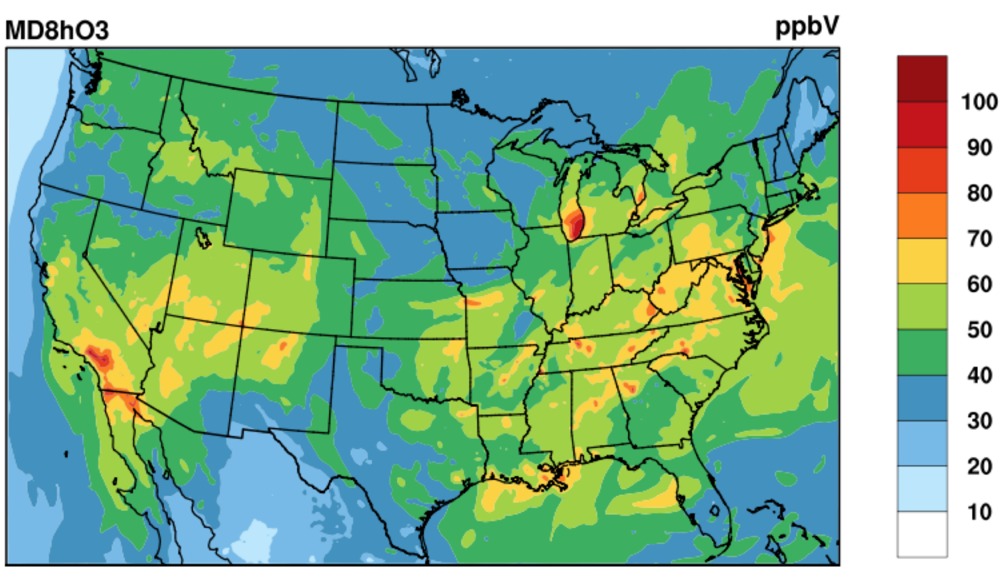
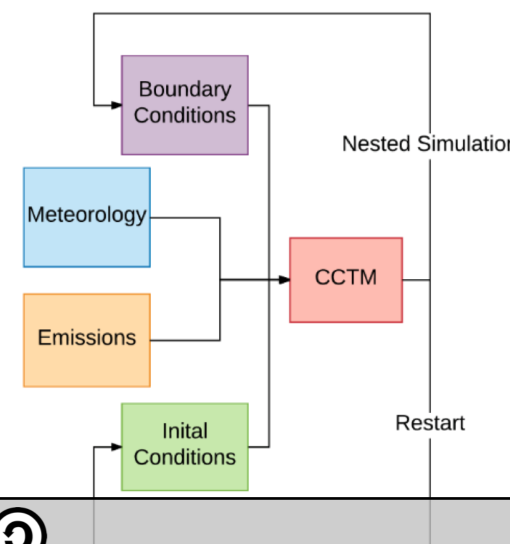
Introduction to Air Quality Modeling

Zac Adelman
LADCO Executive Director

December 2018

LADCO | LAKE MICHIGAN AIR DIRECTORS CONSORTIUM





Course Objectives

- Present key concepts of regional air quality modeling
- Provide audience with a background on the design, implementation, and operation of regional models
- Understand the strengths and limitations of the models
- Gain familiarity with the types analyses used to evaluate and present regional air quality modeling
- Present and define common modeling jargon
  - Key terms will be **highlighted** throughout the webinar

Charge Questions

Questions to consider during this webinar

- How is regional air quality modeling used at you organization?
- What types of knowledge would be useful to you when interacting with others about modeling?
- What is your role and your organization’s role in the air quality modeling process at LADCO?
- What interactions do you envision between yourself and LADCO with regard to air quality modeling?

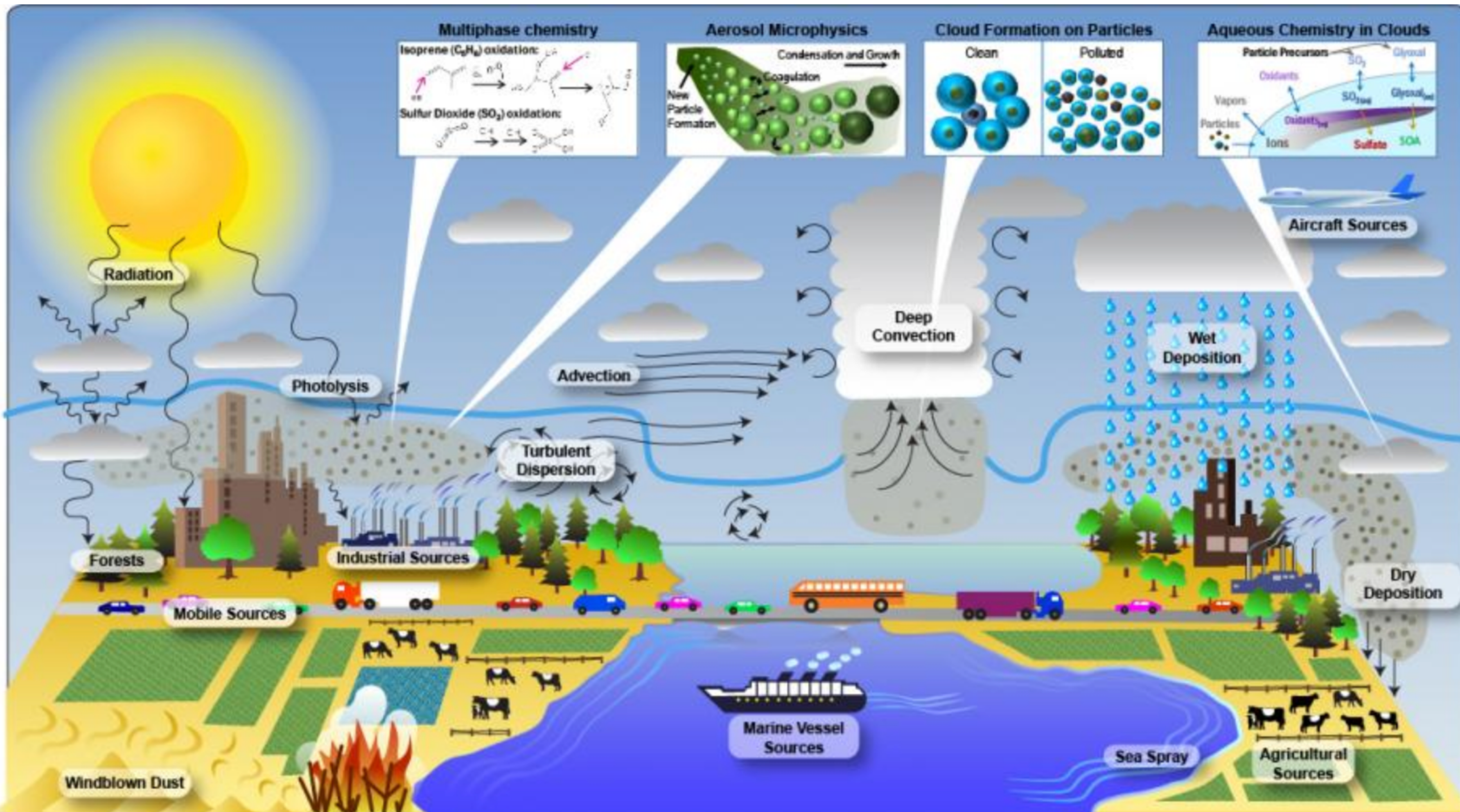
Webinar Outline

- Modeling processes
- Purpose of regional air quality modeling
- Conceptual approach and the continuity equation
- Model components
- Data flows
- Data and file formats
- Accessing data and software
- Operational details
- Model analysis and evaluation
- Limitations
- Future directions

Interchangeable Terms

Regional Air Quality Model =
Photochemical Grid Model (PGM) =
Chemistry-transport Model (CTM) =
Eulerian Model

PGM Processes



Source: www.epa.gov/cmaq

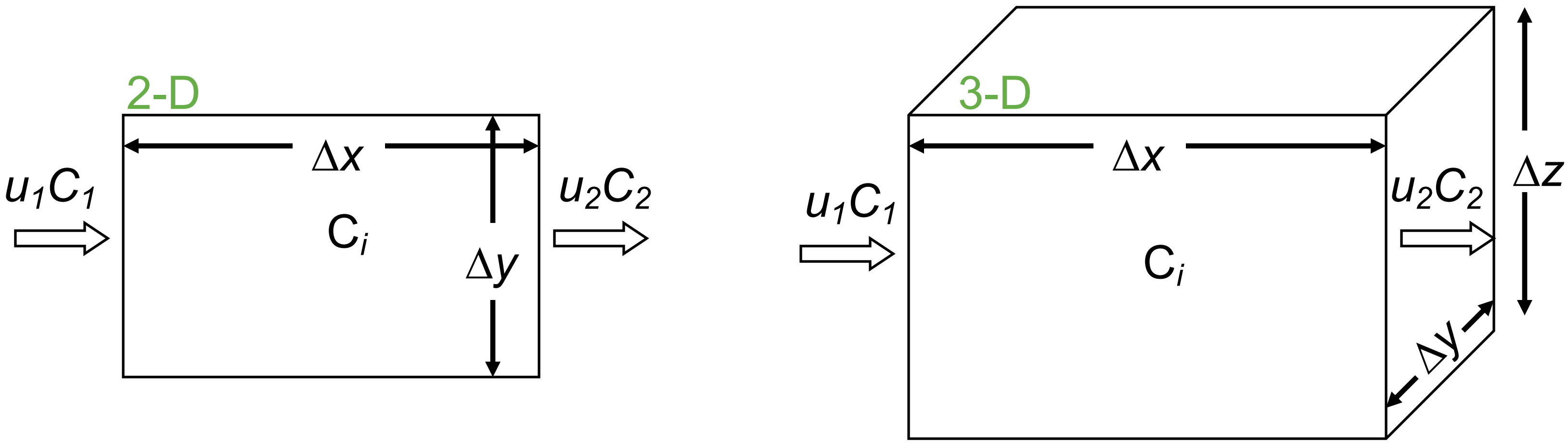
Purpose of Photochemical Grid Modeling

LADCO

- Air quality models integrate our understanding of individual processes into a coherent system
- Air pollution systems are non-linear
  - Need to establish links between emissions sources and ambient concentrations
- Measurements are sparse
  - Models provide a continuous spatial and temporal view of air quality
- Decision support
  - Platforms for testing the effectiveness and impacts of pollution mitigation policies
- Experimental
  - Identify knowledge gaps, quantify drivers, source-receptor relationships
- Deterministic
  - Randomness/noise is not considered in the solution: consistently reproducible
  - Bottom Up: Link processes together to a solution

