DEVELOPMENT, IMPLEMENTATION, AND APPLICATION OF AN IMPROVED MODEL PERFORMANCE EVALUATION AND DIAGNOSTICS APPROACH

Byeong-Uk Kim, William Vizuete and Harvey E. Jeffries*
Department of Environmental Sciences & Engineering, University of North Carolina, Chapel Hill, NC, USA

1. INTRODUCTION

For the air quality management in the United States, the use of photochemical air quality models (PAQM) is mandatory for states violating the National Ambient Air Quality Standards (NAAQS) to develop their State Implementation Plans (SIPs) and demonstrate that their SIP will “more likely than not” lead them to attain the ozone NAAQS in the future. This SIP submission, including the attainment demonstration, is required by the Clean Air Act (CAA) and each state should get the approval of the United States Environmental Protection Agency (US EPA) on their submitted SIP by the statutory deadline.

In practice, application and testing of PAQMs to SIP development is beyond most states’ technical capability. Therefore, US EPA (1991, 1996, 1999, 2005) developed modeling guidance documents that list recommendations for episode selection, model input preparation, model performance evaluation (MPE), and related topics to assist states in their SIP modeling. Among the tasks involved in a practical SIP modeling, MPE is the core step to ensure the model’s reliability for testing the effectiveness of control options described in a SIP to reduce ozone concentrations.

Often, more than one simulation with one or more PAQMs is done until model performance meets the US EPA’s recommended criteria. MPE follows each simulation with improved model inputs and/or configurations. Thus, most SIP modeling becomes iterative and to conduct effective MPE is critical to meet the statutory deadline. That is, if a MPE does not help modelers judge and improve the reliability of modeling results, a state may face the delay of their SIP submission.

Since the US EPA’s publication of the modeling guidance in 1991, MPE practices have not been improved much. Most MPE practices done by states for 1-hr SIP modeling showed several shortcomings. The most apparent problem was the heavy dependence on three statistical measures: normalized bias, gross error, and unpaired peak prediction accuracy. Because US EPA still recommends that the performance of 8-hr SIP modeling be evaluated with 1-hr performance measures as well as new 8-hr measures, the heavy dependency on statistical measures will not likely go away with the mechanical adoption of 8-hr SIP modeling guidance.

In this presentation, we will (1) review the recent advances in MPE principles along with our findings on the problems posed in the MPE practices for past 1-hr SIP modeling, (2) propose our improved MPE protocol and suggest an enhancement of MPE practice for 8-hr SIP modeling, and (3) introduce a set of tools we designed to implement our MPE protocol. These tools are computer software programs that aim to assist states in practicing our improved MPE protocol in an efficient manner.

2. VINDICATION OF MODEL RESULTS

The concept of “vindication of model results” was introduced by Jeffries (1995). Recently, Kim (2006) provided a practical guidance on accomplishing the proposed vindication concept. The next section is the summary of a part of work in Kim (2006). Even though his work is based on the 1-hr modeling, the principles should be applicable to 8-hr modeling.

The ultimate goal of this vindication is for modelers to claim that their models are indeed “best”. In the following section, we present three questions whose answers would lead modelers to claim their vindication statement. These questions were originally posed by Beck (2002) for general environmental modeling and modified for the regulatory ozone modeling by Kim (2006).

2.1 Is a model acceptable and sound in general?

The first question is “Is the formulation of a model scientifically acceptable in general?” This question can be divided into two corollary questions: (1) “Is the science encoded in the model ‘sound’, as explained in Crawford-Brown

* *Corresponding author:* Harvey E. Jeffries, Department of Environmental Sciences and Engineering, UNC-Chapel Hill, 120 Rosenau Hall, CB#7431, Chapel Hill, NC 27599-7431; e-mail: harvey@unc.edu

(2005), and working?" (2) "Is the implementation of scientific knowledge achieved through properly applying modeling procedures of generalization, distortion, and deletion to the more complex reality?" The first question focuses on the ability of a PAQM for a potential SIP modeling application in general. We believe this question is generally answered when the US EPA approved a specific PAQM for a state's SIP modeling exercise. In practice, examination of the general applicability of a PAQM requires intensive and well-designed field studies. In addition, the current regulatory framework requires states to get an approval from the US EPA on their choice of PAQMs for their SIP. Generally speaking, an approved PAQM is believed to be *capable of showing the generally understood behavior of ozone formation in urban and regional episodes*. Often, however, some improvements on a model's formulation may be required based on the findings of a SIP modeling.

