Development, implementation, and application of an improved protocol for the performance evaluation of regulatory photochemical air quality modeling

Byeong-Uk Kim

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Environmental Sciences and Engineering.

Chapel Hill
2006

Approved by:
Professor Harvey Jeffries
Professor Lawrence Band
Professor Douglas Crawford-Brown
Professor Donald Fox
Professor Mort Webster
ABSTRACT

Byeong-Uk Kim: Development, implementation, and application of an improved protocol for the performance evaluation of regulatory photochemical air quality modeling
(Under the direction of Prof. Harvey Jeffries)

Ozone is a secondary pollutant resulting from complex reactions of two precursors: nitrogen oxides (NOX), and volatile organic compounds (VOCs) under ozone-conducive meteorological conditions. Thus, the ozone modeling becomes complex and needs rigorous model performance evaluations (MPE) before the modeling results are used for air quality decisions. In the past regulatory ozone modeling, however, virtually all MPE practices were over-simplified by following the EPA’s current MPE method. That is, modelers cannot answer the most important question in applying air quality models for ozone decision-making processes with the EPA’s MPE method: “why should I believe this modeling?”

In this study I investigated a solution by integrating the theoretical advances of MPE for environmental modeling with my practical knowledge in regulatory ozone modeling. As a result, I developed an MPE method with which modelers must (1) gather and examine graphical/statistical measures in a systematic manner, (2) conduct in-depth analyses with respect to potential ozone control options, and (3) report their performance assessments explicitly in light of policy questions. Because the existing analysis tools showed significant shortcomings in implementing the new MPE method, a new tool was developed to exercise the new MPE method efficiently. With the new tool, modelers can accomplish MPE tasks necessitated by the new MPE method in a timely manner.
The Houston-Galveston Mid-Course Review (HGMCR) modeling was re-evaluated as the case study to demonstrate the advantages of new MPE method. I could reveal that the HGMCR modeling showed significantly low reliability even though the model could pass the majority of EPA’s simple statistical tests. That is, the model showed significantly high biases in winds, NO\textsubscript{X}, and VOCs. Two major roots of high biases were identified: (1) the highly reactive VOCs (HRVOC) adjustment that was not scientifically defensible and (2) the insufficient modeling grid resolution with respect to the nature of ozone problems in Houston. Ultimately, the application of new MPE method led me to develop an alternative modeling case with which I showed that the alternative case could be used in a limited way to test a certain type of HRVOC control strategies by reducing VOCs biases.
To my father, Mr. Boo Lim Kim (1942 – 2002)
ACKNOWLEDGEMENTS

I appreciate my advisor, Professor Harvey E. Jeffries, for his advice and support during my journey of doctoral degree. I am grateful to the rest of my doctoral committee members, Professors Lawrence Band, Douglas Crawford-Brown, Donald Fox, and Mort Webster, for their invaluable advices. I wish to thank to Professor Michael Flynn for stimulating me to explore various aspects of environmental flows.

I am greatly indebted to the members of Jeffries research group for their help with my research and dissertation. I enjoyed discussion about science-policy issues with my colleagues. I also express my special gratefulness to my friend, Sangdon Lee, and his family for making the life of my family in Chapel Hill filled with fun.

I thank to my mother and my sister for their constant supports and for their priceless efforts of having taken care of my father when I could not do anything but praying across the Pacific Ocean. I am grateful to my wife, Eun Young Kim, for her steady love and dedication. I also wish to thank the sisters of my wife for their supports. I thank my God for allowing me to have two precious kids, Rebecca Hyun-Hee and Paul Hyun-Woo. I also wish to say “thank you” to my daughter and my son for their joy and love.

This study is partially supported by the Houston Advanced Research Center under Contract H-13 and Contract H12.8HRB. Partial support was also made by the 8-hour ozone Coalition Group under Contract HC-1. I appreciate their support.
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
<td></td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Air quality modeling</td>
<td>3</td>
</tr>
<tr>
<td>1.3</td>
<td>Model performance evaluation</td>
<td>7</td>
</tr>
<tr>
<td>1.4</td>
<td>Goal and objectives</td>
<td>16</td>
</tr>
<tr>
<td>2. DEVELOPMENT OF AN IMPROVED PROTOCOL FOR EVALUATING THE PERFORMANCE OF REGULATORY PHOTOCHEMICAL AIR QUALITY MODELS</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>22</td>
</tr>
<tr>
<td>2.2</td>
<td>Review of SIP modeling and MPE practice</td>
<td>26</td>
</tr>
<tr>
<td>2.2.1</td>
<td>SIP modeling</td>
<td>26</td>
</tr>
<tr>
<td>2.2.2</td>
<td>MPE practice</td>
<td>29</td>
</tr>
<tr>
<td>2.2.3</td>
<td>The Courts’ view on the role of MPE</td>
<td>30</td>
</tr>
<tr>
<td>2.2.4</td>
<td>Review of selected studies on improving MPE</td>
<td>31</td>
</tr>
<tr>
<td>2.3</td>
<td>Development of PROMPT</td>
<td>33</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Rethinking MPE for SIP modeling</td>
<td>33</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Design and scope</td>
<td>40</td>
</tr>
<tr>
<td>2.4</td>
<td>PROMPT Implementation</td>
<td>47</td>
</tr>
<tr>
<td>2.4.1</td>
<td>PROMPT Implementation - Phase 1 (PI-P1)</td>
<td>47</td>
</tr>
<tr>
<td>2.4.2</td>
<td>PROMPT Implementation - Phase 2 (PI-P2)</td>
<td>52</td>
</tr>
</tbody>
</table>
2.4.3 PROMPT Implementation - Phase 3 (PI-P3) .................................................. 57
2.4.4 PROMPT Implementation - Phase 4 (PI-P4) .................................................. 65

2.5 Discussion and conclusion ............................................................................... 70

3. PYTHON-BASED PERFORMANCE ANALYSIS SUPPORT SYSTEM:
SOFTWARE FOR PERFORMANCE ANALYSIS OF REGULATORY
PHOTOCHEMICAL AIR QUALITY MODELING ............................................. 73

3.1 Introduction ..................................................................................................... 75

3.2 Background of pyPASS development ............................................................ 78

3.2.1 Improved analyses of graphical performance measures ......................... 78
3.2.2 Object-oriented production of graphical measures .................................. 80
3.2.3 Difficulties in using existing tools ............................................................... 81
3.2.4 Design goals and choices for implementation ........................................... 83

3.3 Overview of pyPASS .................................................................................... 86

3.3.1 Supporting libraries .................................................................................. 86
3.3.2 pyPASS structure .................................................................................... 90
3.3.3 Inputs to pyPASS ..................................................................................... 93
3.3.4 Operational procedures of pyPASS application .................................... 97
3.3.5 Outputs from pyPASS ............................................................................. 100
3.3.6 Resource demands of pyPASS ................................................................. 103

3.4 Illustrative examples ..................................................................................... 105

3.4.1 Using pyPASS for site-by-site and day-by-day analysis ....................... 107
3.4.2 Comprehensive tile plots ........................................................................ 114
3.4.3 Integration of pyPASS with Geographical Information Systems ........ 118
3.4.4 Comparison of model predictions with aircraft measurements ........... 120

3.5 Summary and future improvement .................................................................. 124

3.5.1 Summary .................................................................................................. 124
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5.2</td>
<td>Future improvement</td>
<td>125</td>
</tr>
<tr>
<td>4.</td>
<td>ALTERNATIVE PERFORMANCE EVALUATION OF THE HOUSTON-GALVESTON MID-COURSE REVIEW PHASE 2 MODELING</td>
<td>128</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>131</td>
</tr>
<tr>
<td>4.2</td>
<td>Performance evaluation methodology</td>
<td>135</td>
</tr>
<tr>
<td>4.3</td>
<td>Implementation and results</td>
<td>139</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Evaluation Phase One (P1)</td>
<td>139</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Evaluation Phase Two (P2)</td>
<td>158</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Evaluation Phase Three (P3)</td>
<td>167</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Evaluation Phase Four (P4)</td>
<td>183</td>
</tr>
<tr>
<td>4.4</td>
<td>Conclusions</td>
<td>186</td>
</tr>
<tr>
<td>5.</td>
<td>SUMMARY AND RECOMMENDATIONS FOR FUTURE STUDIES</td>
<td>190</td>
</tr>
<tr>
<td>5.1</td>
<td>Summary of this study</td>
<td>190</td>
</tr>
<tr>
<td>5.2</td>
<td>Recommendations for future studies</td>
<td>192</td>
</tr>
</tbody>
</table>

APPENDIX A. CONCEPTUAL MAP OF THE PROTOCOL FOR REGULATORY OZONE PERFORMANCE TESTS 195

APPENDIX B. MODEL PERFORMANCE EVALUATION PROTOCOL FOR HOUSTON-GALVESTON MID-COURSE REVIEW PHASE 2 MODELING 196

REFERENCES 211
**LIST OF TABLES**

Table 1.1. Most frequently used measures to determine a PAQM performance evaluation in a SIP application. .................................................................................................................... 11

Table 2.1. Summary of PROMPT procedural questions. ................................................................. 46

Table 3.1. List of major site-packages used in pyPASS. Note that each library listed in this table may require other libraries. For example, PyTables requires the HDF5 library. ........ 87

Table 3.2. List of pyPASS codes and utility programs with the summary of their functionalities.......................................................................................................................... 92

Table 3.3. Important pyPASS inputs. These files are necessary to operate pyPASS in the full functional mode and are prepared independently from pyPASS modules....................... 94

Table 4.1. Summary of the Protocol for Re-Evaluating HGMCR modeling......................... 137

Table 4.2. Summary of observational database used in this study. Both WILT and LAPT were non-AIRS research monitors and are included to enhance MPE. The geographical information about each site, such as neighborhood monitors, can be found in Figure 4.2... 147

Table 4.3. Summary results of comparison of Base5b.psito2n2 with the conceptual model (P1.3) on 2000-08-25........................................................................................................... 156

Table 4.4. Selected scientific factors considered in re-evaluating Houston-Galveston Mid-Course Review modeling.................................................................................................. 187
LIST OF FIGURES

Figure 1.1. Conceptual representation of environmental processes in a typical air quality model: (a) presents the process dynamics in a virtual cell represented as a continuous stirred tank reactor, (b) displays the interaction of each cell with its neighbor cells, (c) describes the time marching scheme used in solving PDEs (Jeffries, 1995a). ................................................................................................................................ 5

Figure 1.2. Emission inventories needed for State Implementation Plan modeling. .......... 9

Figure 2.1. A conceptual map of some ozone characteristics that makes MPE difficult. ........................................................................................................................................... 25

Figure 2.2. A conceptual map of SIP modeling and the role of MPE. Shadowed boxes with bold fonts represent concepts that PROMPT includes as primary materials; Traditional MPE approaches were weak in incorporating these. The number in ovals indicates two major modeling steps in SIP modeling: (1) base case modeling and (2) future (and future control) case modeling. .............................................................................. 27

Figure 2.3. An example bar chart showing daily peak ozone. Data used for this chart is from Houston-Galveston Mid-Course Review 1993 modeling case and all monitors are sorted by its location from west to east with observed ozone and model predicted ozone. As shown in the figure, there are large spatial discrepancies in peak ozone in model and real world at monitor sites. Even with these differences, this specific modeling case could pass EPA’s three statistical tests. .................................................................................. 51

Figure 2.4. Illustrative examples of wind scatter plots (top), wind time series plot (bottom left), and wind error time series plot (bottom right) at a monitor site. In these plots, times of a day are encoded with different markers and colors. Model prediction and observation are in different colors. This specific case shows gross (> 60 degree) wind direction differences between modeled winds and observed winds from 1300 to 1700. A series of questions should be asked and answered to investigate if these discrepancies will affect control strategy developments. .............................................................56

Figure 2.5. Hypothetical sources (the grey area and the box ‘C’) and receptor (‘R’) in a PAQM. Suppose two meteorological inputs are provided, S winds (MET-A) and E winds (MET-B). If emission intensity from the grey area is homogeneous and other conditions are identical, except the wind directions as shown above, (a) the policy question about the effectiveness of control of emissions from the grey area on the receptor ‘R’ can not be different by MET-A and MET-B even if MET-A is identical to observation. (b) However, the same type of question for the effectiveness of control of emissions from ‘C’ will be answered very different by two meteorological cases. In fact, MET-B will lead policy makers to wrong control by show effects of control of source ‘C’ at the receptor ‘R’ if there is a coincidental agreement between observation and model prediction with MET-B. ............................................................................. 66
Figure 2.6. Example outcome of PI-P4. Areas are color-coded by an evaluator’s confidence on a model performance. The red circles represent the peak ozone observed at monitor sites and values in legends for ozone are in ppb.

Figure 3.1. Overview of pyPASS operation. Box 1 depicts the general pyPASS operation necessary once for an episode. Box 2 represents the general pyPASS operation needed for each simulation. Once a simulation is performed for an episode, users may run multiple pyPASS operations with visualization modules and package materials for subsequent analyses and communication with documentation modules.

Figure 3.2. An example of spatially paired and temporally unpaired daily peak ozone bar chart. The predicted ozone data were made by TCEQ’s HGMCR modeling with CAMx (red) and University of Houston’s modeling with CMAQ (blue) for 2000-08-25. Monitoring sites are sorted by location from west to east of HGMCR modeling domain. The x-axis label denotes the four-letter site codes and the y-axis shows ozone concentration along with one-hour ozone NAAQS (depicted as the purple line) with the label ‘Exceedance’.

Figure 3.3. Example of a wind speed scatter plot at a single site. X-axis is the observed wind speeds and Y-axis is the predicted wind speeds in units that reflect the size of the model grids (e.g. 4 km on each side). The magenta spline curve is to help modelers track each data point sequentially. Daylight hours are numbered in LST. Diagonal lines represent 2:1 (dotted), 1:1 (solid), and 1:2 (dotted), correspondence between predictions and observations. Four different colors are used to distinguish four time periods of a day: 0000-0600 (“Midnight to Morning”), 0700-1200 (“Morning to Noon”), 1300-1800 (“Noon to Evening”), and 1900-2400 (“Evening to Midnight”). Left facing and right facing triangles represent pairs of the observed and modeled vector resultant wind speeds in the morning and afternoon.

Figure 3.4. An example of graphical measures for surface wind analyses: (a) hodograms (i.e. wind time series plots) and (b) wind error time series plots. For hodograms, the origin is the station locations, the radial axis shows wind speed in km/h, and the angular axis shows wind direction in degrees with 0 degrees as winds from north. Each filled circle is the start point of a wind vector with its arrow head at the origin where the station is located (the vectors are not drawn, but the “tails” are connected by a colored line). Magenta is for observations and cyan is for predictions. Triangles are for resultant winds; left-facing one is for morning and right-facing one is for afternoon. Some hours of day data points are labeled with its hour of day such as “13”. For wind error plots, the radial axis is the ratio of modeled wind speeds to observed wind speeds on a log scale. The angular axis is the degree of wind directional differences between predictions and observations. The green circle is a ratio of 1.0 or when observed wind speeds are same as predicted wind speeds. In an ideal case, all points are located at the cross point of 0 degree line and the green circle.

Figure 3.5. An example of a time series plot for chemical concentrations. Solid lines are for observations. Dashed line represents prediction. Three chemicals are plotted.
with different colors. This plot clarifies when the large underestimation of NO that
can be found in scatter plots occurred during the early morning. It is clear that there
was also a large ozone underestimation for 1200-1600.................................................. 113

Figure 3.6. Example of pyPASS tile plot. This type of tile plots is unique pyPASS
outputs that are not available by existing tools. Bottom X-axis and left Y-axis are for
x and y coordinates in projected coordinate system. Top X-axis and right Y-axis are
for cell coordinates. These axes are optional and users can produce a tile plot without
these auxiliary axes, if necessary. All monitors are represented as diamonds filled
with different colors depending on the observed ozone concentrations and arrows that
has a length equal to the distance of wind traveled for an hour. If there is no observed
wind and no chemical measurement, diamonds are replaced with two different sizes
of circles superimposed. If there is no observed wind at a monitor, an arrow is
replaced with a circle. If there is no measured chemical concentration at a monitor,
the diamond is replaced with a circle. The tile plot also holds important geographical
features such as highways and coastal lines. Predicted winds are drawn with light
grey arrows...................................................................................................................... 115

Figure 3.7. Series of tile plots for O3 with predicted surface winds as well as
observed winds and ozone. All time in this plot is hours of 2000-08-25 in LST. ........ 117

Figure 3.8. An example pyPASS application to GIS maps. This example is the result
of overlaying Figure 3.4 (a) on a GIS map containing important emission sources
(normal triangles) and monitors (squares with four-letter site codes). Other
geographical features included in this figure are area sources such as airports (gray
filled polygons), major roads (light brown and yellow), water bodies (sky blue)......... 119

Figure 3.9. Example of comparison of aircraft observation with model predictions.
The top plot is the pyPASS output for comparing aircraft measurements with model
predictions. The bottom plot was made from a GIS map and a part of Flying Data
Grabber outputs (McNally, 2005). The black and red arrows in the bottom plot
indicate observed winds and predicted winds used for modeling at sampling locations
of aircraft flight path. The purple box is the time window corresponding to the purple
circle in the bottom map. ................................................................................................ 123

Figure 4.1. Domains for the HGMCR modeling. BPA domain was considered in the
early phase of modeling but excluded in the HGMCR modeling. More detailed
information such as map projection parameters can be found on TCEQ’s HGMCR
modeling web site (TCEQ, 2006c). ................................................................. 141

Figure 4.2. GIS map of HG 74 km by 74 km domain. Green squares depict ground or
research monitors and provide four-letter labels. Triangles are major VOC point
sources and they are colored depending on their annual VOC emissions in 2000 (red
> yellow > green > blue > white). Gray filled polygons include major airports and
other major area sources. Cyan areas represent water bodies. The Ship Channel runs
from east of C35C to south of H07H and below highway 146................................. 142
Figure 4.3. Example package needed for daily graphical performance analyses as part of P1.3 in which a model behavior is compared with the conceptual model. Each package consists of four graphical measures: a bar chart of unpaired peak ozone concentrations (first from top), a bar chart of unpaired maximum hourly ozone changes (second), a bar chart of unpaired peak NO (third) and NO2 (last). The X-axis of each bar chart shows monitors sorted from east to west. Model data were depicted in red in bar charts.

Figure 4.4. Example package needed for site-by-site and day-by-day graphical performance analyses as part of P1.3. Each package consists of two graphical measures: surface wind speed scatter plots (left), and time series plots (right). For detailed guidance on how to use each graphical measure, refer to the pyPASS article (Kim and Jeffries, 2006c) and to a partial implementation of PROMPT for this case (Jeffries et al., 2005). In the wind speed scatter plots, yellow right-facing triangle and red left-facing triangle represent morning resultant wind speed (RWS) and afternoon RWS. Model data were depicted with dotted lines in time series plots.

Figure 4.5. Example package needed for graphical performance analyses required by P2.2. Each package consists of a set of graphical measures: wind speed scatter plots (top), hodogram (bottom left), and wind error plots (bottom right). This example represents one of the worst cases. For example, observed wind speeds were invariant and close to 4 km/h from 1100 to 1300 while predicted wind speeds changed from 4 km/h to 12 km/h. The afternoon wind speeds were mostly twice as fast as observed wind speeds.

Figure 4.6. Example package needed for graphical performance analyses required by P2.3. Each package consists of a set of graphical measures: a time series (top left) and scatter plots for NO, NO2, and O3.

Figure 4.7. Example graphical measures used for P2.4: CO at WILT (top), ETH and OLE at C35C (middle and bottom).

Figure 4.8. Comparison of Psto2n2 and RegEvnt1 for ETH (top) as well as O3, NO, and NO2 (bottom) at C35C. The fine dotted lines are for Psto2n2. The dashed lines are for Regevent1. Observational data are depicted with solid lines with filled squares.

Figure 4.9. Effects of horizontal grid resolution on event simulation. X-axis is for hours of August 30, 2000 and Y-axis shows domain-wide peak ozone concentration for each hour of day. The amount of event was 1,450 lb ETH and 10,188 lb OLE for an hour and the event duration was two hours as denoted with the purple box. Same amount of event emission was added to each cell in 4 km grids and 1 km grids. In other words, 4 km grid cell where HRVOC event was added is the cell contains 1 km grid cell. As shown in this figure, fine grid simulation is necessary to simulate transient high ozone events (THOEs) that are the most unique ozone behavior in the HG area.
Figure 4.10. Example of graphical analyses for P3.2. Wind time series plot for H08H (top left) and for H03H (top right) overlaid on GIS map. The observed winds are in magenta and the predicted winds are in cyan. The O3, NO, and NO2 time series at LAPT are shown at the bottom.  

Figure 4.11. Ozone concentrations of RegEvnt1 from 1400 to 1500 on 2000-08-30 at the 9th layer. The dots flowing diagonally at lower left corner of tile plots are an aircraft track from 1451 to 1453. Black squares are cells of very high concentrations (> 180 ppb) of ozone. Each tile represents 1 km by 1km grid cells. The estimated ozone plume width is approximately 6 km.  

Figure 4.12. Ozone concentration predicted by three models at 1500 on 2000-08-30: Psito2n2 ran at 4 km (left), Psito2n2 ran at 1 km (middle), RegEvnt1 at 1 km (right).  

Figure 4.13. Results of statistical performance tests of Psito2n2 and RegEvnt1 for 2000-08-30. All of statistical measures were estimated for each monitoring site using the equations proposed by EPA (US EPA, 1991).  