Conceptual approach to PGMs

Extend the 2-D box model to three dimensions



u = wind vector

C_i = concentration of species *i*

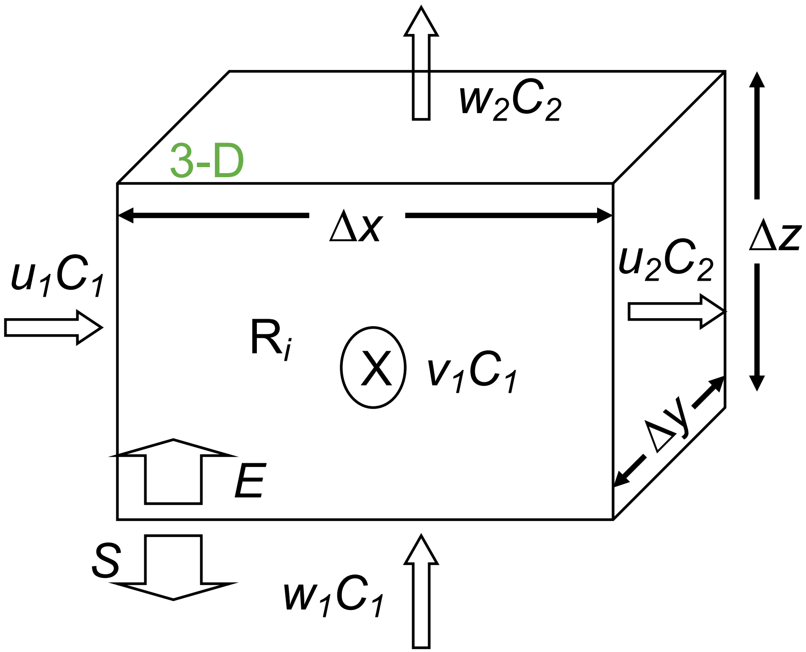
Basic Continuity Equation (flux in 1 direction):

$\Delta C\ \Delta x\ \Delta y\ \Delta z = u_1 C_1 \Delta y\ \Delta z\ \Delta t - u_2 C_2\ \Delta y\ \Delta z\ \Delta t$

Divide by Δt and volume: $\partial C / \partial t = - \partial (uC) / \partial x$

Expanded Continuity Equation

LADCO



Expanded Continuity Equation Derivation:

Expand to flux three dimensions:

$$\partial C/\partial t = -\partial(uC)/\partial x - \partial(vC)/\partial y - \partial(wC)/\partial z$$

$$= -\nabla \cdot (vC) \qquad \text{(flux divergence form)}$$

Add additional production and loss terms:

$$\partial C/\partial t + \nabla \cdot (vC) = D\nabla^2 C + R + E - S$$

u,v,w = wind vectors

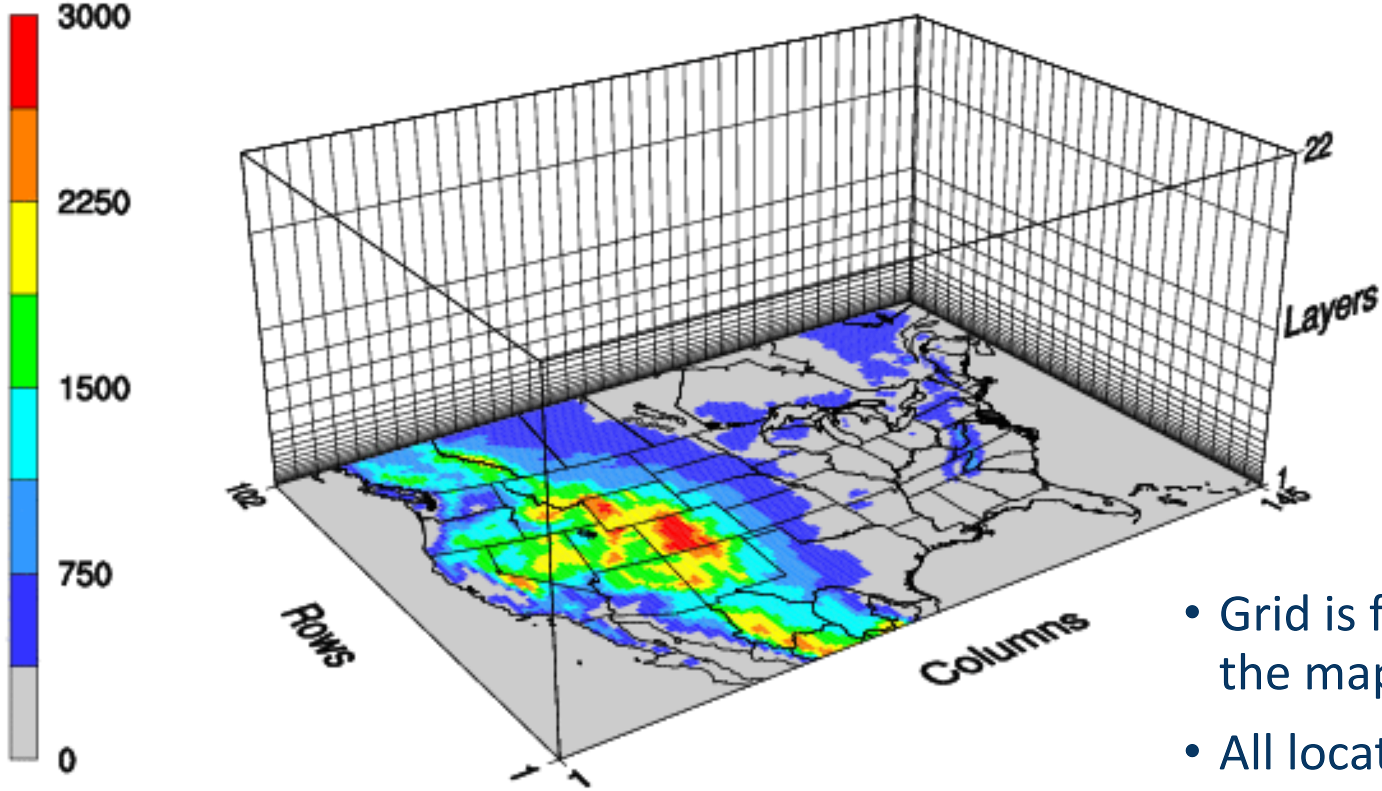
E = emissions

S = loss processes

R_i = Chemical formation of species *i*

D = Molecular diffusion coefficient

Overlay 3-D boxes on a grid



- Grid is fixed relative to the map coordinate
- All locations at all times
- Materials move through all cell faces

Full Continuity Equations

- Gas Continuity Equation

$$\partial C/\partial t + \nabla\cdot(vC) = D\nabla^2 C + R_{chemg} + R_{emisg} + R_{depg} + R_{washg} + R_{nucg} + R_{c/eg} + R_{dp/sg} + R_{ds/eg} + R_{hrg}$$

- Particle Continuity Equation (number)

$$\partial n/\partial t + \nabla\cdot(vn) = D\nabla^2 n + R_{emisn} + R_{depn} + R_{sedn} + R_{nucn} + R_{washn} + R_{coagn}$$

- Particle Continuity Equation (volume concentration)

$$\partial V/\partial t + \nabla\cdot(vV) = D\nabla^2 V + R_{emisv} + R_{depv} + R_{sedv} + R_{nucv} + R_{washv} + R_{coagv} + R_{c/ev} + R_{dp/sv} + R_{ds/ev} + R_{egv} + R_{qgv} + R_{hrv}$$

