Atmospheric Model Evaluation

Vaishali Naik, Gabriele Pfister, & Simone Tilmes

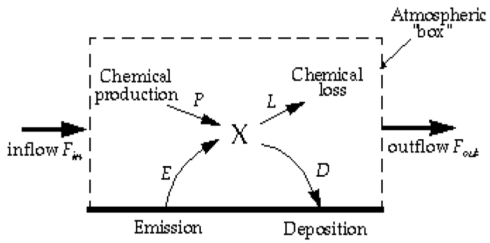
NCAR/ACOM workshop on Fundamentals of Atmospheric Chemistry and Aerosol Modeling
July 13-15, 2018

Lecture Outline

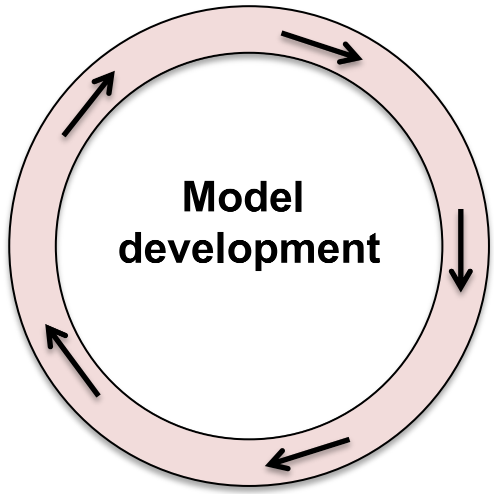
- ❖ What is model evaluation and why we need it?
- ❖ Consideration of Observational Uncertainty in Model Evaluation
- ❖ Approaches for Global and Regional Model Evaluation
    - Model - Observation Comparisons
    - Model - Model Comparisons
    - Process-oriented Evaluation
- ❖ What is a Good Model Performance?

Model Evaluation: What and Why?

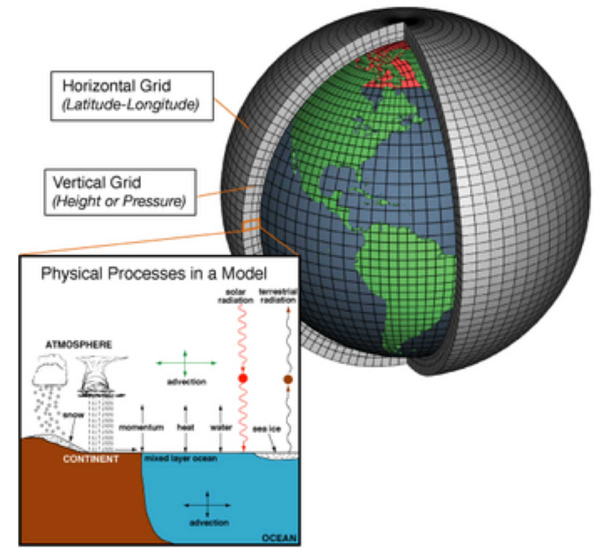
Box and Process Models



- Derive physical concepts
- Develop parameterizations for large scale



Regional and Earth System Models



Models are numerical approximations of a wide range of processes in the atmosphere

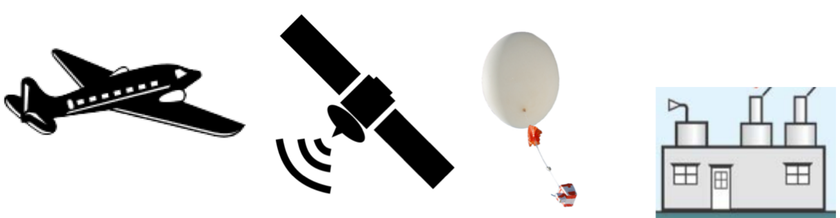
How well do models represent real-world processes?

Model Evaluation

Use of lab measurements, field campaigns, long-term observational datasets, satellites

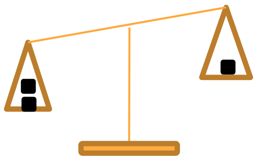
- Assessing skill of a model
- Gain confidence in model results
- Improved process understanding

-> Improved model approximation towards real world processes



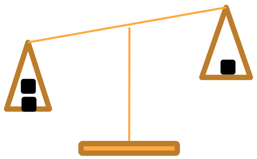
Model evaluation is an important part of model development and improvement

Model Evaluation: what and why?



- Model evaluation **is a quantitative measure of model fidelity/skill** in representing a specific real-world process/system; either the state of the atmosphere, a specific process or sensitivities.
- Evaluation **helps to characterize model errors and identify missing processes.**
- Evaluation provides a means to **improve model process/system representation.**
- Evaluation provides a **measure of our confidence in model future predictions.**

Model Evaluation: what and why?



- Model evaluation **is a quantitative measure of model fidelity/skill** in representing a specific real-world process/system; either the state of the atmosphere, a specific process or sensitivities.
- Evaluation **helps to characterize model errors and identify missing processes.**
- Evaluation provides a means to **improve model process/system representation.**
- Evaluation provides a **measure of our confidence in model future predictions.**

- How to perform like-with-like comparisons?
- How to ensure that model compares well with observations for the right reason?

Consideration of Observational Uncertainty in Model Evaluation

The following uncertainties in observations pose challenges to model evaluation:

1. **Sampling uncertainty**: sparse spatial and temporal resolutions of in-situ monitoring stations, coarse vertical resolution of remote sensing, poor observational constraints
2. **Systematic errors in measurements**: instrumentation error, drifts in satellite retrievals, change in instruments during observation record, model information included in retrievals
3. **Representative errors**: comparisons of different temporal and spatial scales, point measurements at a given day vs. model grid average of the background atmosphere

Understanding the range of uncertainties in observations is critical for useful model evaluation

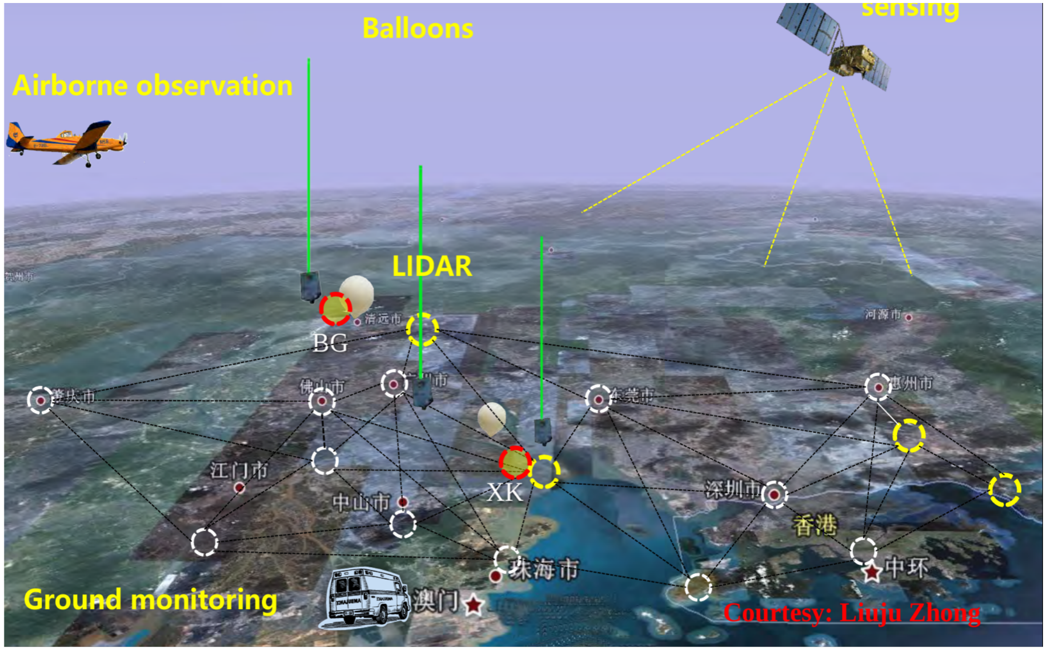
Different types of Observations

In-situ (sondes, aircraft, surface data)

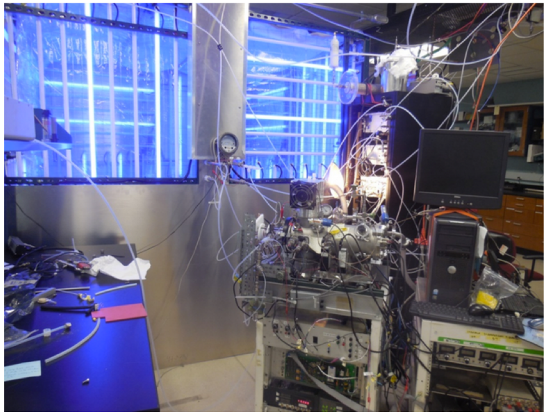
- Direct measurement
- Uneven and incomplete coverage in time/space
- Localized measurements vs. broad model scale (scale mix-match)

Remote sensing: satellites, lidar

- A retrieval includes some degree of model information
- Comparisons to satellites (different averaging kernels provide different answers)



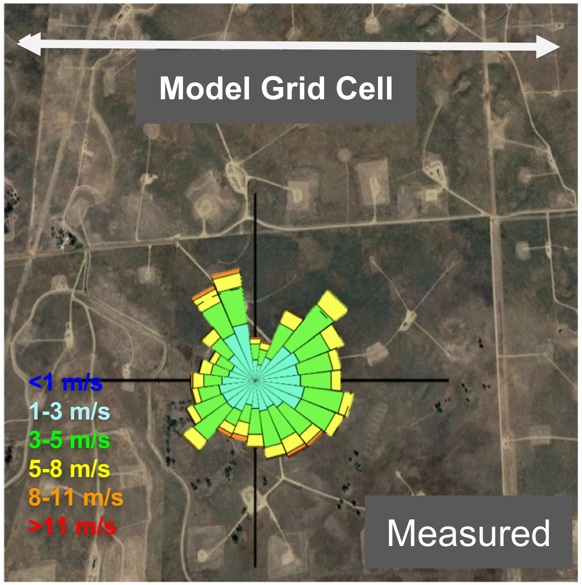
Lab Measurements

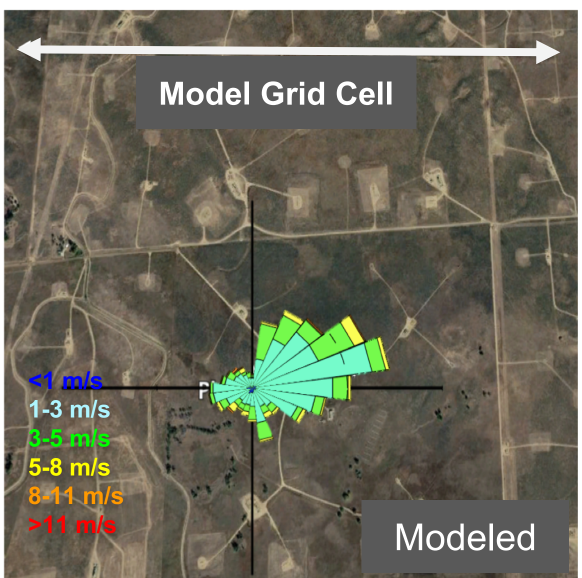


2015 ACOM Annual Report

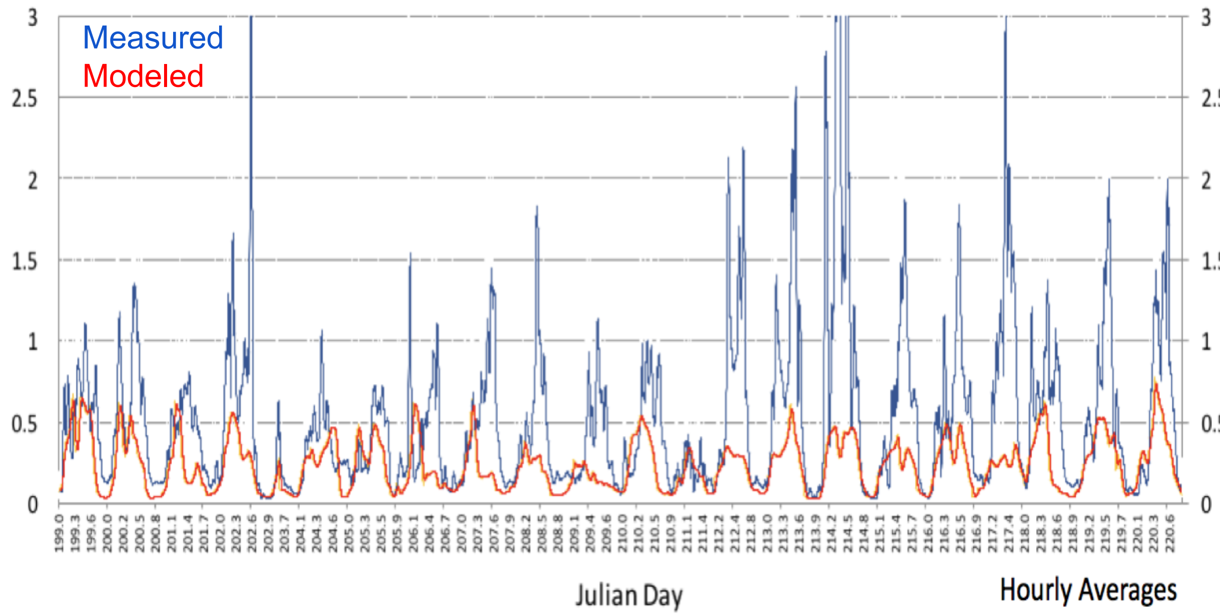
Representativeness: Model Grid Scale

Windrose at Platteville, CO





Surface Benzene Concentrations, Platteville, CO



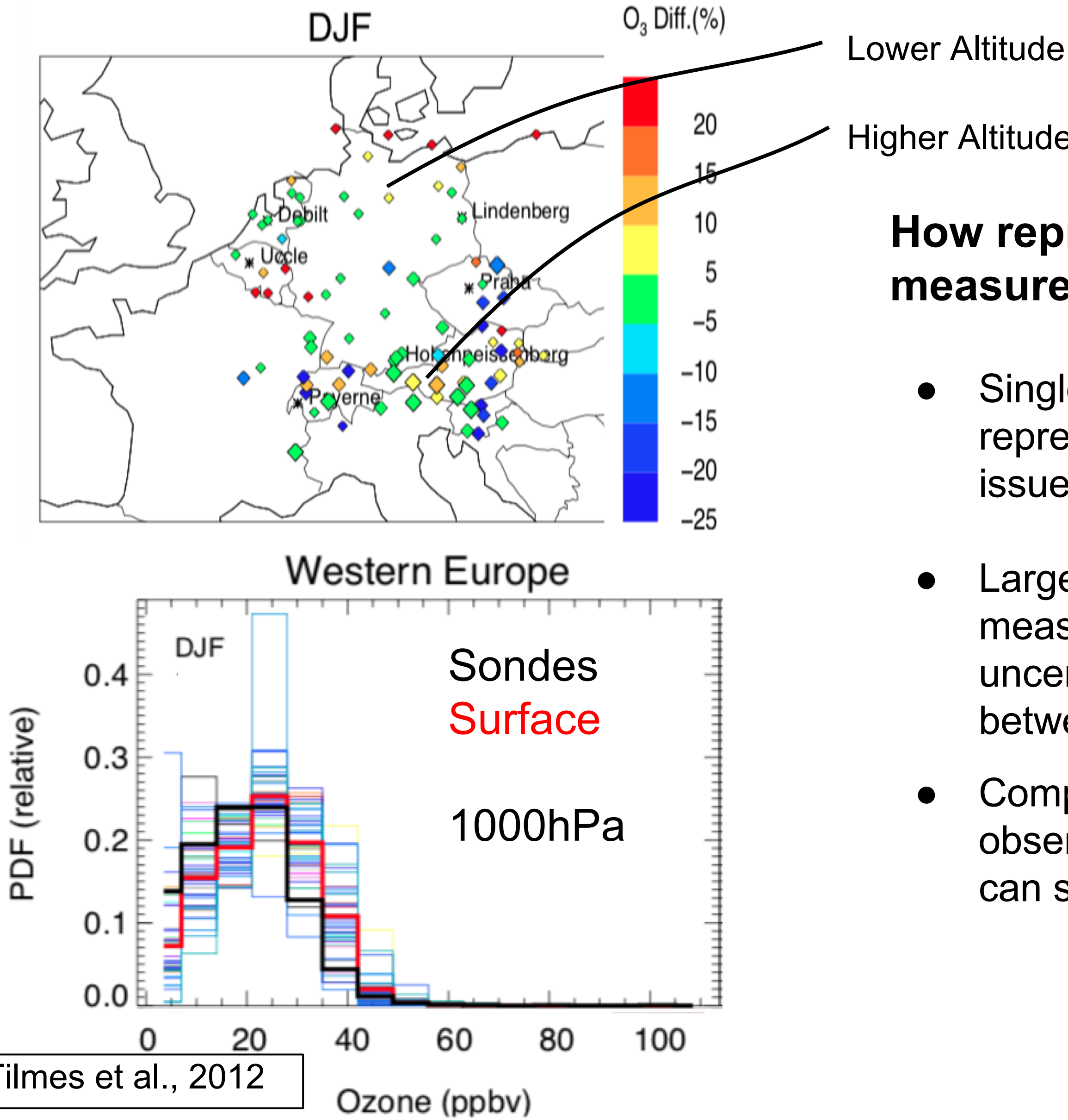
Do we have a fair comparison?

- True model biases or grid resolution issue?
- Transport error?
- Model input error (emissions)?
- Representativeness of observations for larger scale?