Figure 4.14. Suggested reliability of RegEvnt1 on evaluating the efficacy of short-term HRVOC rule based on the results of P4.1 conducted for 2000-08-30. Red circles depict the daily peak ozone concentrations observed at each site. Green, gray, and purple areas represent Reliable, Ambiguous, and Unreliable model performance. The response of RegEvnt1 to HRVOC short-term rule application can be properly assessed at most two sites based on the performance of model on 2000-08-30.
1. INTRODUCTION

1.1 Background

The Clean Air Act (CAA) names “criteria” pollutants that must be regulated due to public health and welfare concerns. The criteria pollutants are carbon monoxide, nitrogen dioxide, ozone, lead, PM2.5 and PM10 (particulate matter < 10 micrometers), and sulfur dioxide. The CAA requires the Environmental Protection Agency (EPA) to set National Ambient Air Quality Standards (NAAQS) for these six principal pollutants. There are two types of NAAQS, primary standards for public health and secondary standards for welfare. If an area in a state violates any of the primary standards, that area is classified as a non-attainment area. Each state having a non-attainment area is required to develop a State Implementation Plan (SIP) to reduce each pollutant that exceeds the standard to levels equal to or below the NAAQS for that area. The primary standard for ozone was a daily maximum one-hour average of 0.12 ppm (parts per million). Those states in violation of this standard in one or more areas are required to develop a SIP for ozone reduction and to perform a model-based attainment demonstration to show that the proposed plan is “more likely than not” effective in producing attainment at a future date.

Photochemical ozone results from the chain reactions of two important precursors, nitrogen oxides (NO\textsubscript{X}) and volatile organic compounds (VOCs), in the troposphere with intense sunlight (Jeffries, 1995a). NO and a small amount of NO\textsubscript{2} are created by virtually all combustion processes because air is used as an oxidant, and high temperature (over \(\sim 2500 \, ^\circ\text{C}\)) can oxidize stable nitrogen molecules in the air. When released to the troposphere,
NO will be oxidized into NO$_2$ quickly by ozone and other oxygenated radicals. This is why NO$_X$, the sum of NO and NO$_2$, is often considered to be the actual precursor to ozone.

VOCs will also undergo a series of oxidation processes that depend on the reactivity of each VOC with OH$^-$ radicals and ozone. Even though ozone is formed during this series of oxidation processes, none of the atoms in the precursors end up in the ozone molecules. Ozone is formed by the reaction of O$_2$ and O($^3$P) that is released from the photolysis of NO$_2$. The NO$_2$ molecules are formed in the process of NO oxidation by ozone or peroxy radicals such as RO$_2^-$ and HO$_2^-$. These peroxy radicals are created in the oxidation process of VOCs by OH$^-$ radicals. OH$^-$ radicals are recycled and new OH$^-$ radicals are created throughout photochemical reaction systems by photolytic processes of oxygenated organic compounds and ozone (Jeffries and Tonnesen, 1994). Ozone formation is complex because of the feedback of recycled radicals and the fact that a major species in the radical propagation chain can also serve as a radical terminator.

The control of photochemical ozone is essentially a matter of reducing one or both of the two major precursors, NO$_X$ and VOCs. However, reductions of a precursor do not always lead to the reduction of ozone. A specific control strategy becomes a complex problem because there are various sources of NO$_X$ and VOCs in non-attainment areas and their emissions vary in space and time. Moreover, controls of some sources are beyond a state’s authority. Two examples are: (1) automobiles on interstate highways are major NO$_X$ sources, but their emissions are under the control of the federal government, and (2) trees are large sources of biogenic hydrocarbons but can not be “regulated” under the CAA. Besides this control authority issue, the specific control strategy development faces two difficult
issues related to ozone chemistry and meteorology: non-linearity of ozone formation and ozone transport.

The non-linearity of ozone formation results from positive and negative feedback characteristics of ozone chemistry. Depending on the environment, a reduction of NO\textsubscript{X} or VOCs may achieve a small reduction of ozone or produce more ozone. Also, reductions of both precursors may not lead to the same proportional reduction of ozone even though it will not cause more ozone formation. This nonlinearity can also result in the same ozone concentration being produced by different precursor conditions. Thus, linear control of precursors will not guarantee the desired ozone reduction. While the non-linearity is likely a local chemistry issue, ozone transport is the phenomenon that ozone and precursors are carried from upwind areas to downwind areas. As a consequence, a local control strategy alone may not be sufficient to meet the NAAQS in some cases and this phenomenon sometimes results in multi-state problems (Farrell and Keating, 2002). Due to these complexities of the ozone control problem, the CAA Amendments of 1990 require any state preparing a SIP for future ozone attainment to demonstrate the effectiveness of the control strategy by using a three dimensional photochemical air quality model (PAQM) or an equivalent analysis tool.

1.2 Air quality modeling

To simulate ozone formation, a PAQM must have adequate representations of several environmental processes. Modern PAQMs represent not only gas-phase chemistry but also horizontal advection and diffusion, vertical advection and diffusion, wet/dry deposition, emissions, and, sometimes interaction with other media such as particulate matters. All of
these processes are translated into a mathematical expression: a set of nonlinear partial
differential equations (PDEs) such as,

\[
\frac{\partial C_i}{\partial t} = -\nabla (\bar{U} C_i) + \nabla (D_T \nabla C_i) + R_i + S_i ;
\]

(1)

where \( C_i \) is the mean concentration of species \( i \), \( \bar{U} \) is the mean wind vector; \( D_T \) is the
turbulent diffusion coefficient; \( R_i \) is the removal rate that includes wet/dry deposition rate and
chemistry loss rate; \( S_i \) is the source term that includes chemistry production rate and emission
rate.

Because this original set of mathematical problems can not be solved analytically,
solutions are usually obtained by numerical methods. As illustrated in Figure 1.1, the most
frequently used numerical methods in PAQMs are formulated on the concept of a collection
of “well-mixed boxes” for spatial integration and the concept of time marching for temporal
integration.
Figure 1.1. Conceptual representation of environmental processes in a typical air quality model: (a) presents the process dynamics in a virtual cell represented as a continuous stirred tank reactor, (b) displays the interaction of each cell with its neighbor cells, (c) describes the time marching scheme used in solving PDEs (Jeffries, 1995a).
In each box in a model, the chemical transformation is estimated. Even a semi-explicit representation of complex chemistry would require the modeling of thousands of reactions involving hundreds of species. The equation for this chemistry is known as a “stiff” problem and leads to high computational costs. Moreover, not all reaction rate constants are known and the chemistry of some species has not been studied sufficiently. Therefore, PAQM developers often design compressed or approximate chemical mechanisms to balance the computational burden against details of the current best atmospheric chemistry knowledge. In general, virtually all important inorganic species are explicitly represented in these compressed chemical mechanisms while most VOCs are classified based on their structure-reactivity relationship or grouped into a few model species (Dodge, 2000).

The most widely used chemical mechanism in PAQMs is the Carbon Bond IV (CB4) mechanism that consists of about a hundred reactions of approximately 40 model species (Dodge, 2000). The limitation of this approach includes model compounds that cannot be directly compared with measured real species. For example, PAR in CB4 represents saturated hydrocarbon bonding (C-C) and many real species contribute to the concentration of PAR in the model.

The operation of a PAQM requires some inputs from other models or observations (Russell and Dennis, 2000). The two most important types of models that provide PAQM inputs are the meteorological models and the emissions models. Meteorological inputs include wind fields, surface temperature, and other parameters. Some of these are also used to generate emission inputs because some emissions depend on meteorological conditions. For example, biogenic emissions depend on ambient air temperature. Emission inputs are also influenced by land use/land cover (LULC). The deposition calculation in a PAQM
requires LULC information and the meteorological model uses LULC information for its surface roughness estimation.

Each auxiliary model that generates PAQM inputs is, by itself, a complex model that is subject to a performance evaluation and quality assurance/quality controls (QA/QC). As a consequence of this operational complexity, interdisciplinary team efforts and comprehensive knowledge are required to exercise the PAQM effectively and to use the modeling results appropriately in the decision making process.

1.3 Model performance evaluation

The judgment of the quality of the modeling results is important. The results of SIP modeling are intended to inform policy decisions regarding the control of high ozone in non-attainment areas. Some of those decisions will be implemented as part of the SIP and may result in the placement of large burdens on social resources and significant influences on daily life, e.g., restriction of construction hours, changes in speed limits, or extra automobile fees. The PAQM performance evaluation is to assess the quality of the modeling results used in decision making processes. The PAQM performance evaluation for SIP modeling is different from the PAQM evaluation for scientific research that examines the correctness of the PAQM formulation and tests whether a PAQM is generally operational (Russell and Dennis, 2000). If the formulation of a PAQM is generally acceptable to the air quality modeling community or it is peer-reviewed by the EPA, the PAQM is classified as an operational PAQM.

In general, the PAQM evaluation for scientific research is not part of the SIP application because it requires well-designed field studies and intensive evaluations. For a SIP development, a state selects an operational PAQM, and then the EPA reviews the selected
PAQM to determine its suitability for use in the SIP development. Upon the EPA’s approval, the state uses the PAQM as its SIP development tool.

For a SIP development, a state completes a series of steps. The state attempts to reproduce high ozone concentrations over non-attainment areas with episodic meteorology and historic emissions using the selected PAQM. This modeling case is known as the “base case”. If the model performance is considered to be good, the state simulates future ozone levels with the episodic meteorology and future estimated emissions. The future estimated emissions are a combination of projected current emissions that include growth and existing Federal and State regulations already required. The future ozone simulation is known as the “future case”. If the predicted ozone level is over the NAAQS, the state will need to propose new control requirements, apply them to the future case emissions, and run the PAQM to test the effectiveness of the control plans. This modeling case is known as the “future control case”. Figure 1.2 shows emission inventories needed for a SIP modeling. Often, the ‘future control case’ is called the ‘future case’ because it is highly unlikely to meet the NAAQS with just the initial future case involving only Federal level controls. Hereafter, ‘future case’ means ‘future control case’ unless specifically noted. The attainment demonstration is to show that the non-attainment areas will “more likely than not” be in attainment with the proposed control plans. During a SIP development, therefore, the PAQM performance evaluation is a vital process because most policymakers are not willing to make decisions with a poorly performing PAQM.
Figure 1.2. Emission inventories needed for State Implementation Plan modeling.
The PAQM performance evaluation for the SIP application focuses on the matching history of the PAQM results in a base case modeling and the utility of the PAQM for the SIP development task in a future case modeling. Since 1991, the EPA has issued three guidance documents to assist states in operating a PAQM, evaluating a PAQM performance, and using the modeling results in their attainment demonstration process (US EPA, 1991; US EPA, 1996; US EPA, 1999).

The 1991 EPA guidance for the model performance evaluation mainly focused on how to determine the pass/fail status of the PAQM performance for the attainment demonstration purpose. Even though the 1991 EPA guidance contains several measures for the model performance, in practice, virtually all SIP applications use the most basic measures; unpaired peak accuracy, bias, and gross error. For details about how to compute these measures, refer to Table 1.1. Consequently, in spite of the 1991 guidance, the attainment demonstration has been very difficult for states to conduct as shown in recent experiences in two of the largest states, Texas and California (Jeffries, 2005). Sometimes, it has even been hard for the EPA modeling development group to conduct the model performance evaluation properly. For example, the operational use of the EPA’s newest model, the Models-3/Community Multi-scale Air Quality modeling system (Models-3/CMAQ), was delayed more than a year even developed over a decade-long effort because emission inventories were inadequate to demonstrate that the model could be used to perform SIP modeling with the 1991 guidance (Jeffries, 2005).
Table 1.1. Most frequently used measures to determine a PAQM performance evaluation in a SIP application.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Expressions</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized bias</td>
<td>[ D = \frac{1}{N} \sum_{i=1}^{N} \frac{C_p(x_i,t) - C_o(x_i,t)}{C_o(x_i,t)}, t = 1,24. ]</td>
<td>±5%~±15%</td>
</tr>
<tr>
<td>Gross Error†</td>
<td>[ E_d = \frac{1}{N} \sum_{i=1}^{N} \frac{</td>
<td>C_p(x_i,t) - C_o(x_i,t)</td>
</tr>
<tr>
<td>Unpaired peak prediction accuracy</td>
<td>[ A_o = \frac{C_p(x,t)<em>{\text{max}} - C_o(x',t')</em>{\text{max}}}{C_o(x',t')_{\text{max}}}, t = 1,24. ]</td>
<td>±15%~±20%‡</td>
</tr>
</tbody>
</table>

Where N is the number of monitoring stations, \( C_0(x_i,t) \) and \( C_p(x_i,t) \) denote observed and predicted concentration of ozone at \( i \)th monitoring station at time \( t \), respectively.

†For hourly observed values of \( O_3 > 60 \text{ ppb} \).
‡Note that this measure does not capture temporal nor spatial distribution. It already assumes that unpaired peak of prediction and observation will be placed near by each other. The Fifth Circuit Court, however, permitted peaks to be matched as much as 55 miles apart because it says “unpaired”.
The 1996 guidance provided two approaches for attainment demonstrations: the Deterministic Approach and the Statistical Approach. The Deterministic Approach was almost the same as the attainment demonstration requirement in the 1991 guidance, except that the standard of passing the test is raised from 120 ppb to 124 ppb and the ‘weight-of-evidence’ (WOE) analysis is recommended if the attainment demonstration by a PAQM is “close enough”. The WOE concept is the tool that EPA offered states to account for the model uncertainties for the attainment demonstration. The 1996 EPA guidance essentially attempted to deal with inherited uncertainties in the modeling results by PAQMs available at that time and problems with estimating future emissions. On the other hand, the Statistical Approach has not been pursued much. This approach requires more ‘burden of proof’ for its ‘weight-of-evidence’ formulation than the Deterministic Approach, which may explain the infrequent use of the Statistical Approach.

The 1999 EPA guidance, only four pages in length, provided a way to estimate additional emission reduction requirements for attainment demonstrations by combining modeling results and observations to overcome some of the regional transport problems without doing actual modeling as part of the ‘weight-of-evidence’ arguments. This guidance was intended to solve a “one time” problem that EPA had in the resolution of a DC Circuit Court Case, but resulted in SIPs submitted from 13 states (Jeffries, 2005).

Even though EPA made efforts to account for the uncertainties of the modeling results by ‘weight-of-evidence’ analysis in the 1996 guidance, no specific guidelines or protocols were given to states about ‘how-to’ do this appropriately. Consequently, current protocols proposed by many states for SIP modeling only contain the most basic measures introduced in the 1991 guidance for the performance evaluation. Also, the 1991 guidance specifies that
the EPA has the authority to approve modeling protocols and attainment demonstrations
proposed by states. The problem is that the EPA has not established criteria explicitly on
evaluating the results of the WOE analyses. Instead, the EPA has accepted the WOE
arguments on a case-by-case basis. Therefore, the current process of accepting the model
performance is too subjective and sometimes even controversial considering the past decision
the EPA made. For example, the EPA approved a SIP submitted by Georgia that used the
linear rollback approach to fill a large gap of ozone control requirement as part of the WOE
analysis (66 FR 63972, 2001). But the linear rollback approach is a method that the EPA
does not recommend as a major ozone control strategy. Another example of the absurd uses
of WOE analysis for attainment demonstration is New York’s SIP. The model used by the
state predicted peak ozone as high as 171 ppb in its future case, then the state conducted
linear ozone reduction tests as part of the WOE analysis to argue that they can decrease the
peak ozone level in the future as low as 118 ppb, which the Court ruled as acceptable (2nd
Circuit, 2003).

While the model performance evaluation always involves subjective judgments, a model
performance evaluation that lacks good objective analyses can mislead policymakers into
making arbitrary and capricious emission reductions that have nothing to do with ozone
control. Moreover, the legal system interprets the completion of the minimum requirements
in the EPA guidance documents (such as the creation of time series without further analysis)
as satisfying the legal requirement even if the model performance is vague or the modeling
results are wrong due to flawed modeling (5th Circuit, 2003).

In the past, the EPA model performance evaluation practice has focused mainly on the
model’s overall ability to reproduce the observed ozone concentration, i.e. ‘matching history’,
by using basic statistical measures and graphical measures. It has been noted that this statistical approach to evaluate overall model performance does not tell much about how a model gets its answer and it forces a user to accept or reject the modeling results as a whole (Beck, 2002). This is the obvious weakness of the current EPA’s model performance evaluation approach because a PAQM may get similar results from different sets of conditions including different inputs and a PAQM’s partial functionality may be sufficient to assist decision makers even though a PAQM does not show good overall performance.

It is important to know whether the matching history is obtained via compensating errors that would preclude the proper use of PAQM results. For example, two different sets of model inputs (assuming one of the sets is correct) could make the PAQM show similar matching history performance, for example, one of the model inputs could be lower mixing heights coupled with less emissions than a second set of inputs but both sets produce predictions are close. However, the PAQM’s responses to the future scenario from these two input sets will be different and the policymaking based on one of the input sets will be in error and may not be effective. The problem is that it is almost impossible to detect compensating errors by just looking at ozone prediction alone because ozone is a secondary pollutant. Consequently, the evaluation of the PAQM performance should examine if the PAQM shows good ‘matching history’ of precursors as well as ozone. Further, the magnitude of the processes in the model that lead to the critical ozone production that dominates the decisions should be investigated, visualized and contrasted to similar processes in other modeling scenarios.

Evaluation of ‘matching history’ of precursors and ozone should also include evaluation of meteorological fields because a correct ozone prediction can only be obtained if the right
amount of precursors moves in and out of a place at the right time. In the current EPA performance guidance documents, meteorological inputs are evaluated with meteorological observations before the PAQM performance evaluation. The meteorological inputs are generated by a meteorological model outside a PAQM. Also, the actual meteorological model outputs are modified with special processors to align grids of the meteorological model to grids of the PAQM. While the meteorological inputs may result in an acceptable overall performance against the meteorological observation, they may not be acceptable for ozone prediction at some places for certain times because of less acceptable performance for mass transport and dilution, which are not evaluated with the typical meteorological observations. Therefore, it is desirable to evaluate the PAQM performance conditionally upon the performance of the meteorological inputs. This is not required in the EPA guidance.

Even if a PAQM produces answers free from compensating errors, it is hard to determine one ‘optimal’ model prediction because some acceptable input sets can result in very similar ozone prediction and we are not sure which input set is more close to reality due to our lack of knowledge and uncertainties; critical aspects of our environmental system are essentially unknowable given current measurement capabilities. Also, given uncertainties in policymaking other than the modeling uncertainties, it is not necessary to have the “most” optimized model prediction for good policymaking unless non-optimal model prediction is very different from the optimal prediction. Rather, it will be more important to examine whether the PAQM results can provide directionally correct information (e.g. \(\text{NO}_x\) control or \(\text{VOCs}\) control), how biased the PAQM results might be, and what the effects of the biases will be on the policymaking. The evaluation of a PAQM as a tool needs to examine whether it will accomplish a specific task (Beck, 2002).
Few research investigations have been done regarding the performance evaluation conditionally upon the quality of the model inputs and how to evaluate a PAQM with its performance to fulfill its designed tasks (Roth, 1999; Fine et al., 2003; Roth et al., 2005). Dramatic increases in computer performance and the widespread use of commercial off-the-shelf components to create a cheap super-computer such as Beowulf clusters have made it easy to operate the PAQM repeatedly. While more people can run the models more cheaply than ever before, no protocol has been developed to guide users in judging the performance of the PAQM in a more comprehensive way. The lack of timely judgment in using and evaluating the inputs and operational choices available in the modeling system, coupled with the failure to integrate the policy components of the problem with the technical modeling components, have resulted in failed modeling efforts (Keating, 1997). Some of these failed modeling efforts have required significant legal actions to remedy them (Texas, 2004), while others have resulted in significant delays in cleaning up the atmosphere and providing a healthier environment (69 FR 8126, 2004; 69 FR 16483, 2004). For more reliable air quality management, we need a model performance evaluation protocol that permits users to examine the PAQM performance conditionally upon the quality of the model inputs and to assess the PAQM utility for the decision making given biases of the PAQM outputs.

1.4 Goal and objectives

The long term goal of my work is to establish a process that can produce “serviceable truth” in environmental management depending on modeling studies. By definition, serviceable truth is “a state of knowledge that satisfies tests of scientific acceptability and supports reasoned decision-making, but also assures those exposed to risk that their interests have not been sacrificed on the altar of an impossible scientific certainty” (Jasanoff, 1990).
The development of process for serviceable truth will require much broader understanding of various fields and interdisciplinary efforts in a holistic way. In this study, I only attempt to resolve issues found in the scientific realm. That is, I focused on how to evaluate environmental models aptly for supporting decision-makers.

In ozone air quality modeling field, a concept of ‘vindicating the use of PAQMs’ (Jeffries, 1995b) emerged a decade ago. This study is a realization of that notion and can be considered as a case-study; the framework and notions developed here can be used for achieving the ultimate goal. My intention was to develop a way of bringing science into policy decision making processes by using PAQMs appropriately so that modelers help policy makers avoid arbitrary and capricious decisions under given modeling uncertainties and resource constraints. MPE is the critical process for using PAQMs appropriately (Roth, 1999; Russell and Dennis, 2000; Roth et al., 2005) and I found there is no good study on this subject at present.

The specific goal of this study was to develop an alternative protocol to the EPA’s current performance evaluation protocol and suitable tools that permits thoughtful modelers to answer the following questions by conducting comprehensive performance analyses systematically: To what extent can I accept the PAQM predictions at face value for a SIP development? And if I cannot, then how should I make judgments about the effectiveness of ozone control options?

These questions cannot be answered by following the EPA’s current protocol without performing many ad hoc diagnostic analyses. Often, these analyses require a lot of time and resources to be completed. In the past, they have been performed in a random, case-by-case
manner. Without good systematic guidance, many of these analyses can be ineffective because some analyses often turn out to be irrelevant to the given problems.