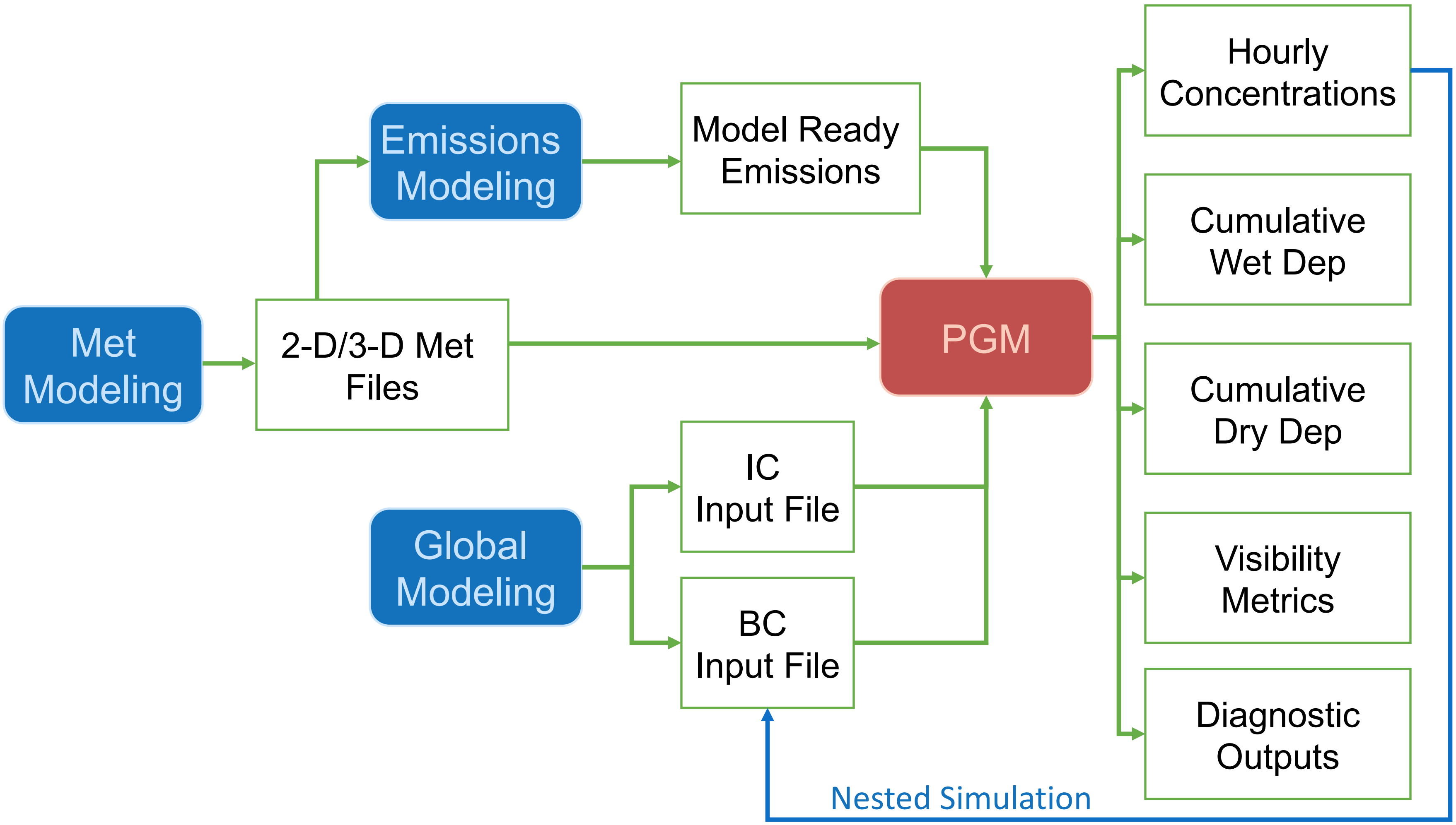
where,

C = pollutant concentration of species

R_i = loss/production by process i

Processes: chem=chemical production/loss, hr=heterogeneous reactions, nuc=nucleation, c/ev=condensation/evaporation, dp/s=depositional growth/sublimation, ds/e=dissolution/evaporation, wash=washout, dep=deposition, emis=emissions

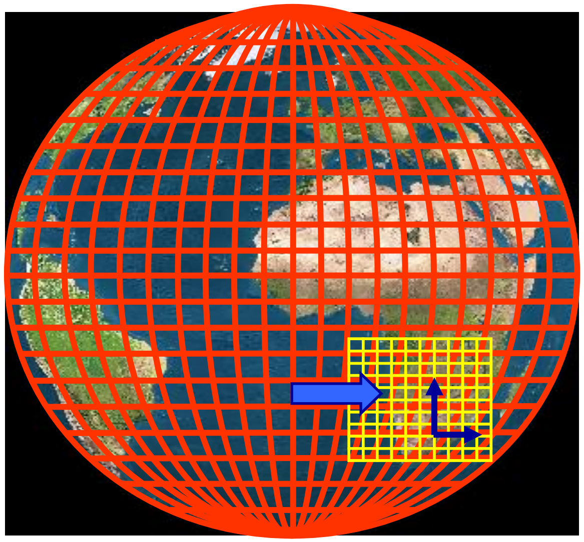
PGM Components



PGM Components: Meteorology

LADCO

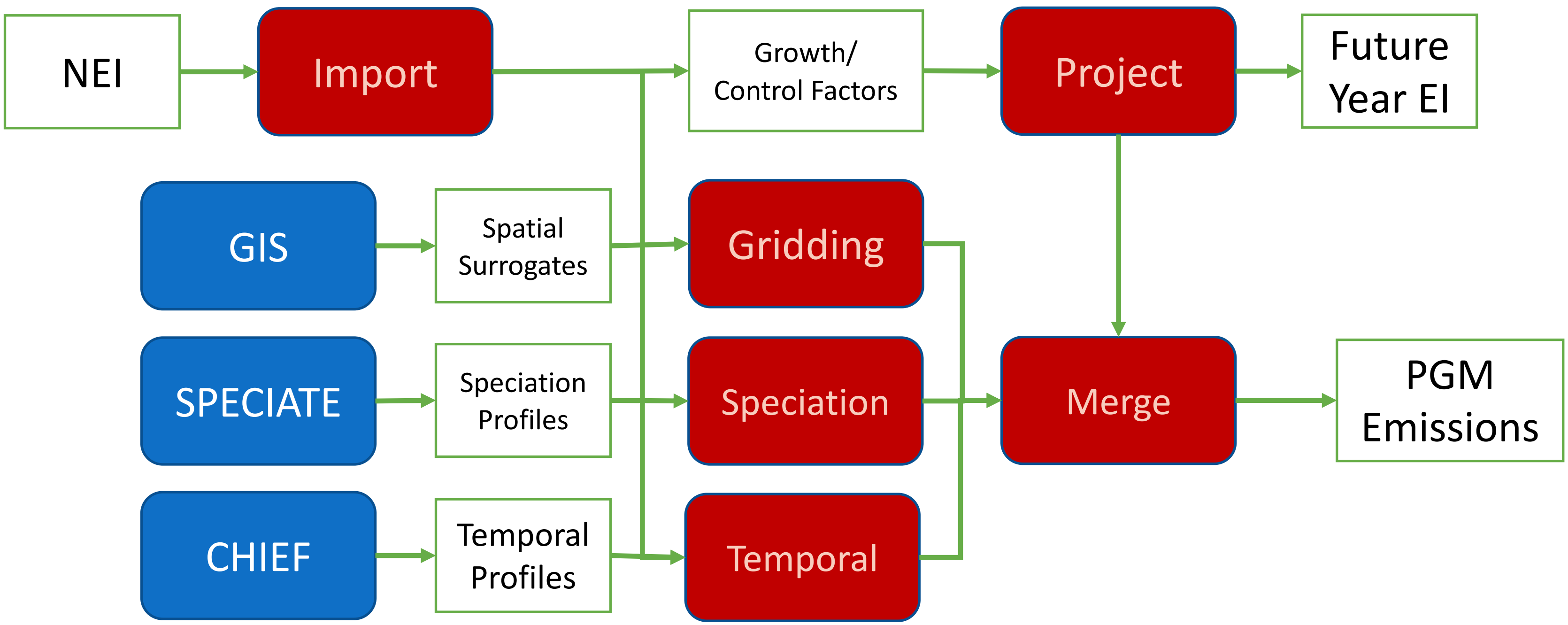
- The Weather Research Forecast (WRF) model is a numerical weather prediction model
- Meteorology dictates the temporal and spatial extent of a PGM simulation
- Dynamical Downscaling
  - WRF uses global General Circulation Model (GCM) data that has been fused with observations, aka reanalysis, to simulate weather



- Key Meteorology Modeling Processes Related to Air Pollution
  - Boundary Layer Dynamics: motions of the atmosphere near the earth's surface
  - Radiation: flux and transfer of solar energy
  - Microphysics: cloud formation, precipitation, and impacts on radiation/heating
  - Land-Surface Modeling (LSM): model of surface coverage, impacts heat and radiation exchange, turbulence, and deposition
  - Four Dimensional Data Assimilation (FDDA): aka "nudging" is the processes of assimilating observations to add value to the simulation

PGM Components: Emissions

- The Sparse Matrix Operator Kernel Emissions (SMOKE) processor converts inventory data to PGM-ready emissions files



PGM Components: IC/BC

- Initial Conditions (IC): concentrations at the first time step of the model simulation
  - Decay exponentially with time as the model simulation proceeds
- Boundary Conditions (BC): concentrations along the outer boundaries of the modeling domain
  - Have a continuous impact on the model solution, particularly from the upwind boundary
- Key Concepts
  - Model Spin Up: amount of time required to minimize the influence of initial conditions on a model simulation; applies to met and PGM, convention is 12-24 hours for met and 10 days for PGM
  - Nesting: placement of progressively smaller and finer grid resolution modeling domains within an outer parent domain
  - Downscaling: transfer of information from a course global scale PGM to a finer scale regional PGM; can also apply for meteorology data (global→regional)

What Does the PGM Do?