2.2 Is a model "working?"

The second question is "Does a model replicate the observations adequately? (i.e. does it make predictions that match history?)" The primary focus of this question is if "a model works." Often, this question was considered being answered with so-called 'operational evaluation' in the air quality modeling community. Traditionally, modelers mainly concentrate on comparison of the predicted ozone concentrations with the observed ozone concentrations. In judging if predictions match well observations, the most frequently used criteria are summary statistics. These statistics, however, do not provide much information about how a model gets its predictions. Moreover, the traditional MPE based on these summary statistics forces a modeler to accept or reject the modeling results as a whole.

Recently, the need for extension of operational evaluations to a model's performance on the various important precursors was recognized. Additionally, as Fine et. al. (2003) summarized, because ozone predictions made by a PAQM is based on various inputs that are highly uncertain, there has been increasing research interest in performing evaluation on model inputs with respect to the potential impacts of model input uncertainties on the model's final predictions.

Besides this narrowness of evaluation coverage, the traditional MPE practice also suffers from the lack of flexibility in its MPE procedures. Frequently, SIP modeling is conducted in a 'waterfall' fashion; that is, once model inputs are considered 'quality-assured', there is no systematic way to review these inputs unless many *ad hoc* analyses point to serious issues after exhaustive model simulations are done.

The current MPE practice also lacks systematic guidance on the implementation of advanced analyses in a specific SIP modeling case and does not utilize high-resolution datasets. We believe this is simply because there is no protocol or general guidance on what to do with these analyses and observations.

In summary, it is very important to evaluate PAQM inputs and outputs simultaneously. In addition, the evaluation on inputs should be conducted in a flexible way so that modelers can review the possibilities of input errors in any step during their MPE.

2.3 Is a model fulfilling its goal?

The last question is "Is a model usable for answering specific (e.g. policy) questions? (i.e. does the model fulfill the designed task?)" This question is rarely asked in past SIP studies. Beck (2002) and Reichert and Borsuk (2005) noted that the lack of absolute accuracy of the model predictions does not preclude the usefulness of modeling for policy development. Moreover, as PAQMs becomes complex, an empirical rejection begins to be harder.

The results of PAQM contain a certain amount of uncertainties due to the fact that model inputs represent part of the past status of environmental systems that are essentially *not knowable*. All environmental systems are open-systems; that is there are always "unknown" factors that are not controllable. In addition, an environment holds its unique 'landscape' such as the composition of industrial sources in a specific area. Therefore, a certain degree of tolerance on model's uncertainty is required when the model's predictions are used. This is, we believe, probably the most important aspect of environmental modeling, as Beven (2002) recognized. At the same time, however, we recognize that there are unacceptable errors for model applications to decision making processes. The typical example of these types of error is the compensating errors that may lead to wrong directions in policy decisions.

Examining model outputs only, as is the case in the use of US EPA's minimal set of standard statistical tests, does not help modelers detect these unacceptable errors. Due to the non-linearity of ozone formation mechanism, different combinations of precursor emission inputs and meteorological inputs can result in similar ozone concentrations. Often, this leads to conclusions that require controlling the "wrong" precursor. Thus, compensating errors are an important issue in photochemical air quality modeling. Some models may show "good" performance in terms of the traditional MPE but should not be used for the policy-making process because of their unacceptable errors.

In summary, MPE should be able to identify (or at least to send a signal about) modeling results are possibly incorrect or, equally important, to indicate that the model may not be reliable. MPE for SIP modeling should not be a series of tasks mechanically comparing predictions with observations. Interestingly, the same problem of insufficient MPE methodologies exists in virtually all environmental modeling communities and a clear solution has not been found. In this presentation, we hope to show a practical solution to the problem.

3. Development of an improved MPE protocol for regulatory ozone modeling

Based on the review of the past MPE practice following US EPA's guidance and the recent advances in MPE researches, Kim (2006) could identify that the several shortcomings in the past MPE practices and proposed an improved protocol, the Protocol for Regulatory Ozone Modeling Performance Tests (PROMPT). PROMPT is a meta-protocol; that is, PROMPT itself does not work as an actual MPE protocol. Model evaluators can utilize PROMPT structure and design goals with accompanying descriptions for the underlying principles when they construct a MPE protocol for a specific SIP modeling.

PROMPT is based on four major guiding questions composed of several subsequent questions as shown in Table 1. Questions 1 and 2 in Table 1 are primarily for the question: "Does a model replicate the observations adequately?" Questions 3 and 4 in Table 1 are mainly for the question: "Is a model usable for answering specific (e.g. policy) questions?" Answers to these questions will formulate rationales for the vindication of modeling results when combined with US EPA's approval on the model selection. Ultimately, modelers may be able to answer the question by policy makers: "Why should I believe this model?"

Table 1. Summary of PROMPT procedural questions.