Three objectives are set to achieve the goal of this study:

1. Formulation of an improved MPE protocol that will help modelers answer ozone policy relevant questions effectively,

2. Development of computerized tools suitable for implementing the MPE protocol developed in this study so that modelers can utilize the new MPE method for real SIP modeling and achieve the goal of MPE in a timely manner, and

3. Application of the new MPE protocol and computerized tools to a real SIP modeling case to demonstrate that the new MPE method and tools developed in this study can essentially resolve issues found in the past MPE practice.
2. DEVELOPMENT OF AN IMPROVED PROTOCOL FOR EVALUATING THE PERFORMANCE OF REGULATORY PHOTOCHEMICAL AIR QUALITY MODELS

Abstract

In the application of photochemical air quality models (PAQMs) for State Implementation Plan (SIP) development, appropriate model performance evaluation (MPE) is critical and mandatory. The traditional largely statistical-based MPE protocols must be often used by state modelers, but those protocols have important disadvantages in generating useful information for supporting ozone decision-making. These disadvantages include allowing model users (1) to accept modeling results that may lead to directionally incorrect emission controls or (2) to reject, as a whole, partially useful modeling results for policy decisions. In this paper, we introduce the Protocol for Regulatory Ozone Modeling Performance Tests (PROMPT), a meta-protocol to improve regulatory air quality model performance evaluation. We derived the underlying principles formulating PROMPT from discussions appearing in the recent literature, emphasizing graphical evaluations and the direct assessment of model performance with regard to ozone control policy questions. We developed the structure and details of PROMPT based on these principles and on our practical experience with real-world SIP modeling analyses. PROMPT contains four major sets of procedures that are specifically designed to evaluate the usefulness of models for SIP development and to provide more explicit information aimed at assisting decision makers effectively. Each set of procedures is composed of a statement of analysis goal, the required information for proposed analyses, a list of the core tasks, and the expected outcomes of each task. Also included are the relationships among different procedures and documentation specifications about reporting analysis results. We conclude that PROMPT can serve the regulatory photochemical ozone modeling community better than traditional approaches by supporting state modelers to develop a case-specific protocol with explicit guidelines for
more systematic and comprehensive performance evaluation. PROMPT will result in cleaner policy-relevant scientific answers to the posed policy questions more directly than the traditional approaches.
2.1 Introduction

Ozone is a pure secondary pollutant: it is not emitted directly from sources but is formed in the atmosphere when two major precursors, NO\textsubscript{x} (=\text{NO}+\text{NO}_2) and volatile organic compounds (VOCs), react under conducive ozone formation conditions such as weak winds and intense sunlight (Jeffries, 1995a). The precursors are released from multiple sources including industrial facilities, cars, and natural sources like soils and trees. Precursor emissions are, however, a necessary but not a sufficient condition for ozone formation. Meteorological factors including winds and mixing heights are important because these determine the transport and mixing of precursors, and thus, eventually control the concentration-dependent chemistry that leads to ozone. Certain meteorological conditions can cause ‘ozone transport’ (OTAG, 1997) that results in multi-state problems (Farrell and Keating, 2002) in which local controls can become ineffective. Because any effective intervention requires causal explanations (Pearl, 2000), the ability to explain the causes of ozone problems at a particular locale is a critical key to finding solutions to prevent occurrences of similar ozone problems in the future and requires insight into complex relationships among meteorology, emissions, and chemistry. At the same time, it is noteworthy that an ozone problem in a locale at a specific time results from very specific reasons (e.g. the combination of a particular meteorological condition with spatiotemporally unique emissions). Therefore, it is highly desirable to conduct explanatory studies for more than one ozone episode to obtain representative insights of causes of any frequent ozone problems at a locale.

Because eulerian photochemical air quality models (PAQMs) were considered the most suitable for simulating the effects of multiple sources on ozone with various meteorological
conditions (National Research Council., 1991), the 1990 CAA Amendments (CAA 1990) highly recommended the use of PAQMs as the primary investigation tool for regulatory ozone problems. The CAAA 1990 requires the use of PAQM as a legally-binding apparatus to seek solutions to a given ozone problem for moderate and above non-attainment areas (NAAs) to the old 1-hour ozone National Ambient Air Quality Standard (NAAQS), as well as all of NAAs to the new 8-hour ozone NAAQS (US EPA, 2003). In the use of PAQMs, the Environmental Protection Agency (EPA) also mandates conducting model performance evaluation (MPE) following an EPA approved protocol. MPE is the process of gauging the reliability of PAQMs as tools for testing the effectiveness of possible emission control options (National Research Council., 1991). Some researchers argued that the new 8-hour standard is expected to be more difficult to attain than the old 1-hour standard because violations are likely to extend to rural areas from urban areas (Chameides et al., 1997). Therefore, it is not hard to imagine that the operation, evaluation, and application of PAQMs will be more complicated, difficult, and resource-demanding in meeting the new 8-hour standards.

The correct identification of causes of ozone problems in terms of precursor contributions is especially important in the MPE of PAQMs. As shown in Figure 2.1, however, MPE of PAQMs becomes a very difficult task because the ozone formation mechanism includes positive and negative feedback processes. These two-way feedback processes result in the nonlinearity of ozone formation, e.g. that excessive $\text{NO}_x$ can inhibit ozone formation and some VOC reduction may have no effect. Therefore, when the condition of a locale is $\text{NO}_x$-rich, $\text{NO}_x$ emission controls can result in increased ozone concentrations. If a model over-predicts $\text{NO}_x$ sufficiently to show different ozone response
to NO\textsubscript{x} control from the response of real world ozone, then this type of flawed modeling can result in directionally incorrect control recommendations such as irrelevant NO\textsubscript{x} control when VOC control is necessary. *Failure to conduct a proper MPE can lead technical staffs in state agencies to provide wrong information to the policy makers.* A consequence of misleading policy makers could be serious given that (1) the compliance cost for ozone is over a billion US dollars (US EPA, 1997), (2) more than 100 million people in the United States live in areas of poor ozone air quality as of 2003 (US EPA, 2004), and (3) ozone still remains the most persistent air pollutant in the United States even after more than two decades of control strategy developments and implementations to solve ozone problems (OTA, 1989; National Research Council., 1991; Georgopoulos, 1995; US EPA, 2004).

Improving MPE methods has been one of the most difficult research areas in the photochemical air quality modeling community (National Research Council., 1991; Georgopoulos, 1995; Russell and Dennis, 2000; Fine et al., 2003), especially for MPE methods suitable for a peer-review conducted by a third-party of a regulatory PAQM application (Roth, 1999). The purpose of this paper is to introduce the Protocol for Regulatory Ozone Modeling Performance Tests (PROMPT) which is a meta-protocol that state SIP modelers and third-party model evaluators can utilize as a guideline MPE protocol to develop their own specific MPE protocol. This process will subsume and improve the MPE based on guidelines provided the US EPA.
Figure 2.1. A conceptual map of some ozone characteristics that makes MPE difficult.
2.2 Review of SIP modeling and MPE practice

Figure 2.2 conceptualizes how SIP modeling is initiated and conducted. Also this figure shows who the major players are in the process and what components are involved in the whole process. This figure should be consulted for the rest of this paper as a road map for the SIP modeling process. The shadowed boxes in Figure 2.2 represent concepts that we adopt or enhance in our PROMPT development.

2.2.1 SIP modeling

In this section, we discuss the SIP modeling process (refer to Figure 2.2 as the map of this section). A state with an area in violation of the NAAQS must develop and submit a SIP that includes a future attainment demonstration; this much be done before the statutory deadline. Otherwise, the state might face sanctions or EPA may impose a Federal Implementation Plan. ‘SIP modeling’ is the modeling process that a state undertakes for the attainment demonstration in the SIP. Before conducting ozone SIP modeling, the state selects one or more PAQM(s) and at least one ozone NAAQS violation historic episode. The model and episode selections are subject to EPA’s approval. Once approved, the state conducts a series of three major modeling tasks with each meteorological condition of the ozone episodes.
Figure 2.2. A conceptual map of SIP modeling and the role of MPE. Shadowed boxes with bold fonts represent concepts that PROMPT includes as primary materials; Traditional MPE approaches were weak in incorporating these. The number in ovals indicates two major modeling steps in SIP modeling: (1) base case modeling and (2) future (and future control) case modeling.
The first major modeling task is ‘base case’ modeling (shown in the box marked with an oval of ‘1’ in Figure 2.2) that attempts to replicate an historic ozone episode using adjusted historic emissions and simulated meteorological fields for the time period of the episode. If the base case is acceptable, the second major modeling task is to simulate a ‘future case’ (shown in the box marked with an oval of ‘2’ in Figure 2.2) that predicts the future ozone state with the base case meteorology and with projected emissions based on controls that are already “on the books” such as existing ‘Rate-Of-Progress’ and on federal programs such as mobile source controls. Note that modifying mobile source controls is not available as a control option to the state; i.e. they are prescribed by US EPA. If the ‘future case’ does not show attainment with these mandatory controls, the third major modeling task is to create a ‘future control case’ (also shown in the box marked with an oval ‘2’ in Figure 2.2) that simulates effects of any additional controls needed for the ‘future case’ to show attainment. These controls come from the ‘catalog’ of controls suggested by policymakers. Frequently, a simple future case, without additional controls, does not show attainment (Russell and Dennis, 2000), thus, a ‘future control case’ is often considered as the real ‘future case’.

Hereafter, the term ‘future case’ means the ‘future control case’ unless otherwise be noted.

From this description, we can find two important characteristics of SIP modeling. First, SIP modeling is constrained by a policy timeline and framework. Typical modeling done for scientific purposes is rarely constrained by such external factors. Second, we recognize that all future cases in the SIP modeling can be thought of as merely sensitivity test cases of the base case because the base case meteorology is used for all future cases and the future emissions are projected from the base case emissions. Thus, we see that the quality of future case modeling, which is of most interest to the policy maker, is heavily dependent on the
quality of base case modeling. This point is often misunderstood by policy makers, and even some state modelers (Smith, 2004b). Note that the new 8 hour modeling may introduce a breakage in the consistency between the base case emissions and the future case emissions by adopting ‘base line’ emissions. For details, refer to the EPA’s 8 hour modeling guideline (US EPA, 2005b). At this point, we do not know yet how to resolve this inconsistency in emission estimations with respect to proper model evaluations.

2.2.2 MPE practice

For PAQM evaluation in regulatory applications, EPA has developed a series of modeling guidance documents (US EPA, 1991; US EPA, 1996; US EPA, 1999; US EPA, 2005b). These documents contain the recommended measures for the MPE, the criteria of the MPE, and the criteria for demonstrating attainment. The guidance documents also recommend performing corroborative analyses and graphical tests along with the three necessary ‘statistical tests’: normalized bias, gross error, and unpaired peak prediction accuracy (for a detailed description of how to compute the test statistics, see US EPA, 1991). In addition to these deterministic evaluations, EPA has more recently developed the ‘weight-of-evidence’ (WOE) determination (US EPA, 1996) as a corroborative analysis for judging the possibility of attainment under modeling uncertainties when the future case is close to attainment. For the new 8-hr standards, EPA introduced the ‘relative reduction factor (RRF)’ and some additional statistical measures (US EPA, 2005b) for attainment demonstrations as a supposedly improved way of accounting for potential uncertainties of PAQM results. As the modeling community is still gaining experience with these new guidelines, these latter approaches and measures will not be discussed here.

Lack of detailed criteria for corroborative analyses and graphical analyses, however, results in the current MPE (and the attainment demonstration) being too subjective. There is
no clear guidance on what to consider or how to judge acceptance; the guidance merely states a requirement of making graphs and of conducting general analyses. We agree that MPE is indeed a human intellectual activity that should involve many subjective judgments. We insist, however, that there must be *rational criteria* for these judgments. The description for how to do a WOE determination is not clear and has been questioned in comments submitted to EPA on SIPs proposed for acceptance. It is not surprising that most of the PAQM applications in SIPs only follow EPA’s MPE approach in a very limited way (ENVIRON et al., 2002; TCEQ, 2004). For example, the three statistical tests proposed in the 1991 guidance document were the primary procedures used in many recent SIPs, even though other analyses (including graphical analyses) were also recommended in the guidance. The statistical test criteria for the EPA’s MPE were derived from model performance practice prior to 1990 (Tesche et al., 1990; US EPA, 1991). These criteria are still used, however, to judge the performance of modern PAQMs and interestingly there is little difference between the performance of old models and that of new ones (Russell and Dennis, 2000).

### 2.2.3 The Courts’ view on the role of MPE

The over use of summary statistics may be in part due to the US Appeals Court’s view on the guidance documents and EPA’s recommendations (5th Circuit, 2003). As discussed in the previous section, the statistical performance criteria were provided as guidance or as suggested performance goals. The legal power of these criteria, however, is beyond that of suggestion. Following are illustrative examples showing how the US Appeals Courts view MPE differently from scientists.

The recent US Appeals Court rulings make it legitimate to accept SIPs if states literally followed EPA’s MPE guidelines (5th Circuit, 2003). For example, the 5th Circuit Court considered it is satisfactory to create a time series without stating any analysis results
explicitly because EPA required time series creation but did not enforce explicit assessment of the analysis of time series. Moreover, EPA approved a WOE determination based on the reduction of 53 ppb from the PAQM predicted 171 ppb peak ozone in the future without performing any serious analysis of air quality modeling with PAQMs and this was acceptable to the Appeals Court (2nd Circuit, 2003). Given the fact that ozone formation is highly non-linear and 171 ppb is not in the range of concentration we can find in our ambient air routinely, a 53 ppb reduction from 171 ppb may end up being an unnecessary extra emission control that may not be defensible scientifically. Part of the reason that the Appeals Courts accepted this weak analysis for the major component of the attainment demonstration is: (1) the Appeals Court’s view that ‘the reviewing court must remember that the agency is making predictions at the frontiers of science.’ (2nd Circuit, 2003) and (2) EPA considers a WOE analysis based on the linear rollback approach (e.g. see 66 FR 63972, 2001) no matter how high modeled ozone concentrations are as long as PAQM predicted ozone values are starting points of the linear rollback approach.

2.2.4 Review of selected studies on improving MPE

While there were lawsuits and conflicts in the regulatory arena regarding applications of MPE and the proper use of PAQMs for decision making, most state modeling staffs and some in the air quality modeling research community have merely followed EPA’s guidance as part of research methods without challenging the current MPE practice (Russell and Dennis, 2000). At the same time, others in the scientific community have recognized that the current practice of MPE was not sufficient in part because statistical tests should not be the sole basis for model performance judgment (Willmott, 1984; Tesche et al., 1990).

Even though the need for better MPE has been often expressed, few studies have been conducted. The existing studies have included a formal uncertainty analysis to resolve
performance issues within a probabilistic framework (Hanna and Davis, 2002) and studies on improving MPE with different performance evaluation methods other than the traditional statistical tests (Hogrefe et al., 2001; Sistla et al., 2001; Fuentes et al., 2003; Sampson and Guttorm, 1999). The number of studies is small and the studies still contain significant shortcomings for use in regulatory applications. These will be examined in more detail below.

Some have argued that approaches utilizing a probabilistic framework is consistent with the current EPA’s efforts to incorporate the modeling uncertainties in using PAQM results and to judge the attainment demonstration with RRF and WOE determination (Hanna et al., 2001). The old 1-hour NAAQS and the new 8-hour NAAQS, however, are set as a ‘bright line’; that is, the standard only allows rounding-off error as a quantitative uncertainty tolerance in an attainment demonstration. For example, the NAAQS for 1-hour ozone is 0.12 ppm so that 124 ppb meets NAAQS while 125 ppb does not (US EPA, 1996). In other words, the current SIP modeling framework does not explicitly allow a formal probabilistic evaluation in the attainment demonstration. As we noted in the previous section, it is important to keep in mind that regulatory modeling is constrained by the statutory framework to which it belongs.

Even though approaches attempting different evaluation methods have potential advantages over the traditional practice, such alternative approaches are still not mature and they share some common problems with the traditional approach. For example, some suggested alternative methods used a performance measure such as the coefficient of determination, $R^2$, (Sistla et al., 2001), which is often considered inappropriate for the purpose of performance evaluation due to its insensitivity to additive errors (Legates and
McCabe Jr., 1999). Some proposed approaches are geostatistical methods (Fuentes et al., 2003; Sampson and Guttorp, 1999) including mapping technologies, which are considered impractical for routine evaluation without ample monitored data because of the spatial and temporal scale issues of ozone formation (Diem, 2003).

The newer formal uncertainty studies (Hanna and Davis, 2002), time series decomposition studies (Hogrefe et al., 2001), and new evaluation measure studies (Taylor, 2001; Legates and McCabe Jr., 1999), all still exhibit a problem commonly found in traditional MPE practices: they focus on how to better conduct MPE specifically for ‘ozone performance’ but ignore the fact that the same ozone concentration can result from many different combinations of precursor concentrations. Thus they permit “getting the right answer for the wrong reason”.

This phenomena - getting similarly good answers for different reasons - is called ‘equifinality’ (Beven, 2002), and is one of the general attributes of any environmental model. Nevertheless, the excessive emphasis on final state variable evaluation is not just a problem of regulatory photochemical modeling community. Most of the MPE practice in other application fields also focuses on measuring the matching history of target outcomes with summary statistics such as mean bias, which does not provide insights into model performance, and admits apparently good modeling performance that actually arose due to compensating error or non-linear relationship among products and precursors. This issue may stem from an outdated and narrow view of the concept of MPE.

2.3 Development of PROMPT

2.3.1 Rethinking MPE for SIP modeling

At the abstract level, various environmental modeling communities acknowledge the role of MPE in a similar way (Fox, 1981; Russell and Dennis, 2000; Beck, 2002; McAvaney,
even though the term, ‘evaluation’, is still mixed with other terms such as ‘validation’ (Roache, 1998) or ‘quality assurance’ (Canepa, 2002). The most succinct expression of the expected outcomes of a MPE in an environmental modeling application can be summarized by answering the following three questions (modified from the original questions in Beck, 2002):

- Is the formulation of a model scientifically acceptable in general? (i.e. what is the adequacy and quality of model formulation for this use?)
- Does a model replicate the observations adequately? (i.e. does it make predictions that match history?)
- Is a model usable for answering specific (e.g. policy) questions? (i.e. does the model fulfill the designed task?)

The first MPE outcome question also includes two corollary questions: (1) is the science encoded in the model ‘sound’ (Crawford-Brown, 2005) and working? (2) Is the implementation of scientific knowledge achieved through properly applying modeling procedures of generalization, distortion, and deletion (GDD) to the more complex reality? In SIP modeling cases, the first question can be answered when a specific PAQM is selected for a study. In general, a state cannot ‘just pick’ a PAQM. EPA must issue an approval of a particular PAQM and the approval process requires a state agency to show that a candidate PAQM is as reliable as one of the PAQMs that EPA has used or one of those that has been through an evaluation process by another state. Ironically, this very specific reason was part of the arguments that the California made for its choice of the Comprehensive Air Quality Model with eXtensions (CAMx) (ENVIRON et al., 2002) over the US EPA’s own Community Multiscale Air Quality Modeling System (CMAQ). Sometimes, the evaluation
of model formulation adequacy involves rigorous tests of numerical algorithms, i.e. ‘verification of numerical solvers’ (Roache, 1998). Because most states do not have the resources to perform this type of analysis, they generally accept a model already recognized by EPA as suitably formulated. In other words, the PAQMs used in SIP applications are believed by the EPA to be capable of showing the generally understood behavior of ozone formation in urban and regional episodes. That is, EPA has reasons to believe that an approved model’s predictions can be reasonably accurate if the model’s inputs are correct.

The ‘reasonable accuracy’, of course, should be balanced against practical need to apply the GDD principles. Model formulators generally assume that better science in PAQMs will achieve better predictions. Nevertheless, they always have to generalize, distort, and delete the details of reality for practical reasons such as reducing computational time when they construct PAQMs, model inputs, and model parameters. Therefore, some degree of inaccuracy in model’s predictions is unavoidable due to the application of GDD in model formulation, including decisions on parameter values as well as input preparation. Note that, even though evaluation of a model’s formulation is beyond the scope of regular SIP modeling studies, a good MPE result needs to be able to lead modelers to inspect the possibility of model formulation issues or to select alternative modules available via a model’s runtime option such as selection of specific numerical solvers for their individual case if justified.

The second MPE outcome question above is the primary focus of most of MPE tasks in SIP modeling. This effort is often called ‘operational evaluation’ in the air quality modeling community. The traditional way of conducting operational evaluations is to compare the predictions with the observation of ozone and to evaluate the model performance with
summary statistics as discussed above. Again, it was noted that summary statistics for evaluating overall model performance do not reveal much about how a model got its answers and it forces a user to accept or reject the modeling results as a whole (Beck, 2002).

The air quality modeling community recently recognized that operational evaluations should be extended to cover history matching of the various important precursors (Russell and Dennis, 2000). At the same time, they also recognized that expanding the evaluation range of chemical species is still not enough (Russell and Dennis, 2000). This is because ozone predictions with PAQM requires various inputs that are themselves highly uncertain (Fine et al., 2003). PAQM inputs include many values from outputs generated by other complex modeling systems such as meteorological modeling systems. Therefore, it is important to evaluate the performance of these input generating models, especially in terms of their effects on ozone prediction.

EPA guidance documents for MPE offer little guidance for evaluating these auxiliary model inputs. Modeling each case requires case-specific inputs from these other complex models such as emission modeling systems and meteorological models. Thus, all modeling systems used in SIP modeling need evaluations, but these may not produce meaningful assessment results until the whole modeling process that utilizes them together is done, i.e. the prediction of concentrations of ozone and precursors over space and time. Most commonly, the evaluation of these other models used in SIP modeling have been conducted in a ‘waterfall’ (McConnell, 1996) fashion under the term of ‘quality assurance’ of inputs (US EPA, 1991); once the input files are ‘quality-assured’, no systematic review is performed unless exhaustive ad hoc analyses indicate there may be serious problems in these
inputs. Also, most of the MPE efforts are made for the base case performance evaluation in the current regulatory framework, i.e. evaluation of PAQM results only.

In our work here, we have recognized the need for evaluating PAQM inputs and outputs simultaneously; a proper MPE needs to evaluate model inputs with similar weights as those given to model outputs. The problem is that the performance of models used to produce inputs for PAQM is hard to evaluate in light of their success for ozone predictions prior to actual PAQM runs because a proper evaluation requires using them to produce PAQM results. For example, it has been recognized that a new evaluation method is necessary to assess meteorological model performance for ozone predictions (Seaman, 2000), but how to do an appropriate evaluation has not been addressed.