LADCO

PRIMARY PROCESSES

- Transport and Dynamics
  - Advection and diffusion algorithms simulate horizontal and vertical transport
- Photochemistry
  - Gas phase chemistry and photolysis, primarily for simulating ozone
- Aerosol Chemistry
  - Aerosol thermodynamics, chemistry, secondary formation, and optical properties
- Heterogenous Chemistry
  - Phase changes and aqueous chemistry
- Deposition
  - Dry and wet deposition of gases and particles

SECONDARY PROCESSES

- In-line Processes
  - Processes that were at some point done independent of the PGM
  - Emissions and photolysis rate calculations can be done in the PGM simulation
  - Can include meteorology-chemistry coupling
- Probing Tools
  - Algorithms to track/trace individual components of the PGM solution
  - Source apportionment
  - Process analysis
- Diagnostic Tools
  - Output extra variables on the details of the model equations

A Note on Gas-Phase Chemistry

LADCO

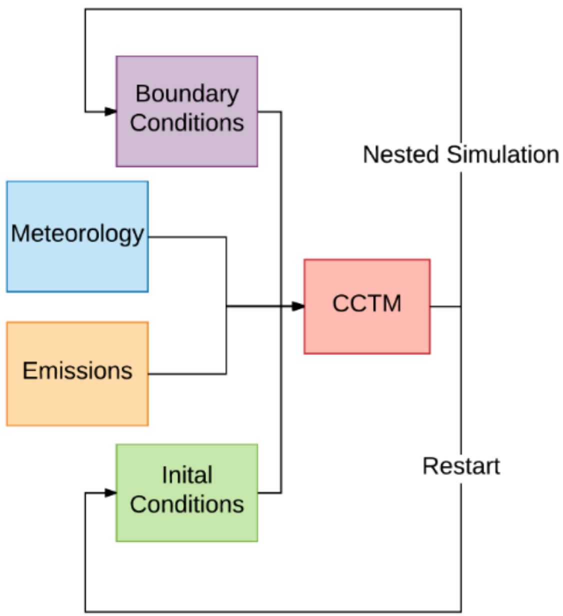
- The majority of the PGM solution time (>75%) goes to solving the chemistry equations
- Gas phase chemistry uses parameterized ("lumped") mechanisms to represent similarly reacting compounds
  - Inorganic reactions (i.e., no carbon involved) are the same across most lumped mechanisms
  - Carbon Bond is a structural parameterization that is based on the number bonds between carbons
  - SAPRC is a molecular parameterization that uses surrogates of molecular classes (i.e., C2-C6 paraffins) to reduce the solution
- Lumped mechanisms reduce thousands of compounds and reactions to mechanisms with a few dozen compounds and 100-200 reactions

Carbon Bond Lumping

PAR = single carbon bond. Propane (3C): C-C-C is 2 PAR

OLE = double carbon bond. Pentene (5C): C-C-C-C=C is 3 PAR + 1 OLE

PGM Data Flows



INPUT

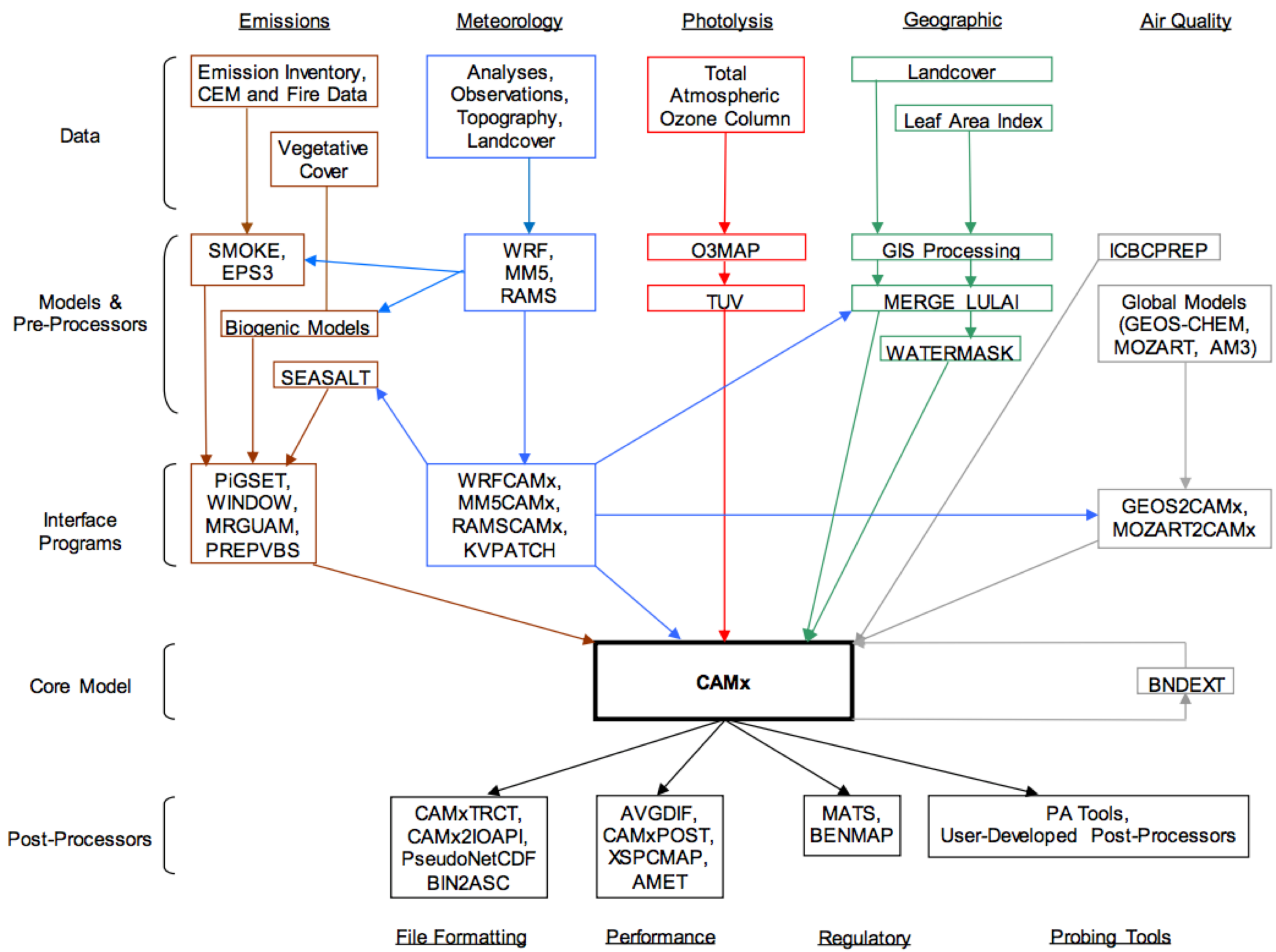
- Meteorology
- Emissions
- Initial Conditions
- Boundary Conditions
- Photolysis Rate Tables
- Landuse/Landcover Data
- Ozone and Aerosol Columns

OUTPUT

- Hourly Concentrations
- Hourly Wet/Dry Deposition
- Reconstructed Visibility
- Initialization/Restart File

- Source Apportionment
- Diagnostic Files
- Process Analysis

PGM Data Flows: CAMx



Data and File Formats

All models use prescribed formats for input/output data to facilitate data processing and quality control

- No two models are the same, leading to complexities in exchanging data and the need for data processors/translators

Data and File Formats

LADCO

- **ASCII/Text**: CSV or fixed column formats, typically used for inventory data and static look-up tables
- **Binary:** non-text, computer readable files; some are self-describing with headers/metadata, some are just a loosely organized block of data

- NetCDF: Network Common Data Form
  - General binary format of self-describing data arrays; portable across computers/operating systems
  - Many different variations of netCDF: WRF, I/O API (CMAQ), CF-compliant
- HDF: Hierarchical Data Format
  - Similarities to netCDF, designed for large amounts of data
- GRIB
  - Compressed, self-describing binary format used by meteorology community
- UAM
  - CAMx, legacy Urban Airshed Model binary format

Accessing PGM Data and Software

- All of the regional modeling and analysis software is open-source and accessible
- All PGMs and components are written in Fortran and tested on Linux systems
  - Fortran remains the best engineering language for complex math and multi-processor computing
  - Linux minimizes operating system overhead, provides better tools for accessing memory and processors
- Common Distribution Approaches
  - Tarballs: archives of data, source code, scripts
  - Git: open source repository of source code and scripts (https://github.com), git command line interface for downloading/updating codes

Operational Details

- All PGMs and WRF need to be compiled before use
  - Download source code, not executables
  - Requires a compiler: a software tool that builds source code into programs that can be run at Linux command line
    - Upside: programs can be optimized for a particular hardware configuration
    - Downside: requires > novice Linux skills; if things go wrong it can be difficult to troubleshoot if you don't know the programming language
- Implementation
  - Interpreted Language (e.g., Python, R, NCL): access functionality of the language without needing to compile anything; typically for analysis/processing
  - Compiled Language (e.g, Fortran, C, C++): implementation requires building executables before accessing the language functions
  - Libraries: provide application programming interfaces (APIs) to connect/access file formats and programming languages

Operational Details

- What's a modeling script vs a program?
  - The actual modeling software is a compiled program (executable, binary)
  - Fortran programs require variables (i.e., input/output file names, configuration settings) to be set when they are executed
  - Scripts are text files, typically written in a shell language related to a Linux user environment (bash, C-shell) that set variables can call an executable.
  - Example C-shell script:

```
#/bin/csh

setenv INPUT /home/camx/input.txt
setenv OUTPUT /home/camx/output.txt
set PROGRAM = /apps/CAMx_v5.1/camx.exe

$PROGRAM
```

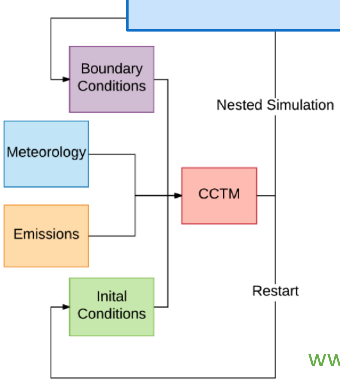
Operational Details
Typical Workflows

Meteorology
- Define time periods/domains
- Modeling Protocol
- Download reanalysis and observational data
- Compile and test WRF
- Develop operational scripts
- Shakeout modeling
- Evaluation
- Production modeling
- Evaluation
- MPE Report
- Post-process for PGM

Emissions
- Modeling Protocol
- Download emissions data
- Generate spatial surrogates for new domains
- Download SMOKE Platform
- Develop operational scripts
- Generate biogenic, seasalt, dust emissions
- Shakeout modeling
- Evaluation
- Production Modeling
- Evaluation

PGM
- Modeling Protocol
- Download/compile PGM
- Test PGM installation
- Develop operational scripts
- Shakeout modeling
- Evaluation
- Production modeling
- Post processing
- Evaluation
- MPE Report

Boundary Conditions
- Request/download global modeling data
- Downscale to PGM inputs
- Evaluation?



