Definition, objectives, and types of models

Daniel Jacob, Harvard University
Atmospheric chemists are interested in a wide range of issues:

**LOCAL**
- Plume releases
- Urban smog
- Industrial pollution

**REGIONAL**
- Visibility
- Regional smog
- Acid rain

**GLOBAL**
- Ozone layer
- Climate forcing
- Biogeochemical cycles
To address these issues we need models: simplified representations of complex system amenable to analysis

- Minimize number of variables
- Ignore minor processes
- Draw on empirical relationships

- Interpret observations
- Gain understanding of processes
- Make future projections

“All models are wrong, but some are useful” (G. Box, 1976)
What kind of models?

- **Physical models**: simplify the physical equations describing the system
  - The most fundamental quantitative approach

- **Heuristic models**: draw mental inferences based on knowledge/experience
  - Your brain is a powerful computer. Use it!

- **Statistical models**: use empirical relationships to make projections
  - Need physical basis, but can be very powerful
Statistical model of extreme winter haze events in Beijing

Point process fit of DJF 2009-2017 daily PM$_{2.5}$ to wind velocity and relative humidity

Expect lower RH over China in future climate

less frequent extreme air pollution events

Pendergrass et al., in prep.
Physical modeling by solution of continuity equation

Solve continuity equations for number densities \( \mathbf{n} = (n_1, \ldots n_K) \) of ensemble of \( K \) species:

Eulerian flux form

\[
\frac{\partial n_i}{\partial t} = - \nabla \cdot (n_i \mathbf{U}) + P_i(n) - L_i(n)
\]

Eulerian advective form in terms of mixing ratios \( \mathbf{C} = (C_1, \ldots C_K) \):

\[
\frac{\partial C_i}{\partial t} = - \mathbf{U} \cdot \nabla C_i + P_i(C) - L_i(C)
\]

Lagrangian form (moving frame of reference):

\[
\frac{dC_i}{dt} = P_i(C) - L_i(C)
\]

where

\[
\frac{d}{dt} = \frac{\partial}{\partial t} + \mathbf{U} \cdot \nabla
\]
Break down dimensionality of continuity equation by operator splitting

Solve for transport and local terms separately over time steps $\Delta t$

$$\frac{\partial C_i}{\partial t} = -\mathbf{U} \cdot \nabla C_i + P_i(\mathbf{C}) - L_i(\mathbf{C})$$

Advection (and other transport):  
Chemistry (and other local processes): 

$$\frac{\partial C_i}{\partial t} = -\mathbf{U} \cdot \nabla C_i \quad \text{(Eulerian)}$$  
$$\frac{dC_i}{dt} = P_i(\mathbf{C}) - L_i(\mathbf{C})$$

Advection equations:  
PDEs with no coupling between species  

Chemical equations:  
$K$-dimensional system of ODEs
Eulerian models partition atmospheric domain into gridboxes

This discretizes the continuity equation in space

- Present computational limit \( \sim 10^8 \) gridboxes

- In global models, this implies a grid resolution \( \Delta x \) of \( \sim 10\text{-}100 \) km in horizontal and 0.1-1 km in vertical

- Courant number limitation \( u \Delta t / \Delta x \leq 1 \); in global models, \( \Delta t \sim 10^2\text{-}10^3 \) s
Eulerian models often use equal-area or zoomed grids

Equal-area grids: avoid singularities at poles
- icosahedral triangular
- cubed-sphere

Zoomed grids: increase resolution where you need it (or when, in an adaptive grid)
- nested
- stretched
Regional models: limited domain, boundary conditions at edges

WRF domain with successive nests

1-way nesting  2-way nesting
Hybrid sigma-pressure vertical coordinate system

Midpoint pressure at level $k$: $p_k = A_k p_0 + B_k p_s$

$A_k = 0$

$B_k = 0$
Lagrangian models track points in model domain (no grid)

- Transport large number of points with trajectories from input meteorological data base ($U$) over time steps $\Delta t$
- Points have mixing ratio or mass but no volume
- Determine local concentrations in a given volume by the statistics of points within that volume or by interpolation