Pfister et al., FRAPPÉ 2014 WRF-Chem Analysis

Sampling Uncertainty: Spatial Scale

Ozonesondes versus Surface Observations



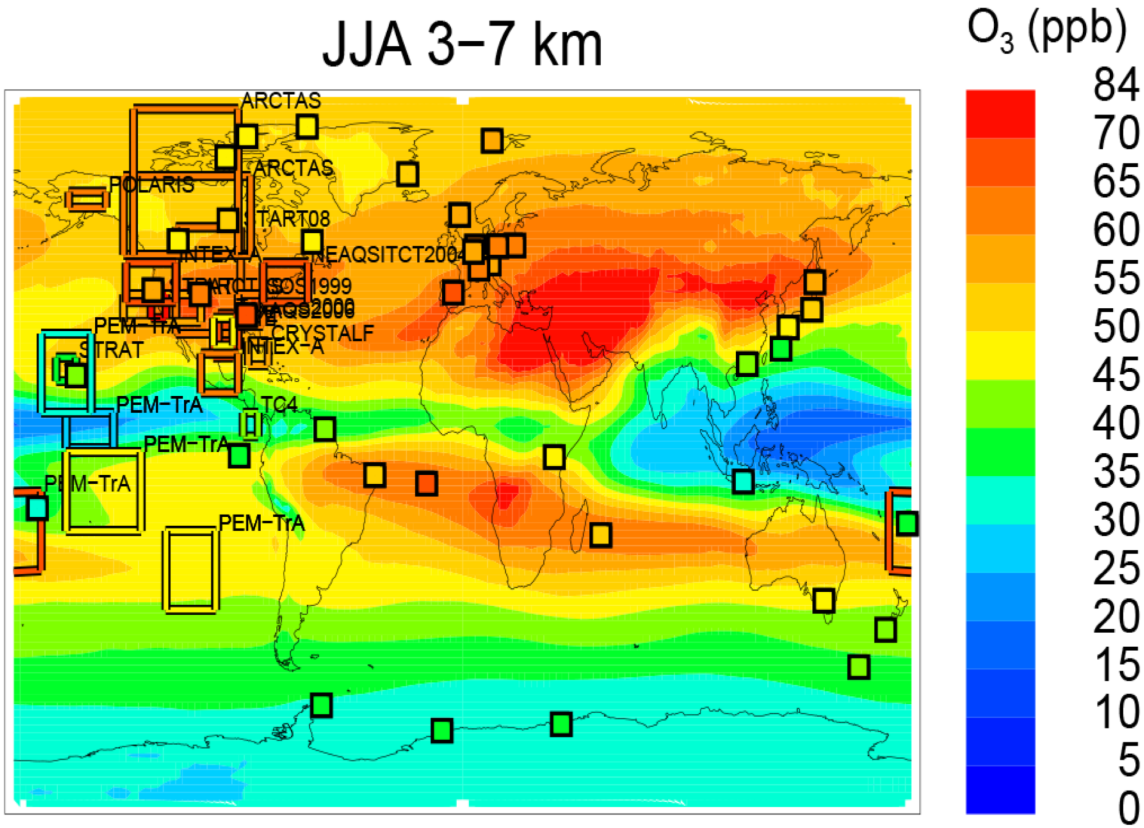
Tilmes et al., 2012

How representative are sparse measurements?

- Single stations or datasets may be not representative, potential calibration issues for different stations
- Large coverage of surface measurements can help reduce uncertainties through differences between single stations
- Comparison between different observations (taken at the same time) can still lead to different answers

Representativeness: Temporal and Spatial Scale

Model climatology (colored background) versus climatologies derived from ozonesondes (symbols) and aircraft (boxes)



Long-term observations

- High temporal resolution and continuity of surface and ozonesonde observations generally make them more representative

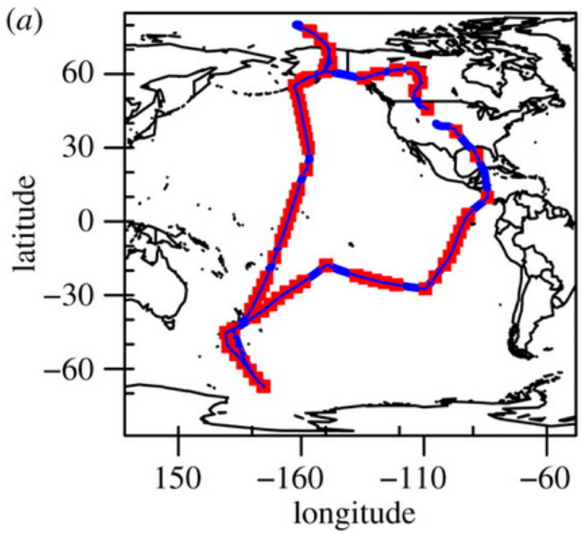
Single aircraft campaigns often target specific questions

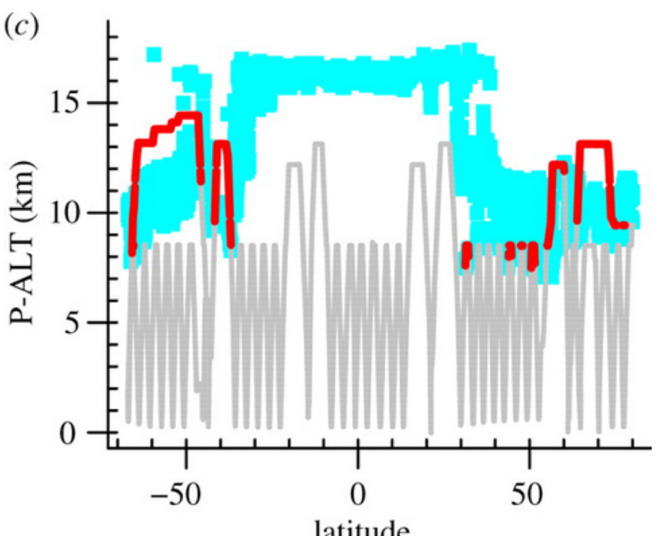
- A number of different species are measured
- Climatological evaluation requires filtering of data
- Simulate exact location/time and co-sample model with observations in space and time for like-with-like comparison

Tilmes et al. (2016)

Sampling of Background Atmosphere with in-situ Measurements

HIPPO1 Deployment





Wofsy et al. (2011)

Recent aircraft measurements target observations of background atmosphere -> climatological evaluation

Sampling of different chemical species using commercial aircraft:

- MOSAIC (2005-2014) / IAGOS (2014-present) CARIBIC on Lufthansa Airbus (2004-present)

Aircraft campaigns designed to sample the background atmosphere:

- HIPPO: 2009-2011 four seasons over the Atlantic
- ATom: 2016-2018 four seasons; Atlantic, Arctic, Pacific, SH high latitudes

ATOM Flight Path

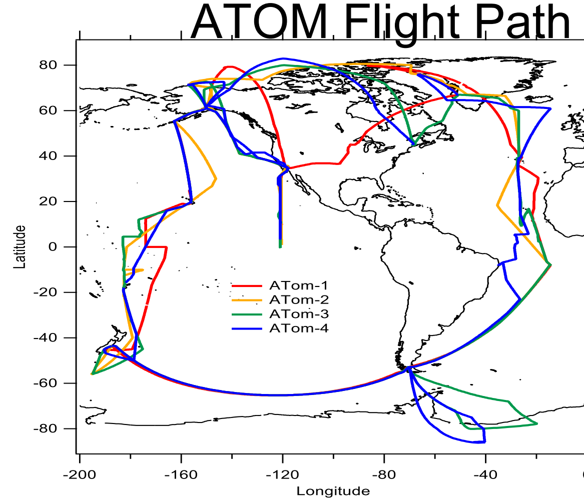


Figure courtesy Rebecca Hornbrook

Observations from Satellites

- **Satellites continuously measure radiation in various wavelength bands** including ultraviolet (UV), visible (Vis), infrared (IR) and microwave (MW)
- **A satellite product is not a true measurement of the derived quantity**
- **Satellite retrieval** of trace species **depends on knowledge (assumptions) of the state of the atmosphere**, e.g. presence of clouds and aerosols, the vertical distribution of the species, and surface properties including topography and albedo.
- **Inconsistencies between the assumptions used in the retrieved data and modelled distribution lead to inflation of errors.**
- **Coverage depends on measurement method,** larger coverage and low vertical resolution (nadir viewing) vs. high vertical resolution but limited spatial resolution (limb viewing)

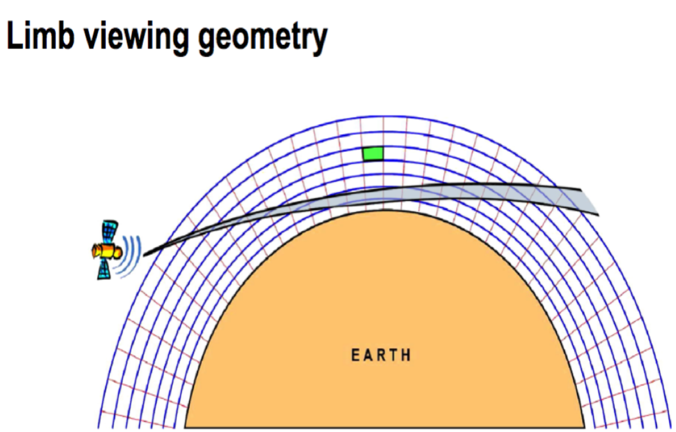


Figure courtesy Gabriele Stiller

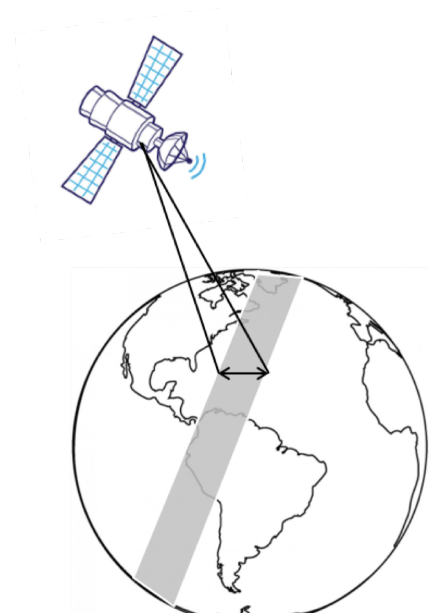
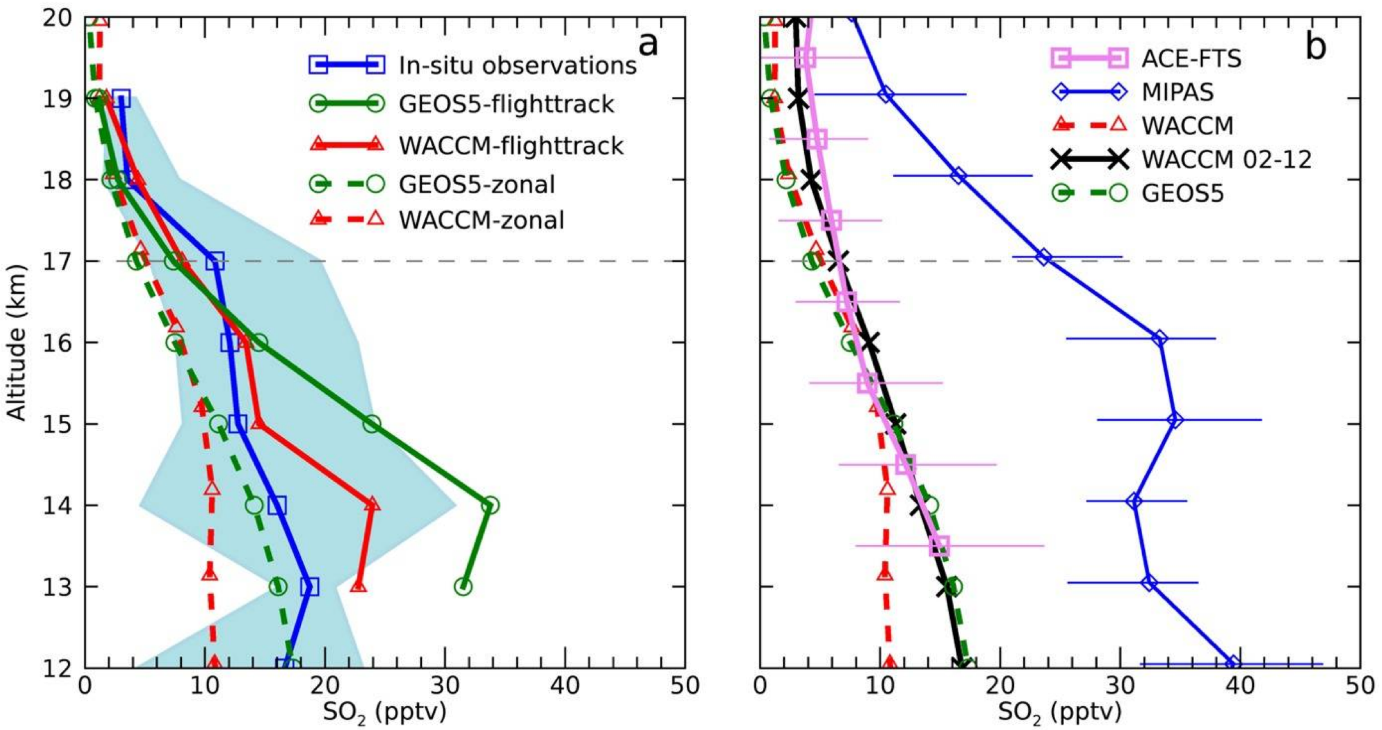


Figure courtesy Gabriele Pfister

Challenges in deriving Satellite Products from Retrievals

Measurements and model estimates of SO_2 in the UTLS



Different satellite observations can show very different results of the same quantity. In-situ aircraft measurements reveal large bias of satellite observations of SO_2 in the upper troposphere lower stratosphere.

-> Improved understanding of the sulfur budget in the stratosphere.

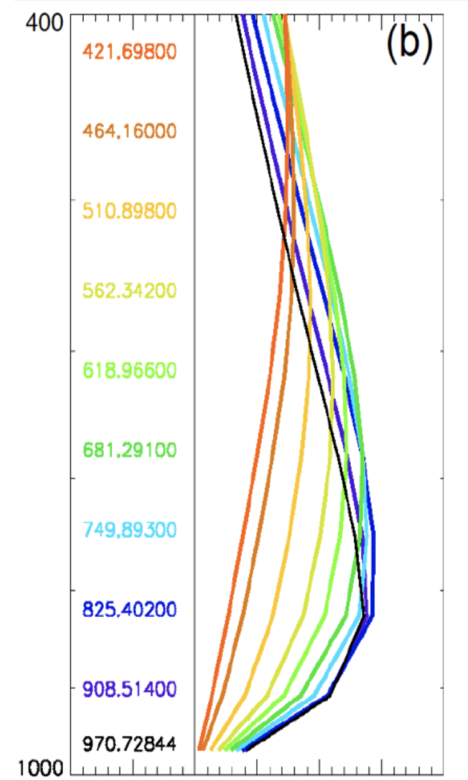
Rollins et al., 2017

Overcoming Challenges for Model-Satellite Comparisons

Inconsistencies between satellite data and model output can be minimized with **careful consideration of the representativeness** (horizontal coverage, temporal sampling, vertical information) of satellite data for model-satellite comparisons

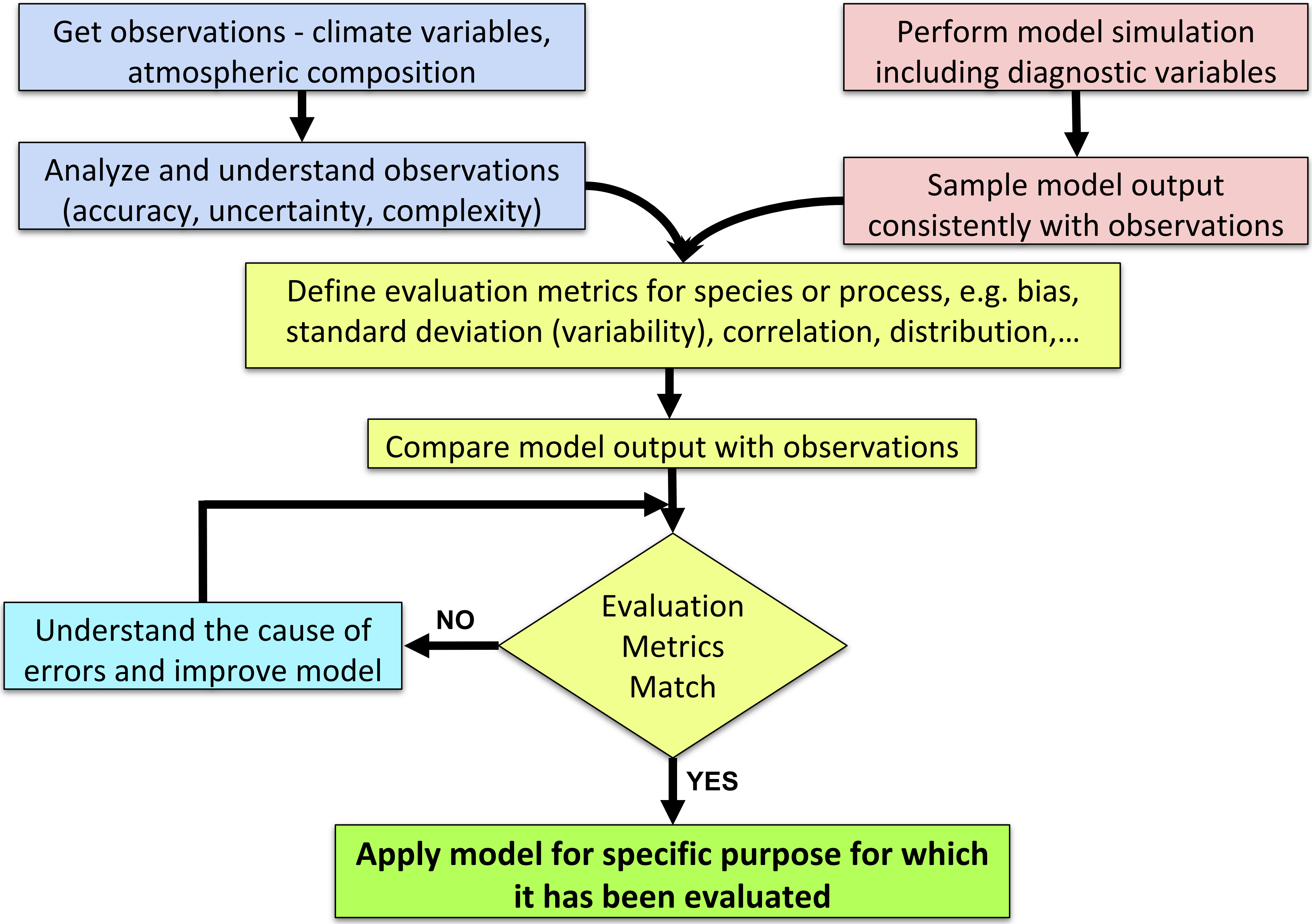
- Apply **appropriate averaging kernel on the model output** to obtain consistent model vertical distribution for comparison with satellite retrieval
- **Sample model data as consistently as possible to the satellite retrieval** in space and time (e.g., overpass time), and under similar atmospheric conditions (e.g., clear-sky vs. cloudy sky, day vs. night)
- **Use consistent definitions of atmospheric state** (e.g., definition of tropopause)

A Satellite Retrieval is NOT an In-Situ Observation



Averaging Kernel
Sensitivity of retrieved to true profile

Steps in Atmospheric Model Evaluation



Approaches for Global and Regional Model Evaluation

Model Evaluation against Observations

Evaluate the capability of a model to realistically represent observed features

- Mean climatology, long-term trends and variability, extremes,...