Model Analysis Techniques

- Air quality models output hourly, 3-D (x,y,z) concentrations of gas and aerosol species
- Analysis begins with concatenating/extracting variables from hundreds of files into a few manageable datasets
  - Example: Daily maximum layer 1 O_3, NOx ($NO+NO_2+HONO$), and total $PM_{2.5}$ ($SO_4 + NO_3 + NH_4 + EC + OC + PM_{Other}$)
- Common Analyses
  - Averaging, daily maximums
  - Pairing observations with model grid cells
  - Visualization
  - Bivariate statistics (error, bias, correlations)
- A robust set of tools are available for post-processing/analysis
  - Scripts for driving interpreted languages (R, NCL, NCO, Python)
  - Visualization software: VERDI, IDV, R, NCL, Panoply
  - Analysis packages: AMET, I/O API Tools, Metstat

Model Performance Evaluation

LADCO

- MPE = Model Performance Evaluation
- Question why a model is doing what it is doing
- What are the inherent uncertainties and how do they impact the model results?
- Evaluation Techniques
  - Comparisons against observations/measurements
  - Sanity checks
  - Looking for known trends (diurnal/seasonal patterns, chemical signatures, source signatures)
- Comparisons to measurements
  - Pair in space and time
  - Compare predicted vs observed maximums (means)
  - Paired in space but not time
  - Statistical metrics

Model Performance

LADCO

When has a model demonstrated acceptable performance

- Not a simple question
  - A pollutant may consist of many components (e.g., PM consists of sulfate, nitrate, ammonium, organic carbon, elemental carbon, soils, coarse mass, etc.)
  - The PGM may simulate some components well, others poorly
  - Right answer for the wrong reasons?
- What model performance metrics should be used to facilitate inter-comparisons between applications and models?
- What metrics should regulators use for assessing attainment?

Model Performance

Model performance goals

- What level of accuracy is considered to be close to the best a model can be expected to achieve?
  - Regulatory vs. scientific applications
- Different performance goal and criteria developed for:
  - Different components of each pollutant
  - Different seasons of the year
  - Different parts of the country
  - Urban vs. rural sites
  - Complex vs. simple terrain
  - Clean vs. polluted days

Model Performance

Performance Goals and Criteria

- Not a pass/fail metric, used to help interpret model skill
- Based on literature reviews and past experience

Fractional Bias	Fractional Error	Comment
≤±15%	≤35%	O_3 model performance goal that would be considered very good model performance for PM species
≤±30%	≤50%	PM model performance Goal, considered good PM performance
≤±15%	≤35%	PM model performance Criteria, considered average PM performance. Exceeding this level of performance for PM species with significant mass may be cause for concern.

MPE Challenges

- Modeling scales are growing both spatially and temporally
  - Datasets becoming larger: more difficult to manage/move around
  - Need to process and digest voluminous amount of information
- Heterogeneous nature of observational datasets
  - Vary by network, by quality, by format, by frequency
- Measurement or model artifacts
  - What is modeled is not always measured
  - Need adjustments before comparisons
- Problem of incommensurability
  - Comparing point measurement with volume average

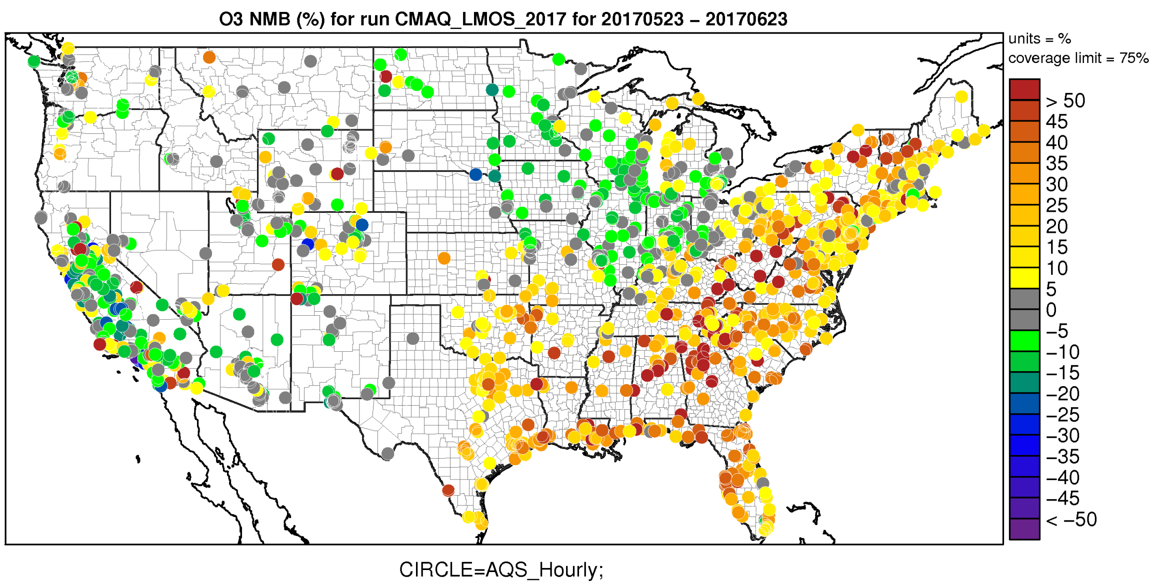
Operational Evaluation

- Mostly quantitative
- Compute suite of statistical measures of performance
  - Peak Prediction Accuracy, bias metrics (MB, MNB, NMB, FB), error metrics (RMSE, FE, GE, MGE, NMGE), etc.
  - "Goodness-of-fit" measures (based on correlation coefficients and their variations)
  - Various temporal scales
- Time-series analyses
  - Hourly, weekly, monthly
- Grid (tile) plots
- Scatter plots
- Soccer Plots

