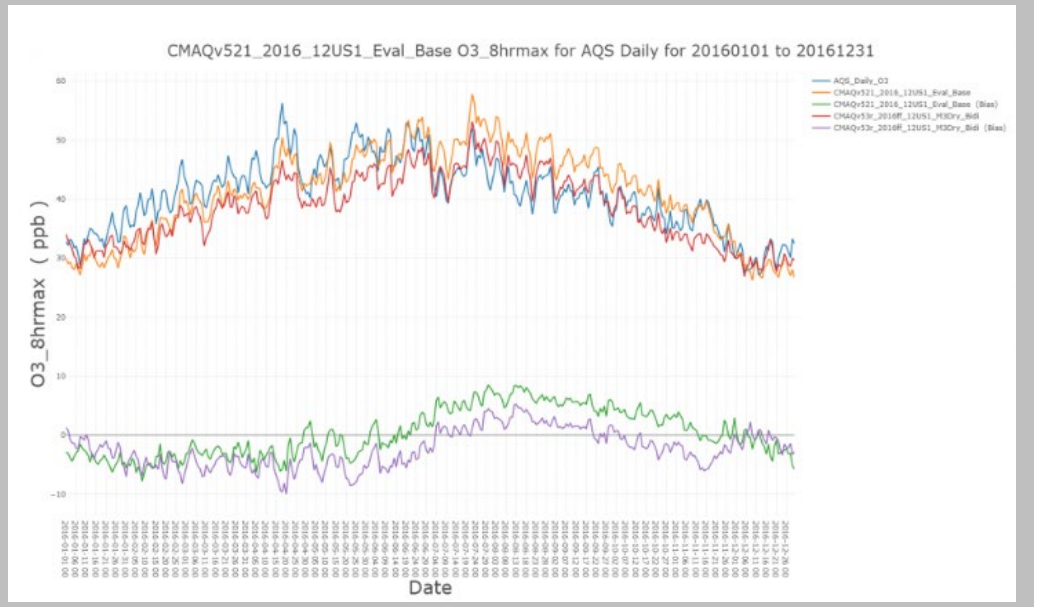
Developed over a decade ago in PHP for internal use.

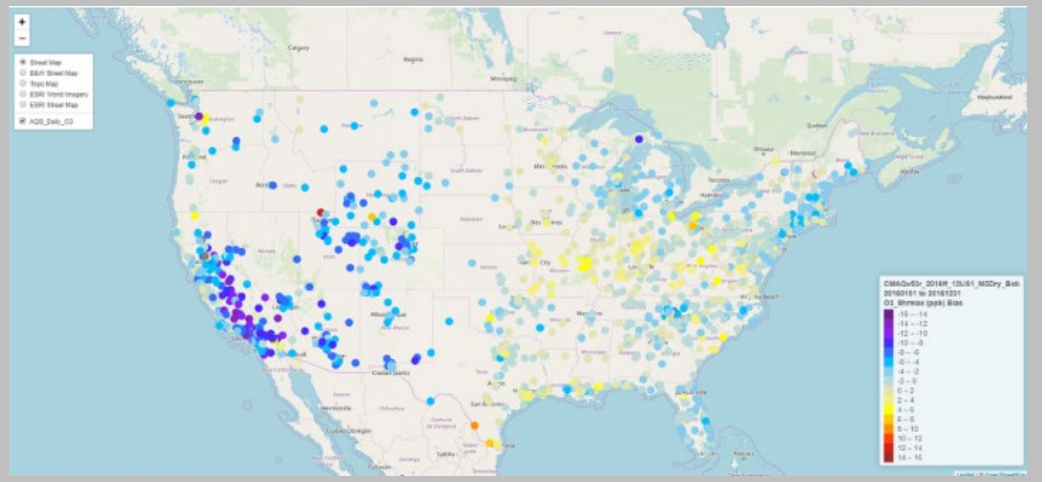
Web based interface available on EPA intranet



New Release: AMETv1.4

- Leveraging R's interactive plots for model evaluation
  - leaflet: maps
  - dygraph: time series
  - plotly: time series, bar charts, boxplots and scatter plots
- Next steps
  - New user interface for evaluation plots – R Shiny?
  - Other ideas from this week's workshop!





Contact and More Information

Contact

- Kristen Foley: Foley.Kristen@epa.gov
- Wyat Appel: Appel.Wyat@epa.gov

More Information

- CMAQ site: https://www.epa.gov/cmaq
- CMAQ GitHub Repository: https://github.com/USEPA/CMAQ
- AMET GitHub Repository: https://github.com/USEPA/AMET