Model-to-Model Evaluation

Evaluate the range of uncertainty inherent in model representation of different processes

- Multi-model evaluations, benchmarking and data assimilation, regional versus global models, community diagnostics and performance metrics,...

Process-Oriented Evaluation

Experiment and evaluation designed to focus on a specific process

- Multi-model process-oriented, process-oriented diagnostics,...

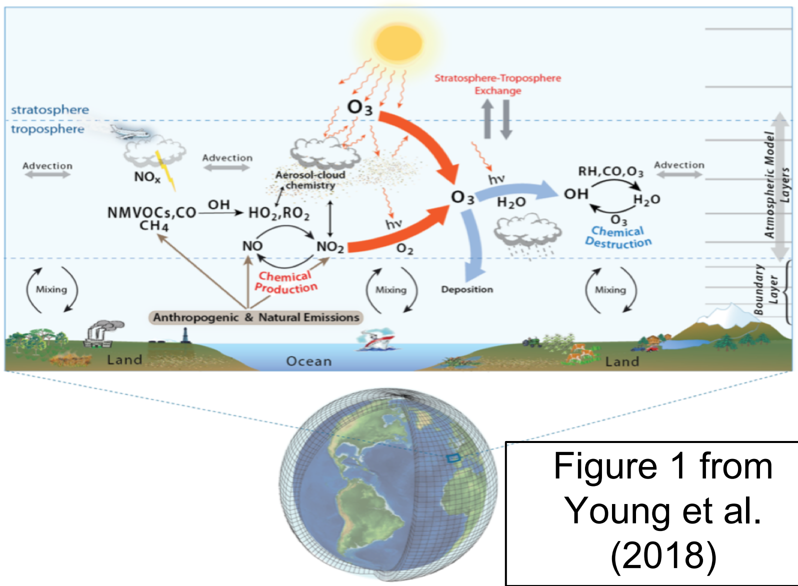


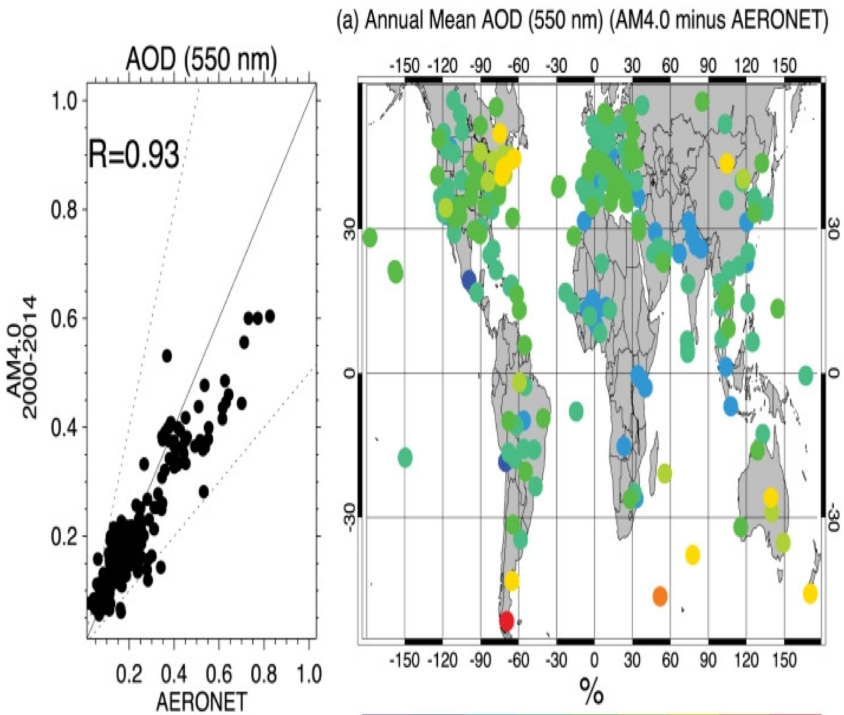
Figure 1 from Young et al. (2018)

Model vs. Observation

Evaluation of Mean Climatology

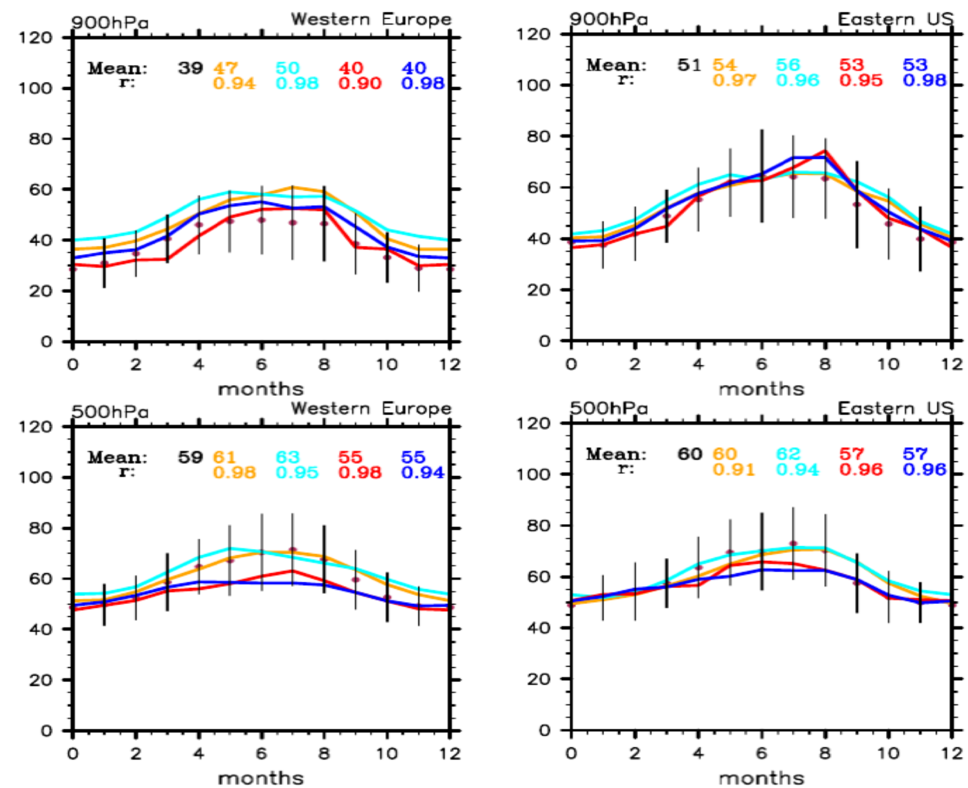
- Provides a measure of model's skill to accurately represent the mean state
- Climatology reduces uncertainties in observations

Climatological Mean AOD (2000-2014)



Zhao et al. (2018)

Seasonal Cycle: Model versus Ozonesonde Climatology (1995-2011)



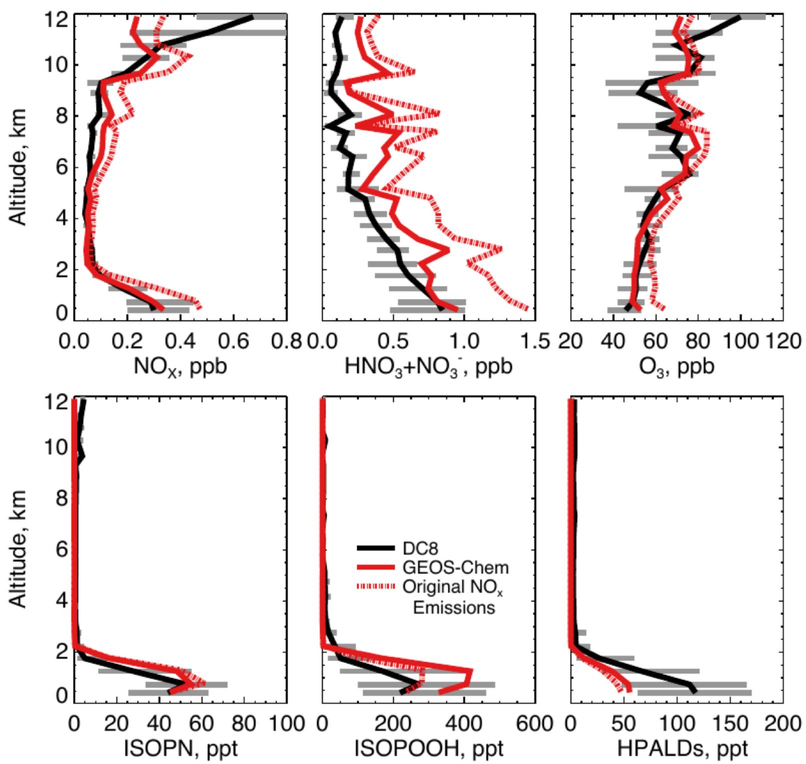
Tilmes et al. (2015)

Model vs. Observation

Evaluation against Aircraft Data

Provides clues on drivers of specific model biases, if various chemical species are co-measured

Too high anthropogenic NO_x emissions partly explain model overestimate of surface ozone over the Southeast U.S. SEAC^{4}RS



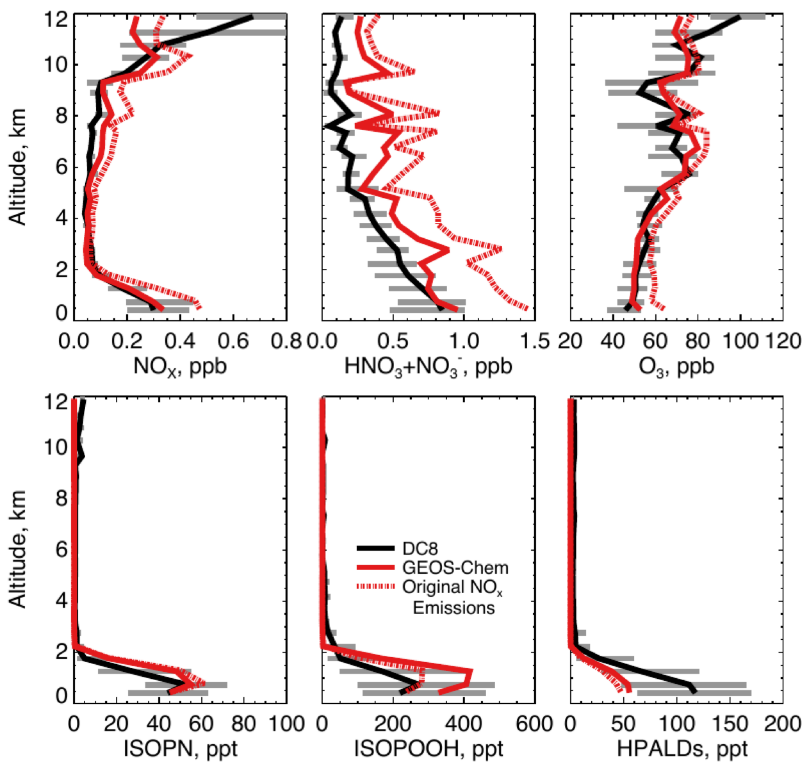
Travis et al. (2016)

Model vs. Observation

Evaluation against Aircraft Data

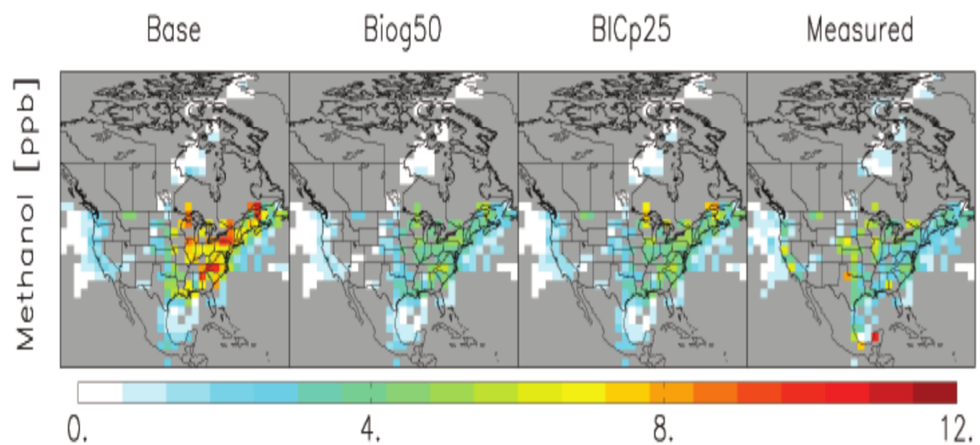
Provides clues on drivers of specific model biases, if various chemical species are co-measured

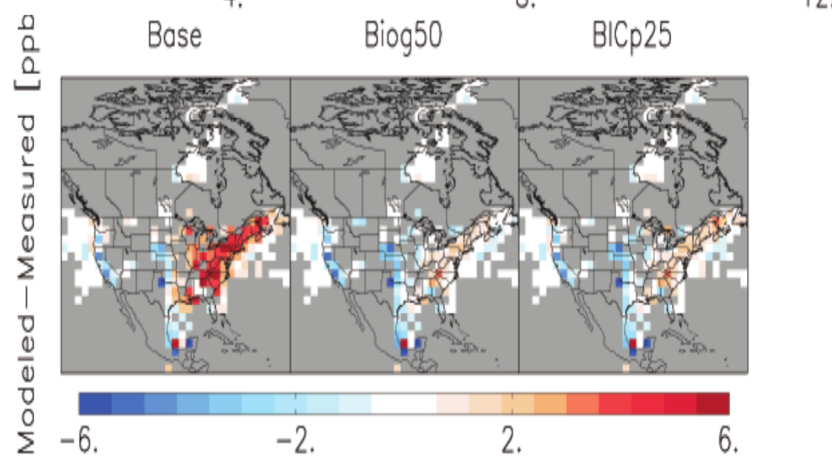
Too high anthropogenic NO_x emissions partly explain model overestimate of surface ozone over the Southeast U.S. SEAC[4]RS



Travis et al. (2016)

Model overestimates methanol compared to an aircraft composite over eastern North America because of too high biogenic emissions from broadleaf trees and crops





Millet et al. (2008)

Model vs. Observation

Long-term Trends and Variability

- Evaluation against timeseries observations necessary to understand model sensitivity
- Builds confidence in projections and attributions, however consideration of representativeness and natural climate variability is important