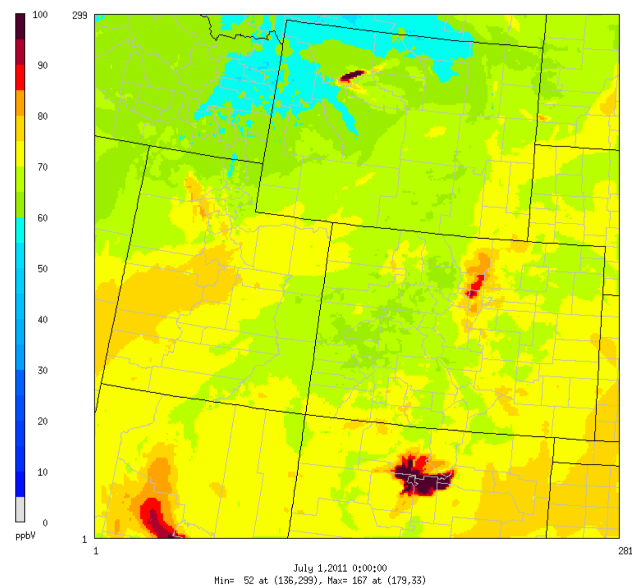
Spatial Plots

Statistical spatial plots: MPE stats at monitors

Tileplots: gridded model output concentrations

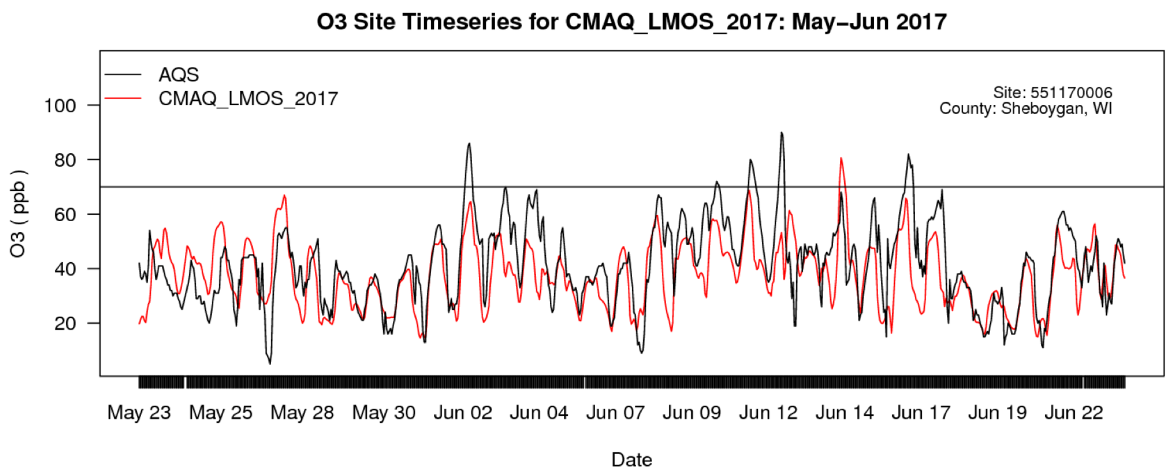


May 23 – June 23, 2017 O3 Normalized Mean Bias

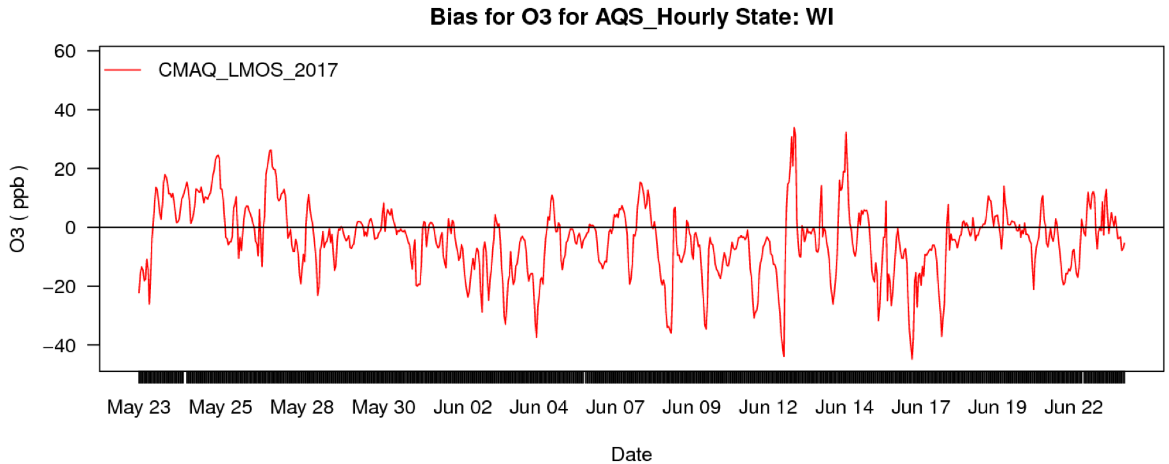


July 2011 Monthly Max O_3

Timeseries Analysis



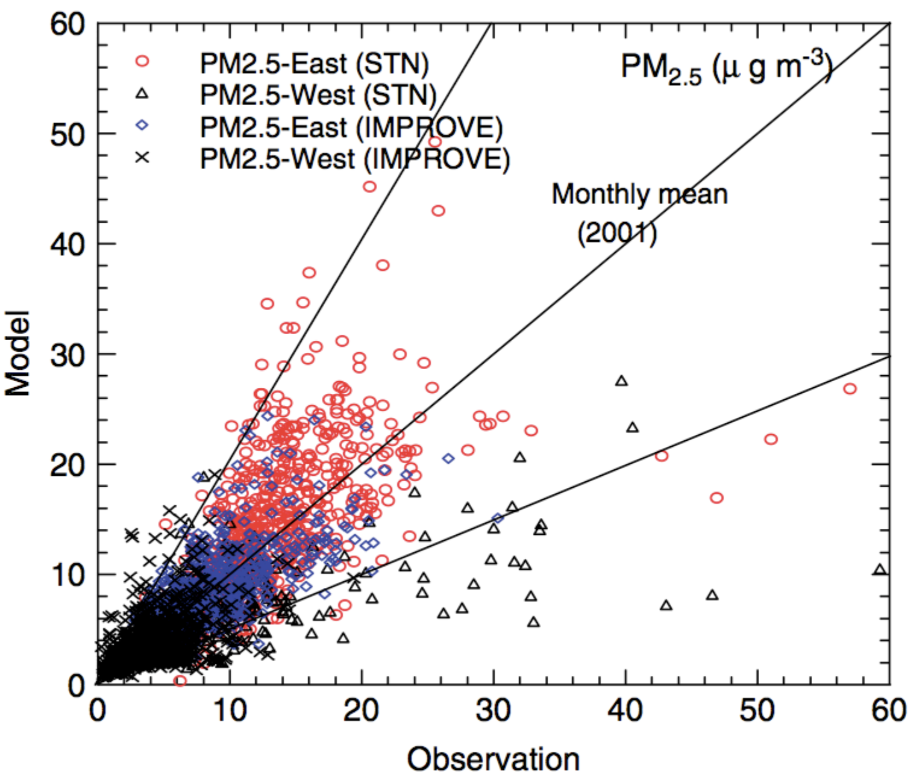
May – June 2017 Daily Max O_3 at Kohler Andrae, WI



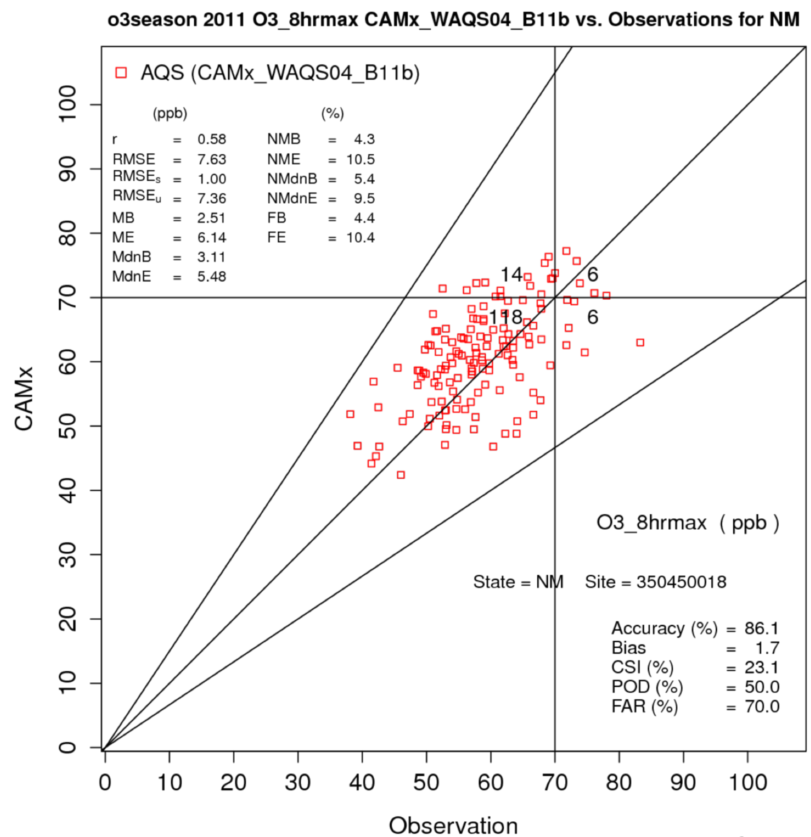
CAMx Bias in Daily Max O_3

Scatter Plots

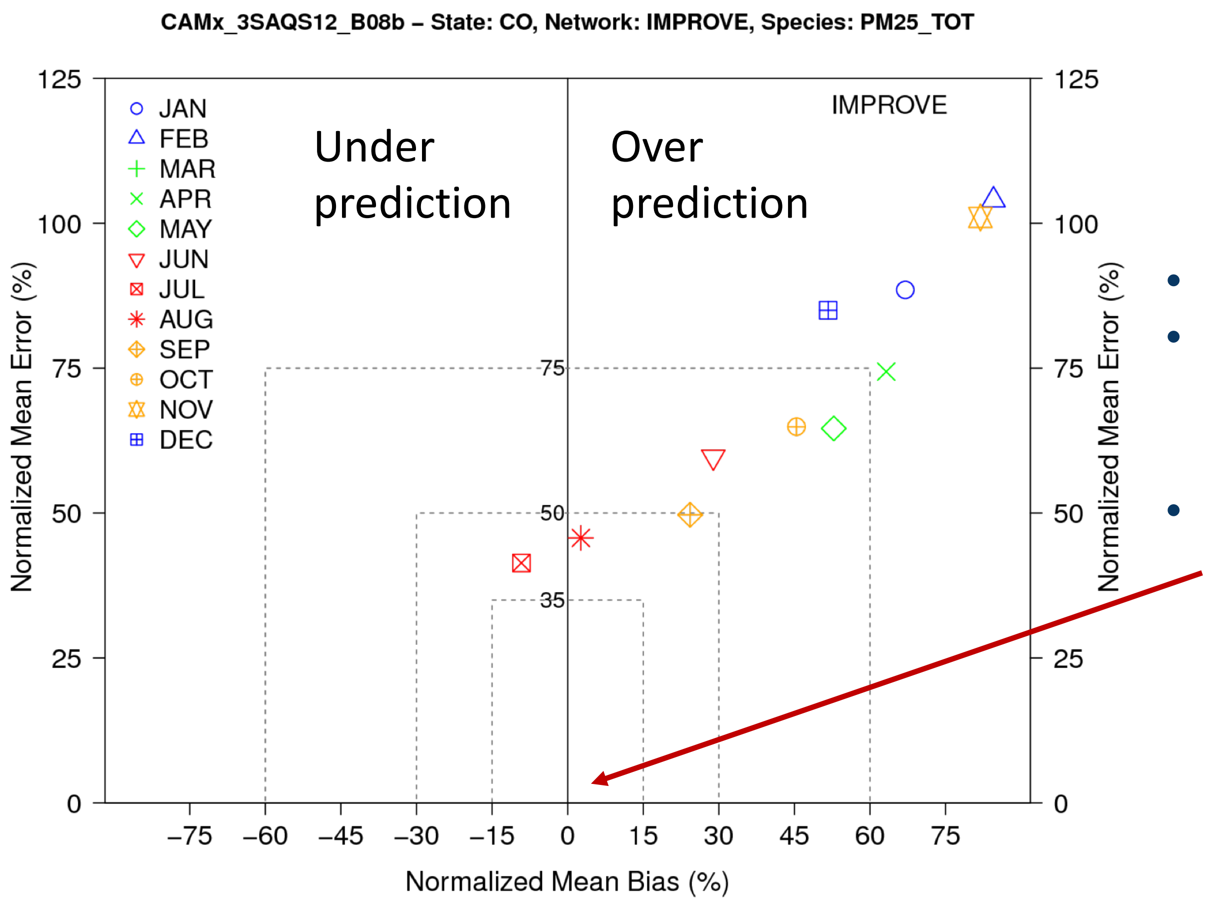
2001 Monthly Mean $PM_{2.5}$ at STN and IMPROVE monitors in the East and West U.S.