Often modelers conduct various advanced analyses on a case-by-case basis, i.e. ad hoc manner to resolve problems when performance questions are not answered by operational evaluations commonly practiced in the past. These advanced analyses include Process Analysis (PA), Direct Decoupled Methods (DDM), Ozone Source Apportionment Technology (OSAT), probes such as indicator species, Pseudo-Steady State modeling (PSS), etc. PA inspects the component contributions of modeled processes in terms of their contribution to the predicted ozone (Tonnesen, 1995; Wang, 1997; Lo, 1995; Jang et al., 1995; Jeffries and Tonnesen, 1994). DDM calculates the sensitivity of a region to various conditions in the model as well as regular model predictions simultaneously (Dunker et al., 2002). OSAT crudely estimates the contributions of emission sources to the predicted ozone at a location (Yarwood et al., 1996). The approach using probes (Arnold et al., 2003) and PSS (Kleinman et al., 2002) assess the ozone formation characteristics at a location based on observational data. The presence of these analyses themselves, however, is not a solution to
MPE issues due to many practical reasons. Some analyses require significant resources (e.g. DDM) and some are not implemented in all modern PAQMs (e.g. OSAT). Some are so complex to use (e.g. PA) that only their developers have applied them to SIP applications in meaningful ways. Some methods are currently under debate about their validity and applicability, especially those methods based on ‘indicators’ or ‘observational models’. Therefore, providing good guidance regarding how to identify the necessity of these advanced analyses is also essential to conduct better SIP modeling studies. Note that, except for PA, most of these advanced analyses still emphasize the analyses of chemical signals without considering meteorological signals directly.

Additionally, the current operational evaluation for SIPS typically only utilizes ground monitors even though some states have high-resolution, three-dimensional datasets such as aircraft measurements. We believe this is because there is no guidance by EPA on use of observation for MPE other than ground monitors. In general, to utilize these data properly, we need to consider the quality of observational data such as spatiotemporal resolution and prepare model prediction close to the quality of observational data. It may not be possible to list all possible measurements, however, but we can describe what to consider in general terms as a guidance on how to utilize observation other than routine ground monitoring information.

The third MPE outcome question given above, i.e. fulfillment of the designed task, has been asked only rarely in past SIP studies, and when it has been considered, the answers are often superficial. The lack of absolute accuracy of the model predictions does not necessarily preclude using environmental models for policy development (Morton, 1993; Beck, 2002; Reichert and Borsuk, 2005). Also, an empirical rejection of model performance
becomes harder as the complexity of models grows (Beck, 2002). Moreover, PAQM results are naturally uncertain; PAQM inputs that represent part of the past status of environmental systems are essentially *not knowable*. Environmental systems are all open-systems (Oreskes et al., 1994), such that there is always the presence of ‘unknown’ factors in the system that are not controllable. Further, each environment has its unique ‘landscape’ (Beven, 2002) such as the composition of industrial sources in a specific area. Therefore, answering the third question by allowing some tolerance in using the model’s predictions is probably the most important aspect of environmental modeling. But, while it is true that there is an acceptable error in using environmental models for policy applications, some types of errors are not acceptable regardless of magnitude, especially when they lead to wrong directions in the policy.

For example, erroneous modeling results can result in the recommendation of VOCs reduction when NOx control might be a more necessary. The trouble is that these errors are not likely detected by inspecting outputs only nor are they likely revealed by the normal statistical tests recommended by EPA. Similar ozone concentrations can be estimated with different combinations of input precursor emissions and input meteorological conditions. Consequently, there are cases where predictions match observations well but for different reasons (Russell and Dennis, 2000), especially when errors are compensating each other. Often, this ‘compensating error’ leads to conclusions that require controlling the wrong precursor. These compensating errors are probably the most important issue in photochemical air quality modeling because ozone response to precursor changes is not monotonic for either precursor.
In summarizing our discussion about rethinking the notion of MPE, we conclude that a good MPE for SIPs needs to appraise a model’s ability to match historical observation (including non-routinely monitored data) in accordance with the sufficiency of model’s inputs especially the meteorological inputs and emission inputs. Also the appraisal should consider the allowable tolerance of a model’s performance in light of the posed policy questions to be tested with the model. MPE should be able to identify or at least signal the possibility that modeling results are directionally correct or, equally important, to indicate that the model may not be reliable at even this task. In other words, a MPE for SIP use should not be a series of tasks merely comparing predictions with observations. The same problem of insufficient MPE methodologies exists in virtually all environmental modeling communities and a clear solution has not been found. The research on evaluating model performance in terms of model adequacy within regulatory framework simply has a very short history (Russell and Dennis, 2000; Beck, 2002). Here, we respond to these needs by offering a meta-protocol for developing improved MPE protocol in the area of ozone SIP modeling.

2.3.2 Design and scope

2.3.2.1 Design goals

As we reviewed in the previous section, assuming that the first MPE outcome question (i.e. model formulation) has been resolved satisfactorily, a good MPE for SIP use should provide answers to the following two questions: (1) does the model make predictions that match history?, and (2) does the model fulfill the designed task? The first question seeks to determine (a) if the model can show reasonable agreement with observation for the right reasons and (b) what level of reliability we can achieve with the model. This second question seeks to determine how the level of reliability of the model influences the use of
modeling results in creating SIP policy. Thus, these two questions can be summarized into one question along with complementary questions in parallel: 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 judgments about the effectiveness of ozone control options? The goal of PROMPT is to provide guidelines for constructing a proper MPE protocol for state modelers to follow when they attempt to provide answers about these questions to policy makers for their specific SIP modeling.

These questions cannot be answered by following the EPA’s current protocol without performing many \textit{ad hoc} diagnostic analyses. Often, these analyses require a lot of time and resources to be completed; in the past, these were performed without systematic guidance. Many of these analyses can be ineffective without taking a systematic approach because some analyses often turn out to be irrelevant to the given problems. For example, running DDM is not likely useful if meteorological inputs have serious wind speed or direction errors.

Based on our rethinking of MPE for SIP modeling and to resolve issues found in the EPA’s existing protocol, we designed PROMPT to have four desirable features so that PROMPT can (1) provide a systematic guideline for what to examine in various graphical analyses and how to perform analytical procedures including guidance on when to perform advanced diagnostic analyses, (2) extend its scope beyond the traditional range of observation such as ground monitors by providing guidance on the use of high resolution data sets such as aircraft measurements, (3) incorporate explicit ways of taking into account policy relevant tolerance in the model evaluation framework, and (4) appraise the possible impact of model input biases on choices of ozone control options. The first two features are
designed for improved history matching. The latter two features are designed to test if a model can fulfill its designed tasks.

2.3.2.2 Scope and limitations

The application of PROMPT requires an operational PAQM for at least one episode. This means an operational modeling system including all the proper model inputs for the episode. We recognize, however, that the application of PROMPT may generate needs for reviewing the setup process of the PAQM. PROMPT does not address issues regarding general acceptability of PAQMs; these can be judged better by special evaluation studies specifically designed as part of model development. Thus, PROMPT is focused on addressing how to evaluate the performance of a PAQM used in a specific SIP modeling case.

A generally acceptable PAQM may not work on a specific case due to a limited range of meteorological conditions on which the PAQM was tested and/or other factors that were not resolved while the PAQM was developed. At the same time, it is risky to use a PAQM that is not generally acceptable. Thus, PROMPT presumes that a state selects, upon the EPA’s approval, a generally acceptable PAQM including selection of the run-time options for the specific science modules in the PAQM and has selected a proper episode for its SIP modeling case. Again, we recognize that the findings from the application of PROMPT may lead to some changes to the model configuration (including possible model formulation changes via selecting alternative sets of runtime option or even modifying source codes) or result in selecting alternative episodes.

The outcome of examination may bring up the need for formal uncertainty analysis such as Monte Carlo analysis but a protocol for conducting these formal uncertainty analyses is beyond the scope of this current study. Adopting the formal uncertainty analysis may require
more caution; the current state of these approaches is not considered sufficient enough to deal with epistemic uncertainties systematically consistent with aleatory uncertainties (Ferson et al., 2004). Thus, PROMPT does not include any guidance on formal uncertainties analyses combined with the outcome of PROMPT because the primary focus of this study is to develop a protocol that help the evaluator identify the models’ epistemic uncertainties.

2.3.2.3 Structure of PROMPT

As describe in the previous section, one of the significant issues in EPA’s MPE protocol is that it takes a ‘waterfall’ (McConnell, 1996) approach to evaluation of model inputs with respect to effects of input biases on ozone prediction. That is, there is no systematic feedback to model input evaluation after output evaluation. In addition, the MPE practice following the EPA protocol frequently leads modelers to do trial-and-error changes on inputs until the model performance is satisfactory in terms of statistical measures. This iterative process often requires a lot of resources and becomes fine tuning processes by losing the focus of base case SIP modeling: constructing a model that is sufficient to show the ‘right causes’ of ozone problems in an area. Also, in iterations exercised in past MPE practices, a performance measure is often used only once during the evaluation process. For example, time series plots are made and only a general description of the evaluator’s judgment about the plot is made in very short paragraphs, often only a simple statement such as shows ‘reasonable agreement’.

To overcome these shortcomings, we argue that proper MPE should be conducted in a progressive manner, i.e. a multi-phase evaluation is needed. Each phase needs to give a different degree of information with regard to model’s ability to replicate historical observation correctly and the implication of model biases on policy questions. Throughout
all of the evaluation phases, performance measures need to be used multiple times but with higher degree of concern for details in each consecutive phase. For example, at the first phase, time series may be examined to see if peak ozone in the model at monitor location occurs close to when observed peak ozone did. In later phase, the time series should be inspected to see if there is co-related changes to nitrogen oxides changes or if there is any irregular changes in ozone signals compared to typical urban ozone signals. Therefore, we structured PROMPT to consist of four sets of analytical procedures that will be taken in a sequence. Thus, PROMPT’s design goals are pursued incrementally at each evaluation step. This approach is possible by constructing PROMPT procedures to use the same analysis material multiple times but with different degrees of inspection corresponding to the phase of evaluation while adding more material to the already evaluated material as the evaluation advances further. The advantage of progressive analysis is that it provides (1) quick screening at the beginning of evaluation, (2) chances of finding deficiencies by inspecting same material repeatedly, and (3) fast feedback to evaluators to re-visit previous evaluation phases. Therefore, the model evaluators can conduct more guided analyses in sequential phases. As shown in Table 2.1, we constructed four major questions and subsequent corollary questions to each major question that evaluators have to answer with PROMPT by increasing the level of analyses gradually.

To incorporate the four desirable features that we identified in the previous section (two features for the matching history question and the other two features for fulfillment test), for each set of procedures, PROMPT 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.
Table 2.1. Summary of PROMPT procedural questions.

<table>
<thead>
<tr>
<th>Question</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Does this model show or have all necessary components to produce the</td>
<td>a. What are the model setup and justification? What amounts and kinds of</td>
</tr>
<tr>
<td>phenomena that we can expect from the current best perceptual/conceptual</td>
<td>observation are available for evaluation? How are model inputs prepared</td>
</tr>
<tr>
<td>model?</td>
<td>for model operation?</td>
</tr>
<tr>
<td></td>
<td>b. Is the overall ozone behavior in the model consistent with the</td>
</tr>
<tr>
<td></td>
<td>conceptual model?</td>
</tr>
<tr>
<td></td>
<td>c. If not, what are the possible causes? Is there any alternative</td>
</tr>
<tr>
<td></td>
<td>model inputs or configurations?</td>
</tr>
<tr>
<td>2. Can this model distinguish which precursor(s) to control for ozone</td>
<td>a. Does protocol for graphical measure construction exist?</td>
</tr>
<tr>
<td>reduction?</td>
<td>b. Does model show correct source-receptor relationship?</td>
</tr>
<tr>
<td></td>
<td>c. Does model have biases in surface winds, NOX, and O3 (plus CO if</td>
</tr>
<tr>
<td></td>
<td>available)?</td>
</tr>
<tr>
<td></td>
<td>d. Which precursors are important for potential policy options?</td>
</tr>
<tr>
<td>3. How precisely can the model estimate control requirements?</td>
<td>a. How does model perform at locations where observations are available?</td>
</tr>
<tr>
<td></td>
<td>b. How does model predict at locations where no observation exists?</td>
</tr>
<tr>
<td></td>
<td>c. What are the resolution of control options in space and time?</td>
</tr>
<tr>
<td>4. What are the possible biases in the prediction and the impact of</td>
<td>a. Where does the future ozone problem occur in the model? How does the</td>
</tr>
<tr>
<td>biases on the policy choice?</td>
<td>model perform and/or predict those locations?</td>
</tr>
<tr>
<td></td>
<td>b. Do the biases found in model predictions affect the choices of</td>
</tr>
<tr>
<td></td>
<td>possible control options?</td>
</tr>
<tr>
<td></td>
<td>c. What is the evaluator’s confidence on the reliability of model</td>
</tr>
<tr>
<td></td>
<td>performance in supporting proposed policy options?</td>
</tr>
</tbody>
</table>
2.4 PROMPT Implementation

2.4.1 PROMPT Implementation - Phase 1 (PI-P1)

The goal of PI-P1 is to confirm that a model is potentially capable of describing a specific ozone problem dealt with in a SIP. If a model shows ozone behavior consistent with what a conceptual model describes, we can say that the model is at least usable for the SIP modeling, i.e. testing a model’s descriptive power. The required information for PI-P1 is an operational PAQM and a conceptual model. Note that a model’s precision is not really a main subject at this phase yet.

We recommend two core tasks for evaluating a model’s descriptive power. The first task includes documenting and reviewing details of the modeling system setup including grid structures and available observations including the nature of measurements such as spatiotemporal resolution. Also, it is highly recommended to plan how to use each class of observation can be used in further performance analyses. The second task includes comparing the model’s behavior with the conceptual model based on observations.

Relevant model setup and inputs should be reviewed to ensure they are the best available or to find if there is an alternative. It is not clear how to define the appropriateness of using the term ‘best’. Like our scientific activities, however, the claim of what is ‘best’ is a result of dialogical consensus, not necessarily a unanimous conclusion (Crawford-Brown, 2005). We do not propose that every choice of the model setup should be right at this point because it is impossible to examine whether it is ‘right’ with limited information, which is almost always true to SIP modeling. Rather, we mandate to record what the rationales are for picking a specific set of model configurations so that we can check their appropriateness whenever we suspect part of model configuration causes performance problems. For inputs, we consider they are best if there are no alternative inputs or that will not likely be any
additional ones given resources. Again, the term ‘best’ is much more a practical term and is not about being accurate or correct scientifically.

One of the most important things in examination of inputs is (1) to check how raw outputs of meteorological models and emission models were converted and manipulated for simulations, if any, and (2) to prescribe and justify how to process those data in consistent resolution to observational data sets made by other than ground monitors. Any adjustments such as interpolating wind fields should be documented with justification, tools, and instruction of how to do adjustments to ensure reproducibility. Also, the location information should be compiled precisely, especially with respects to monitor locations, along with detailed information on what geographical coordinate systems were used. Moreover, the resolution of observational data needs to be considered for the purpose of proper prediction-observation comparison to achieve the data quality consistency in MPE. It seems to be trivial at a glance but it is very critical when we do comparison of observation with prediction and try to figure out any root of discrepancies in the context of geographical relations among receptors and sources. We may conclude a model passes this part of the test if anyone who has a reasonable level of technical skills to operate PAQMs can reproduce the model simulations under the evaluation process with the provided documentation and inputs including raw meteorological models.

A conceptual model is a description of our understanding of the ozone problem inferred from all available observations and current theories of ozone formation and transport. Unfortunately, even though EPA guidance requires a conceptual model for a SIP modeling, the regulatory air quality modeling community does not have a formal framework for building a conceptual model yet. Thus, we simply adopt results of intellectual interactions of
meteorologists, inventory staffs, atmospheric scientists, field scientists/engineers, and stakeholder groups, if possible. Note that modelers are not necessarily in charge of developing the perceptual/conceptual model. If there is no conceptual model, it will be almost impossible to proceed further even though eventually the conceptual model can be revised depending on evaluation results and other new findings during a SIP modeling. Therefore, we simply compel modelers to acquire the conceptual model information or to request it from the groups in charge. This is a very important task that is often overlooked, under-funded, and performed late in most SIP works.

The main outcomes of PI-P1 are primarily identifying characteristics of ozone signals produced by a model and comparing them with the conceptual model’s description. Description of characteristics should have temporal and spatial components such as when an ozone peak occurs at a monitor and where the monitor is located at in the study domain. The rate of ozone concentration change and dominant wind pattern are also desirable components of ozone characteristics description. Also, evaluators need to obtain information with regard to precision and accuracy of measurements because measurements are the ultimate material used throughout the MPE process. Unless specific information is provided, evaluators may use federal guidelines for referencing accuracies and precisions of measurements such as ‘Meteorological Monitoring Guidance for Regulatory Modeling Applications’ (US EPA, 2000) for ground monitors.

Even if a model may be able to pass some tests, especially statistical tests, any apparent anomalies should be marked for further examination. Evaluators may ask modelers (1) to identify and explain any illustrated weakness, and (2) to carry this list forward to be addressed in subsequent evaluation phases. If necessary, evaluators need to request
alternative episodes, improvement of meteorology, etc. Visualizing peak ozone for each day after sorting the monitoring sites by direction such as east to west is a simple graphical test that may help evaluators detect obvious geographical ozone biases such as shown in Figure 2.3. Depending on the timeline of the SIP submission and available resources and unless it is very clear that the model is not usable, evaluators may need to proceed to the next phase of MPE while keeping in mind the apparent weakness of the model.
Figure 2.3. An example bar chart showing daily peak ozone. Data used for this chart is from Houston-Galveston Mid-Course Review 1993 modeling case and all monitors are sorted by its location from west to east with observed ozone and model predicted ozone. As shown in the figure, there are large spatial discrepancies in peak ozone in model and real world at monitor sites. Even with these differences, this specific modeling case could pass EPA’s three statistical tests.
2.4.2 PROMPT Implementation - Phase 2 (PI-P2)

The goal of PI-P2 is to clarify the possible issues in model’s ability to estimate ozone response to precursor control. The emphasis is on the correctness of the source-receptor relationship and on how ozone forms, i.e. the precursor signals in the model predictions. In other words, we are interested in whether a model has emissions that are consistent with observations, and whether the modeling system delivers emitted precursors to where real world winds do. Time series plots and scatter plots should be examined briefly at this stage and more in-depth in the next phase of PROMPT. In generating these plots, special attention should be paid to attributes of the plots such as scales of magnitude (e.g. concentration). One important mistake that often makes these plots ‘dequantifying’ is using improper scales, labels, colors, and more (Tufte, 1997). One good example of dequantifications is over use of contour plots using too many colors. In these plots, users often interpolate spatial data to very fine scales that do not correspond to model cells. Even though these plots seem to be esthetic, they distort the information necessary for high precision analysis. In other words, it may help to get a general or overall impression of model behavior, but it does not help show how spatial gradients are modeled using different grid configurations.

In general, it is good to be aware of ‘overly smoothing’ model outputs to appear to have more spatial and temporal resolution than they actually have. By choosing proper scales and other plot attributes, these graphical measures can provide more information even in quantitative ways. Unfortunately, there are no commonly accepted methods for setting these plot attributes. Instead, we recommend that these basic settings should be settled upon in the early phase of MPE (or preferably before SIP modeling) among all related groups involved in the SIP modeling processes. In this way, we can at least have an internally consistent graphical language to exchange information effectively.
The minimal set of variables for these plots is a collection of plots for surface winds, ozone, nitrogen oxides (plus CO if available), and VOCs (continual measurement data such as auto-GC preferred). When analyzing time series plots, we recommend evaluators divide time series for a day in three temporal sections such as midnight-sunrise, sunrise-noon (or peak ozone hour), noon (or peak ozone hour)-sunset, sunset-midnight. At PI-P2, evaluators need to acquire continual VOCs measurements such as those at Photochemical Assessment Monitoring Stations (PAMS) sites and to convert them properly to compare them with model predictions (i.e. real species converted to model species). If there are no PAMS or any continuous VOCs measurement is available, we recommend evaluators acquire at least canister VOC measurements. One can proceed without continual VOC data because (1) wind errors can indicate whether the chemical signal comparison can be meaningful or not, (2) assessment of NOx bias will narrow the possible issues of VOC bias, and (3) canister VOCs data are useful for comparison of emission inventory with ambient composition in terms of modeled VOCs species. Depending on the results of analyses suggested below, evaluators can roughly classify the reliability of model in developing control strategies into four classes: “None”, “NOx only”, “VOCs only”, or “Both NOx and VOCs”.

In the case of large biases of wind speed and/or direction near important sources or persistent error during retention time (i.e. the time that average winds take to cross a modeling domain), ‘none’ will be the likely answer. However, answers are not solely based on scientific assessment because it is not possible. There are no rigid criteria that we can set for our acceptance level of a model since each situation requires different tolerance. For example, a wind direction error of some degree can be acceptable if the error occurs near large area sources of low intensity while the same magnitude of error may not be acceptable
when the same wind direction error happens at very high intensity point sources. Moreover, the acceptance depends on whether that error can influence policy choices. Therefore, we recommend evaluators consult policy analysts or whoever is in charge of developing policy questions for a SIP to get proper information for evaluators’ acceptance claim.

The quality of wind inputs is especially important at this stage because wind biases may lead to the conclusion of a model’s inability to distinguish precursors for ozone control depending on the apparent source-receptor relationship. Biases of chemical concentrations with good wind fields indicate possible biases in emission inputs, but biases of winds likely make emission biases ambiguous. Due to wind biases and error, evaluators may need to request reanalysis of meteorological model results or model episode selection. For testing wind inputs, we recommend inspection of hodograms for observed and modeled winds and hodograms for wind differences and wind speed scatter plots as shown in Figure 2.4. The wind errors shown in the example figure for daytime is over 60 degrees, which may prevent one from performing the proper comparison of model outputs and observations unless the surrounding emissions were very homogeneous. The important attributes of these plots are what the sizes of biases are during each hour window. Large differences in wind speeds during morning hours when ozone is being formed or around peak ozone hours will be very important sources of error especially when the predicted ozone is due to near by precursor sources. If we are dealing with transported ozone problems, wind errors across the modeling domain are important; persistent but tolerable local wind error may end up being unacceptable because the modeled airmass may undergo a different chemical environment in the model than in real world. Thus, wind errors can influence cause-and-effect and change the control response.
For chemical signals, we propose evaluators focus on the overall features such as when the ozone rise occurs during the day, what the shape of the ozone time series looks like at each monitor on each day, and the associated NO and NO₂ time series, e.g. does the model over-predict NO and NO₂ or VOCs. Other important aspects of time series are whether a model shows NOₓ inhibition, NOₓ changes during traffic hours if a monitor is near major line sources, and so on. For spatial scale, the past MPE often includes 8 cell values surrounding a monitor to show possible small misalignment of ozone plume. This is based on older practice when there were few monitors in the domain. One problem is that now the size of a model cell varies from 36 km to 1 km or even smaller in a single episode. Thus, we recommend evaluators use the value of hourly wind speed for an hour as a radius for including cells that might be near by monitors.