**PROS over Eulerian models:**
- stable for any wind speed
- no error from spatial averaging
- easy to parallelize
- easily track air parcel histories
- efficient for receptor-oriented problems

**CONS:**
- need very large # points for statistics
- inhomogeneous representation of domain
- individual trajectories do not mix
- cannot do nonlinear chemistry
- cannot be conducted on-line with meteorology
Lagrangian receptor-oriented modeling

Run Lagrangian model backward from receptor location, with points released at receptor location only

- Efficient quantification of source influence distribution on receptor ("footprint")

flow backward in time

(\(x, y, z\))

receptor

source
Representing non-linear chemistry

Consider two chemicals $A$ and $B$ emitted in different locations, and reacting by

$$A + B \rightarrow \text{products}$$

**Eulerian model**

A and B react following the mixing of gridboxes

**Lagrangian model**

A and B never react
Reducing model dimension
...because a model shouldn’t be more complicated than it needs to be

- 2-D models used in stratosphere where zonal gradients are weak
  - Turbulent diffusion parameterization of eddy meridional transport
- 1-D models used for boundary layer mixing when horizontal gradients are weak
  - Turbulent diffusion and non-local (convective) transport parameterizations
- O-D (box) models when flux divergence can be neglected in continuity equation
  - Chemical steady-state models for short-lived species
  - Global box models for long-lived species
  - Chemical mechanism diagnostic studies
Gaussian plume modeling of point source dispersion

Transport in cross-wind direction is parameterized as diffusive process:

\[
\frac{\partial C}{\partial t} = -u \frac{\partial C}{\partial x} + K_y \frac{\partial^2 C}{\partial y^2} + K_z \frac{\partial^2 C}{\partial z^2}
\]

for steady wind, inert plume

Turbulent diffusion coefficients

Steady state solution with suitable boundary conditions:

\[
C(x, y, z) = \frac{q}{4\pi(K_y K_z)^{1/2}} x \exp\left[ -\frac{u}{4x} \left( \frac{y^2}{K_y} + \frac{z^2}{K_z} \right) \right]
\]
On-line and off-line approaches to chemical modeling

**On-line: coupled to dynamics**

GCM conservation equations:
- air mass: $\partial \rho_a / \partial t = \ldots$
- momentum: $\partial \mathbf{u} / \partial t = \ldots$
- heat: $\partial \theta / \partial t = \ldots$
- water: $\partial q / \partial t = \ldots$
- chemicals: $\partial C_i / \partial t = \ldots$

**Off-line: decoupled from dynamics**

GCM conservation equations:
- air mass: $\partial \rho_a / \partial t = \ldots$
- momentum: $\partial \mathbf{u} / \partial t = \ldots$
- heat: $\partial \theta / \partial t = \ldots$
- water: $\partial q / \partial t = \ldots$

---

**PROs of off-line vs on-line approach:**
- computational cost
- simplicity
- stability (no chaos)
- compute sensitivities back in time

**CONs:**
- no fast chemical-dynamics coupling
- need for meteorological archive
- transport errors

Chemical data assimilation, forecasts best done on-line

Chemical sensitivity studies may best be done off-line
Inverse modeling and data assimilation

Quantify selected variables driving a physical system (state vector $\mathbf{x}$, dim $n$) by using:
- the observable manifestations of the system (observation vector $\mathbf{y}$, dim $m$)
- a physical model $\mathbf{y} = F(\mathbf{x})$ (forward model)
- a prior estimate $\mathbf{x}_A$ before the observations have been made
- Statistics for the errors $\varepsilon$ in the different components of the problem

