**October 25, 2016 Meeting Model Performance Evaluation Issues/Topics**

Below is a summary of the issues and topics pertaining to model performance for regulatory modeling for ozone, PM2.5 and regional haze that we discussed during our ad hoc meeting on October 25, 2016. Note that (1) we did not discuss the extent to which we agree or disagree on the importance or relevance of any particular issue and (2) the photochemical modeling guidance already addresses some, but not all of these issues. We may agree on appropriate changes that are needed to the guidance in some cases, but in other cases the guidance may already be adequate.

*Issues Raised During the Meeting*

1. What is the purpose of doing a model performance evaluation?

2. How can model performance information be synthesized to inform policy decisions?

3. Is there be a common set of key model performance statistics and graphical representations that should be created routinely to enable a consistent comparison of model performance across model applications and to track model performance over time as models and modeling platforms evolve?

4. Under what circumstances are model performance benchmarks or goals useful for judging the utility of a particular modeling platform?

5. Is there a technical basis for numerical “pass/fail” performance criteria?

6. Is base year model performance an appropriate indicator of the credibility of the modeling-based response to emissions changes?

7. Are there ways to evaluate model-predicted contributions since contributions are not a measured quantity?

8. How can model performance information be used to help select appropriate days/episodes to use in calculating future year design values and/or contributions?

9. Should EPA’s modeling guidance contain recommendations on procedures for quality assurance of model inputs for base year and future year model runs?

10. Photochemical modeling systems (i.e., model inputs, model construct, model representation of physical and chemical processes) are “evolutionary”. That is, the regulatory and research community is continuing to develop new data and techniques intended to improve the credibility of model applications. In this regard, at what point is a modeling platform considered “stable” and “reliable” for use in regulatory applications?

11. Dynamic evaluations can provide useful information on the ability of the modeling system to predict the effects of changes in emissions. However, because of the time, resources, and technical expertise needed to conduct a dynamic evaluation, it may be difficult for individual states to perform this type of evaluation to support SIP modeling.

12. Is there an efficient framework or process for states and EPA to discuss regulatory modeling problems and performance issues?

13. It is important, but resource intensive, to conduct and evaluate global models that provide boundary conditions for EPA and state regulatory modeling. Can EPA in conjunction with NOAA and NASA evaluate global models in a transparent manner and provide the results to MJOs/states?

14. In pairing observations with gridded model predictions, should consideration be given to the location of the monitor within the grid cell? That is, if a monitor is near the edge of a grid cell, should it be paired with the predictions from that grid cell or would it be more appropriate to apply some type of spatial interpolation technique?

15. Is there value in conducting multiple model runs with varying inputs to judge the stability of Relative Response Factors?

16. Is there value in evaluating the model predictions of precursor concentrations against routine ground-level precursor measurements? How should model performance results for precursors be portrayed in view of limitations in precursor measurement techniques/instrumentation?

17. Can data from personal monitors be included as part of a model performance evaluation?

*Issues Raised by Georgia EPD*

1. MPE on subset of days used in RRF calculations and contribution calculations.
2. Detailed site-specific MPE for all monitors in the domain, especially current nonattainment and maintenance monitors and projected nonattainment and maintenance monitors.  This would include time series plots and monthly statistics shown on bubble (dot) plots to show the spatial trends across the domain.
3. For interstate transport assessments, need to quantify the overall MPE from the contributing states (e.g., over-predictions at the majority of monitors in a state may lead to significant contribution linkages that are not real).
4. Adjustments of contributions by removing poor performing days and/or adjusting contribution results (up/down) based on model biases.  The contribution adjustment approach should consider bias at the receptor monitor (#1 and #2 above) and from the contributing state (#3 above).
5. Discussion on use of 1x1 cell vs. 3x3 max cell array for RRF and contribution calculations.  GA EPD feels it is inappropriate to use any grid cell besides the grid cell containing the monitor for these calculations since the non-monitor grid cells have not been evaluated against observations.  If higher concentrations exist in nearby grid cells, they will be identified with the unmonitored area analysis (UAA).

*Issues Raised by NC DNER*

1. Model performance metrics - monitor cell vs max 3x3 array

* + EPA bases model performance only on the grid cell containing the monitor.  However, RRFs and future design values are computed using the maximum cell in the 3x3 array around the grid cell.  In some cases (i.e. many coastal monitors), the bias and error using the max 3x3 is much worse than the monitor grid cell or the maximum of land cells within the 3x3 array.  Model performance can help inform whether to use the monitor cell, the maximum cell in a 3x3 array, or a modified 3x3 array that omits cells with high bias/error.

2. Fine scale (4km or finer) modeling

* + Numerous studies show that 12km WRF modeling does not perform well along coast lines.  A resolution of 4km or less is needed to properly resolve sea breeze circulations.

3. Model performance considerations for Interstate contributions

* + Remove poor performing model days from contribution calculations.  A single poor performing model day can lead to the incorrect conclusion that a state is contributing.  There will be a need to determine what is "poor" model performance.  In the North Carolina Transport SIP, we used a very high threshold of +/- 40% normalized bias.  Lower thresholds should be considered and are likely more appropriate.

4. Alternative methods for determining contributions

* + A. Compute contributions on days with ozone that was actually above the standard in the base year, and

B. Using daily RRFs on observed ozone, compute contributions on days projected to have ozone above the standard in the future.  In a perfect world, method “B” should match the methodology within CSAPR2.

5. Model multiple years

* + Wind patterns in a particular year may not represent the typical wind pattern in most years.  EPA should consider modeling 2 or more years for transport.

6. Adjust future maximum design values if they are not centered on the base year.

* + The future max DV for a max DV centered before the base year will not account for a year of emission reductions, since the RRF is calculated from the base model year.  This will skew the future max DV higher.   Alternatively, the future max DV based on a max DV centered after the base year will skew the future max DV lower.

7. Trajectory analysis / residence times

* + Backward trajectory analyses that factor in residence times and the amount a monitor exceeds the standard (i.e. higher ozone days given more weighting for residence times).

8. Canada Emissions

* + "Ballpark" projections of Canada emissions are important for projecting future design values in the northeast US.  At the very least, make conservative adjustments to mobile emissions.

9. How do you handle a situation where you have models that have similar model performance but predict different future DV and contributions?