Quadrant (Skill) Plot of 2011 Summer Season MDA8 O_3 @ New Mexico AQS Sites



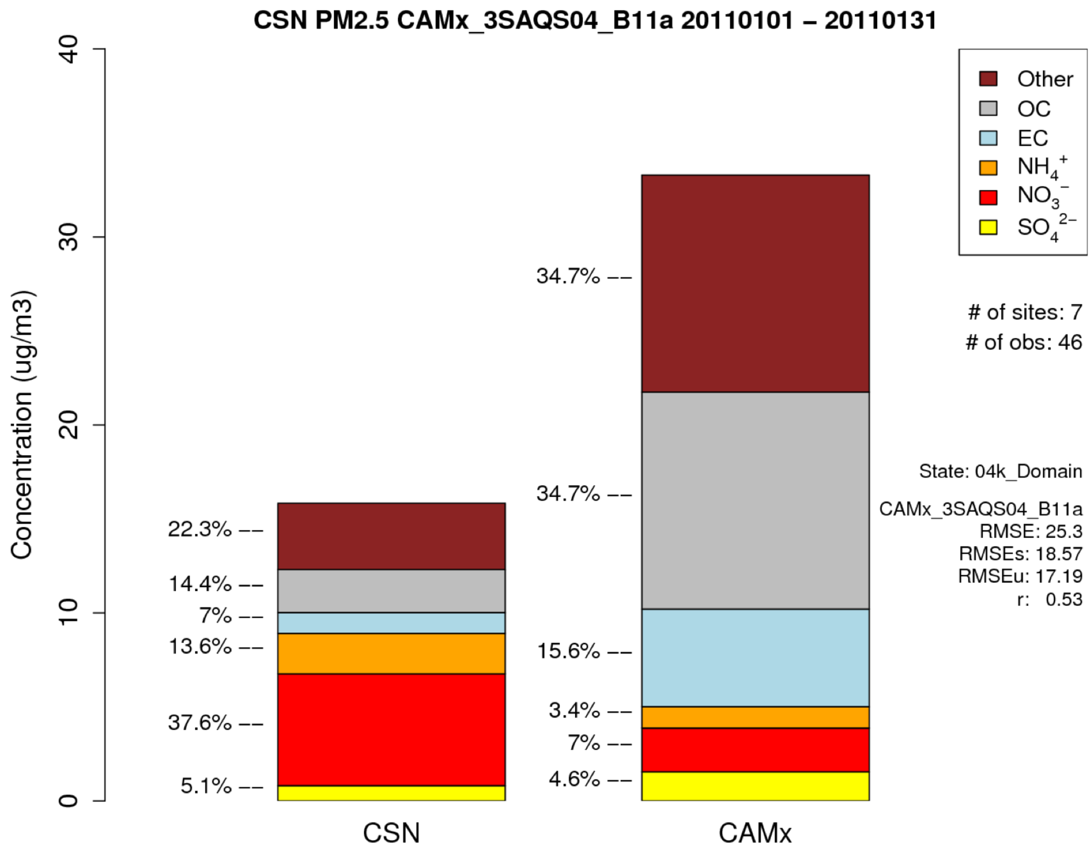
Soccer Plots



- Error vs Bias Plot
- Dashed lines indicate model performance goals and criteria
- The closer to the origin, the better the model performance

2008 Monthly Average PM2.5 at IMPROVE sites in Colorado

PM Speciation Stacked Bars



- Period averaged PM species at a site (or multiple sites) in a network
- Compares observations and model

January 2011 Monthly Average $PM_{2.5}$ species at all CSN sites in the Intermountain West modeling domain

Diagnostic Evaluation

LADCO

- Qualitative and Quantitative
- Compute pollutant ratios
  - Metrics different for each problem being diagnosed / studied
    - O_3/NO_z,, H_2O_2/HNO_3 for NO_x versus VOC limitation
    - NO_z/NO_y for chemical aging
    - PM species ratios such as NH_3/NH_x, NO_3/(total nitrate) for gas-particle partitioning, NH_4/SO_4, NH_4/NO_3, etc.
- Innovative Techniques
  - Principal Component Analyses
  - Process Analyses
  - Tagged species tracking
    - Source Apportionment: CAMx OSAT & PSAT

Sensitivity Analysis

LADCO

- How does the model respond to a change in an independent parameter?
- In air quality modeling sensitivity analysis is used to:
  - Quantify the atmospheric response to an emissions control
  - Determine the uncertainty in the response
  - Perform inverse modeling and data assimilation
- Methods
  - Brute Force Method (BFM)
  - Decoupled Direct Method (DDM)
  - Green's Function Method
  - Automatic Differentiation in Fortran (ADIFOR)
  - Adjoint (inverse model) Method

Brute Force Method (BFM)

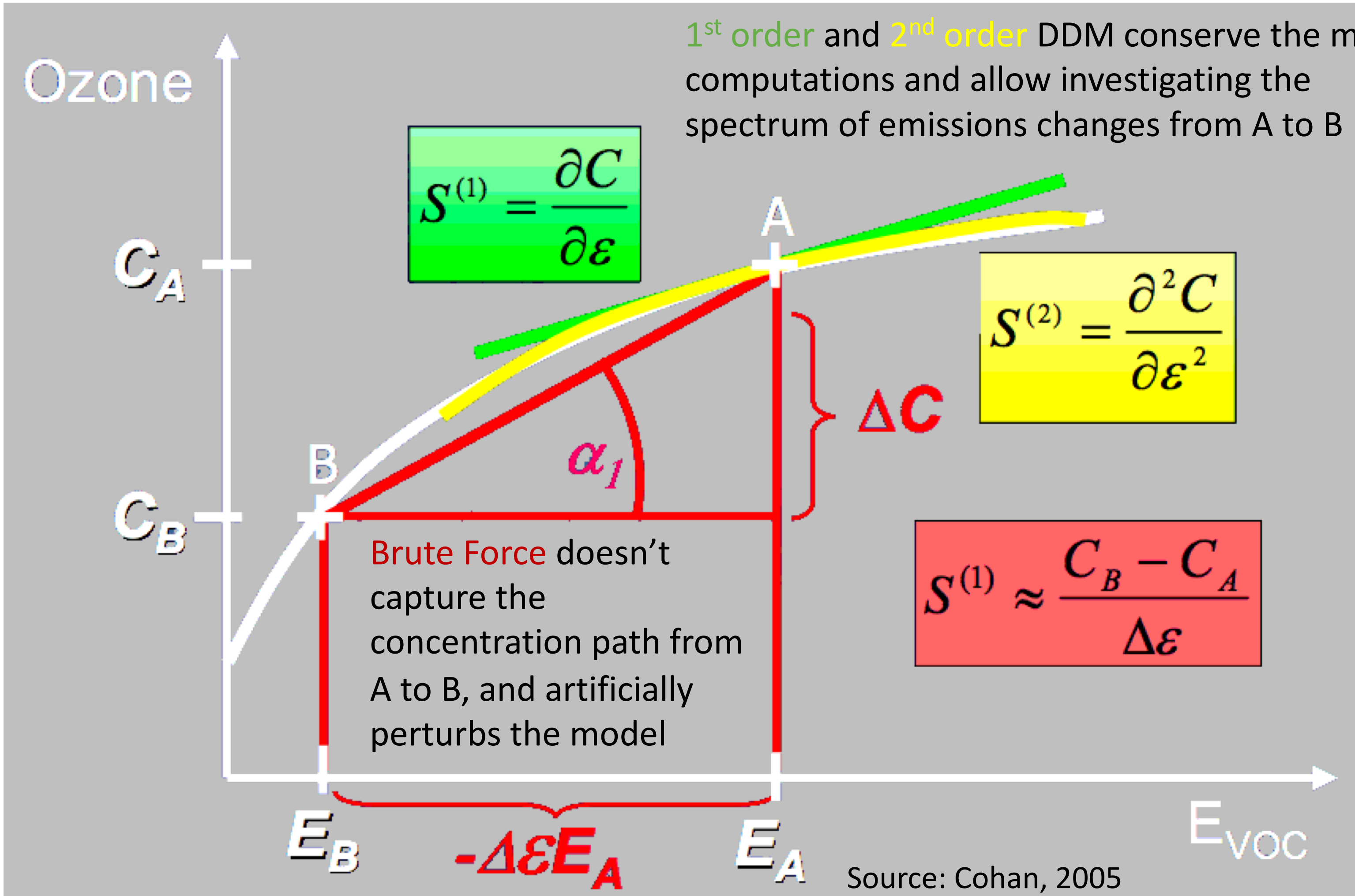
LADCO

- Vary the input parameters one-by-one in separate model simulations and evaluate the change in the predicted concentrations
- Pros
  - Easy to apply
  - One additional run for each sensitivity parameter
  - Straightforward to interpret the results
- Cons
  - Noisy for small perturbations
  - Unrealistic for large perturbations - computationally expensive