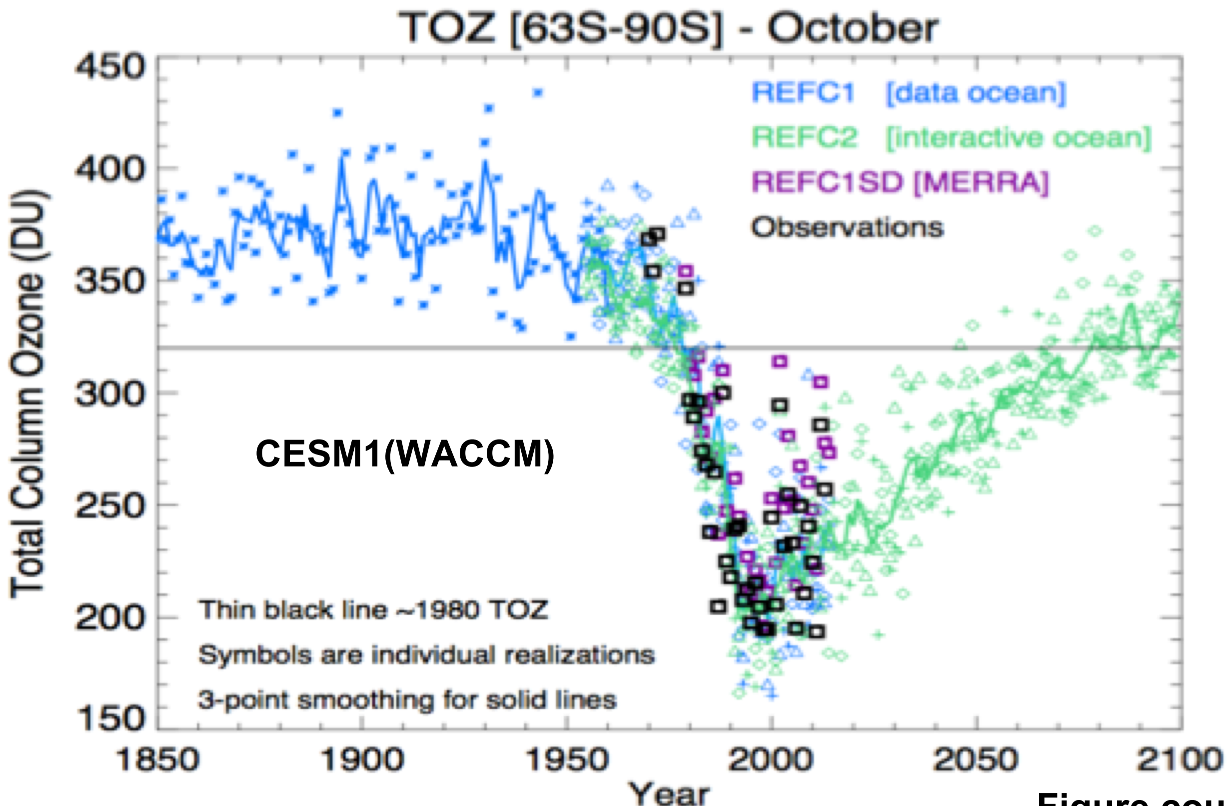
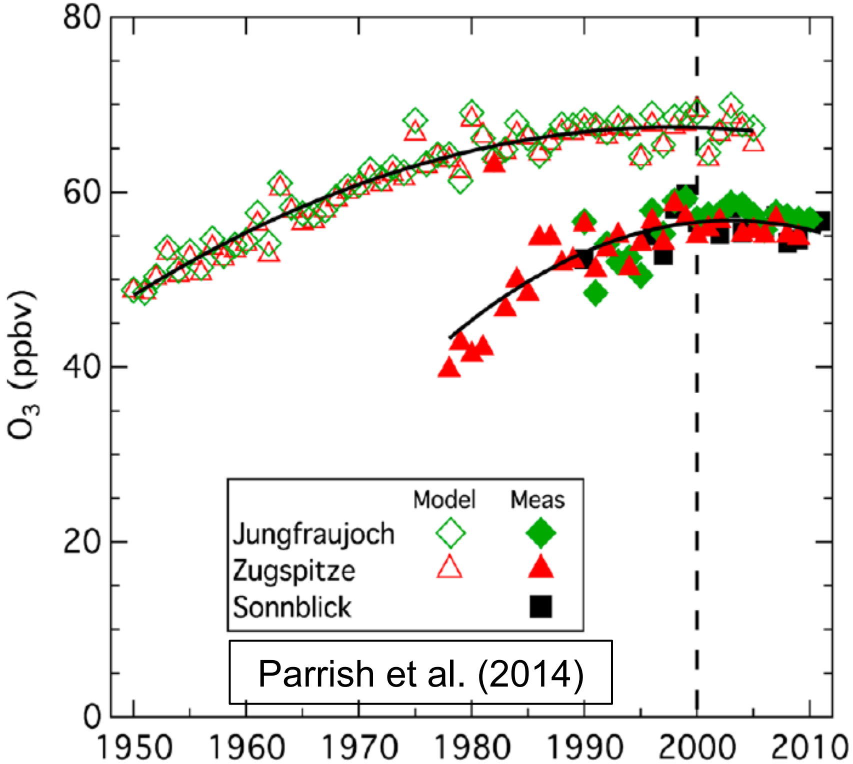


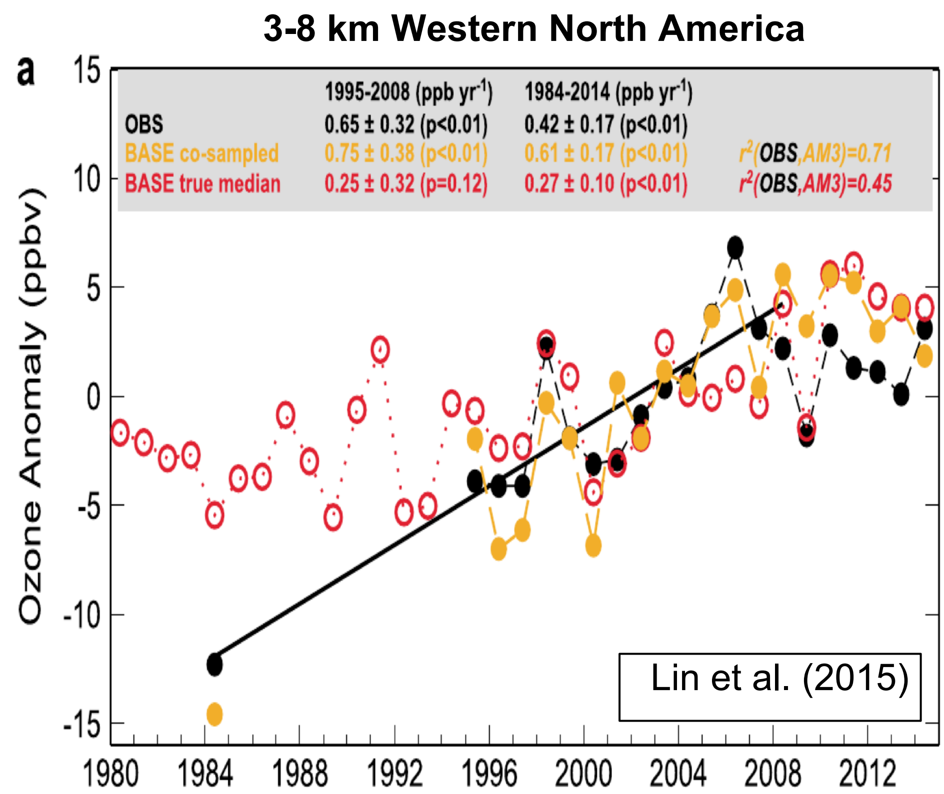
Figure courtesy Doug Kinnison

Meteorological variability generated by **free-running chemistry-climate models (CCMs) may not capture variations** seen in observations → run with meteorology nudged to observations

Model vs. Observation

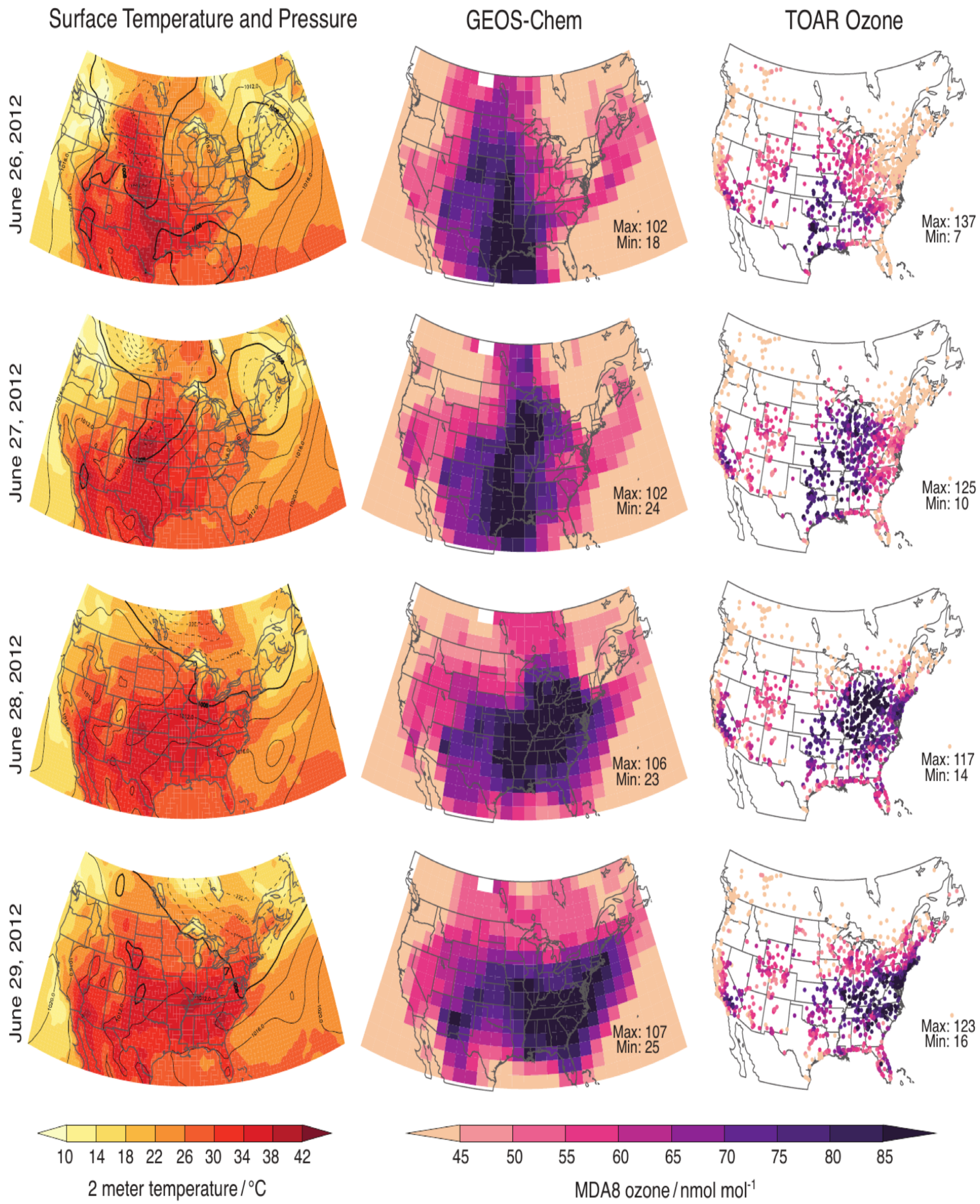
Sparse in-situ measurements and natural climate variability
complicate evaluation of model simulated trends and variability





Important to co-sample model in space and time with available observations
in addition to nudging the meteorology, spatial and temporal averaging necessary
to detect significant trends

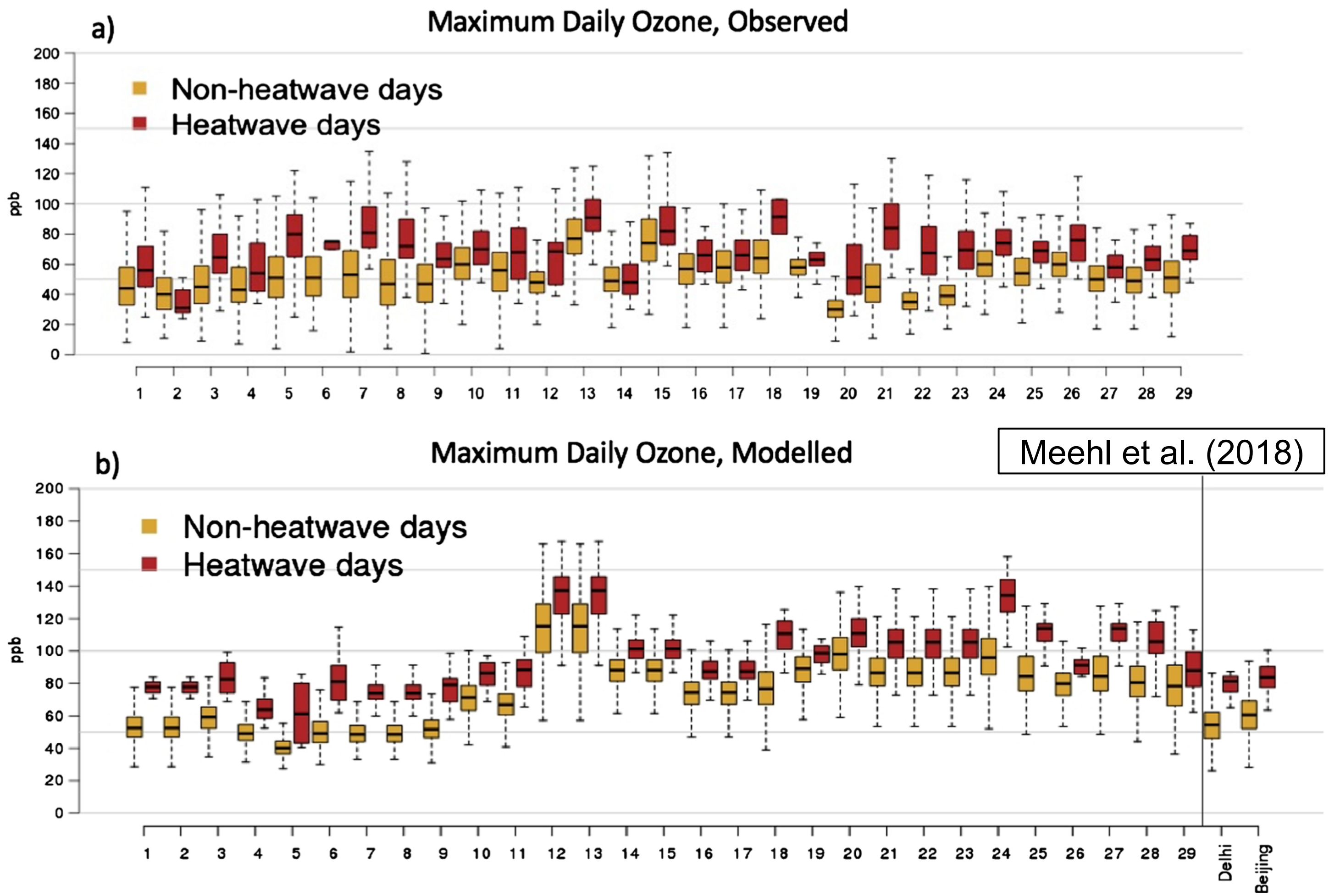
Evaluation of Extreme Air Pollution Events



- Evaluation of extreme events requires **clear definition of "extreme"**
- Evaluation of underlying **synoptic-scale meteorology and local emissions** necessary for building confidence in modeled extremes
- **Dense, high frequency, long-term, and reliable measurements** necessary for evaluating model skill in representing frequency, intensity and duration

Young et al. (2018)

Qualitative Comparison between Models and Observations



Ozone increase during heat waves over different cities around the globe

- Differences in absolute ozone between models and observations, but
- consistent behavior between models and observations

⇨ Confidence in specific process

Model-to-Model Evaluation

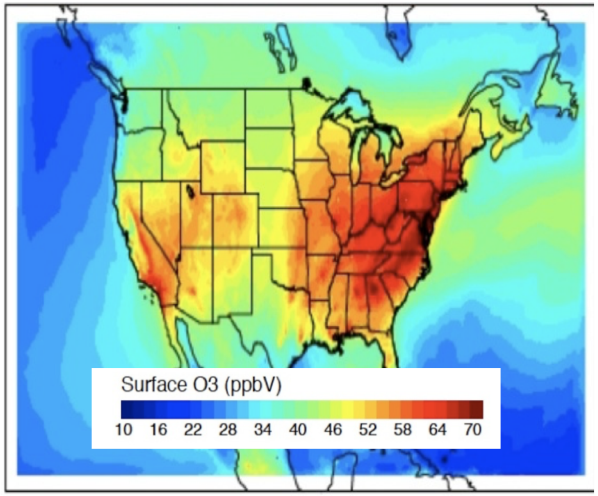
Evaluation of Long-Term Projections

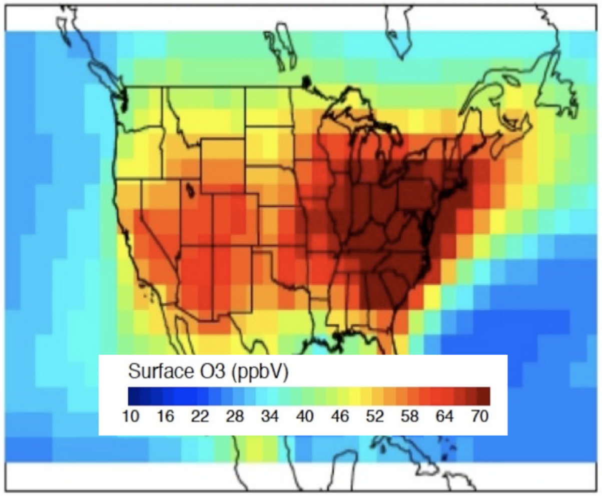
Coarse resolution models tend to **overpredict surface ozone**

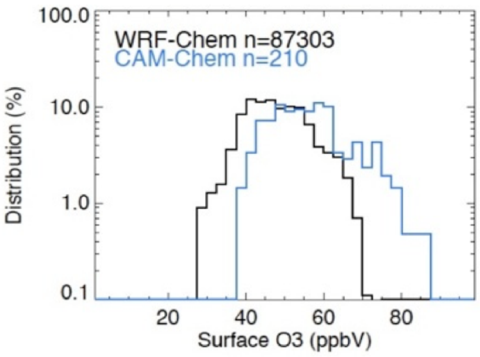
Regional Model: (12km)

Global Model: (2.5 deg)

Present







JJA Surface Afternoon Ozone

Pfister et al. (2014)

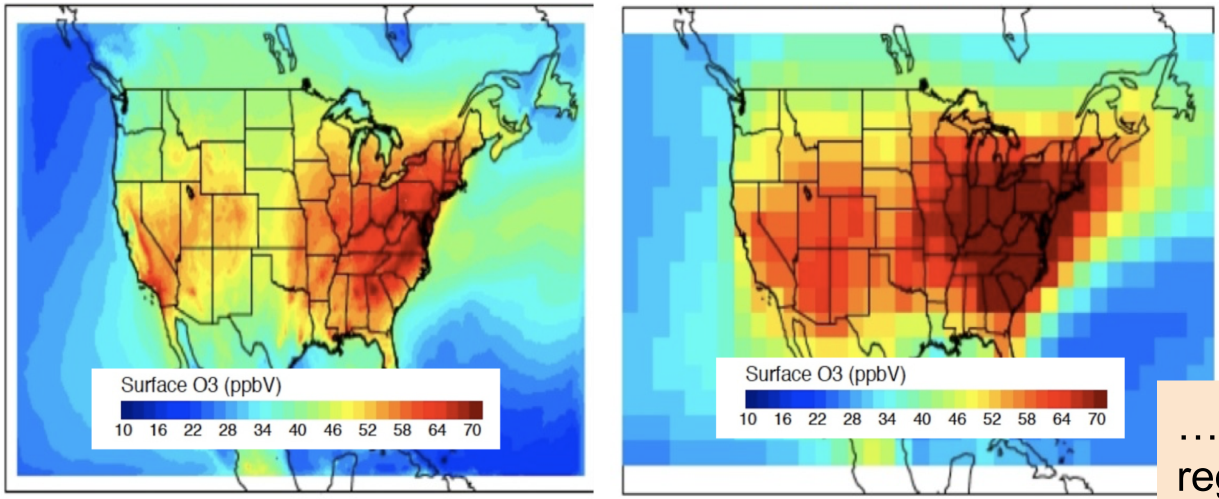
Model-to-Model Evaluation

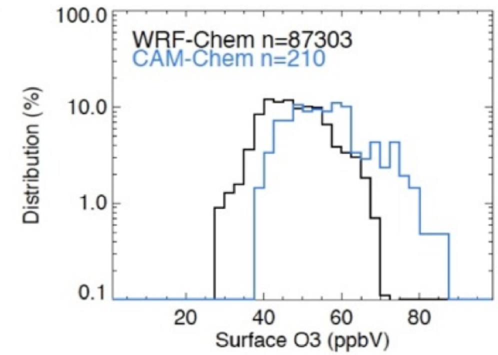
Evaluation of Long-Term Projections

Coarse resolution models tend to **overpredict surface ozone**

Regional Model: (12km) | Global Model: (2.5 deg)