General approach:

$$
\begin{align*}
\text{prior estimate} & \quad \text{forward model} \\
\mathbf{x}_A + \varepsilon_A & \quad \mathbf{y} = F(\mathbf{x}_A) + \varepsilon_M \\
\text{observations} & \quad \text{mismatch} \\
\mathbf{y} + \varepsilon & \quad \text{statistical fit of } \mathbf{x} \\
\text{best (posterior) estimate} & \quad \hat{\mathbf{x}} \pm \hat{\varepsilon}
\end{align*}
$$
Bayes’ theorem: general basis for optimal estimation

\[ P(x) = \text{probability density function (pdf) of } x \]
\[ P(x,y) = \text{pdf of } (x,y) \]
\[ P(y \mid x) = \text{pdf of } y \text{ given } x \]

\[ P(x,y)dxdy = P(x)dxP(y \mid x)dy = P(y)dyP(x \mid y)dx \]

\[ P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} \]

Normalizing factor (unimportant)

Optimal estimate solution for \( x \) given \( y \) is \( \text{max}[P(x \mid y)] \)

\[ \Rightarrow \text{solve for } \nabla_x P(x \mid y) = 0 \]
Solution of inverse problem minimizes cost function $J(x)$

corresponds to finding maximum of posterior pdf $P(x | y)$

$$J(x) = (x - x_A)^T S^{-1}_A(x - x_A) + (y - F(x))^T S^{-1}_o(y - F(x))$$

Prior terms

$- \ln P(x | y) = - \ln P(x) - \ln P(y | x)$

Error covariance matrices
Model adjoint

If a model can be linearized as a product of matrices,

\[ n(t) = A C E n(t - \Delta t) \]

\( A \) = advection
\( C \) = chemistry
\( E \) = emissions
(for example)

Then the adjoint is the transpose

\[ (A C E)^T = E^T C^T A^T \]

Application of the adjoint to a unit forcing \( \mathbf{v} = (1, 0, \ldots 0)^T \) gives the sensitivity of \( n_{1,x}(t) \) to the concentration field (and other model parameters) at the previous time step:

\[ E^T C^T A^T \mathbf{v} = \partial n_1(t) / \partial n(t - \Delta t) \]

Useful for:
- source attribution in receptor-oriented problems
- variational minimization of cost function
Increasing computational performance of models

Moore’s law is slowing down, clock speeds are flat

Since we cannot count on faster cores we need to increase the number of cores
Shared-memory vs. distributed-memory model architectures

**Shared memory architecture (OpenMP):**
- Easy to code but does not scale well with more than ~20 cores

**Distributed-memory architecture (MPI):**
- Harder to code but scales much better

Local operations $\frac{dC_i}{dt} = P_i(C) - L_i(C)$ scale perfectly column-by-column in massively parallel architectures;

Makes chemistry comparatively cheaper as #cores increases
The better a model becomes, the less accessible it is

- Increase model resolution to capture finer-scale phenomena
- Use more advanced schemes to incorporate better scientific knowledge
- Use advanced software (e.g. Earth System Modeling Framework) for better parallelization and model interoperability

- Need to download more input data
- Need much more computing power
- Compiling and configuring models become much more difficult

The better a model becomes, the less accessible it is
A solution: cloud computing

- Swipe your credit card and rent a supercomputer when you need it

GEOS-Chem code has been ported to the AWS cloud with loaded input data, libraries, etc. Just configure input file and run!

- Advantages for users:
  - No need to purchase an expensive system
  - No need to have systems experts to set up the model
  - No need to understand the model
  - Reproducibility of results from standard model is guaranteed
  - Download fully configured model to your own system

Zhuang et al. [BAMS, in prep.]
Atmospheric chemistry modeling has a bright future!

- Many problems to choose from: fundamental, applied, integrative
- Frontier of knowledge is near: many zeroth-order problems remain
- Next frontier in Earth system modeling
- Continual stream of new data sets to challenge us
Chapter 1: The concept of model
Chapter 2: Atmospheric structure and dynamics
Chapter 3: Chemical processes
Chapter 4: Model equations, numerical approaches
Chapter 5: Radiative, chemical, aerosol processes
Chapter 6: Numerical methods for chemical systems
Chapter 7: Numerical methods for advection
Chapter 8: Parameterization of small scales
Chapter 9: Surface fluxes
Chapter 10: Model evaluation
Chapter 11: Inverse modeling

To know more!