1. Does this model show or have all necessary components to produce the phenomena that we can expect from the current best perceptual/conceptual model?
   a. What are the model setup and justification? What amounts and kinds of observation are available for evaluation? How are model inputs prepared for model operation?
   b. Is the overall ozone behavior in the model consistent with the conceptual model?
   c. If not, what are the possible causes? Is there any alternative model inputs or configurations?
2. Can this model distinguish which precursor(s) to control for ozone reduction?
   a. Does protocol for graphical measure construction exist?
   b. Does model show correct source-receptor relationship?
   c. Does model have biases in surface winds, NO_X, and O_3 (plus CO if available)?
   d. Which precursors are important for potential policy options?
3. How precisely can the model estimate control requirements?
   a. How does model perform at locations where observations are available?
   b. How does model predict at locations where no observation exists?
   c. What are the resolution of control options in space and time?
4. What are the possible biases in the prediction and the impact of biases on the policy choice?
   a. Where does the future ozone problem occur in the model? How does the model perform and/or predict those locations?
   b. Do the biases found in model predictions affect the choices of possible control options?
   c. What is the evaluator's confidence on the reliability of model performance in supporting proposed policy options?

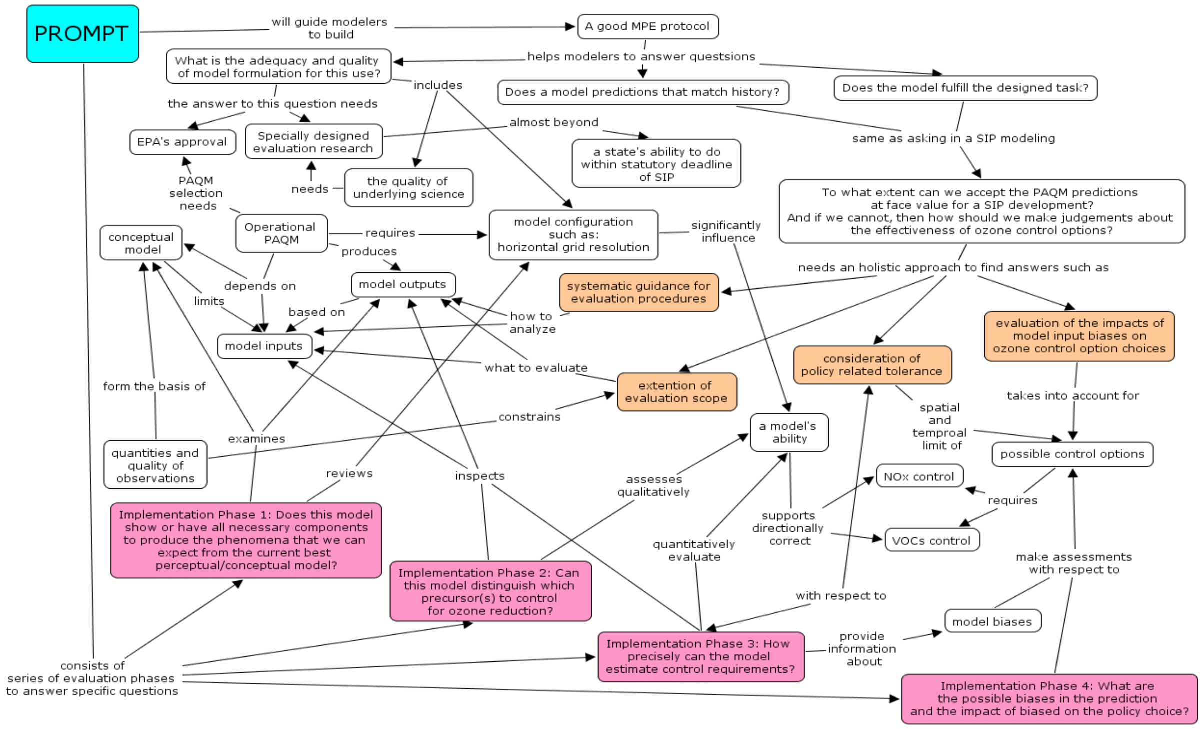
PROMPT
will guide modelers to build
A good MPE protocol
helps modelers to answer questsions
What is the adequacy and quality of model formulation for this use?
Does a model predictions that match history?
Does the model fulfill the designed task?
includes
the answer to this question needs
EPA's approval
Specially designed evaluation research
almost beyond
a state's ability to do within statutory deadline of SIP
same as asking in a SIP modeling
PAQM selection needs
needs
the quality of underlying science
To what extent can we accept the PAQM predictions at face value for a SIP development? And if we cannot, then how should we make judgements about the effectiveness of ozone control options?
conceptual model
Operational PAQM
requires
produces
model configuration such as: horizontal grid resolution
significantly influence
needs an holistic approach to find answers such as
depends on
limits
model outputs
systematic guidance for evaluation procedures
based on
how to analyze
evaluation of the impacts of model input biases on ozone control option choices
model inputs
what to evaluate
consideration of policy related tolerance
extention of evaluation scope
form the basis of
constrains
takes into account for
spatial and temproal limit of
a model's ability
examines
quantities and quality of observations
possible control options
reviews
inspects
assesses qualitatively
NOx control
requires
Implementation Phase 1: Does this model show or have all necessary components to produce the phenomena that we can expect from the current best perceptual/conceptual model?
supports directionally correct
quantitatively evaluate
VOCs control
Implementation Phase 2: Can this model distinguish which precursor(s) to control for ozone reduction?
make assessments with respect to
with respect to
model biases
Implementation Phase 3: How precisely can the model estimate control requirements?
provide information about
consists of series of evaluation phases to answer specific questions
Implementation Phase 4: What are the possible biases in the prediction and the impact of biased on the policy choice?