Decoupled Direct Method (DDM)

- Sensitivity equations are derived by taking the derivative of the CTM model concentration equations relative to a given sensitivity parameter
- These equations are integrated decoupled from concentrations using the same numerical operators as the CTM - easy to implement
- Pros
  - Computational and space efficiency
  - Accurate, avoids numerical error for small perturbations (local slope)
  - Can examine non-linearity with HDDM
- Cons
  - Complex to implement in a model
    - Aqueous & aerosol chemistry especially difficult
    - Must update each time PGM model changes
  - Inaccurate for large perturbations
    - This can be addressed by using high-order coefficients

Brute Force vs DDM



Ozone Source Apportionment

LADCO

- Tagged Species: Chemically active tracers follow ozone precursors through the chemistry mechanism
  - Tracers are spectators to the model chemistry and track the chemical conditions (NOx or VOC limited) under which O_3 is formed

Simplified Ozone Chemistry Mechanism

$$VOC\cdot + NO \rightarrow NO_2 + VOC\cdot\cdot$$

$$NO + O_3 \rightarrow NO_2 + O_2$$

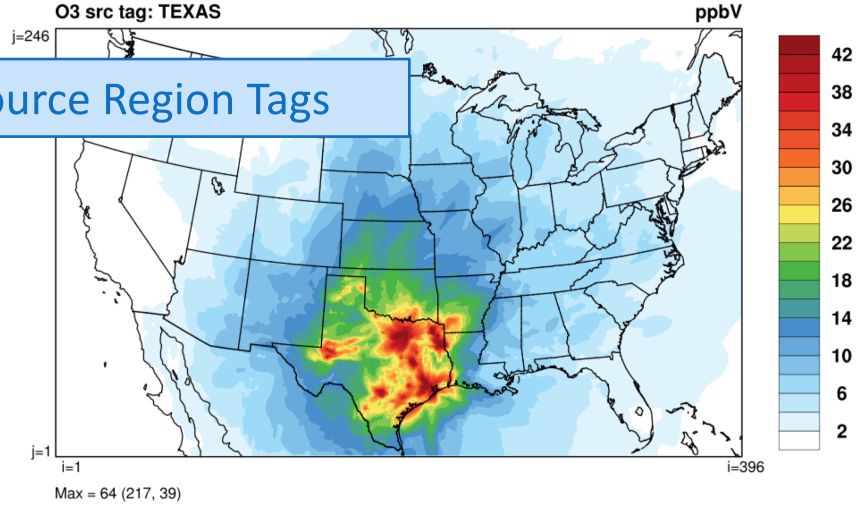
$$NO_2 + hv \rightarrow NO + O$$

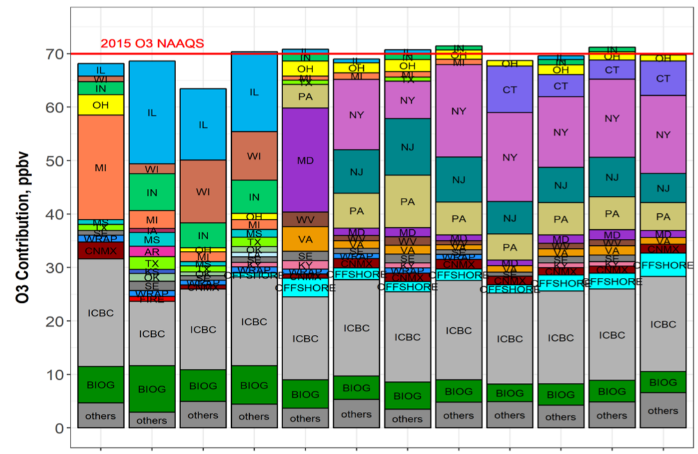
$$O + O_2 + M \rightarrow O_3 + M$$

- CAMx Ozone Source Apportionment Tool (OSAT)
  - Attributes O_3 formation to biogenic or anthropogenic VOC without distinguishing the "controllable" nature of the emissions
- CAMx Anthropogenic Culpability Assessment (APCA)
  - Considers "controllable" emissions by attributing O_3 formation to anthropogenic precursors
  - Example: in VOC-limited conditions where biogenic VOC and anthropogenic NOx react to form O_3, OSAT would attribute the O_3 to biogenic sources; APCA attributes the O_3 to the anthropogenic NOx

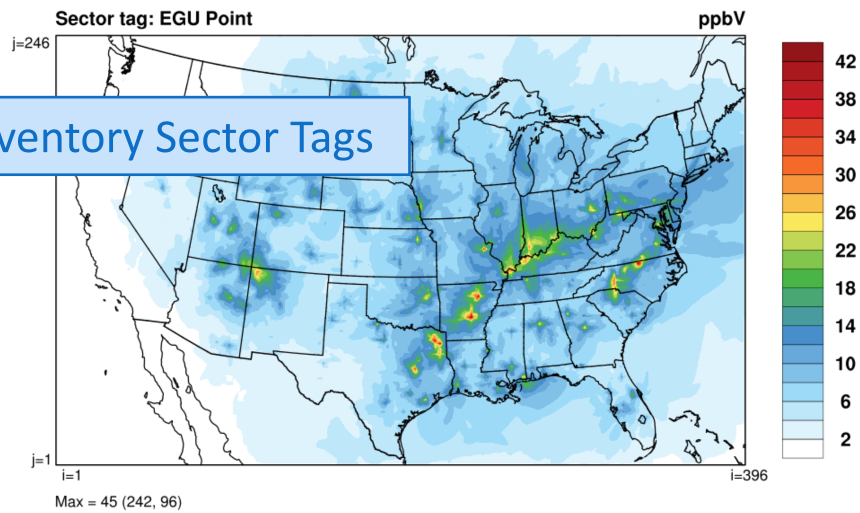
Source Apportionment
APCA Example

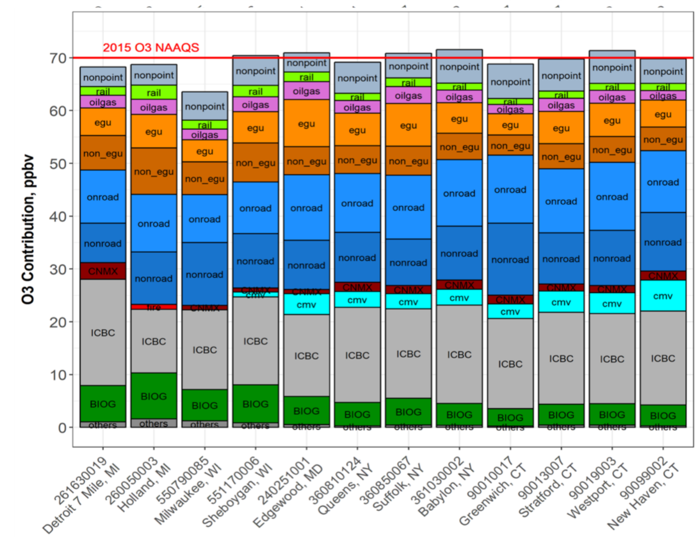
Source Region Tags





Inventory Sector Tags

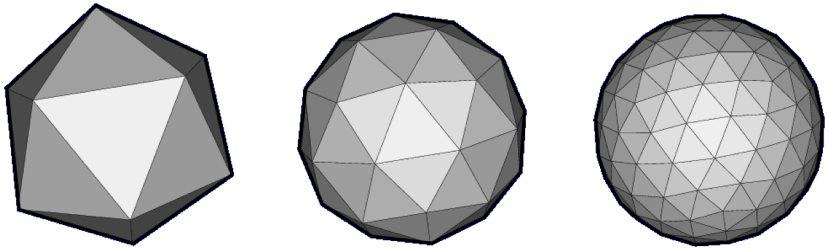




Limitations of PGMs

- Quantifying uncertainty is difficult because of the modularity of the PGM platforms
  - How do you estimate uncertainty in met and then propagate to the PGM?
- Abstraction/Distortion/Deletion
  - Parameterization is needed to balance accuracy with computing economy
- Algorithms and evaluation focus on conditions around the mean
  - Air quality planning is concerned with outliers/peaks
- Model algorithms are limited at fine spatial and temporal scales
  - WRF is a mesoscale model, PGM's are regional models; localized pollution issues challenge the capabilities of the PGM system
- Lag between research and implementation of science
- We don't know what we don't know
  - Organic aerosol, gas-phase aromatic, heterogeneous chemistry are evolving

Future Directions



- Alternative model grid structures
- Adaptive grids and time scales
- Probabilistic applications
- One-Atmosphere+: Expand the representation/integration of all atmospheric processes
- Meteorology←→Chemistry: expand two-way coupled modeling to planning
- Earth-System Modeling: Expand the two-way coupled meteorology-chemistry model to all media (soil, water)
- Data assimilation: Improve PGM skill with nudging to different data platforms (in-situ, remote sensing, sensors)
- Artificial Intelligence: User AI and neural networks to optimize model solutions to novel problems

Final Note

LADCO

- PGMs are the best platform that we have for integrating all of our knowledge about air pollution
  - They do amazingly well given that they parameterize a system with a lot of chaos and uncertainty
- PGMs are decision support tools
  - Understand their strengths and limitations
  - They aren't designed to provide **THE** answer, just an answer, or a set of answers
- To Learn More
  - EPA Modeling Guidance for Demonstrating Air Quality Goals (Nov 2018)
  - CAMx and CMAQ User's Guides
  - LADCO modeling protocols and MPE reports
  - Dive In! Download model code and compile, run a test case