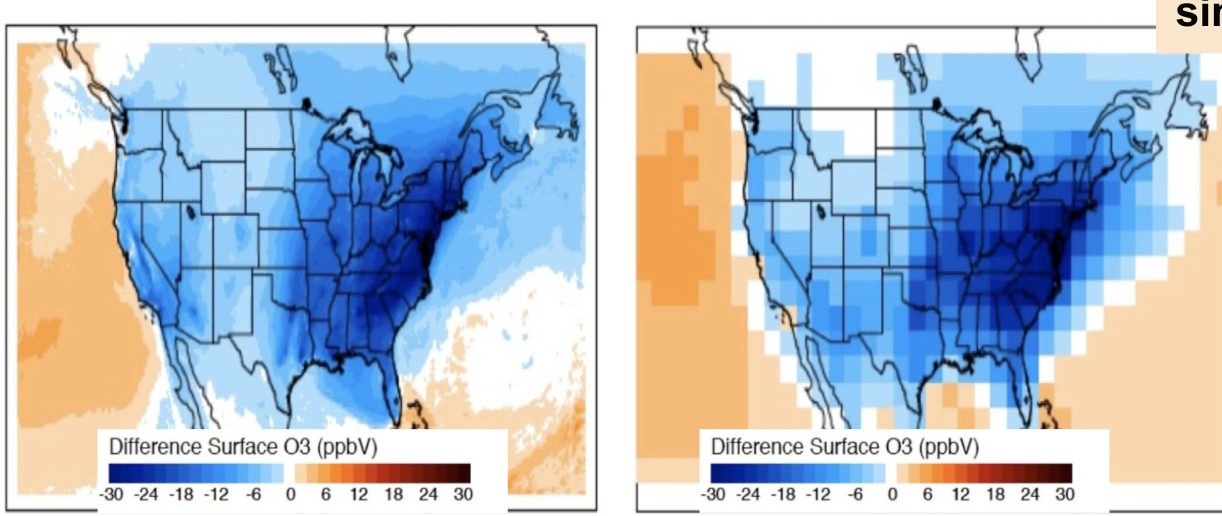
Present

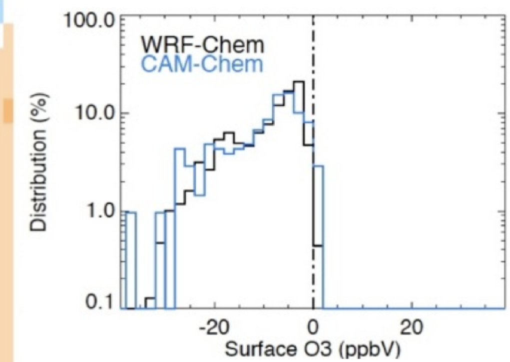




… but both global and regional predictions show **similar change** in ozone

Future *minus* Present





JJA Surface Afternoon Ozone

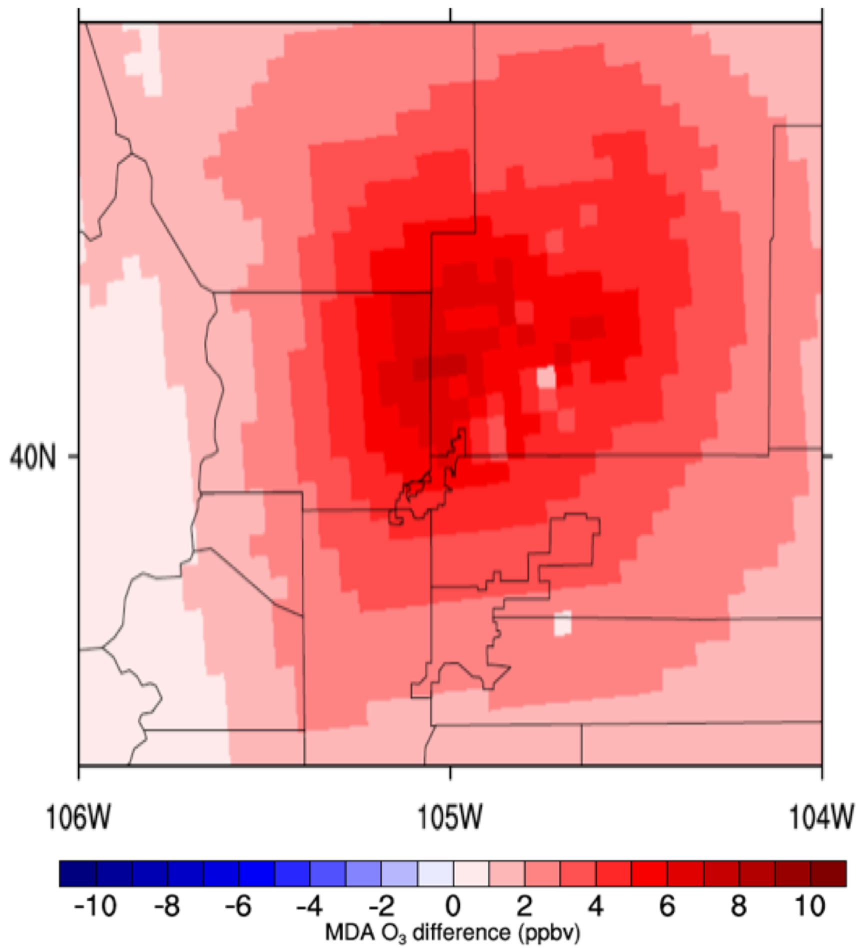
Pfister et al. (2014)

Model-to-Model Evaluation

Comparing **models of different complexity** facilitates independent evaluation of parameterizations or conclusions

Regional CTM

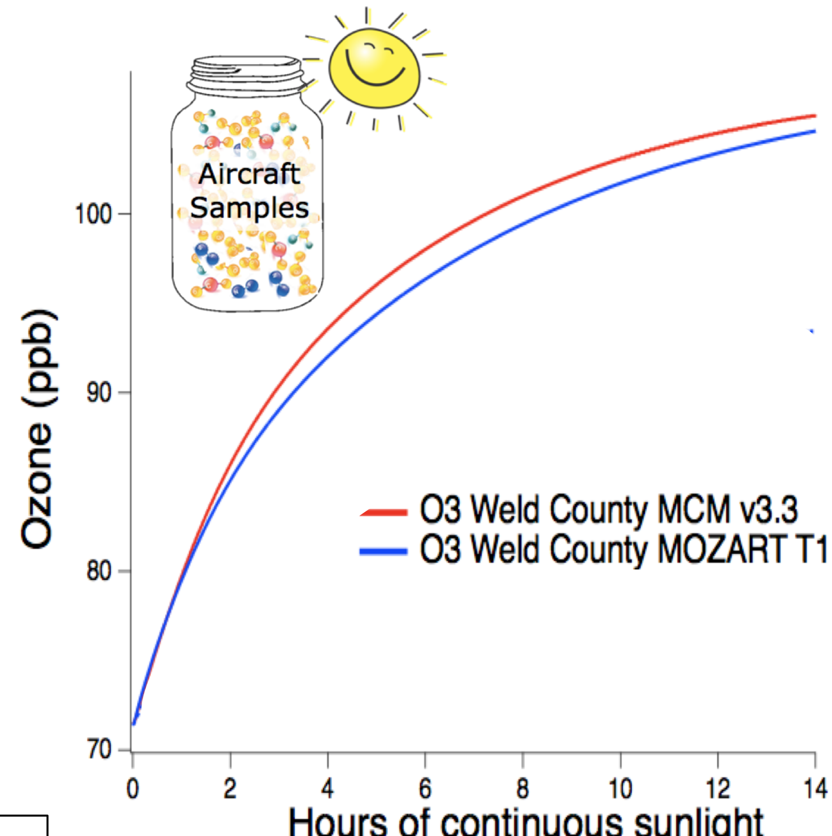
Zero-out OG emissions



Chemical Box model

driven by aircraft Observations

(1) Evaluate lumped chemical mechanism



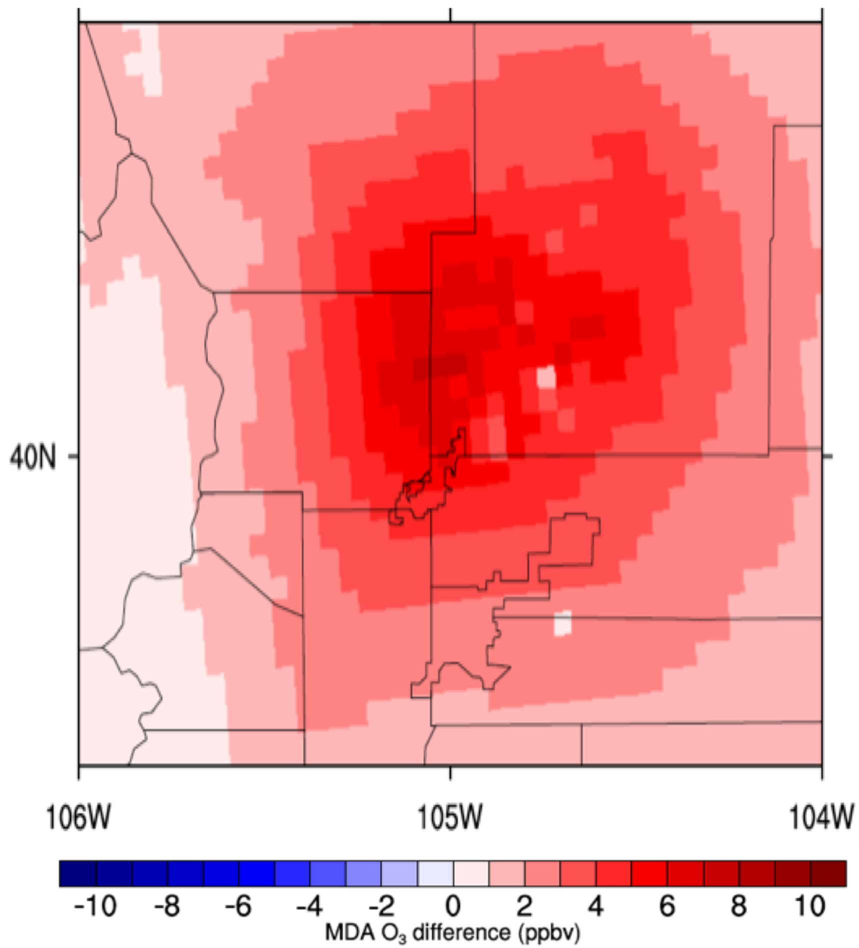
Pfister, Flocke and Lee, FRAPPÉ Final Report (2017)

Model-to-Model Evaluation

Comparing **models of different complexity** facilitates independent evaluation of parameterizations or conclusions

Regional CTM

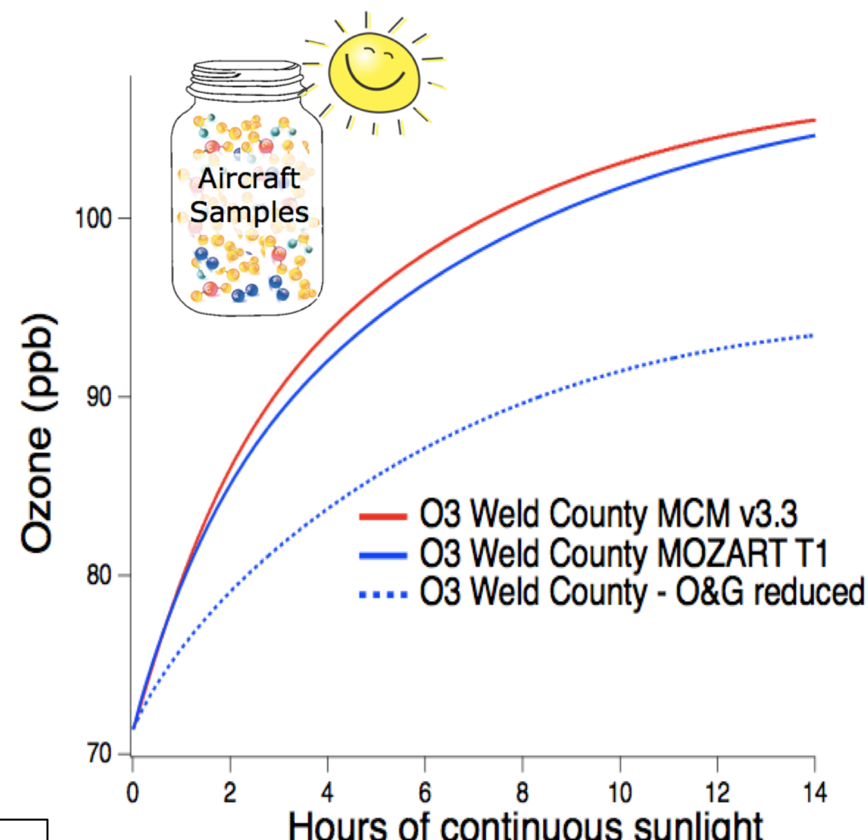
Zero-out OG emissions



Chemical Box model

driven by aircraft Observations

(1) Evaluate lumped chemical mechanism
(2) Reduce excess concentrations for OG species - Evaluate CTM conclusions

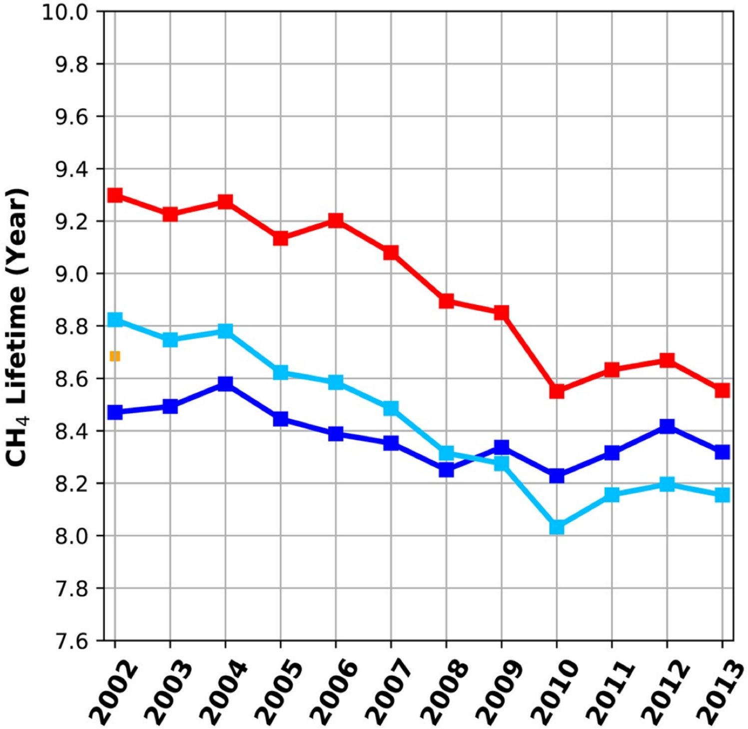


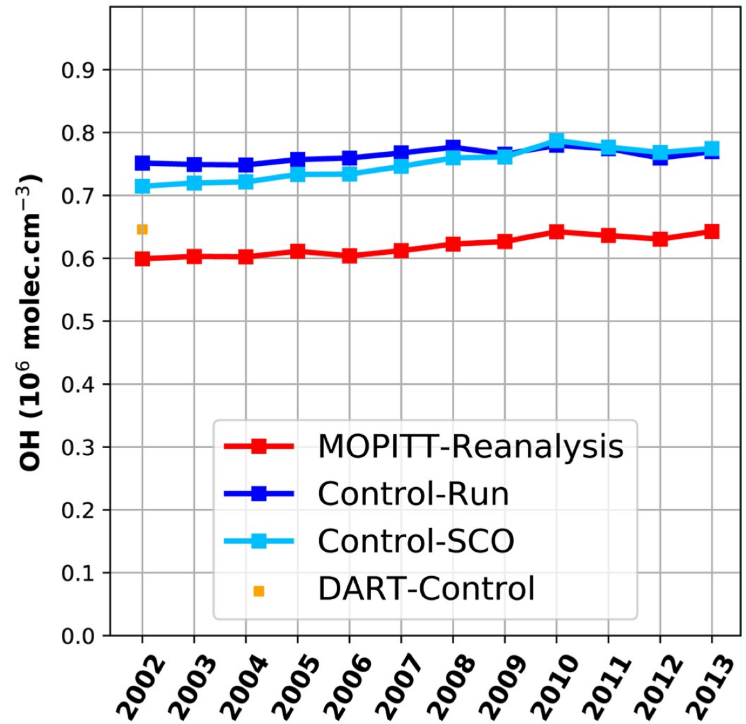
Pfister, Flocke and Lee, FRAPPÉ Final Report (2017)