Figure 1. Conceptual map of the Protocol for Regulatory Ozone Performance Tests

The current PROMPT consists of four implementation phases and each PROMPT implementation phase is designed to answer each question in Table 1 with a set of analysis procedures. Each procedural set contains the statement of analysis goals, the required information (including characteristics of information) for following procedures, the list of proposed analyses with recommended material, and the suggested procedures to follow. PROMPT also includes the relationship among different tasks and the documentation requirement. The conceptual map of PROMPT is depicted in Figure 1.

4. SUPPORTING TOOLS FOR IMPLEMENTING PROMPT

In implementing PROMPT approach, Kim and Jeffries (2005) found the existing tools are inefficient for permitting an implementation of PROMPT. Therefore, Kim and Jeffries (2005) developed the Python-based Performance Analysis Support System (pyPASS) to facilitate the application of the new MPE approach for regulatory photochemical modeling. Since its introduction to the regulatory air quality community by Kim and Jeffries (2005), pyPASS has been enhanced to accommodate needs for the 8-hr modeling and users' requests. In this presentation, I will summarize the recent advances of PROMPT and pyPASS. For details of pyPASS advances, readers are encouraged to refer to the presentation of Leiran (2006).

5. REFERENCES

Beck, B., 2002: Model evaluation and performance. In *Encyclopedia of Environmetrics*, Eds. A. H. El-Shaarawi and W. W. Piegorsch, John Wiley & Sons, Ltd., Chichester.

Beven, K., 2002: Towards a Coherent Philosophy for Modelling the Environment. Proceedings of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences 458, 2465-2484.

Crawford-Brown, D., 2005: The Concept of Sound Science in Risk Management Decisions. Risk Management 7, 7-20.

Fine, J., Vuilleumier, L., Reynolds, S., Roth, P., Brown, N., 2003: Evaluating Uncertainties in Regional Photochemical Air Quality Modeling. Annual Review of Environment and Resources 28, 59-106.

Jeffries, H. E., 1995: Science and Policy Interaction in the Air Quality Decision Process. In *Policy and Air Quality*, Eds. P. Solomon and M. Rodgers, Air and Waste Management Association, PA.

Kim, B. U., 2006: Development, implementation, and application of an improved protocol for the performance evaluation of regulatory photochemical air quality modeling, PhD Dissertation, University of North Carolina at Chapel Hill.

Kim, B.U. and Jeffries, H.E., 2005: Python-based Performance Analysis Supporting System (PyPASS): A software tool set for the performance analysis of regulatory photochemical air quality modeling, Extended Abstracts, *4th Annual CMAS Models-3 Users' Conference*, Chapel Hill, NC, CMAS Center.

Leiran, B., Kim, B.U., Vizuete, W., Jeffries, H.E., 2006: Improvements to the Python-based Performance Analysis Support System (PyPASS): New tools to support air quality model evaluations, Extended Abstracts, 5*th Annual CMAS Models-3 Users' Conference*, Chapel Hill, NC, CMAS Center.

Reichert, P., Borsuk, M. E., 2005: Does high forecast uncertainty preclude effective decision support? Environmental Modelling & Software 20, 991-1001.

US EPA, 1991: Guidance for Regulatory Application of the Urban Airshed Model.

US EPA, 1996: Guidance on Use of Modeled Results to Demonstrate Attainment Of the Ozone NAAQS.

US EPA, 1999: Guidance for improving weight of evidence through identification of additional emission reductions, not modeled.

US EPA, 2003: Appendix W to: Part 51. Guideline on air quality models.

US EPA, 2004: http://www.epa.gov/air/airtrends/2003ozonereport/intro.html

US EPA, 2005: Guidance on the use of models and other analyses in attainment demonstrations for the 8-hour ozone NAAQS.