In PI-P2, the most difficult class among the four categories of model’s reliability can be “VOCs only” because it indicates there are some issues in NOₓ predictions. Depending on the significance of wind errors and the reactivity of VOCs with hydroxyl radicals, however, we can utilize relative compositions of some VOC species in ambient measurements to compare with compositions of those VOC species in the model outputs and in emission inputs. Since we know that the model’s NOₓ prediction has issues in this case, model evaluators may want to see some VOC species that have low reactivity with OH⁻ and are not sensitive to NOₓ in the model. If there are no VOC measurements available, the model’s performance for VOCs predictions will significantly depend on other factors such as the quality of VOCs emission inventory, and will remain an open question at some level.
Figure 2.4. Illustrative examples of wind scatter plots (top), wind time series plot (bottom left), and wind error time series plot (bottom right) at a monitor site. In these plots, times of a day are encoded with different markers and colors. Model prediction and observation are in different colors. This specific case shows gross (> 60 degree) wind direction differences between modeled winds and observed winds from 1300 to 1700. A series of questions should be asked and answered to investigate if these discrepancies will affect control strategy developments.
2.4.3 PROMPT Implementation - Phase 3 (PI-P3)

The goal of PI-P3 is to test if a model estimates the necessary precision for the control requirement estimation, depending on the precision demanded by policies. Some control options may require less precise answers than others. There are cases where the imprecision can mislead policy makers, especially when the imprecision can result in an ambiguous assessment of ozone response to emission changes. Consequently, the issue of ‘precision’ becomes a context-dependent question that cannot be answered by modelers alone. At the same time, modelers cannot get the necessary information without communicating with policy makers. Therefore, modelers should consult with policy makers concerning the ‘precision’ question. For example, our expectation of the model’s desired accuracy should be different when policy makers want to see the effectiveness of two different controls: a specific point source control versus domain wide control. The specific point source control needs much more local precision and accuracy than the domain wide control, which can relieve high accuracy requirements at local scale. Thus, PROMPT requests modelers to have the following information for evaluating the anticipated precision: 1) the description of proposed control options, 2) the range of precursors that the selected modeling system can distinguish for control option tests, and 3) the potential direction of biases in precursor predictions.

In traditional approaches, different types of biases are often treated similarly. Model biases are a set of one-way errors. For example, predicted ozone at a site for an hour cannot be overestimated and underestimated simultaneously, yet the ozone field can have very small overall error with positive and negative biases at different places over a modeling domain. At the same time, more than two biased processes can also form compensating errors if their
magnitudes are similar while their contribution to ozone concentration change is opposite. Therefore, PROMPT defines two kinds of compensating errors.

Type I compensating errors are those that influence summary statistics. For example, one of the commonly used statistics in the EPA’s existing protocol is ‘normalized bias,’ an average of the sum of the difference of observed ozone and predicted ozone normalized to the observed ozone. This measure can mislead modelers. For example, two large biases with different signs can produce a small normalized bias. Therefore, to reveal Type I compensating errors, the evaluator should report summary statistics with temporal and spatial distribution of the error for ozone and other variables. Potential compensation should be described and documented for use in judging its impact on decisions.

Type II compensating errors are the errors caused by the model’s internal compensation. For example, a PAQM can match history with NOx emissions lower than in reality if the winds are slower or the mixing height is lower than in reality. Type I error analysis may lead to investigation of Type II errors, however, Type II errors cannot be examined only through investigating chemical concentrations because these concentrations are the product of all model processes involved in ozone formation. We strongly recommend applying PA (Tonnesen, 1995; Wang, 1997; Jang et al., 1995) if a Type II compensating error seems to occur because PA is specifically designed and implemented to investigate Type II compensating errors.

For observations, there is no good tool to segregate process contributions to final ozone and developing such tools are beyond the scope of this study. Nevertheless, we recommend two types of analyses: (1) the observation-driven constrained steady state (CSS) box model approach (Kleinman, 2005) and (2) diagnostic indicators such as NO/NO2 ratio (Arnold et
al., 2003), if tools for these approaches are implemented or can be implemented for a SIP modeling given resources. The CSS box model approach is promising because it can provide ozone production rate, \( P(O_3) \), that is directly comparable with the PA’s \( P(O_3) \) output.

The focus of the procedures described in the following two sections is how to reveal the types of compensating errors and to provide guidance on further analyses. Note that the outcome of evaluating the anticipated precision will likely vary spatially and temporally. Application of the following two analyses to each monitor for every day for the selected episode period will provide necessary information to generate spatiotemporal resolved answers.

2.4.3.1 Performance analysis at observed locations

The main goal of this analysis is to assess the ‘matching history’ of the PAQMs. The most basic graphical presentations of history matching are time series plots for all available species and physical variables at each monitor. Matching history of a chemical concentration can be reliable if the matching history of the meteorological inputs is also acceptable. Matching the history of surface winds is a good surrogate for matching the history of the meteorological inputs because wind observations are more frequently made than other meteorological factors such as the ventilation factor or mixing height. Considering surface wind speed and direction, the preferred “time series” plot is a hodogram (or hodograph), a plot showing a set of vectors sharing vector head (or tail) positions. Time series plots and hodograms are supplemented with scatter plots of predicted values verses observed values. These are used to detect the overall bias or extreme values. Other more specialized plots such as quantile-quantile plots are also used as needed.
A similar but more complex graphical analysis would be performed with segments of aircraft measurements following the general approach for monitored sites. However, extra tasks are needed to prepare data for the matching history assessment because the nature of aircraft data is different from routine ground monitor data in spatial/temporal resolution and the number of species observed.

The mere creation of the plots mentioned above is not sufficient. Agreements and disagreements must be identified and explained. At a minimum, the following two sets of questions must be asked conditionally on the quality of ‘history matching’ exhibited:

- If model predictions generally match history, is there any way this might be due to compensating errors among processes such that the apparently good match occurs for the wrong reasons? Are the process rates from the modeling system used in the study consistent with those from modeling systems that are apparently working well?
- If model predictions do not match the history, what are the likely causes of the failure? Are the physical conditions correctly simulated by the model? Is the wind speed and direction approximately correct? Is the volume of the mixed layer approximately correct? Is the vertical mixing process too slow or too fast? Are emission and deposition processes or magnitudes atypical? Are the chemical rates as expected?
- After reviewing all the monitors for each day, the results can be divided into three categories:
  - Category MH-R: Those monitor sites where the model matches history reasonably well and there are no indications of compensating error.
• Category MH-A: Those monitor sites where the history matching is ambiguous and there is little evidence as to the cause.

• Category MH-U: Those monitor sites where the physical conditions simulated by the modeling system preclude a good history match for chemical concentrations, especially for secondary products like ozone.


The monitors within Category MH-R are candidate sites and days for evaluating the future prediction of the model for their response to the various control strategies that are to be considered. Those monitors within Category MH-A should also be examined in the future case condition, but they would be considered as merely supportive and would not be used to actually decide if a particular control strategy would be successful. The monitors within Category MH-U would not be used in the future case evaluation.

Having determined which monitors on each day are reliable for assessing the creation of secondary pollutants, more standard EPA statistical tests can be applied to only these monitors according to the type of attainment demonstration being performed. Note that all analyses above are not just relevant to those monitor sites that were exceeding the standards because a reliable model must also accurately predict non-exceedance monitor-days too. If a model is not able to do this, it is a strong indication that any accurate prediction on exceedance days is the result of compensating errors and these monitor sites even on the exceedance days might not be reliable in future ozone predictions.
2.4.3.2 Analysis of model predictions at non-observed locations

One of the primary reasons for using a complex process-based model is to predict environmental state variables (such as ozone) where no observations are available. The model should be able to predict environmental state variables reliably if (1) it computes the contributions of all the important processes to the state variables (including concentrations of primary and secondary chemical species) over a certain period and (2) it estimates changes of the state variables reflecting the local conditions by summing those contributions. What warrant the reliability of the model’s predictions are the trustworthy methods for computing the contributions of the processes to the changes of the state variables and the availability of accurate inputs. If there is a reason to question the reliability of the computational methods or the accuracy of the model inputs, the model’s predictions where no observations are available become suspicious. Assessment of the model’s reliability where no observations are available is especially important to evaluation of a future control strategy because the success of the future control strategy may hinge on predictions where no observation is available. We note that this is especially true in the new 8 hour ozone test because it relies heavily on a designed value at a monitor, yet the model’s highest ozone and smallest change in ozone in the future case may be at a non-monitored location.

If a PAQM is dealing with a space among a set of monitors where the modelers have already assess their performance using the assessment strategies above, and where it could be concluded that the model predictions are matching historical observations reasonably well, then, modelers would have good reasons to accept that the model’s predictions in the non-observed area among them are reliable. On the other hand, if modelers know that a model will not give correct reduction estimations for the future case at observed locations and can
explain the deviation of a model’s prediction from the history, then modelers would have
good evidence that the model’s response in an area among the poorly predicted monitors is
not useful for testing policy options. In fact, it is important to know about unreliable
predictions because modelers can inform policy makers that the model cannot provide
scientifically defensible evidence with regard to certain proposed control options. On the
other hand, if modelers are dealing with a space distant from monitors with a different
physical environment, more intensive analysis efforts would be needed to conclude how the
model would be reliable both in the base case and in the future case. We believe that
knowledge of unreliable model performance is better than ignorance of the model’s
trustworthiness when it comes to setting policy.

Several procedures, including but not limited to the following, will likely be useful in
assessing the reliability of the model prediction at non-observed locations:

- Visualize the model’s inputs for the important processes in the area of interest. This
  might include plotting the vertical diffusivities over land and water; visualizing the
  low level and high level emission inputs of the area; visualizing the model’s predicted
  wind field; and performing dispersion simulations for selected emissions without
  chemistry to determine how various sources are contributing to the focus area.
- Perform process analysis of the focus region to visualize and understand the
  interaction among the physical and chemical processes and to explain the state of the
  chemical transformations.
- Conduct selected sensitivity analyses by varying important inputs or process
  representations and determine the effects these have on the model. Each sensitivity
analysis may require the performance analysis component and the prediction analysis component, i.e. a recursive application of the model evaluation.

- Review the state of the science and the alternative representations available and assess if the current representation in the model is adequate. It is desirable to acquire auxiliary tools for this procedure such as a modified version of the PAQM, if possible.

After reviewing all such areas for each modeled day, the results can be divided into three categories:

- Category MP-R: Those model locations where the model’s performance is more likely than not adequate and one should accept the predictions as reliable.
- Category MP-A: Those model locations where the model’s performance is ambiguous and there is little evidence as to the cause.
- Category MP-U: Those model locations where the model’s performance is either the physical conditions simulated or chemical conditions simulated by the modeling system are more likely than not resulting in biased results.


After determining which non-monitored areas on which days are likely reliable for assessing the creation of secondary pollutants, those areas may be included in the standard EPA analyses according to the type of attainment demonstration being performed. On the other hand, the assessments arrived above can be major components of a weight-of-evidence argument to explain why the model’s results should not be accepted at face value.
2.4.4 PROMPT Implementation - Phase 4 (PI-P4)

The goal of PI-P4 is to assess the potential effects of model biases found in PI-P2 and PI-P3 on the proposed policy options for ozone control. At this stage, all previous evaluation procedures should be completed. Therefore, what we expect as material for performing PI-P4 are (1) whether the model shows conceptually consistent behavior with observations, (2) whether the model can support various precursor controls in sub-domains, i.e., what types of control can be evaluated for effects given the model’s performance in these areas, and (3) how precisely the model can and must estimate control requirements. We recommend in PI-P4 that evaluators assess how the scientific biases found in previous procedures might potentially bias policy choices.

The difference between scientific bias assessment and science-policy bias assessment can be best illustrated with an example. As shown in Figure 2.5, suppose we have two sets of modeling input in which the only difference is the surface wind direction. If one wind set is perfectly matches observation and the other wind set does not match as well, the matched set is better scientifically than the other set. If the sources we want to control, however, are homogeneous around the receptor, the receptor becomes insensitive to the wind direction differences. Subsequently, the correctness of surface winds is not really important with respect to its value for decision making. In other words, both sets of wind can be used as inputs to PAQMs to test some policy options. This is an important aspect of the PI-P4 concept. PI-P4 should be a process for identifying model performance issues in terms of their importance with regard to science-policy questions. To achieve this aspect of PI-P4, evaluators and modelers need to communicate with policy makers in cataloging proposed control options as we discussed in the description of PI-P3.
Figure 2.5. Hypothetical sources (the grey area and the box ‘C’) and receptor (‘R’) in a PAQM. Suppose two meteorological inputs are provided, S winds (MET-A) and E winds (MET-B). If emission intensity from the grey area is homogeneous and other conditions are identical, except the wind directions as shown above, (a) the policy question about the effectiveness of control of emissions from the grey area on the receptor ‘R’ can not be different by MET-A and MET-B even if MET-A is identical to observation. (b) However, the same type of question for the effectiveness of control of emissions from ‘C’ will be answered very different by two meteorological cases. In fact, MET-B will lead policy makers to wrong control by show effects of control of source ‘C’ at the receptor ‘R’ if there is a coincidental agreement between observation and model prediction with MET-B.
Our recommendation is to focus on those monitors with high confidence (i.e., Category MH-R) and/or with moderate confidence (i.e., Category MH-A) with further diagnostics or the assessment of control option effects. At the same time, model behavior at monitors with low level of confidence (i.e., Category MH-U) should be further examined to see if we can make improvements of model performance at those locations. If we cannot improve model performance, we need to explain why it is so and to assess the effects of our inability to enhance model performance on posed policy questions. Another important step we recommend is to inspect locations where the future ozone concentrations are high. If the future ozone is likely happening at locations where the model show moderate confidence (i.e. Category MP-A) or low confidence (i.e. Category MP-U), the whole SIP modeling process should be reviewed seriously, including the modeling episode selection, because we may be dealing with the ozone problem with a model that can not provide reliable answers to our policy questions.

As one of the most important outcomes from the application of PROMPT, we recommend evaluators create a GIS map showing the assessed reliability of model performance on the proposed policy options for each episodic day. As shown in Figure 2.6, evaluators need to summarize their performance evaluation results on maps with detail commentaries including what form their confidence on the model performance. All information used for their evaluation should be accessible and subject to discussion with the use of well-maintained Hyper Text Markup Language (HTML) documents through World Wide Web (WWW) for stimulating this task effectively. HTML pages are especially useful for on-going SIP modeling and there are ample tools to convert HTML pages into publishable document format such as Portable Document Formats if any formal
documentation is concerned later. Because SIP modeling is constrained by non-scientific conditions, such as SIP submission deadlines, policy makers may have to make policy decisions even with SIP modeling results in which they have low confidence. In this case, model evaluators should prepare recommendations about how to use partial useful modeling results. We recommend that evaluators clarify what kind of policies proposed by policy makers should be more “limited” or “constrained”. This process should be iterative and interactive between policy makers and modelers. Also, we argue (1) a SIP based on limited model performance should in the SIP commit to future studies or research to resolve the uncertainties or issues and (2) the commitment needs to be explicit in the SIP documentation. Regardless of how the SIP modeling turns out, we argue that modelers need to provide a vindication of the model results (Jeffries, 1995b) by asserting and defending that “no one knows how to do this modeling any better that was done at the time with available resources and scientific knowledge.”
Figure 2.6. Example outcome of PI-P4. Areas are color-coded by an evaluator’s confidence on a model performance. The red circles represent the peak ozone observed at monitor sites and values in legends for ozone are in ppb.
2.5 Discussion and conclusion

In this study, we developed a meta-protocol, called PROMPT, that a specific MPE protocol for SIP modeling can be created from. Our design of PROMPT was intended to resolve weakness in traditional MPE protocols that merely followed the EPA’s guidance documents. Reviews of MPE practice and literature along with our practical experience of conducting and evaluating SIP modeling were used to design PROMPT. We found that traditional approaches depend primarily on summary statistics; these have already been criticized (Willmott, 1984; Tesche et al., 1990; Beck, 2002) as ways to conduct MPEs that are scientifically defensible.

We identified two possible roots of weaknesses for MPEs. One is the fact that the EPA’s guidance is not science-philosophically sound (at least in its practical implementation). The other root is the failure of EPA’s guidance to address policy concerns directly in MPE processes by providing scientific information with respect to policy questions in a holistic way. The consequence of weak MPE was illustrated through several court decisions with regard to the acceptability of SIP modeling results. It is also clear that the air quality modeling community as well as the air quality management community has recognized the needs of better MPE methods than the traditional approaches (Russell and Dennis, 2000).

When we developed PROMPT, we focused special attention on addressing science-policy concerns explicitly in our new protocol and to utilize graphical analyses and interpretation that have been insufficiently incorporated in past SIP modeling evaluation works.

PROMPT consists of four major sets of analytical procedures to answer the questions listed in Table 2.1. Each set of procedures contains the goal of analyses, necessary information/material to follow procedures, how each set will be related to the other sets of
procedures, and the expected outcomes of each set. Another emphasis in our PROMPT development is that all evaluation results from those four procedural sets should be open to public for review. Evaluators’ comments and opinions are not exceptional. In fact, this corresponds to the request made by Popper for proper falsification of scientific theories and will make evaluation results much more transparent and persuasive than the traditional “black-box” approaches.

PROMPT is effective in its resource demands and the quality of its outcome. PROMPT helps evaluators exclude possible irrelevant analyses from branches of analysis tree at earlier phases of the MPE. PROMPT is also dialogical in its construction and nature, i.e. PROMPT forces evaluators and modelers to interact with policy analysts and/or policy makers by providing open questions that can only be solved with inputs from policy analysts and/or policy makers. PROMPT is also comprehensive in its scope, systematic in its structure, and practical when exercised with given resources and within the regulatory framework while the fundamental basis of the protocol is consistent with the recent notion of MPE and ‘good’ science. Consequently, evaluators using PROMPT can obtain important policy relevant performance information more quickly than before.

MPE processes for one-hour SIP modeling formed the basis for PROMPT development. Interestingly, the section about MPE in the new 8-hour modeling guidance is virtually same as the MPE chapter of 1-hour modeling guidance (US EPA, 2005b). Also, for the new 8-hour SIP modeling, modelers will likely use the same modeling systems used in the one-hour modeling studies and the new modeling guidance requires modelers to evaluate 8-hour SIP modeling at hourly model performance. PROMPT, therefore, will be applicable to the new eight-hour modeling studies with only minor extensions. Also, while the current PROMPT
does not directly treat the issues related to multiple episodes, we envision that evaluators can repeat multiple analyses following a protocol based on PROMPT since we designed PROMPT as a guiding protocol for a single episode application.

The more challenging MPE question for 8-hour SIP modeling is how to interweave performance analyses and how to interpret analyses results in the context of 8-hour ozone NAAQS because 8-hour SIP modeling does not necessarily require a real base case to test the efficacies of control policy options; that is, the new modeling guidance allows a state to use a meteorology that does not necessarily correspond to the actual base case emission (or more precisely the ‘baseline’ emission) for the attainment demonstration modeling. This challenging question should be investigated to its reliability and effect before states begin serious MPE processes and attainment demonstration modeling for the ozone 8 hour NAAQS.

Acknowledgement

This work was supported in part by the Houston Advanced Research Center under Contracts H-12.8HRB and H-13. Also, partial support was provided by the Houston-based stakeholder Mid-Course Correction Group and by the 8-Hour Coalition Group. We wish to thank Dr. Jim Smith at Texas Commission on Environmental Quality for providing modeling inputs and outputs for the Houston Galveston Mid-Course Review modeling.
3. PYTHON-BASED PERFORMANCE ANALYSIS SUPPORT SYSTEM: SOFTWARE FOR PERFORMANCE ANALYSIS OF REGULATORY PHOTOCHEMICAL AIR QUALITY MODELING

Abstract

Model performance evaluation (MPE) for regulatory modeling is the core task to establish the reliability of models used for developing air quality management. Many existing MPE approaches and practices for regulatory ozone modeling have shortcomings in their support for ozone air quality decision making. At the same time, the subject of how to perform these MPEs has been described in the regulatory air quality modeling community as a difficult research area. We have recently proposed a comprehensive MPE approach to enhance the quality of MPE protocols. In implementing these protocols with our approach we found that the existing tools, which were designed to assist MPE practices following traditional approaches, fell short of meeting our needs. That is, existing tools are inadequate for permitting a more comprehensive MPE. Here, we describe the Python-based Performance Analysis Support System (pyPASS) that facilitates the implementation of the new MPE approach for regulatory photochemical modeling. We briefly illustrate advantages of pyPASS by showing some results from an application to a real regulatory modeling case. We also show that pyPASS can provide more focused information for comprehensive model performance evaluation with less resources than can traditional tools.
3.1 Introduction

In the United States, performance evaluation of regulatory photochemical air quality models (RPAQMs) plays a necessary role when RPAQMs are used for finding answers to questions about ozone air quality management. If a state violates the National Ambient Air Quality Standard (NAAQS) for ozone, the state is required by the Clean Air Act Amendments 1990 (CAAA 1990) to develop a State Implementation Plan (SIP) to prevent such violations in the future. Because process-based, three-dimensional air quality models are considered to be the most suitable tools for simulating ozone formation (National Research Council, 1991), the CAAA 1990 mandates the use of these models as the legal method for developing SIPs and for demonstrating that these SIPs can achieve attainment in the future. This legal requirement is applied to states having (1) moderate and above non-attainment areas (NAAs) under the old 1-hour ozone NAAQS, and (2) all of NAAs under the new 8-hour ozone NAAQS (US EPA, 2003). Model performance evaluation (MPE) is the core procedure used to establish the reliability of predictions made by these RPAQMs for application to the development and testing of SIPs.