Model-to-Model Evaluation

Benchmarking using Data Assimilation

Model simulation constrained by MOPITT CO observations



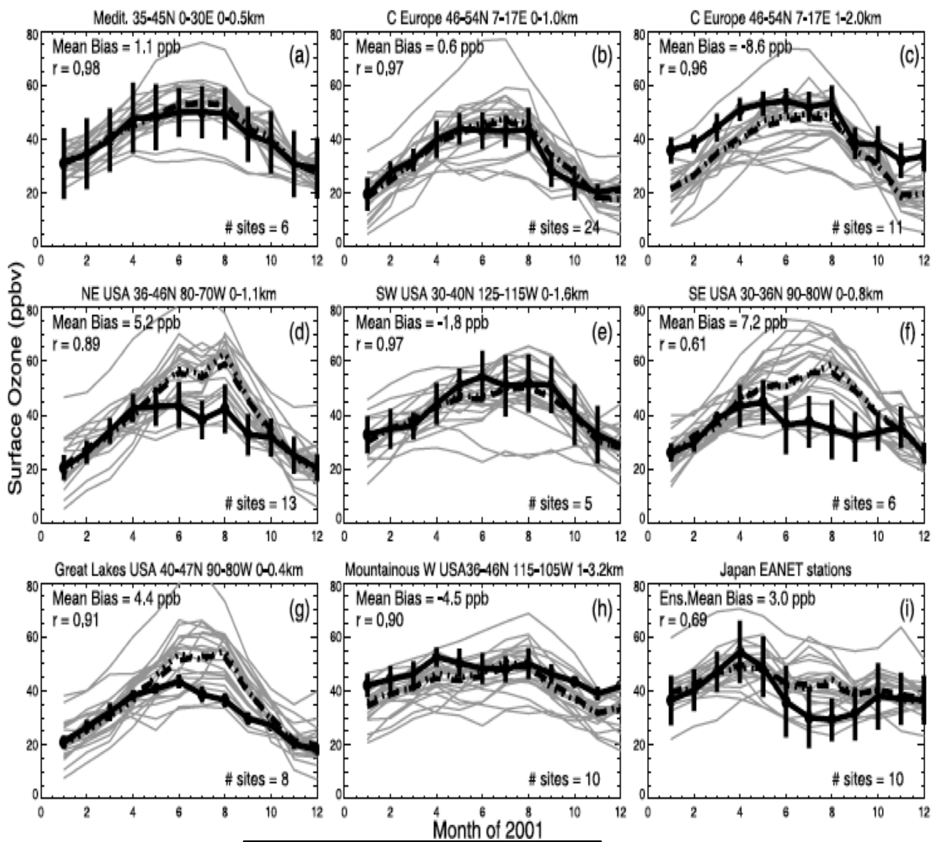


Gaubert et al. (2017)

- **Data assimilation** aims to optimally integrate observations and model simulation to **improve estimates of the atmospheric state.**
- Can help **identify shortcomings in composition and processes** and can be used as benchmark simulation

Multi-model Evaluation

Differences across models can be useful
to identify common problems across models,
explore structural uncertainty, and identify errors

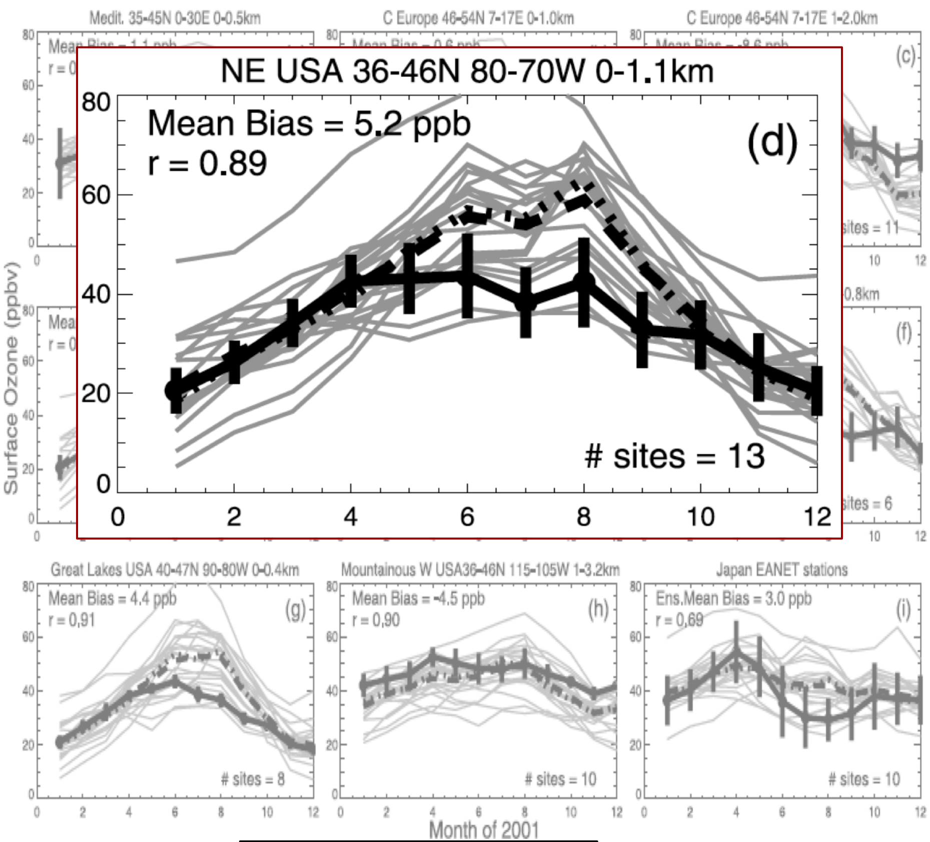


Fiore et al. (2009)

Multi-model Evaluation

Differences across models can be useful
to identify common problems across models, explore structural uncertainty, and identify errors

Persistent high bias in modeled summertime surface ozone over some regions

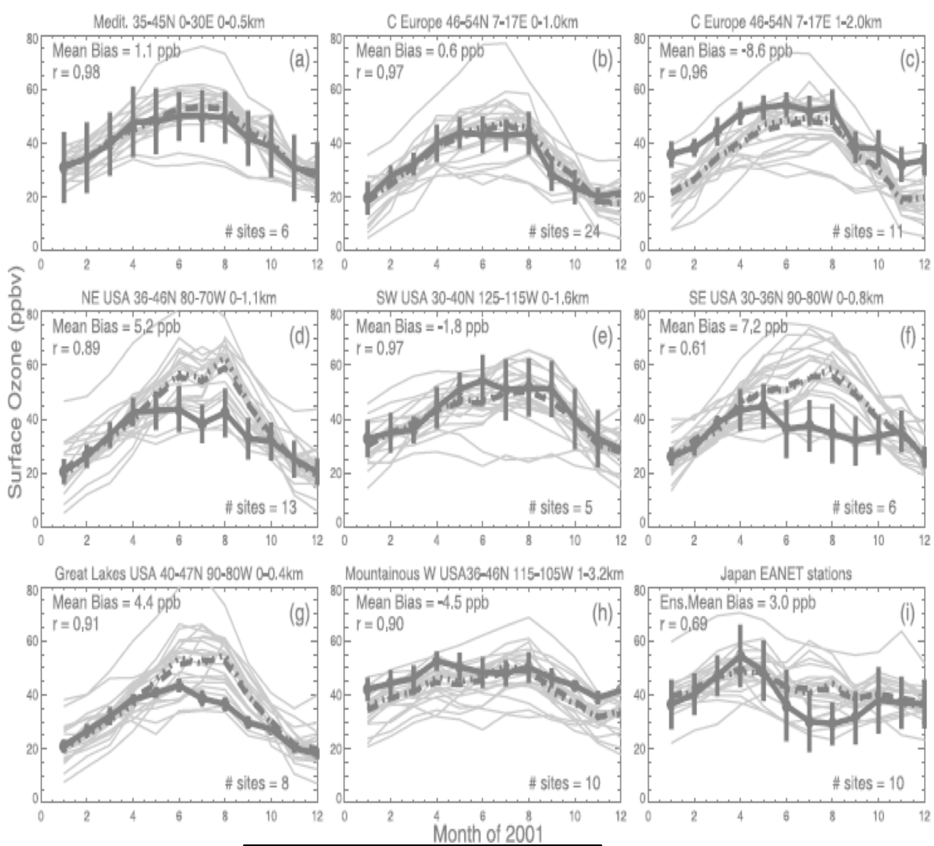


Fiore et al. (2009)

Multi-model Evaluation

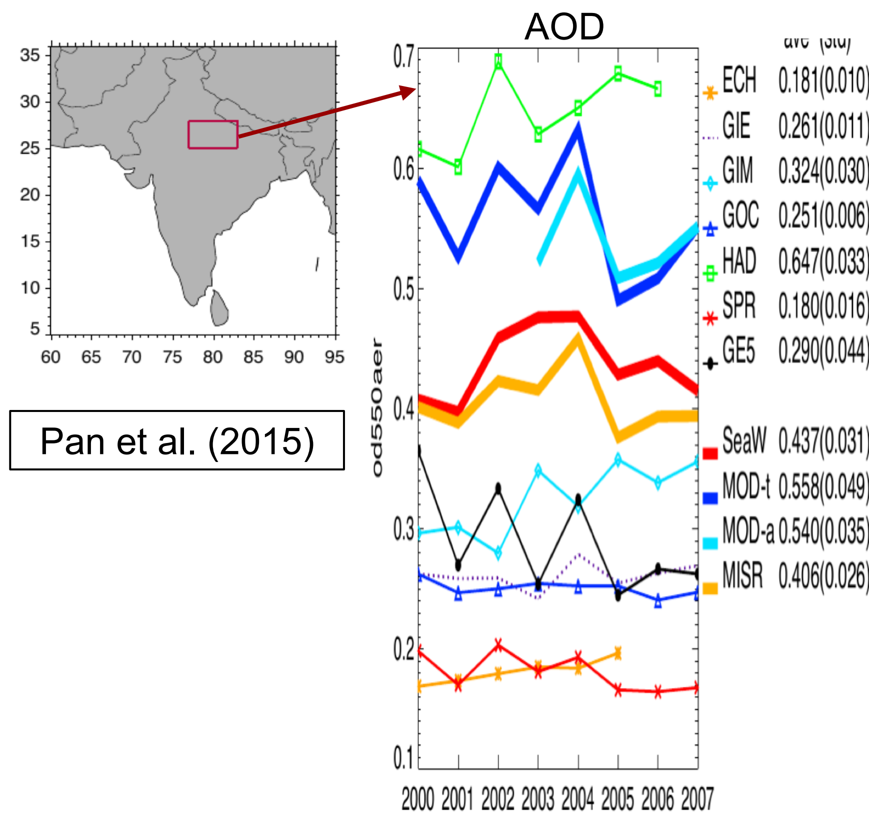
Differences across models can be useful
to identify common problems across models, explore structural uncertainty, and identify errors

Persistent high bias in modeled summertime surface ozone over some regions



Fiore et al. (2009)

Models underestimate AOD over central Indo-Gangetic Plains - attributed to common problems in emissions and nitrate aerosols

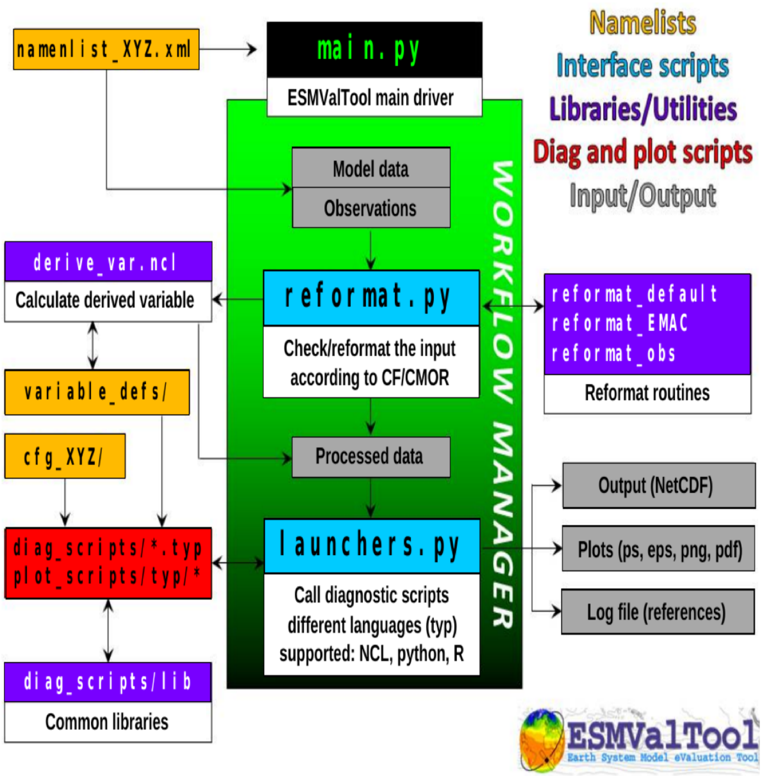


Pan et al. (2015)

Multi-model Evaluation

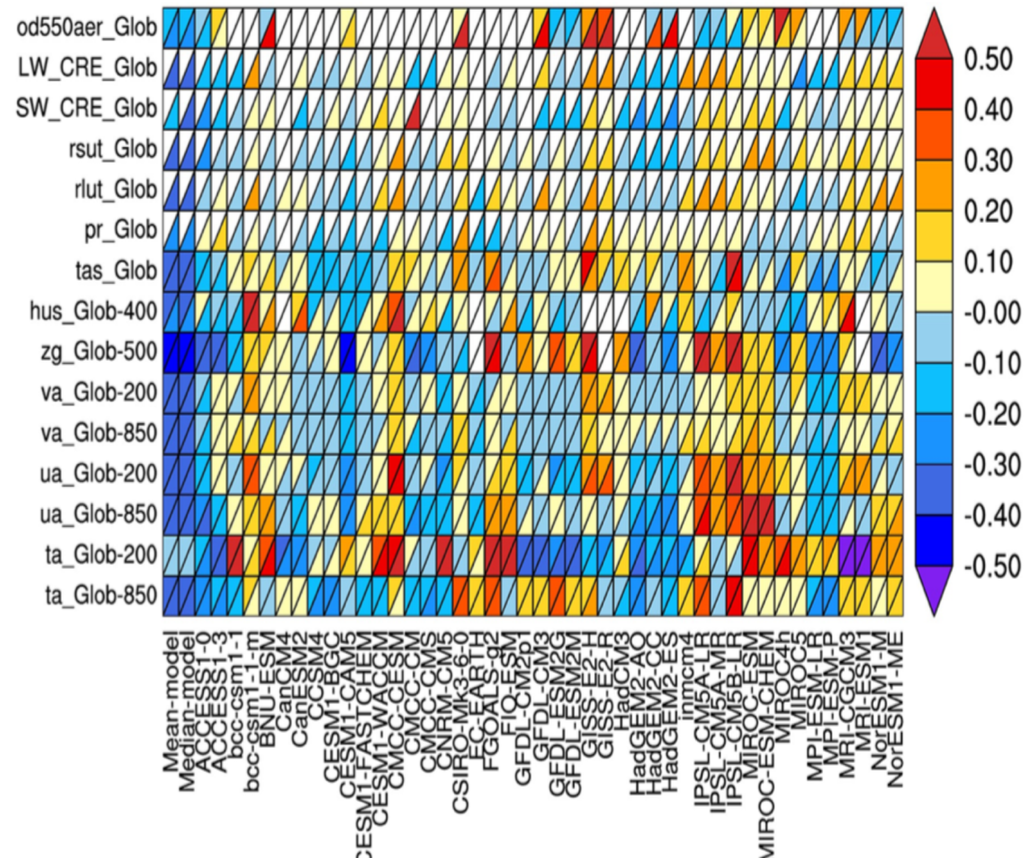
Performance-Based Metrics to compare performance of single or multiple models against observations

ESMValTool Structure



Portrait Diagram

Global RMSE

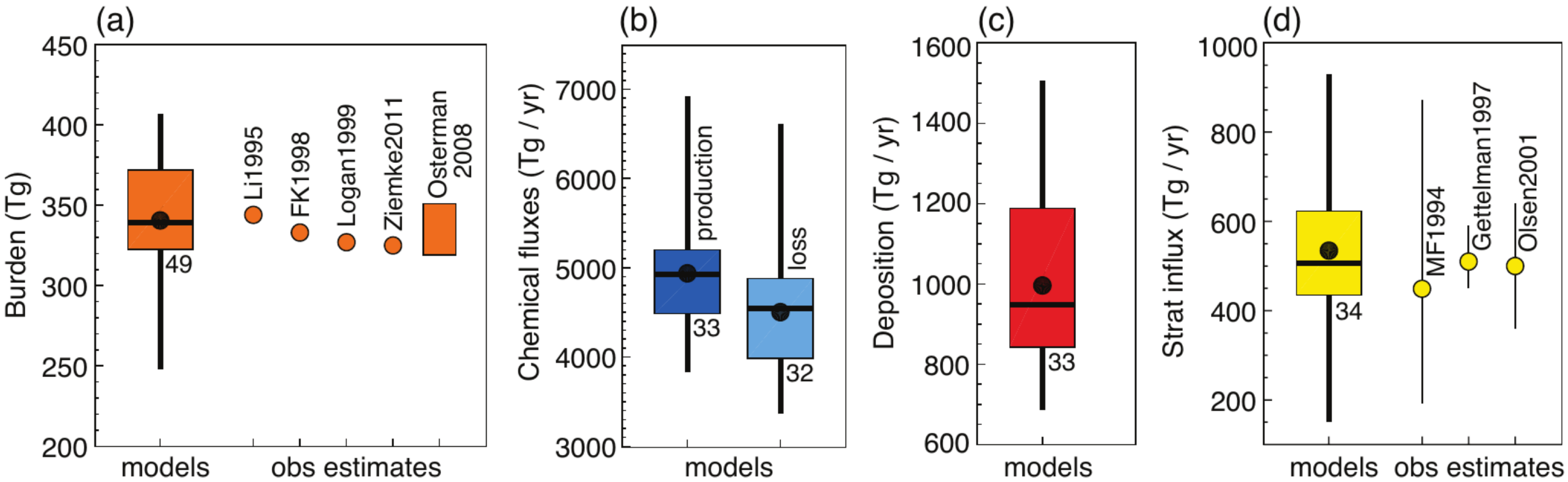


Eyring et al., 2016

Obs4MIP+CCMVal = Earth System Model Evaluation Tool (ESMValTool)

Multi-Model Process-Oriented Evaluation

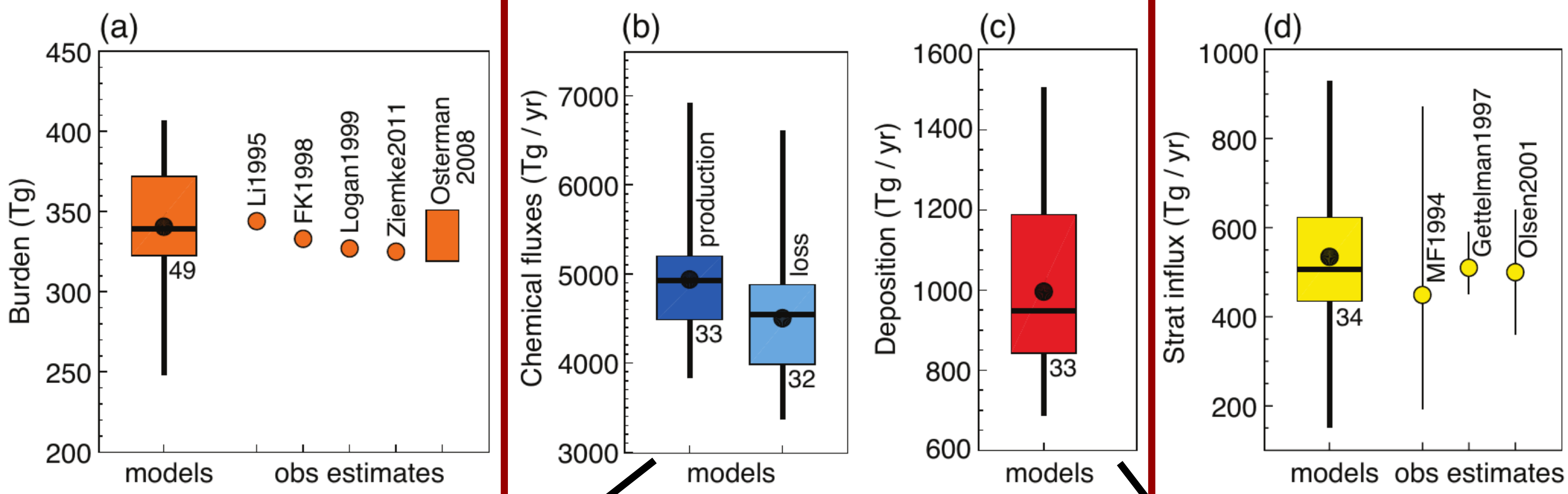
Multimodel Species Burden and Budget - first order metric for intercomparisons (e.g., O_3, CO, aerosols,..)



Young et al. (2018)

Multi-Model Process-Oriented Evaluation

Multimodel Species Burden and Budget - first order metric for intercomparisons (e.g., O_3, CO, aerosols,..)



Young et al. (2018)

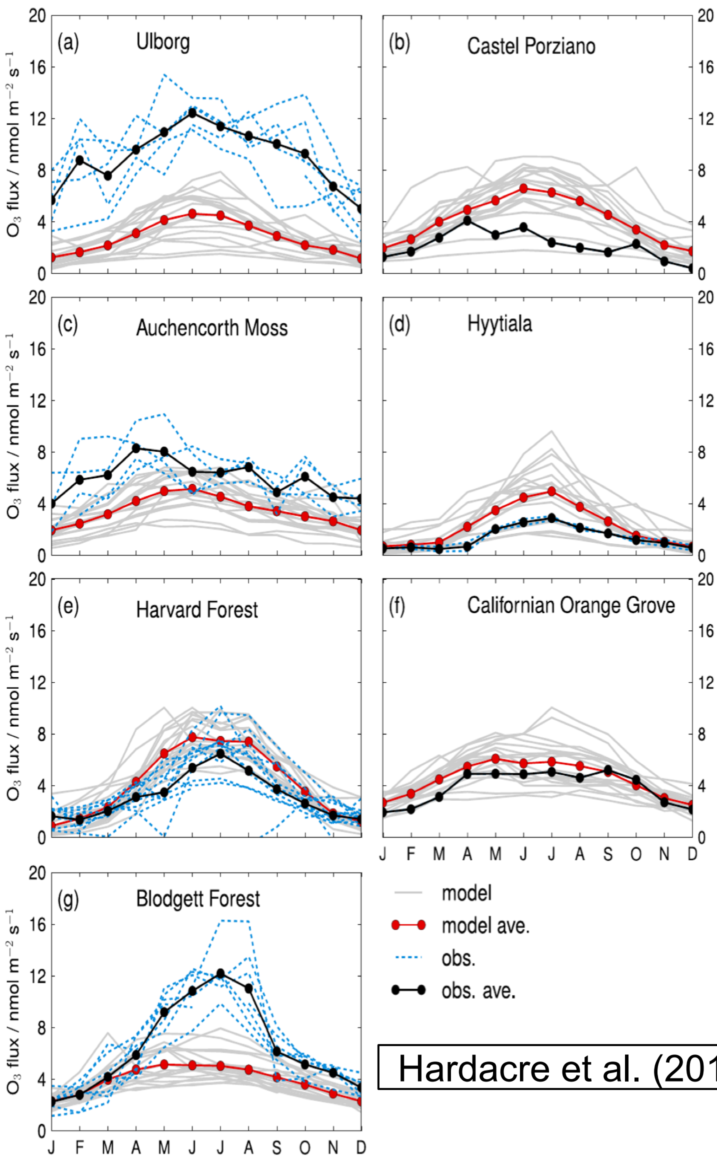
No Global-scale Observational Estimates

Consensus across models that Prod > Loss
Intermodel differences related to different chemical mechanisms

Large intermodel spread indicates considerable uncertainty in dry deposition of ozone → opportunity for improvements!

Process-oriented Evaluation

Multi-model Ozone Dry Deposition Flux versus Obs



Hardacre et al. (2015)

Targeted evaluation of individual processes with a single model using process-level diagnostics **improves understanding and helps refine models**

AM3-DD dry deposition coupled with Land Model

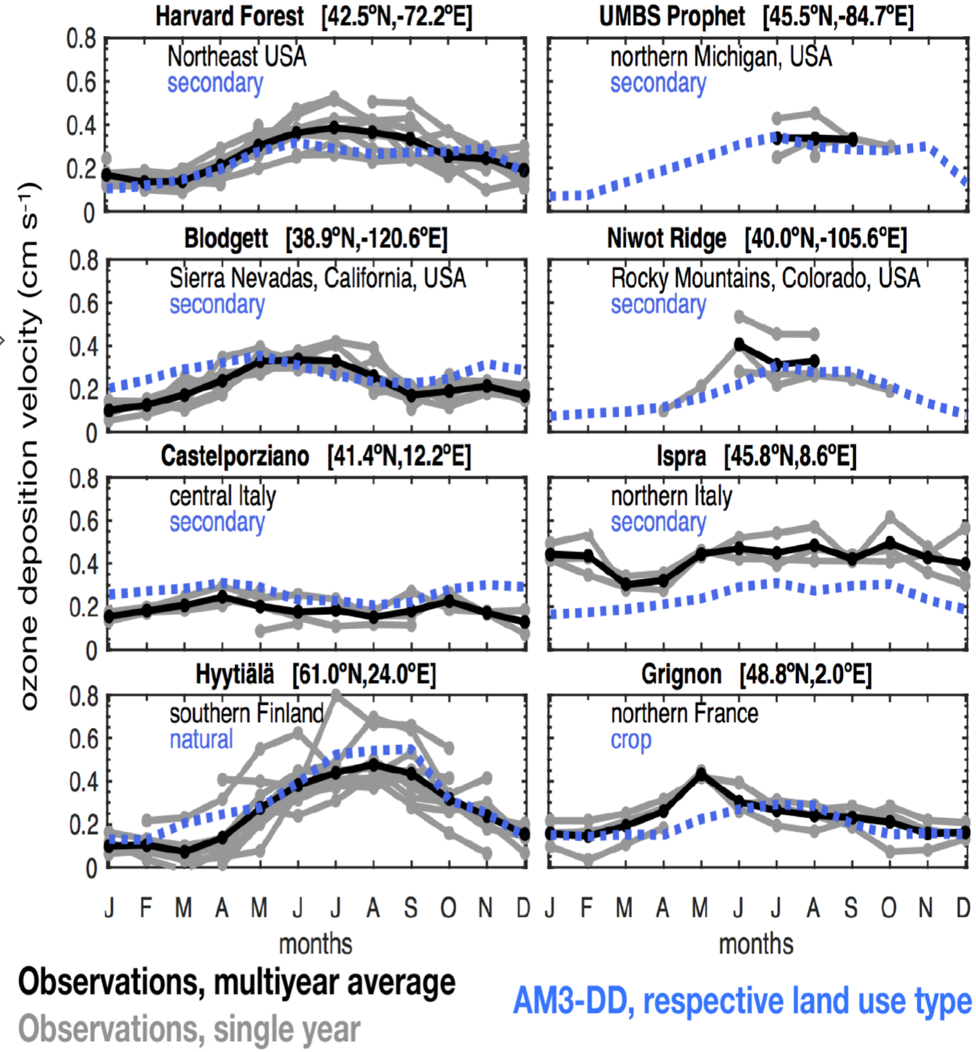
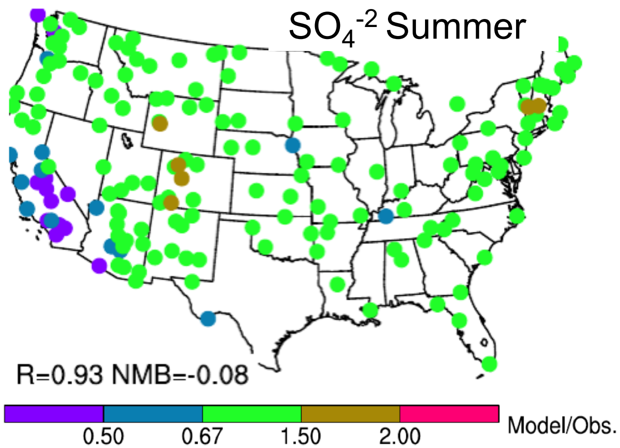


Figure courtesy Olivia Clifton

Process-oriented Evaluation

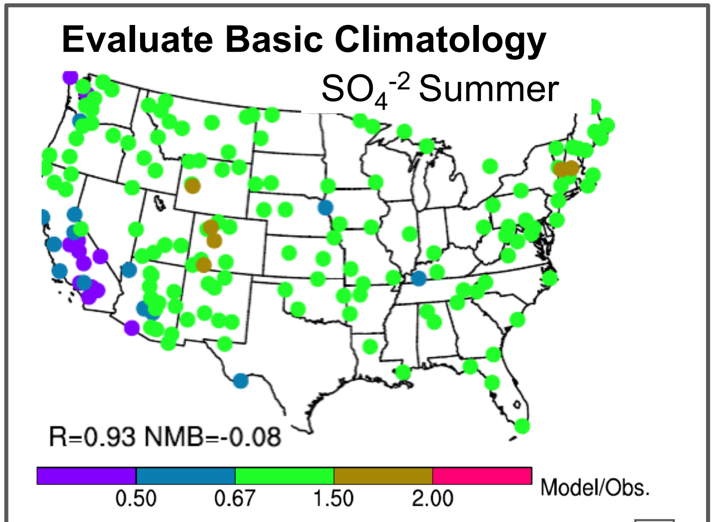
Evaluate Basic Climatology



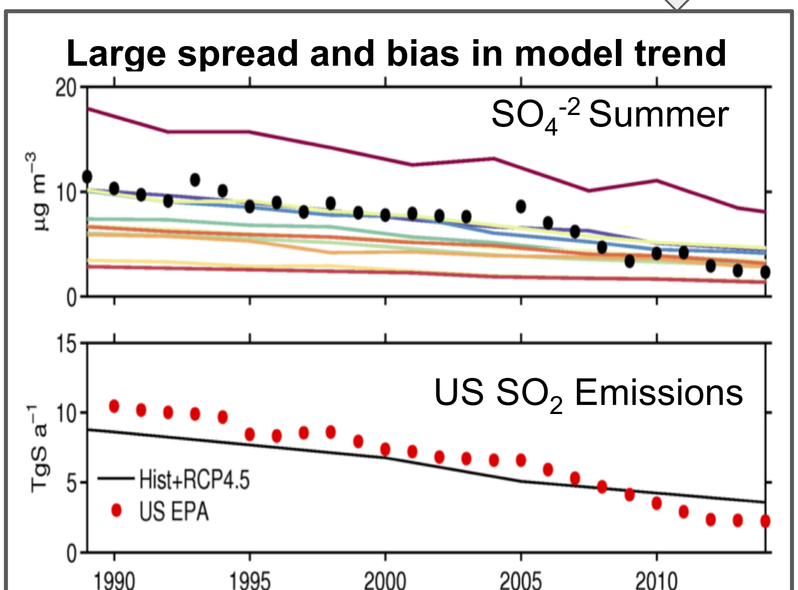
But what about sensitivity?

Paulot et al. (2016)

Process-oriented Evaluation



But what about sensitivity?

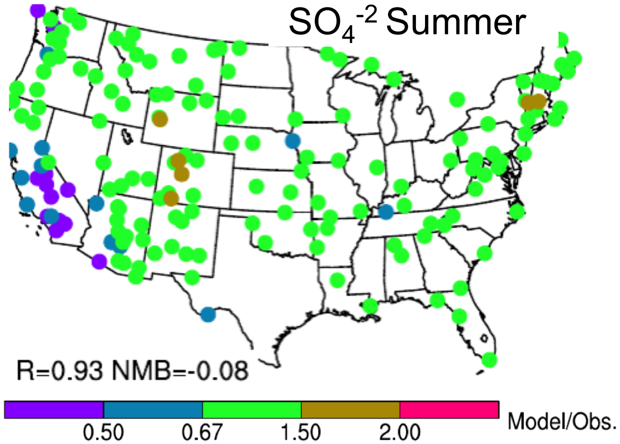


Emissions?, chemistry?, wet deposition?

Paulot et al. (2016)

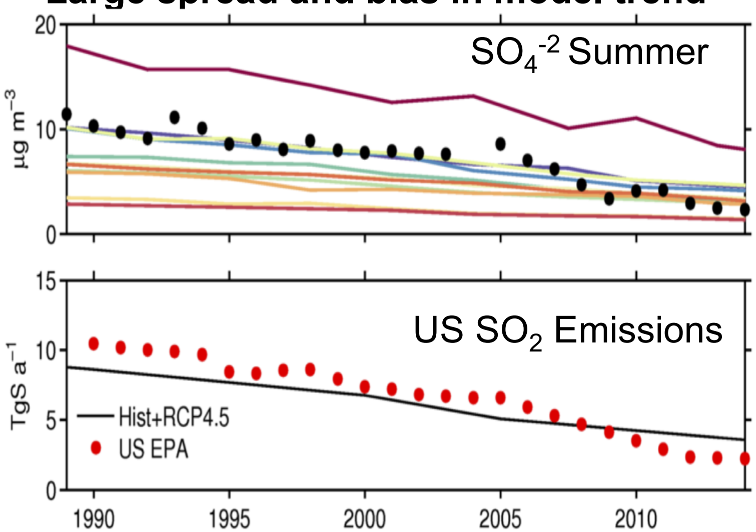
Process-oriented Evaluation

Evaluate Basic Climatology



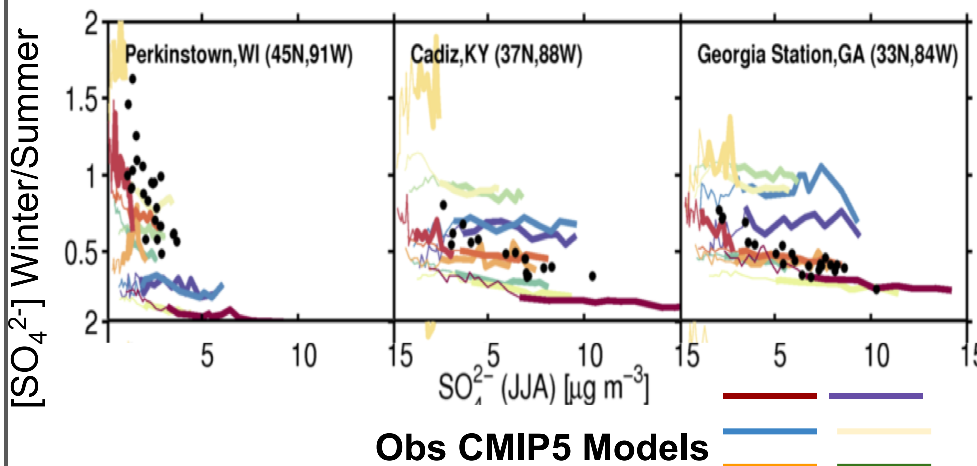
But what about sensitivity?

Large spread and bias in model trend



Emissions?, chemistry?, wet deposition?