In regulatory applications of RPAQMs, a state’s ultimate goal is to develop effective ozone control strategies and to demonstrate that its proposed control strategies will “more likely than not” achieve attainment of NAAQS in future. In general, any SIP modeling comprises three major tasks: (1) replication of historical ozone NAAQS violation events, (2) prediction of the future ozone concentrations after control policies that are required under Federal EPA rules are applied, and (3) development of any additional control strategies and demonstrate their efficacy for attainment if the modeled future ozone concentration by task 2 still violates the NAAQS. For task 3, the state can also utilize corroborative analyses
including Weight-Of-Evidence analyses to complement the modeling-based attainment demonstration. Task 1 is called ‘base case’ modeling and model performance evaluation (MPE) is the process in which modelers judge the acceptability of model predictions in this task and appraise the degree of model prediction reliability for the purposes of task 2 and task 3.

The MPE for management studies are often distinguished from the MPE for scientific studies (Russell and Dennis, 2000); that is, an MPE for management studies emphasizes the acceptability of modeling results for answering how to best manage ozone air quality, while an MPE for scientific studies focuses the accuracy of the science realized in a model to describe environmental systems. SIP modeling is definitely a management study and an MPE for SIP modeling needs to meet the general demands for the management studies, thus ‘fulfillment of designed tasks’ is one of the most important aspects of the model’s reliability assessment (Beck, 2002). Creating and performing an MPE is a complex process and developing a strategy to do SIP modeling including MPE may be a very large burden for many states. Hence, to assist states in the performance of MPE for ozone SIP modeling, the US EPA developed SIP modeling guideline documents that include: (1) proposed model performance evaluation (MPE) methods and (2) information required for each method (US EPA, 1991; US EPA, 1996; US EPA, 1999; US EPA, 2005b). The guidance documents, however, do not clearly state what criteria should be considered and what procedures need to be followed to conduct MPE properly and effectively, especially for important graphical analyses.

Consequently, state regulatory agencies have often literally followed the overly general guidance even for their specific situations. For example, (1) state agencies estimate
statistical performance measures and compare them with suggested performance criteria that are often considered improper (Russell and Dennis, 2000) and (2) they simply create the graphical measures specified in the guidance documents and provide only oversimplified concluding statements such as “showing reasonable agreements” without providing sufficient rationale about how model performance was judged. Moreover, virtually all past MPE for SIP modeling solely focused on “ozone performance”; this has been considered one of the major shortcomings of the past MPE practices following the US EPA’s MPE protocol (Roth, 1999; Russell and Dennis, 2000; Roth et al., 2005). While some research has called for better MPE, the photochemical air quality modeling community has recognized that improvement of MPE methods is one of the most difficult research areas (National Research Council., 1991; Georgopoulos, 1995; Russell and Dennis, 2000; Fine et al., 2003). This is especially true for MPE methods suitable for peer-review conducted by a third-party of a regulatory PAQM application (Roth, 1999; Roth et al., 2005). In other words, there is a lack of appropriate MPE guidance that states might follow to enhance their MPE efforts.

We have recently proposed an improved MPE approach called Protocol for Regulatory Ozone Modeling Performance Tests (PROMPT) (Kim and Jeffries, 2006b). PROMPT is a meta-protocol to construct a MPE protocol instance for comprehensive and systematic evaluations of a particular SIP application. Two distinctive characteristics of PROMPT are (1) it emphasizes “progressive” analyses in which the model evaluator examines same performance measures multiple times during the course of evaluation while inspecting more information as the MPE advances, (2) it requires “day-by-day and site-by-site” analyses as well as the limited overall performance evaluation that has been the mainstay of past MPE efforts.
Partial implementation of a PROMPT-like MPE has already shown promising results. These results include explicitly taking into account for model biases in assessing a model’s reliability to answer policy questions (Jeffries et al., 2005). We found, however, some challenges existed to implement fully PROMPT-like protocols with existing tools such as the Package for Analysis and Visualization for Environmental (PAVE) (CMAS, 2005); this was because existing tools lacked important functionalities to generate the information necessary to conduct PROMPT-like performance analyses. In other words, they were not optimally designed to support the needs of a PROMPT-like protocol.

In this paper, we introduce the Python-based Performance Analysis Support System (pyPASS), a new analysis tool specialized in supporting MPE for SIP modeling. In the following section, we present the rationale for developing pyPASS, the overview of pyPASS, illustrative examples of pyPASS applications, and discuss future improvements.

3.2 Background of pyPASS development

3.2.1 Improved analyses of graphical performance measures

Essentially graphical measures used in MPE for SIP modeling are “information”. They are indeed a type of “representation of knowledge” and it has been recognized that these informative measures can play an important role in finding out possible model errors (Tesche et al., 1990; US EPA, 1991). In the SIP modeling community, however, we found that the informational power of graphical measures was frequently underestimated by considering them as “just qualitative” measures. In the past decade, research on information visualization has grown rapidly with increasing numbers of journal articles focused on seeking better principles, algorithms, and computerized methods related to information visualization (Chen, 2002). One of the key findings in the pioneering research on information visualization were
examples of many cases where users of graphical information (including scientists) made their graphics ‘dequantified’ by choosing improper graphical attributes such as scales and colors (Tufte, 1997). Indeed, we have found many graphical measures that were poorly designed and presented in technical supporting documents prepared as part of SIP modeling.

Graphical measures were dequantified and information was presented more poorly than would have been if these measures were properly designed and implemented. Moreover, to conduct MPE following a PROMPT-like protocol, we argued that sets of multiple graphical measures should be prepared more formally than in past SIP modeling practices. Although it is not an exhaustive list, we present several recommendations:

- Wind speed scatter plots, hodograms (i.e. a type of wind speed and direction time series), and wind error hodograms need to be examined as a group for more complete analyses.

- Daily peak ozone and precursor concentration at monitor locations should be depicted after sorting monitors into a physically meaningful order such as west-to-east or some other spatial ordering that is appropriate for the case (not alphabetical by site name).

- Ozone time series at monitoring sites should be coupled with the precursor species such as nitrogen oxide signals plus carbon monoxide and selected volatile organic compounds, if available, in a plot.

- Spatial plots (i.e. tile plots) of selected species should include more auxiliary information such as geographical features and surface winds.

- All attributes of graphical measures such as coloring (e.g. not Microsoft Excel’s “default” order colors) should be consistent through iterations of MPE. In other
words, different sensitivity tests run results should be presented with identical attributes of graphical elements used for original simulations.

- The representation of data needs to reflect the nature of data, e.g. hourly averaged data have to be displayed as step lines correspondingly and tile plots should plot as squares, not spatially interpolated (which distorts the model’s actual results).
- The format of graphical measures should be flexible so that graphical measures can be easily used with other information such as high resolution maps crafted with geographical information systems (GIS). Also, a compression method for graph files without loss is highly desirable.

In addition, we recommend using complementary information such as resultant wind speeds of morning and afternoon, improved statistical measures proposed in the literature, and others that will be described below. In some important cases, high resolution measurements are available through aircraft observation and automated gas chromatography measurements of VOCs (auto-GC). The recommendation for routine monitored data above can also be applied to these high resolution data. We believe utilizing these non-routine data in MPE will provide important additional insights into the model’s behavior resulting in more meaningful evaluations.

### 3.2.2 Object-oriented production of graphical measures

The graphical measures required by our new analysis method present significant information and the procedures to conduct graphical analyses are comprehensive. Nevertheless, the production of these graphical measures can be automated quite easily. Once we know the desirable attributes of graphical measures, we can create new plots by changing the core information such as legend and graphical representation of data (e.g. lines)
but keeping other properties of plots. For example, if we want to compare multiple modeling
results with observation, we just need to add or insert chemical signals to graphs. At the
abstract level, these activities are essentially modeled as handling “objects” (e.g. a time series
plot of model-observation at a site on a day) or deriving an object from the abstract
description of objects (i.e. class). How to conceptualize and implement these abstract level
tasks is one of the well-known software engineering problems and has a solution: object-
orientation. Therefore, practically, we can implement the automation of graphical measure
production easily and efficiently with available object-oriented programming tools.

3.2.3 Difficulties in using existing tools

Using existing tools for the new MPE approach is non-trivial and some analyses
recommended by the new MPE are impossible to do with them. There is no single tool that
can meet most of our analytical needs. For some analyses, we have had to utilize multiple
tools to make graphical measures and we could not produce some important measures. In
recent MPE efforts, we requested that the original developers modify the existing programs
to help achieve our needs. For example, the recent update of PAVE version 2.3 reflected
these requests. In worse cases, we could not make changes because some tools were
commercial and no source codes were available. The following points are several examples
highlighting the challenges in applying existing tools for a PROMPT-like MPE: (1)
information integration, (2) repetition of same types of graphical measures, (3) customization
of graphical measures, and (4) openness of all tools to public.

First, an MPE practice following PROMPT requires “integrated” graphical measures
holding several entities. The existing tools are, however, incapable of performing this
required integration of information for the new approach. An example is a tile plot of
predicted chemical concentrations with predicted wind vectors, observed chemical
concentrations and observed winds at monitor locations need to be all plotted together. In addition, the new approach demands customized “packaging” to present effectively graphical measures with evaluators’ written judgment. The customized packaging is necessary to reveal and clarify important observed model behavior that is difficult to detect with one graphical measure alone and difficult to review effectively without guiding commentary. Thus, we need a tool that provides information packaging capability.

Second, a collection of graphical measures will be produced repeatedly during the exercise of the new MPE approach. As we discussed above, the contents of each graphical measure can be very integral and complex. The process of making these measures, however, is quite repeatable and the same types of graphical measures are often generated multiple times. For example, we recently conducted MPE tasks for a SIP modeling case in which we ran eight major simulations. To analyze one of those simulations, we produced 320 time series plots, 782 scatter plots, and about 3500 tile plots of surface concentrations of 10 species for a part of the modeling period. The existing tools are, however, ineffective in generating large amounts of information routinely; execution time is too long or faster operation requires auxiliary support even including separate hardware. For example, using PAVE to create tile plots with Intel Xeon 2.8 GHz with 2.5GB RAM machines took several hours to produce a collection of tile plots needed to evaluate a simulation. This type of problem is partly due to the fact that most of existing tools are either (1) general-purpose tools that are not optimized for MPE tasks for SIP modeling or (2) exploratory tools that are most suitable for trial-and-error type of analyses performed in the interactive mode. To overcome these shortcomings, some tools frequently used in SIP modeling processes have
been extended to support script-driven executions. Even so, such script capabilities are often very limited and the computational load demand is still not resolved by adding scripts alone.

Third, the new approach encourages the use of customized graphical measures for each SIP modeling case to accommodate the analytical power to detect case-specific problems. Often, new information is generated during the MPE and modelers need to reflect this new knowledge in their analyses. Consequently, the new approach requires some degrees of flexibility in updating and modifying elements of graphical measures. Existing tools are inflexible at producing these types of customized information. Essentially none of the existing tools allow users can make user-specified extensions and features easily. These include (1) external extensions such as adding new graphical measures, (2) internal extensions such as specializing existing graphical measures.

Last, one of the most important aspects of the new MPE is to make the MPE process very clear for external reviews. This means that all tools used for MPE need to be available to external reviewers for the replication of results or for inspection of the process. Consequently, it is desirable to have MPE tools that are low cost (or free) to the public.

In our practical experience, there is no tool that can satisfy all the needs of the new MPE analysis. Therefore, to properly implement a PROMPT-like MPE, a new performance analysis toolset is necessary.

3.2.4 Design goals and choices for implementation

To overcome the shortcomings of existing tools described in the previous section, we specified several desirable features as guidance for pyPASS development. First, pyPASS needs to provide information that may not be directly available through existing tools. There are cases where some information is not available because of the absence of proper data retrieval and/or information generating tools. Missing information may result in flaws in
judgment of a model’s adequacy for its policy applications. Thus, it is important that users can incorporate all available information for MPE with pyPASS. Second, the operation of pyPASS should be faster and more efficient in generating information than existing tools used in regulatory air quality modeling studies. The performance evaluation for regulatory modeling is often constrained by regulatory timelines and MPE is an iterative process during a SIP modeling. Often the cycle of modeling-evaluation requires changes of graphics and documents. Because more time for focusing analyses is desirable in environmental modeling (Argent, 2004), efficiency in generating and documenting information is important for SIP modeling. Third, pyPASS codes must be flexible and extendable to accommodate future changes of data formats and needs for new analysis measures. That is, source codes of pyPASS modules should be available and the modification of pyPASS should not be an overwhelming or incomprehensible task. Last, pyPASS must be publicly available with little or no cost to acquire and use it.

To implement these features, we made several decisions for the pyPASS development. First, we chose to retrieve only necessary data for each phase of performance analyses. The previous tools require access to the entire input and output files of RPAQMs representing a large amount of model output, most of which is not used in the MPE. Also, there are often many sets of these outputs that are the results of various “trial” simulations or sensitivity runs that require comparison. In our recent study, we needed access to 48 sets of simulations. Each set of files is > 500 GB restricting tool operation to be on the same machine used to run the simulations. This limited the ability of non-modelers (e.g. policy analyst) to conduct MPE on their local machines. Thus, a design element was to make available the important model data needed for effective MPE on typical personal computers such as laptop machines.
This was achieved by creating two data extraction utilities, CAMxSubset and CMAQExtract, and by using heterogeneous hierarchical files. With these designs, we could put data necessary to compare eight simulations for ten days at monitor sites into a file of 5 MB. This permits a much more varied and comprehensive comparison of model outputs as part of MPE. Second, we developed, as a part of standard pyPASS graphics, some graphical measures that are not available via current analysis tools or may require complex operations with multiple tools to obtain the graphics. Second, because batch operations of generating predefined graphics are more suitable for producing large sets of information efficiently than interactive operations, we selected the command-line as the user interface and made pyPASS generate graphics with “pre-defined” configurations via “option” files. Choosing proper labels, scales, and colors of graphics carefully and deliberately can enhance the quality of information delivered by graphical measures (Tufte, 1997). Therefore, we designed pyPASS graphics carefully by considering how to effectively carry information in them for PROMPT-like MPE. Last, we selected Python as the main programming language for implementing platform of pyPASS. Python is an object-oriented language that takes advantage of scripting, i.e. no “compiling and linking” step and automatic dynamic memory management (Python Software Foundation, 2005). Therefore, development cycles (i.e. debugging-fixing-testing) can be faster compared with the use of compiler languages such as C++. Since Python is a script language, all source codes of pyPASS modules are available for users to read and review. In addition, fast code development can be achieved by use of sets of libraries that have been developed to support scientific computing and visualization. Object-orientation in Python can enhance the reusability of pyPASS modules in future
improvement. Because the Python interpreter is free and almost all supporting libraries for pyPASS modules are also free, users can acquire for pyPASS free or at very low cost.

3.3 Overview of pyPASS

3.3.1 Supporting libraries

To build pyPASS, we utilized many existing Python libraries and applications, which are often referred to as “site-packages”. In turn, each site-package may require other libraries and/or applications. Table 3.1 shows all external applications, libraries, and Python-site packages needed to make pyPASS fully functional. Despite trying to make pyPASS run across as many platforms as possible, some pyPASS modules are available only on specific operating systems. This is due to some libraries or applications provide necessary Application Programming Interface (API) through operating system specific developmental environment. A good example is the Microsoft Word™ API through win32com (Hammond, 2005) that is used to build pyPASS modules to automatically generate analysis reports in the Word file format. Alternatives being implemented are to create LaTeX files that are platform independent. TeX processors exist on virtually all common computer platforms to convert LaTeX text (TUG, 2006) and almost any style of graph output into Portable Document Format (PDF) files (Adobe, 2006).
Table 3.1. List of major site-packages used in pyPASS. Note that each library listed in this table may require other libraries. For example, PyTables requires the HDF5 library.

<table>
<thead>
<tr>
<th>Name</th>
<th>Version</th>
<th>Source</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAMxSubset CMAQExtract</td>
<td>8.0</td>
<td><a href="http://ftpozone.sph.unc.edu">http://ftpozone.sph.unc.edu</a></td>
<td>Extraction of binary data from RPAQM inputs/outputs</td>
</tr>
<tr>
<td>ChartDirector</td>
<td>4.0</td>
<td><a href="http://www.advsofteng.com">http://www.advsofteng.com</a></td>
<td>Library for all graphical objects such as XY plots</td>
</tr>
<tr>
<td>mx.DateTime</td>
<td>2.0.6</td>
<td><a href="http://www.egenix.com/files/python/mxDateTime.html">http://www.egenix.com/files/python/mxDateTime.html</a></td>
<td>Library for ISO time string support and date-time objects</td>
</tr>
<tr>
<td>numarray</td>
<td>1.3.3</td>
<td><a href="http://www.stsci.edu/resources/software_hardware/numarray">http://www.stsci.edu/resources/software_hardware/numarray</a></td>
<td>Efficient numeric library</td>
</tr>
<tr>
<td>PyTables</td>
<td>1.1</td>
<td><a href="http://pytables.sourceforge.net">http://pytables.sourceforge.net</a></td>
<td>Python wrapper for HDF5 library (<a href="http://hdf.ncsa.uiuc.edu/HDF5">http://hdf.ncsa.uiuc.edu/HDF5</a>)</td>
</tr>
<tr>
<td>pyRXP</td>
<td>1.0.7</td>
<td><a href="http://www.reportlab.org">http://www.reportlab.org</a></td>
<td>Library for fast XML parser</td>
</tr>
<tr>
<td>GDAL</td>
<td>1.19</td>
<td><a href="http://hobu.biz/software/python_guide_esri/9_10_Grids">http://hobu.biz/software/python_guide_esri/9_10_Grids</a></td>
<td>Library supporting vector and raster object; also including PROJ.4 module</td>
</tr>
</tbody>
</table>
For the data file format, we adopted PyTables (Altet, 2005). This format has many advantages that are important for the pyPASS development. First, PyTables supports a hierarchical data model that enhances the use of heterogeneous data. With PyTables, we can access file data with very similar methods used for standard directories and files on disk file systems. That is, we can use a “natural” way to access data with multiple characteristics and attributes; i.e. path-like access statements can be used to locate and extract the exact data set needed in the file. Further, data from a variety of sensitivity simulations can be conveniently stored in a single file, which facilitates “simulation vs. simulation” comparisons. PyTables is also highly efficient at accessing large (e.g. >2GB) files because the underlying data format of PyTables is the Hierarchical Data Format version 5 (HDF5) that has been used widely in scientific community (NCSA, 2005). PyTables, however, expands the supporting data types beyond what HDF5 supports. That is, users can store not only multi-dimensional arrays (e.g. chemical concentrations) but also various “table objects” (e.g. meta-data such as monitor description) in a single PyTables file. In addition, PyTables provide the consistent interface for accessing various data objects and supports online file compression. Because data storage always becomes an issue in SIP modeling, the file compression in PyTables is practically desirable. Second, PyTables utilizes Python’s numarray library that is an efficient and powerful numerical data processing library (Space Telescope Science Institute, 2005). With the numarray library, users can perform complex computations with large multi-dimensional data arrays without the use of complex indexing calls that are typical of a FORTRAN implementation. If computational performance becomes a significant problem, users can easily replace computationally intense parts of numarray with C or FORTRAN that are callable by Python to improve the
numerical performance. Hereafter, we will call PyTables made by pyPASS as pyPASS Tables.

The library chosen for visualization was ChartDirector™ (Advanced Software Engineering Limited, 2005) because it is a flexible and efficient complex graphics library available for multiple platforms. Users can efficiently customize their graphics in great detail via ChartDirector’s object-oriented API: users can treat graphical elements (such as lines) as objects and can stack objects to populate and produce complex plots with user-defined layouts. For example, to generate a wind field plot with several chemical concentration for different sensitivity runs, users only need to have one wind field plot object and overlay it on chemical plots without re-creating the object of wind field plot. Even though ChartDirector is a commercial product, it inexpensive and can be freely distributed to users depending on the types of end-user license.

There are some utility libraries we used for the pyPASS development. These libraries are used across pyPASS modules to meet our implementation needs. First, we used pyRXP (ReportLab, 2005) for parsing eXtensible Mark-up Language (XML) (W3C, 2005) files. We adopted an XML format to store important meta-data (including meta-data about the meta-data such as modeling domain definition). XML is specifically designed for such data description and storage. We also store important pyPASS input data, such as the description of ground monitors and observational data sets in an XML format. We used pyRXP (ReportLab, 2005) due to its superb parsing performance and XML validation ability.

Second, we utilized a Python interface module to the Geospatial Data Abstraction Library (GDAL) for converting geocoordinate data to and from projected coordinate data for locations. The GDAL library itself focuses on raster data access but it also contains several
libraries for vector data access and coordinate transformation libraries, such as PROJ.4 (USGS, 2005). While the current use of GDAL by pyPASS modules are limited, we chose GDAL to expand pyPASS functionalities in near future such as reading and writing GIS data files (e.g. Shapefile) directly from pyPASS.

Last, we frequently used time objects and chose the ISO 8601 date and time representations from the mx.DateTime library (eGenix, 2005), to efficiently and accurately calculate calendar time and time differences. In air quality modeling, most of the data are temporal in nature. One of the difficulties in temporal data handling in SIP modeling is that models often use different time basis from observational dataset. For example, one RPAQM may use Coordinated Universal Time for its output file while second RPAQM uses Local Daylight Time and ground monitors recorded measurements in Local Standard Time. Dealing with this temporal information is often quite cumbersome and prone to programming bugs. Presenting results in UTC time is often confusing to policymakers. By using an object-oriented time library, we could reduce work efforts and make all temporal data used during MPE tasks with pyPASS conforming to a single time representation, including handling time zones.

3.3.2 pyPASS structure

pyPASS itself is organized as a standard Python site-package so that users can import pyPASS classes and call functions by passing proper arguments from the command-line or via the Python interactive shell. The current version of pyPASS is composed of three major sets of main code modules, one set of utility, and two external utility programs. Each set of pyPASS modules has function modules and classes. The first set is a collection of 7 modules for handling air quality model data and for integrating necessary metadata such as
map projection information. The second set is a collection of 8 modules visualizing model prediction and observation. The third set is a collection of 8 modules documenting results of performance assessment. \texttt{pyPASS} also has a set of 6 utility modules to support the main \texttt{pyPASS} modules by providing functions such as building domain objects from a domain definition file written in \texttt{XML}. Explanation of individual module’s functionality is presented in Table 3.2. Because \texttt{pyPASS} modules are scripts, all \texttt{pyPASS} modules should be easily modifiable and expandable during the course of performance analyses of SIP modeling. Two external utility programs written in the standard C language, \texttt{CAMxSubset} and \texttt{CMAQExtract}, are used to extract values from input and output files of two most commonly operated RPAQMs: the Comprehensive Air quality Model with eXtenstions (\texttt{CAMx}) (ENVIRON, 2005) and the Community Multi-scale Air Quality model (\texttt{CMAQ}) (US EPA, 2005a). These two utilities were necessary because we desired a common data extraction method for various RPAQMs that use different file formats for their inputs and outputs. For example, the current \texttt{CAMx} creates \texttt{FORTRAN} binary files for its output while \texttt{CMAQ} generates IOAPI files (US EPA, 2005a) that are built on the \texttt{netCDF} library (UNIDATA, 2005). Further, we achieve a large reduction in size of the RAPQM files such that the necessary MPE simulation results of 10’s of model runs can easily fit on a user’s laptop computer.
Table 3.2. List of pyPASS codes and utility programs with the summary of their functionalities.

<table>
<thead>
<tr>
<th>Set name</th>
<th>Module name</th>
<th>Description of function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data handling set</td>
<td>AircraftInModel.py</td>
<td>File object representing model data on aircraft tracks for user defined intervals</td>
</tr>
<tr>
<td></td>
<td>AQMSSubset.py</td>
<td>Definition of data object of “layers” of model inputs and outputs</td>
</tr>
<tr>
<td></td>
<td>MakeAircraftFile.py</td>
<td>File object representing observed data on aircraft tracks</td>
</tr>
<tr>
<td></td>
<td>MakeMonFile.py</td>
<td>File object representing observed data at monitor sites</td>
</tr>
<tr>
<td></td>
<td>MakeSimSUBFile.py</td>
<td>File object representing data for “layers” of model inputs and outputs</td>
</tr>
<tr>
<td></td>
<td>MakeSimTSFile.py</td>
<td>File object representing data of model inputs and outputs at monitor sites</td>
</tr>
<tr>
<td></td>
<td>ModelAtAircraft.py</td>
<td>File object representing observed data averaged with user defined intervals</td>
</tr>
<tr>
<td>Visualization set</td>
<td>DrawTDPlots.py</td>
<td>Definition of objects for tile plots</td>
</tr>
<tr>
<td></td>
<td>MakeACTSSSSLots.py</td>
<td>Procedure for aircraft plots</td>
</tr>
<tr>
<td></td>
<td>MakePeakBarPlots.py</td>
<td>Procedure for daily peak bar plots of chemicals</td>
</tr>
<tr>
<td></td>
<td>MakeStdPlots.py</td>
<td>Definition of all objects of pyPASS standard graphical measures except tile plots</td>
</tr>
<tr>
<td></td>
<td>MakeTDPlots.py</td>
<td>Procedure for tile plots</td>
</tr>
<tr>
<td></td>
<td>MakeTSSSSLots.py</td>
<td>Procedure for time series plots and scatter plots</td>
</tr>
<tr>
<td></td>
<td>MakeWindTSSSSLots.py</td>
<td>Procedure for hodograms and error hodograms</td>
</tr>
<tr>
<td></td>
<td>MonModStat.py</td>
<td>Procedure for estimating various statistical measures</td>
</tr>
<tr>
<td>Documenting set</td>
<td>MakeAll3S2STSSMSPlotDocs.py</td>
<td>Procedure for making pages with chemical species plots from multiple scenarios</td>
</tr>
<tr>
<td></td>
<td>MakeAllBPMSPlotDocs.py</td>
<td>Procedure for making pages with peak bar plots</td>
</tr>
<tr>
<td></td>
<td>MakeAllNO2SSMSPlotDocs.py</td>
<td>Procedure for making pages with NO₂ scatter plots</td>
</tr>
<tr>
<td></td>
<td>MakeAllWSSSPlotDocs.py</td>
<td>Procedure for making pages with hodograms</td>
</tr>
<tr>
<td></td>
<td>MakeMSPlotDocs.py</td>
<td>Procedure for compiling pages generated by other modules</td>
</tr>
<tr>
<td></td>
<td>MakeMSWordFmText.py</td>
<td>Script for creating pages with plots and commentary</td>
</tr>
<tr>
<td></td>
<td>Txt2MSWord.py</td>
<td>Parser for user’s configuration of pages to create MS Word files</td>
</tr>
<tr>
<td></td>
<td>pyWord.py</td>
<td>Procedures constructing MS Word file with plots and commentary</td>
</tr>
<tr>
<td>Utility set</td>
<td>ChartSettings.py</td>
<td>Definition of setting objects for pyPASS standard graphics</td>
</tr>
<tr>
<td></td>
<td>ColorScale.py</td>
<td>Definition of color bar object for tile plots</td>
</tr>
<tr>
<td></td>
<td>DomainGrid.py</td>
<td>Definition of objects representing model grids and domains</td>
</tr>
<tr>
<td></td>
<td>PlotUtils.py</td>
<td>Procedures for frequently used functions such as coordinate transformation</td>
</tr>
<tr>
<td></td>
<td>TDPPlotUtils.py</td>
<td>Commonly used procedures to support tile plot production</td>
</tr>
<tr>
<td></td>
<td>XML2Dicts.py</td>
<td>Procedure to convert XML to Python dictionary</td>
</tr>
<tr>
<td>External utility programs</td>
<td>CAMxSubset</td>
<td>Extract user-specified data from CAMx inputs and outputs</td>
</tr>
<tr>
<td></td>
<td>CMAQExtract</td>
<td>Extract user-specified data from CMAQ inputs and outputs</td>
</tr>
</tbody>
</table>
3.3.3 Inputs to pyPASS

MPE tasks for SIP modeling can be grouped into three sets, as shown in Table 3.3. The first set is related to the set-up of a SIP modeling study. Example tasks are to document modeling domains including grid configurations. The second set is relevant to actual simulations and the preparation of observational data including conversion of them into comparison-ready formats. The last set is associated with comparing observation with prediction of a single model run or predictions of more than one model runs, or performing comparison of various model simulations without observation. Consequently, when analyzing model performance with pyPASS, each task set needs a distinctive set of inputs. Note that the task of performing simulations is different from analyzing them. During analyses tasks, evaluators may examine various aspects of a simulation with different graphical measures. That is, evaluators use the same data several times to produce different forms of materials necessary for a given analysis task.
Table 3.3. Important pyPASS inputs. These files are necessary to operate pyPASS in the full functional mode and are prepared independently from pyPASS modules.

<table>
<thead>
<tr>
<th>Input</th>
<th>Content</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain Definition</strong></td>
<td>Definition of geographical coordinate system</td>
<td>XML</td>
</tr>
<tr>
<td></td>
<td>Definition of projected coordinate system</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Definition of horizontal grids and vertical grids including user-defined domain labels</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Modeling period</td>
<td></td>
</tr>
<tr>
<td><strong>Monitor Definition</strong></td>
<td>Observational data description including data sources</td>
<td>XML</td>
</tr>
<tr>
<td></td>
<td>Time including time zone and monitoring interval</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identification number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Location information including county name where the monitor is located</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Measurement Parameters with valid operation period</td>
<td></td>
</tr>
<tr>
<td><strong>Monitor observation data</strong></td>
<td>Observed variables at surface; meteorological variables and concentrations of chemicals</td>
<td>ASCII text</td>
</tr>
<tr>
<td></td>
<td>speciated for the chemical mechanism used in RPAQM</td>
<td></td>
</tr>
<tr>
<td><strong>Aircraft observation data</strong></td>
<td>Aircraft position over time in space and observed variables; aloft</td>
<td>ASCII text</td>
</tr>
<tr>
<td></td>
<td>meteorological variables and concentration of chemicals speciated for the chemical</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mechanism used in RPAQM</td>
<td></td>
</tr>
<tr>
<td><strong>RPAQM outputs</strong></td>
<td>Four-dimensional chemical concentration fields</td>
<td>Specific binary format to each RPAQM files</td>
</tr>
<tr>
<td><strong>RPAQM meteorology inputs</strong></td>
<td>Four-dimensional meteorological variable fields</td>
<td>Specific binary format to each RPAQM files</td>
</tr>
<tr>
<td><strong>pyPASS options</strong></td>
<td>Command-line options for the attributes, scope, and contents of graphical measures</td>
<td>ASCII text</td>
</tr>
</tbody>
</table>

† Similar definition inputs is designed for aircraft flight definition and planned to be implemented in the next release of pyPASS.
The first input group comprises study-specific inputs that should be available before using any pyPASS module. It is desirable not to change these inputs significantly during pyPASS application if users want to keep all information consistent. For example, pyPASS requires users to prepare the domain definition in the XML format that describes the definition of geographical coordinate system, the definition of projected coordinate system, the definition of horizontal grids and vertical grids including user-defined domain labels, and the modeling period. If this information should be changed, modelers have to re-run their SIP modeling with changed conditions. Consequently, modelers have to run pyPASS to generate all graphics again. In practice, information stored in XML files pertains to the whole SIP modeling processes so that they are not likely changed during the course of SIP modeling.

The second input group is composed of simulation-specific inputs that are actually the extracted outputs of RPAQM or pyPASS modules for data conversion. In terms of the origin of inputs, the second input group consists of two types of data depending on whether they are created directly by pyPASS modules. Since some pyPASS inputs are outputs of other pyPASS modules, we only describe inputs that are not directly produced by pyPASS modules. For those inputs that pyPASS makes, we describe them in detail when we discuss outputs in the next section. After running each simulation, users need to extract data from regular model binary inputs and outputs. CAMxSubset and CMAQExtract are provided to do the extraction task based on user supplied input command file. Users can develop their own extraction codes for RPAQMs other than CAMx or CMAQ, as long as their extract binary files meet the relatively simple extracted binary file format that pyPASS requires. This extraction process typically runs on the same computer system on which the air quality
model is operated and returns a small-sized file to the MPE user. This allows users to store the extracted data on a second system for analyses. The format of extracted binary files can be found in the user’s manual distributed with CAMxSubset and CMAQExtract. Types of extraction will vary with the phase of MPE. Typical extraction will be performed to obtain (1) values of cells containing ground monitors during modeling period for monitor-by-monitor analyses, (2) values of cells that permit the calculation of bi-linear interpolated values of each set of four cells including a cell containing a ground monitor and three cells nearby the cell of each monitor or other collections of cells for different interpolation scheme, (3) sliced cell arrays of three-dimensional modeling domain, (4) cell values on aircraft tracks for comparing model prediction with aircraft observation. One of the important requirements of this second input group is ensuring the consistency of data resolution between observation and prediction (e.g. keeping the model on the same time average as the observations.) Users need to acquire proper data handlers to perform model extractions.

The third input group consists of analysis-specific inputs that include all necessary configurations for pyPASS graphics and operational details supplied through command-line options. Examples are command line options to control color-coding schemes for chemical species used in time series plots. Users can supply commonly used options for specific pyPASS modules through files or they can overwrite options in files by passing alternative options at the time of module calling. Also, the third input group includes documentation set up options such as places of plots and evaluators’ commentary statements for model performance.
3.3.4 Operational procedures of pyPASS application

As shown in Figure 3.1, operations of pyPASS can be grouped into three categories: (1) observational data formatting operations that are necessary only once for an episode (Box 1 in Figure 3.1), (2) prediction data processing operations that are needed for every simulation (Box 2 in Figure 3.1), and (3) graphics production and documentation operations that are done for each performance analysis including each revision of analysis materials.

The first group of operations pertains to storing observational data in pyPASS Tables. For ground monitors, pyPASS uses speciated ASCII data files and monitor definition files written in XML. Once these files are prepared, actual measurement data are converted into a pyPASS Tables file that has the header information from the monitor definition files. For aircraft measurements, pyPASS takes ASCII data files and converts them into separate pyPASS Tables files. There is no good standard aircraft measurement file formats at present. Therefore, we created file conversion modules suitable for manipulating NOAA and BNL aircraft observation data available through Texas Air Quality Study 2000 (Daum et al., 2005). This will be further evaluated and improved in future releases of pyPASS.
Figure 3.1. Overview of *pyPASS* operation. Box 1 depicts the general *pyPASS* operation necessary once for an episode. Box 2 represents the general *pyPASS* operation needed for each simulation. Once a simulation is performed for an episode, users may run multiple *pyPASS* operations with visualization modules and package materials for subsequent analyses and communication with documentation modules.
The second group of operation is to convert data extracted from RPAQM files into **pyPASS Tables**. For data conversion, **pyPASS** uses outputs of CAMxSubset and CMAQExtract. These data extractors, written in portable C language, retrieve data for user-specified cells in the model domain from CAMx/CMAQ binary input and output files. Two typical sets are (1) cells where stationary monitors are located and (2) horizontal/vertical layers. The format of the CAMxSubset/CMAQExtract outputs is an IEEE big-endian C-binary file holding arrays of data following an ASCII header. The header portion of files provides metadata such as the model grid IDs, site names, and users’ comments (for details, refer to documents distributed with CAMxSubset and CMAQExtract). Therefore, if there are proper data retriever utilities, files of other RPAQMs can be used with **pyPASS**. Once the extracted files are created, the next step is to convert these extracted binary files into a **PyTables** file and to merge model metadata into the same section of the hierarchical **PyTables** file.

The third group of operations consists of one mandatory subgroup plus an optional subgroup: (1) to create graphics and calculate statistics by invoking proper **pyPASS** modules with command-line keywords and arguments, (2) to document necessary MPE measures and evaluator’s comments. Depending on the plot types, each **pyPASS** module provides various choices to manipulate the details of the graphics, such as a line thickness. However, some of the graphical elements cannot be changed without directly modifying the **pyPASS** module. For example, all hourly averaged values are plotted with stair-step graphs. We fixed or pre-defined these options for **pyPASS** graphics because they were intended to convey accurate information with regard to the nature of data and to minimize confusion over data properties. Nevertheless, users can modify the current **pyPASS** codes easily with the knowledge of the
API used for building pyPASS and interfaces of pyPASS modules. Once MPE measures are generated, the next step is to document analysis results. In the documentation operations, automation of generating analysis report pages is critical for users who have to produce a report containing a large set of graphs that are consistent across whole analysis report documents, especially when plot style modifications are needed after hundreds of pages of documents containing hundreds of graphs are made. To make this documentation process efficient and practical, pyPASS provides automation modules to create assessment documents by inserting plots and texts. Currently, pyPASS supports MS Word Document (Microsoft, 2006) as the format of assessment documents. The documentation operation, however, is optional because users may skip this step until they are confident of analysis results or they may want to conduct more in-depth analysis and to review more materials.

On the other hands, these “reports” compactly organize, on a day-by-day and site-by-site basis, the collection and integration of plots and texts to provide a concise presentation of the model’s ability to reproduce observations. Thus, they should be produced and consulted throughout the model’s iterative operation for a base and future case simulation. The entire collection can help convey the extent to which the modeling system is sensitive to critical inputs and thus help the decision makers understand model’s potential reliability with regard to the decision they are making.

3.3.5 Outputs from pyPASS

The pyPASS produces four major types of outputs: (1) pyPASS Tables holding observational data and/or model data with corresponding metadata, (2) graphical measures such as time series plots, (3) statistical measures such as normalized mean biases, and (4) documentation of all of evaluator’s comments and graphics.
Even though the detailed structure of pyPASS Tables varies depending on the type of analyses they are intended, pyPASS Tables have three major components: (1) “attributes” containing metadata for pyPASS Tables such as data sources, (2) the “data description” table holding all necessary information for individual data such as monitor locations, and (3) “data set” tables storing all actual data such as hourly measurements of ozone.

Types of graphics made available by applications of pyPASS are the following:

- Bar charts for peak concentrations and hourly concentration change rates that are spatially paired and temporally unpaired observation-prediction data at ground monitor locations
- Scatter plots for chemical species and wind speed with guiding curves at ground monitor locations
- Time series plots for a single chemical or multiple chemicals spanning user-defined time periods at ground monitor locations
- Hodograms: time series plots for surface winds in polar coordinates at ground monitor locations
- Tile plots for chemical concentrations with optional components such as background maps, lines, and ground monitors, including observed winds and chemical concentration
- Surface wind vector plots with optional components such as background maps, lines, and ground monitors, including observed winds
- Comprehensive time series plot for comparing aloft model predictions with aircraft measurements

For statistical evaluation, pyPASS provides selected statistics as well as values for a 2x2 contingency table to permit users to compute various skill scores. Whenever pyPASS
produces statistical measures, it also reports the number of valid points for each measure. Users can use the number of valid points to judge the quality of statistical measures. The current version of pyPASS estimates several statistics. In addition, pyPASS codes for computing statistics are intuitive so that any reasonably experienced Python programmer should be able to implement calculation of other statistics as desired. Following is the list of statistics that pyPASS provides by default:

- **Mean Normalized Biases**:
  \[ MNB = \frac{1}{N} \sum_{i=1}^{N} \frac{\{C_p(x_i,t) - C_o(x_i,t)\}}{C_o(x_i,t)}, t = 1, 24 \]

- **Mean Normalized Gross Errors**:
  \[ MNGE = \frac{1}{N} \sum_{i=1}^{N} \frac{|C_p(x_i,t) - C_o(x_i,t)|}{C_o(x_i,t)}, t = 1, 24 \]

- **Unpaired Peak Accuracy**:
  \[ UPPA = \frac{C_p(x,t)_{max} - C_o(x',t')_{max}}{C_o(x',t')_{max}} \times 100\%, t = 1, 24 \]

- **Modified Index of Agreement (Legates and McCabe Jr., 1999)**:
  \[ d_1 = 1.0 - \frac{\sum_{i=1}^{N} |C_o(x_i,t) - C_p(x_i,t)|}{\sum_{i=1}^{N} (|C_p(x_i,t) - C_o| + |C_p(x_i,t) - C_o|)}, t = 1, 24 \]

- **Modified Coefficient of efficiency (Legates and McCabe Jr., 1999)**:
  \[ E_1 = 1.0 - \frac{\sum_{i=1}^{N} |C_o(x_i,t) - C_p(x_i,t)|}{\sum_{i=1}^{N} |C_p(x_i,t) - C_o|}, t = 1, 24 \]
Values for 2x2 contingency table; four numbers representing true positive, false positive, false negative, and true negative.

Where, N is the number of observation-prediction pairs used in statistics calculation, $C_o(x, t)$ and $C_p(x, t)$ are the observed and modeled ozone concentration at the $i$th monitor. $\overline{C}_o$ is the mean observed ozone concentration at the $i$th monitor. Among these statistical measures, the Modified Index of Agreement and the Modified Coefficient of efficiency are relatively new measures in the air quality modeling arena. We included these measures because they are designed to remove concerns about using correlation-based measures that are considered improper to appraise the model’s “goodness-of-fit” (Legates and McCabe Jr., 1999).

### 3.3.6 Resource demands of pyPASS

The following information about resources used by pyPASS is intended as a brief reference for future pyPASS users to plan proper storage acquisition and computational cost estimation even though the actual resource demand may be significantly different based on users’ hardware, operating systems, and the configuration for plots.

The size of each pyPASS module code is 5 KB to 50 KB. The memory usage depends on the size of the file used, i.e. the data set for tile plots is much bigger than time series data at monitor locations. For example, in our past analyses the data of time series at monitor locations is 10 to 20 times smaller than data for tile plots. Still, the memory use of pyPASS is significantly smaller than the size of actual air quality model outputs because we only use the extracted data of model outputs for our analyses.

All pyPASS graphics are in Portable Network Graphics (PNG) format that provides lossless compression with high compression efficiency (Roelofs, 2006).
compression may be less efficient compared with other compression methods allowing image losses such as Joint Photographic Experts Group (JPEG or JPG) algorithm. Keeping original image exactly, however, is especially important because we are dealing with quantitative images. Note that the precise comparison of the compression efficiency for the graphics used in MPE needs more in-depth study because there are many compression parameters that are not directly comparable between PNG format and JPEG format. There are additional advantages to using the PNG format: (1) PNG files can be imbedded directly into PDF documents, MS Word documents, web pages, (2) the PNG format supports text inclusion in the PNG file allowing users to write meta-data of a graphical measure such as the name of simulation, comments, type of post-processing, and so on.

Time and storage resources will vary according to individual computer configuration. We operated on a test machine running Windows XP (SP2) and equipped with a Pentium-4 3.2 GHz processor, 2 GB memory, and other necessary parts such as a video card. The actual run time of each pyPASS operation was measured as clock time. The elapsed time of each run varies with the type of graphics and how much information is stored in a plot. For an example test, sets of concentration tile plots were made. Each tile plot is 408 pixels by 408 pixels with surface wind vector fields and observed winds and chemical concentration. We chose tile plots as an example because they take the longest run time among pyPASS graphics. We ran pyPASS to create 660 benchmark plots and repeated this run three times. pyPASS took an average of 154 seconds as an average of three runs on the test machine, i.e. 0.23 second per benchmark tile plot. Storage use of individual graphics also varies by their contents and resolution. For example, the actual size of a benchmark plot varies from 70 KB to 100 KB while one bar plot of 880 pixels by 350 pixels needs 5 KB. In contrast, PAVE
(CMAS, 2005) would take about 30 minutes to produce a similar number of tile plots that only contained the chemical signals (i.e., no wind data).

3.4 Illustrative examples

In this section, we present various types of graphics and explain important differences of pyPASS outputs compared with the typical graphics found in SIP modeling. Note that our intention is to introduce what pyPASS can produce and not to describe how to conduct an actual performance analysis; that is the subject of another paper (Kim and Jeffries, 2006a). Therefore, we focus on discussing the graphics and the associated features. We present pyPASS graphics in the order of possible analytical steps and if necessary make some comments on the material used in generating graphics. The dataset used in generating the following graphics is primarily from Houston-Galveston Mid-Course Review (HGMCR) modeling conducted as part of developing the Texas SIP. The CMAQ output used for creating Figure 3.2, however, was provided by the University of Houston. For details of the model configuration and other information, please refer to the Texas Commission on Environmental Quality web site (TCEQ, 2006b) and the University of Houston’s site (IMAQS, 2006).
Figure 3.2. An example of spatially paired and temporally unpaired daily peak ozone bar chart. The predicted ozone data were made by TCEQ’s HGMCR modeling with CAMx (red) and University of Houston’s modeling with CMAQ (blue) for 2000-08-25. Monitoring sites are sorted by location from west to east of HGMCR modeling domain. The x-axis label denotes the four-letter site codes and the y-axis shows ozone concentration along with one-hour ozone NAAQS (depicted as the purple line) with the label ‘Exceedance’.
3.4.1 Using pyPASS for site-by-site and day-by-day analysis

Figure 3.2 shows a bar chart of spatially paired and temporally unpaired peak ozone concentrations at all monitor sites on 2000-08-25. Note that the monitors are sorted from west to east to give a sense of possible spatial discrepancies of model prediction. Users can choose to change the sorting direction easily if west-east is not the directional tendency that users want to examine. While two different model outputs are plotted in Figure 3.2, pyPASS can deal with an arbitrary number of cases. This specific figure shows predictions made by two modeling systems, CAMx and CMAQ, using the same emission inventory and meteorological model output on same grid configuration. Possible roots of differences in predicted ozone concentrations between these two models are (1) model’s internal representation of some environmental processes such as vertical diffusion, (2) processors used to create “model-ready” inputs from emission models and meteorological models. Therefore, this kind of plot can provide a “quick” screening when modelers want to do “head-to-head” comparison.