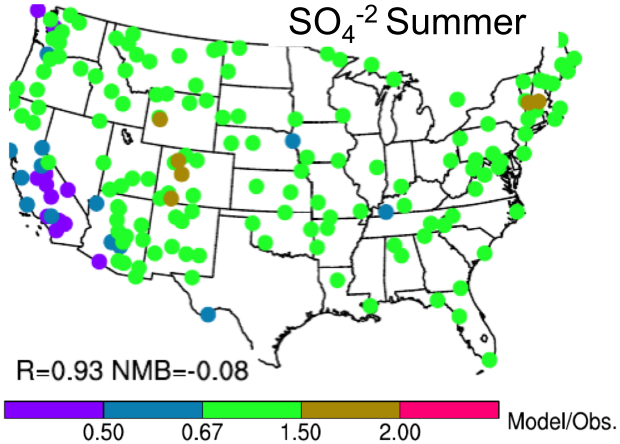
Use observational diagnostics to reduce diversity



Paulot et al. (2016)

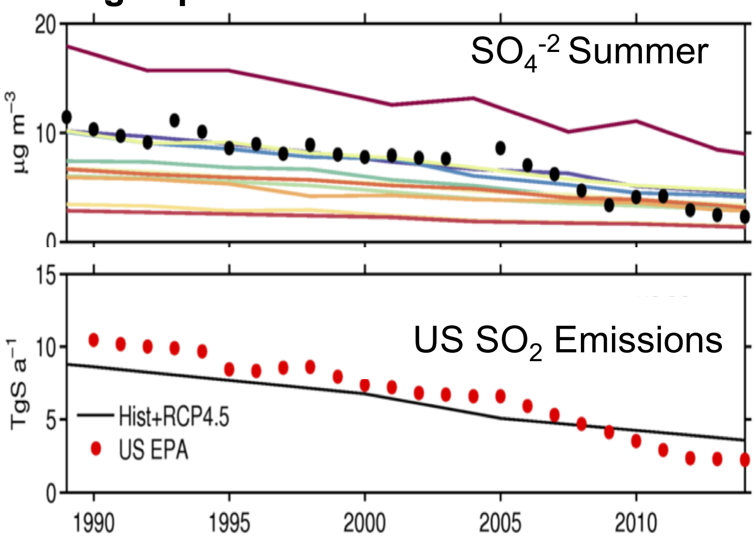
Process-oriented Evaluation

Evaluate Basic Climatology



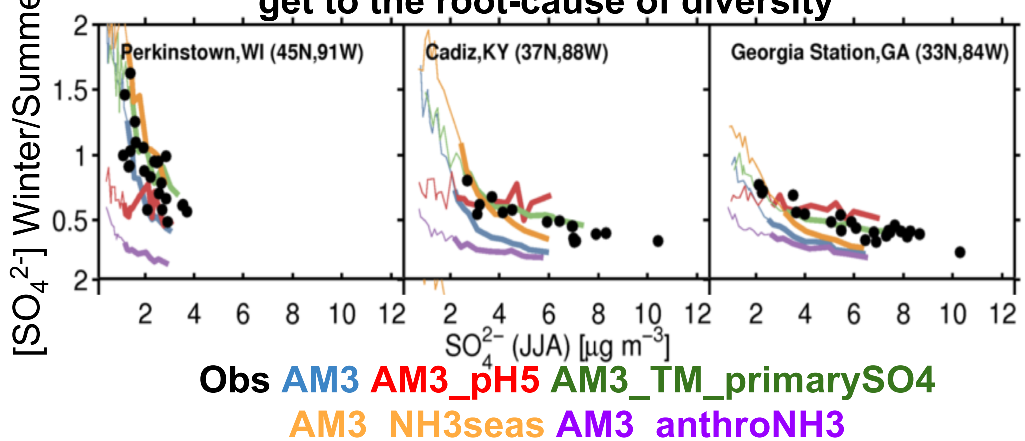
But what about sensitivity?

Large spread and bias in model trend

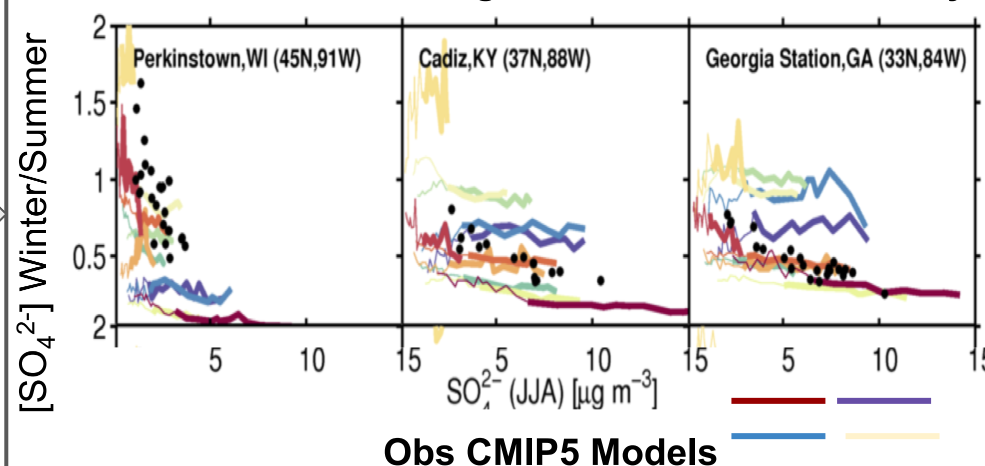


Emissions?, chemistry?, wet deposition?

Perform sensitivity simulations with one model to get to the root-cause of diversity



Use observational diagnostics to reduce diversity

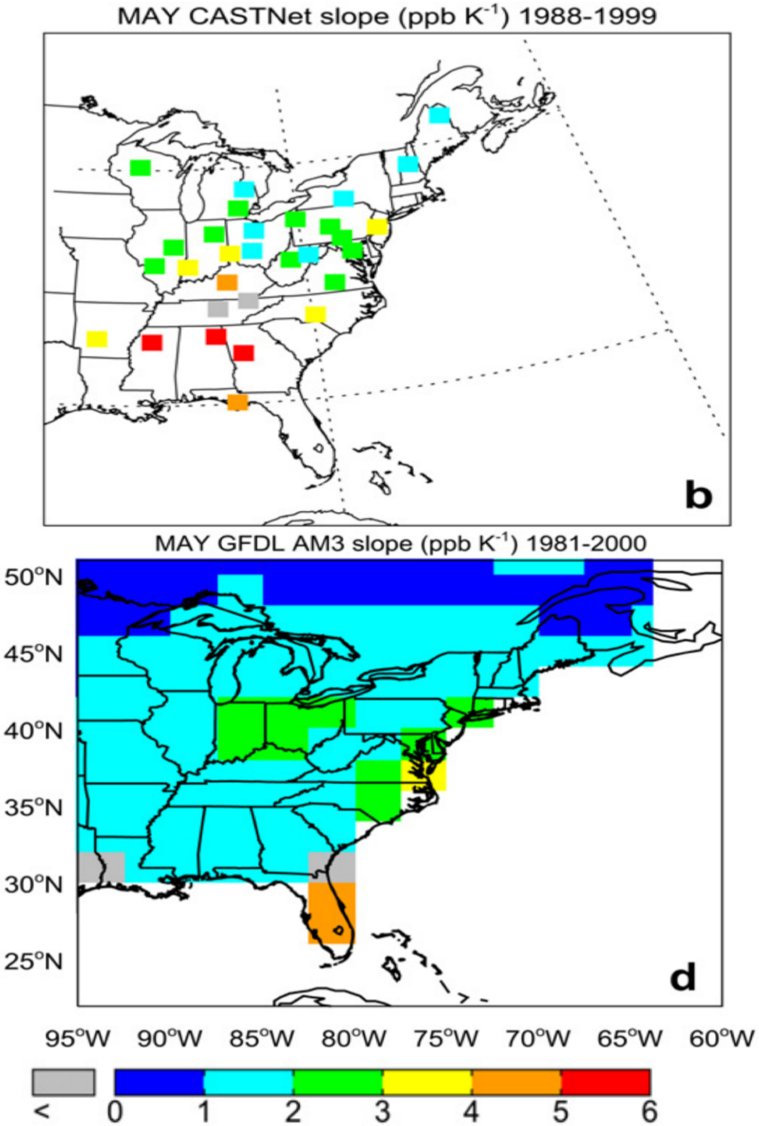


Paulot et al. (2016)

Process-oriented Evaluation

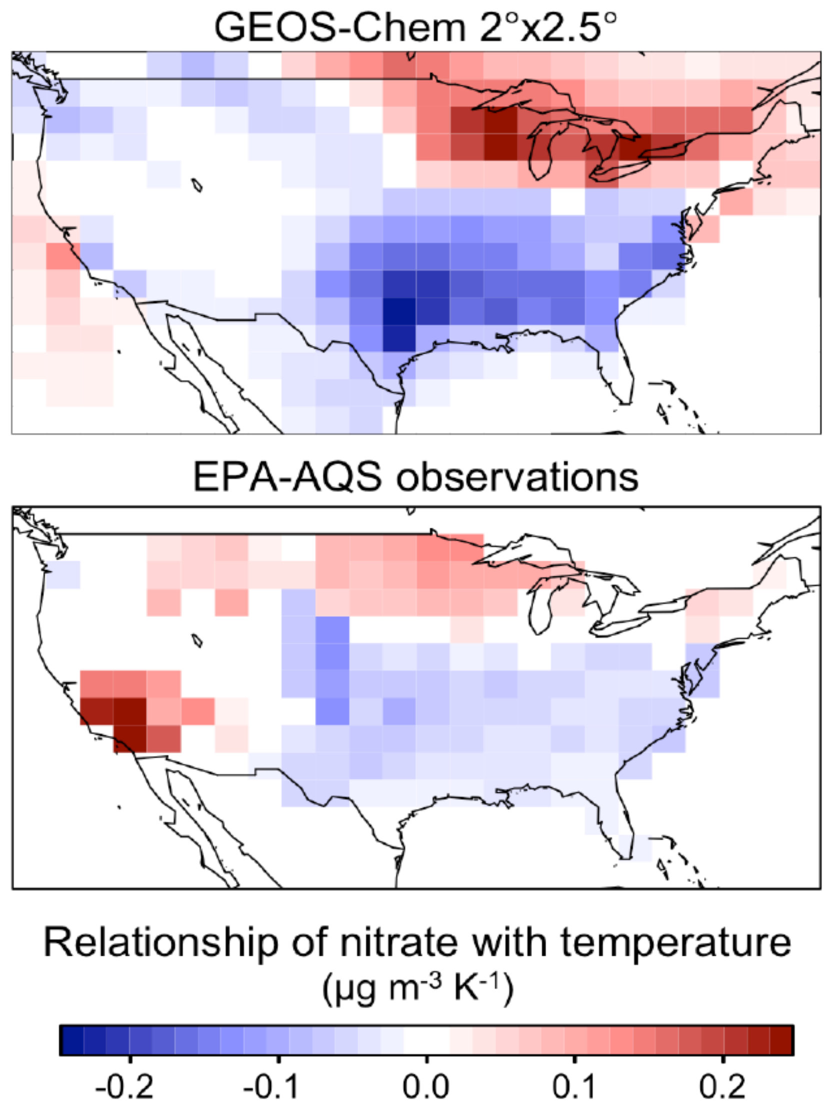
Observed relationships between trace species and meteorology provide a test for model processes

$d[O_3]/dT$



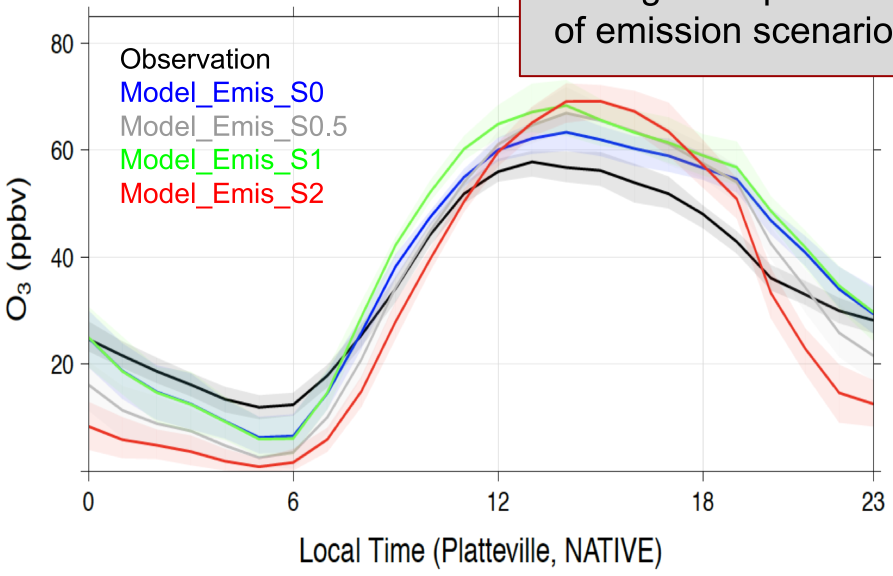
ppbv K^{-1}

Rasmussen et al. (2012)



Tai et al. (2012)

Sources of Disagreement

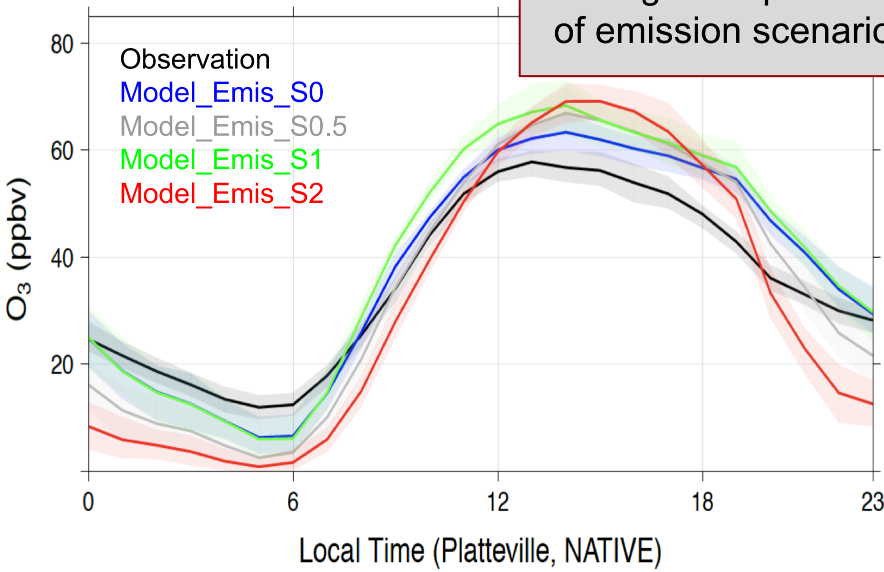


Modeled peak ozone too high independent of emission scenario

Pfister, Flocke and Lee, FRAPPÉ Final Report, 2017

Sources of Disagreement

Modeled peak ozone too high independent of emission scenario

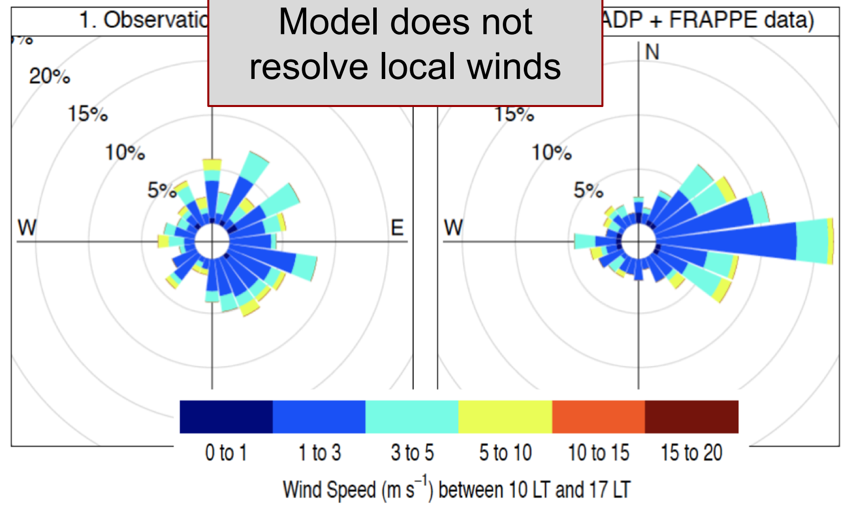


Multiple Factors can Contribute to Model-Observation Differences

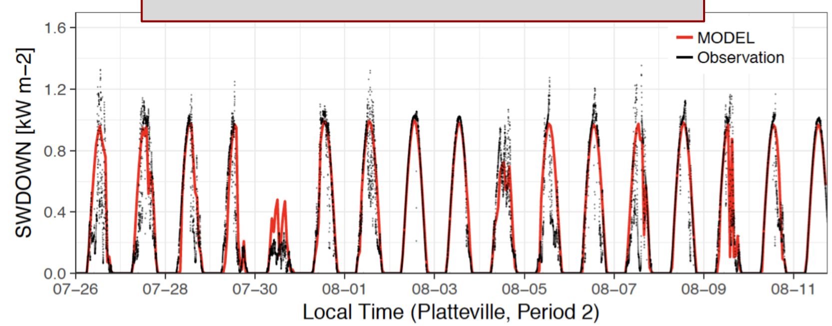
- Model Inputs - emissions
- Chemistry
- Physics - Clouds, Winds, Radiation, Boundary Layer, ...

Pfister, Flocke and Lee, FRAPPÉ Final Report, 2017

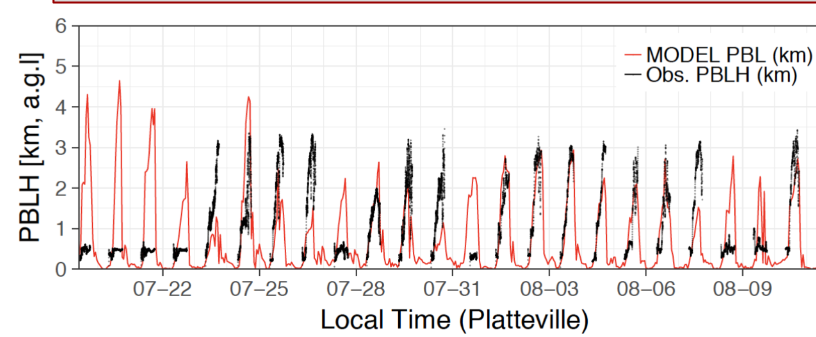
Model does not resolve local winds



Model underestimates clouds



In parts large difference in Boundary Layer



What is a Good Model Performance?

- There is **no single metric that captures model skill**
- Choice of evaluation method(s) depends on model application and available observational constraints

Critical assessment of the model-measurement comparison is needed:

- How representative are the measurements and the model for the specific time period and location?
- Is the evaluation appropriate for the purpose of the study?
- Does the model have the appropriate level of complexity for the specific problem being addressed?
- What is the acceptable level of model performance?