In Figure 3.2, both of the models under-predicted ozone at most of the monitor locations, especially where ozone violations occurred. CMAQ showed much lower daily 1-hour peak ozone concentration at almost all sites, as compared to CAMx output. Given that most of inputs for the CMAQ simulation are same as the CAMx simulation, it would be interesting to investigate this difference. This is beyond the scope of this paper. More in-depth analyses on comparison of these two model simulations will be designed in a subsequent paper. We instead focus on the CAMx simulation results to illustrate how pyPASS can help the modelers enhance their MPE. CAMx showed false negative results at six sites, i.e. the model predicted ozone concentrations lower than the NAAQS while observed ozone concentrations
were over the NAAQS. Those sites are HALC, HSMA, C35C, HOEA, DRPK, and H07H. The modelers or evaluators may want to look at the meteorological predictions and observations at one of these five sites further.

**pyPASS** provides three major graphics for winds: wind speed scatter plots, wind hodograms, and wind error hodograms. These three types of plots for a site should be “packaged” to provide a more comprehensive understanding. A wind speed scatter plot for HALC on 2000-08-25 is shown in Figure 3.3. This specific graph shows several important points about the modeled winds: (1) they were slow from 0700-0800 while the observed winds were ~ 4 km per hour for the same period, and (2) the modeled winds were more than two times faster than the observed wind staring from 1900 on 2000-08-25. In addition, another important phenomena revealed by this plot was that both observed winds and predicted winds were slower than 8 km per hour until 1400. Given that the model grid was 4km, the approximate wind speed was slower than 2 grid cells per hour. Even though the scatter plot does not show wind direction, it helps users to screen what time period and what range of wind speeds that need further examination with a hodogram for wind direction errors.
Figure 3.3. Example of a wind speed scatter plot at a single site. X-axis is the observed wind speeds and Y-axis is the predicted wind speeds in units that reflect the size of the model grids (e.g. 4 km on each side). The magenta spline curve is to help modelers track each data point sequentially. Daylight hours are numbered in LST. Diagonal lines represent 2:1 (dotted), 1:1 (solid), and 1:2 (dotted), correspondence between predictions and observations. Four different colors are used to distinguish four time periods of a day: 0000-0600 (“Midnight to Morning”), 0700-1200 (“Morning to Noon”), 1300-1800 (“Noon to Evening”), and 1900-2400 (“Evening to Midnight”). Left facing and right facing triangles represent pairs of the observed and modeled vector resultant wind speeds in the morning and afternoon.
A hodogram by pyPASS is shown in Figure 3.4 (a). This plot is made from the same speed data used in Figure 3.3, but with wind direction added. It is now clear that the winds from 1300-1600 are not in good agreement in direction even though wind speeds were similar. Also, this plot gives an idea that the dominant wind direction of observations and predictions are not the same. For example, the observed winds for 0700-0800 was easterly when the model winds were very weak compared to the observation. If there is important emission sources near distance within 4-8 km, this difference can cause differences between the prediction and the observation for the time period. Figure 3.4 (b) shows the wind error hodogram based on the same data used in Figure 3.4 (a). This site does not show good agreement for approximately half of the day. The screening criteria for roughly reasonable agreements were ± 50% of wind speed and ± 30 degrees for wind direction. During two nighttime periods, 0000-0600 and 1900-2400, the predicted winds are different from the observed wind speeds. Unless there is a separate report on the uncertainties or imprecision of observed wind, observed wind data in this figure need to be considered to meet the US EPA’s standard (US EPA, 2000); that is, ±0.2 m/s + 5% of observed wind speed and ± 5 degrees for wind direction. Therefore, the criteria we adopted in this example were not smaller than the expected observational error.
Figure 3.4. An example of graphical measures for surface wind analyses: (a) hodograms (i.e. wind time series plots) and (b) wind error time series plots. For hodograms, the origin is the station locations, the radial axis shows wind speed in km/h, and the angular axis shows wind direction in degrees with 0 degrees as winds from north. Each filled circle is the start point of a wind vector with its arrow head at the origin where the station is located (the vectors are not drawn, but the “tails” are connected by a colored line). Magenta is for observations and cyan is for predictions. Triangles are for resultant winds; left-facing one is for morning and right-facing one is for afternoon. Some hours of day data points are labeled with its hour of day such as “13”. For wind error plots, the radial axis is the ratio of modeled wind speeds to observed wind speeds on a log scale. The angular axis is the degree of wind directional differences between predictions and observations. The green circle is a ratio of 1.0 or when observed wind speeds are same as predicted wind speeds. In an ideal case, all points are located at the cross point of 0 degree line and the green circle.
It is often hard to judge whether these differences are important without examining chemical signals. An analysis to explore whether the given wind differences are important will be quite complex and needs investigation at each site on each day. To assist chemical signal analyses, pyPASS provides two types of graphics suitable for site-by-site and day-by-day analyses: scatter plots and time series. Scatter plots are similar to examples often used in traditional MPE and will not be shown or discussed here.

Figure 3.5 is the time series plot of $O_3$, NO, and NO$_2$ concentrations for HALC on 2000-08-25. Users can specify the maximum value on the Y-axis and the time span used for the X-axis. Different chemicals are shown in different colors and symbols. Observations are always plotted with solid lines while model predictions are all shown in different types of dashed lines (more than one simulation sets of predictions can be included.) Combined with scatter plots, we could identify that significant NO underestimation occurred in the morning before 0700. The NO$_2$, however, was overestimated for the same time period. Ozone was also significantly underestimated for the 1300-1400 period when the ozone peak was observed.
Figure 3.5. An example of a time series plot for chemical concentrations. Solid lines are for observations. Dashed line represents prediction. Three chemicals are plotted with different colors. This plot clarifies when the large underestimation of NO that can be found in scatter plots occurred during the early morning. It is clear that there was also a large ozone underestimation for 1200-1600.
Two hours of 2000-08-25, 0700 and 1300, should draw modelers’ attention to examine predicted winds from the wind plots. At 0700, it was clear the predicted winds were very calm while the observed winds maintained speeds of over 4 km per hour. The model and observations do show good agreement after six hours of gross discrepancies. On the contrary, at 1300, the model showed relatively good agreement in wind direction even though the wind speed is about 60% of the observed wind. At this point, modelers may want to examine the emission conditions around this monitor and spatial distribution of chemical concentrations to check if the 60% wind speed error can likely cause a problem.

### 3.4.2 Comprehensive tile plots

One of the major advantages of pyPASS over existing tools is the ability to create tile plots containing predicted surface wind fields and observed wind and chemical concentration on one plot. An example of a pyPASS tile plot is shown in Figure 3.6. This plot shows a “snap-shot” of model behavior and what was observed in the real world. Compared with tile plots from other tools, pyPASS tile plots are unique in MPE processes because it includes all available information and depicts all graphical elements in physically meaningful ways. For example, the size of each tile is the size of grid cell that a model ran even though the plot domain and grid can be any user-specified domain and grid. Figure 3.6 displays spatial ozone and winds distribution of 4-km simulation and of observation in 1-km domain. Users can put individual snap-shot tile plots together to create animations or to produce time series of tile plots to examine the temporal changes in the model and in the observation.
Figure 3.6. Example of pyPASS tile plot. This type of tile plots is unique pyPASS outputs that are not available by existing tools. Bottom X-axis and left Y-axis are for x and y coordinates in projected coordinate system. Top X-axis and right Y-axis are for cell coordinates. These axes are optional and users can produce a tile plot without these auxiliary axes, if necessary. All monitors are represented as diamonds filled with different colors depending on the observed ozone concentrations and arrows that have a length equal to the distance of wind traveled for an hour. If there is no observed wind and no chemical measurement, diamonds are replaced with two different sizes of circles superimposed. If there is no observed wind at a monitor, an arrow is replaced with a circle. If there is no measured chemical concentration at a monitor, the diamond is replaced with a circle. The tile plot also holds important geographical features such as highways and coastal lines. Predicted winds are drawn with light grey arrows.
A collection of tile plots in time is critical when evaluators often need to inspect
dynamic aspects of model behaviors. Figure 3.7 is an example of tile plot time series that
depicts ozone distribution over part of 4km grid domain from 1300 to 1600 on 2000-08-25.
The area used for Figure 3.7 is the same area used for 1-km grid simulation, i.e. users can
make a tile plot for any focus area while preserving the model’s grid resolution. Note that
\texttt{pyPASS} represents the wind fields by incorporating the nature of data. For example, actual
wind inputs for RPAQM can be instantaneous values for each hour. Because the observed
winds are hourly averaged values, \texttt{pyPASS} computes hourly averaged winds internally to
make wind field plots be consistent with the site observation. In addition, the hourly travel
distances of wind vectors match the length of grid size. This area experienced vary rapid
intrusion of SE winds from Galveston Bay that flushed out ozone clouds in the model while
several monitors in downtown Houston still observed high ozone.
Figure 3.7. Series of tile plots for $O_3$ with predicted surface winds as well as observed winds and ozone. All time in this plot is hours of 2000-08-25 in LST.
The HALC site is the upper-leftmost site in tile plots of Figure 3.7. It is clear that HALC experienced rapid ozone concentration change for 1300-1500 with winds from east or south-east while the model did not capture that ozone plume. Figure 3.7 also revealed that there was sharp spatial gradient of ozone concentration between monitors on western Houston and monitors on eastern and southern Houston around 1300-1400. The CAMx simulation, however, could not simulate this ozone behavior. This deficiency is obvious at the H04H site (top-middle diamond with green color in the 1400 tile plot of Figure 3.7). At this point, evaluators have several options to proceed: moving onto follow-up analyses of H04H site or conducting more in-depth analyses of HALC site.

3.4.3 Integration of pyPASS with Geographical Information Systems

pyPASS can generate all plots with a transparent background, which is helpful when users want to overlay and stack multiple plots over GIS maps. Figure 3.8 is an overlaid plot of Figure 3.4 (a) on a GIS raster map showing important emission sources. The GIS raster map is scaled to match the number of pixels used in the radial axis of Figure 3.4 (b), i.e. the number of pixels between grid lines representing the 4km modeling domain is same as the number of pixels for the 4km/h interval on radial axis of Figure 3.4 (a). By matching pixel size, users can perform quick quantitative wind analyses.
Figure 3.8. An example pyPASS application to GIS maps. This example is the result of overlaying Figure 3.4 (a) on a GIS map containing important emission sources (normal triangles) and monitors (squares with four-letter site codes). Other geographical features included in this figure are area sources such as airports (gray filled polygons), major roads (light brown and yellow), water bodies (sky blue).
The most obvious information gleaned from Figure 3.8 is that the HALC monitor is located near high-traffic roads. Given that the modeled winds (cyan line) were slower than observed winds (magenta line) around 0700, the real world winds could bring an airmass from eastern roads nearby. The coincidental NO agreement for 0700 was likely due to biases in combined contribution of various processes in the model compared with the real world. These deviations may be explained in a variety of ways. Physically, it might be the combination of higher mixing with slower advection in the model. Chemically, there might be more NO emissions in the model but they were oxidized by ozone: predicted NO₂ concentrations were twice the observed NO₂ at 0700 while observed and predicted O₃ concentrations were similar. For precise analyses, we need to use more diagnostic tools, such as Process Analysis (PA) (Jeffries and Tonnesen, 1994; Tonnesen and Jeffries, 1994; Jang et al., 1995). These examples are but a small part of work recommended in the PROMPT protocol. For more comprehensive discussion, see the article of PROMPT application (Kim and Jeffries, 2006a).

3.4.4 Comparison of model predictions with aircraft measurements

Recently, some studies using three dimensional air quality models have showed an MPE effort to use aircraft measurements for performing part of MPE because aircraft measurements can be used to examine model performance aloft (Brunner et al., 2003; Jiang and Fast, 2004). A common issue in comparing aircraft observation with model prediction is reconciling spatial and temporal resolution of model predictions and observations (Svensson and Klemm, 1998). The choice of specific data reconciliation methods often depend on the resolution of model outputs (Brunner et al., 2003). At present, a common practice in air quality modeling studies is a linear interpolation in time for model outputs with no spatial
interpolation (Svensson and Klemm, 1998; Fast, 2005). Two previous studies, one by Brunner and the other by Jiang and Fast, used models that produce instantaneous predictions every 30 minutes. There were no clear indications whether these two studies also examined the meteorological inputs aloft with aircraft measurements.

Matching the resolution of model predictions consistently between two models such as CAMx and CMAQ is problematic. Choosing a temporal interpolation scheme is complicated by the fact that CAMx produces hourly averaged outputs while CMAQ provides instantaneous values at each hour. Outputs of both models are averaged values over a cell volume. It is feasible to produce hourly averaged outputs from CMAQ outputs to match the resolution of CAMx outputs, but it is not desirable to do the other way, i.e. derivation of instantaneous outputs from CAMx outputs. Therefore, we decided to use hourly and cell-volume averaged model values to create aircraft plots.

In addition, there are complications in using altitudes of aircrafts estimated by Global Positioning System devices based on Mean Sea Level from an earth model. Frequently, most GPS devices use the WGS84 as a default earth model when they report altitudes. Model cell heights used by both CAMx and CMAQ, however, are based on Above Ground Level. Therefore, when pyPASS estimates which layer aircraft tracks are in, pyPASS corrects predicted cell heights by adding surface elevation data to them after matching horizontal grid resolution of elevation data with model’s grid resolution and then compare the corrected cell heights with GPS readings made by aircrafts.

An example aircraft plot is shown in Figure 3.9. The aircraft measurements used for this plot were from a NOAA aircraft that was operated as part of TexAQS 2000 (Daum et al., 2005). The original aircraft measurements of gaseous species shown in the figure were made
at every second. Users can choose different time intervals for averaging point measurements by aircraft to match the resolution of model prediction. For this example, we set the interval as 40 seconds because the aircraft flying speed was about 100 m/s and flying distance for 40 seconds was comparable to the model’s horizontal grid resolution, i.e. 4km.

What is clear in Figure 3.9, when the aircraft passed over downtown Houston, the model underestimated ozone significantly (over 50 %), especially around 1341. At the same time, CO, NO, and NO₂ were overestimated by factor of 2–3 or more. Given that the wind measurements only represent very short-term wind speed and direction, it is difficult to directly compare model (predicted) winds with aircraft observed winds. Both observed winds and predicted winds were directionally similar from the SE, upwind from a major road. Therefore, modelers may want to investigate some possible causes of the large chemical discrepancies such as on-road emission problems or mixing height problems. This kind of aircraft plot can help clarify how a model performs aloft and design the next phase of analyses.
Figure 3.9. Example of comparison of aircraft observation with model predictions. The top plot is the pyPASS output for comparing aircraft measurements with model predictions. The bottom plot was made from a GIS map and a part of Flying Data Grabber outputs (McNally, 2005). The black and red arrows in the bottom plot indicate observed winds and predicted winds used for modeling at sampling locations of aircraft flight path. The purple box is the time window corresponding to the purple circle in the bottom map.
3.5 Summary and future improvement

3.5.1 Summary

With pyPASS, modelers can do a more comprehensive analyses outlined in MPE procedures mandated by the PROMPT approach. Modelers can holistically identify which parts of the model need more attention because pyPASS provides a variety of information based on all available data. In summary, pyPASS can help enhance the quality of MPE practices for SIP modeling:

- pyPASS creates information that is necessary for a PROMPT-like MPE that could otherwise not be available with existing tools; modelers can view information including geographical features in more integrated ways than past MPE practices.

- It improves the quality of information used in MPE for SIP modeling; the resolution of data is preserved to prevent ‘dequantification’ of graphical measures and all graphical measures are designed to convey focused information specifically for MPE.

- It reduces the resource demands to conduct MPE practices; selective data extraction permit evaluators to carry observational data and several modeling results in a personal computer. Object-oriented design of pyPASS graphics ensures good computational performance and reusability of codes. Additionally, the on-line file compression decreases the storage use more than 30 %.

- The user interface for operation offers efficient information production; users can do various comparative analyses such as observation vs. a single simulation, observation vs. multiple simulations, or a base simulation vs. multiple sensitivity runs efficiently with sets of “run-time options”.

124
• It utilizes observations from non-routine ground monitors; aircraft observation can be compared with model predictions and users can match the model’s data resolution with the observational data.

• It provides more guided information, as shown in our illustrative example, to clarify the priority of MPE tasks; PROMPT-like MPE emphasizes “progressive” analyses that require well-defined sequence of evaluation procedures. Because pyPASS was designed to support PROMPT-like MPE, pyPASS can help modelers implement their advanced MPE effectively.

• It ensures flexibility in improving modules; all source codes are open to public and the cost of using pyPASS is very low or free. Thus, any peer-reviewer can examine possible issues in codes and modify source code to improve pyPASS.

3.5.2 Future improvement

Even though pyPASS has many advantages over traditional MPE tools, the current pyPASS can still be improved to perform more extensive analyses. Following is the list of planned improvements for the next version of pyPASS.

First, there are increasing attention to the use of ‘probes’ such as O$_3$-to-NO$_X$ for diagnostic model performance investigation (Arnold et al., 2003). These probes can provide diagnostic information more than plain concentration plots. Displaying these probes from observation and model predictions is not yet implemented in the current version of pyPASS. It would also be desirable to support various command-line computations such as calculating the differences of two model simulations. These functionalities can be realized easily since Python provides important libraries such as a parser.
Second, the current version of pyPASS has a limit in the type of high resolution measurements it can use. For example, it does not handle profiler observation and Lidar (Light Detection and Ranging). These types of observation have great potential to increase the quality of MPE by providing information aloft in detail. The most difficult issue in implementing pyPASS graphics to utilize these non-routine data are that the spatial and temporal resolutions of the observations. These vary depending on specific measurement techniques with sparse documentation on how to report the observed data properly. Significant amount of studies may need to be conducted to use high resolution data properly for SIP MPE. For example, selecting model grid cells for visualization of vertical slices is quite challenging. Additionally, a good interpolation algorithm is needed for these high resolution data points.

Third, pyPASS is currently unable to incorporate emission information. A graphical measure for emission input inspection can be produced such as tile plots of emission intensities from emission input files directly. Then, we can overlay transparent hodograms at monitor sites on emission plots to evaluate potentially problematic locations due to biases of predicted surface winds in the model without actually running simulations.

Fourth, the current aircraft plots lack information with regard to VOCs. A major issue in observed vs. predicted VOCs comparison is that canisters sampling times are longer than the time required by aircrafts moving across model grid cells. Solution to this issue is under investigation.

Fifth, we need an improved visualization of statistical measures so that modelers can better interpret the statistical measures with pyPASS standard graphics. Additionally, we
only estimate these statistical measures for each site at present, but sub-region analyses can provide more insights in spatial biases of model predictions.

Last, we need to develop a pyPASS module to embed texts in PNG files directly. Even though PNG format supports embedding of metadata in the ASCII format or some values such as statistical measures corresponding to graphs in PNG files, the direct text writing in PNG functionality is not yet implemented in pyPASS. Also, displaying those texts with proper formats will improve the amount and quality of information carried with a pyPASS plot. Users could examine and acquire graphical information, statistical information, and information metadata from a single graphic file. Currently, users may utilize external software such as ImageMagick, but it is desirable to implement the comment addition functionality in the future version of pyPASS.

Acknowledgement

This project was partially funded by the Houston Advanced Research Center through Contract H12 and partially funded by the 8-Hr Coalition through Contract HC1. We wish to thank Dr. Jim Smith at Texas Commission on Environmental Quality for sharing modeling inputs and outputs of TCEQ’s HGMCR modeling for 2000 and Dr. Daewon Byun at the University of Houston for files of the CMAQ simulation equivalent to the TCEQ HGMCR simulation. We appreciate the extraction code work supported by Dr. Robin Dennis and Mr. Todd Plessel at High Performance Computing Group at US EPA/NOAA, RTP. We also thank to Dr. Tom Tesche and Mr. Dennis McNally for sharing their Flying Data Grabber code and outputs as well as testing CAMxSubset and CMAQExtract on their platforms.
4. ALTERNATIVE PERFORMANCE EVALUATION OF THE HOUSTON-GALVESTON MID-COURSE REVIEW PHASE 2 MODELING

*Atmospheric Environment*, To be submitted.
Abstract

In the regulatory air quality modeling community, the problem of weak model performance evaluation (MPE) is significant but few studies have been conducted to resolve this issue. We recently proposed an improved MPE methodology for regulatory ozone modeling. The purpose of this study reported here is to illustrate advantages of the new MPE method by re-evaluating the performance of the Houston Galveston Mid-Course Review (HGMCR) SIP modeling as a case study. Here, we attempted to answer two questions: (1) To what extent can we accept the predictions made by the HGMCR models at face value for the purpose of developing a Texas 2000 State Implementation Plan (SIP) for ozone problems at HG? and (2) If we cannot, then how should policy makers make judgments about the effectiveness of ozone control options that were proposed in the Texas 2000 SIP? To answer these questions, we developed a specific MPE protocol for the HGMCR modeling and utilized performance analysis tools designed and developed to implement the new MPE method. For the first question, we concluded that the HGMCR modeling showed significantly low reliability for developing and testing ozone control options. Four major reasons are: (1) the precision of meteorological inputs used for the model simulation were inadequate with respect to the resolution anticipated by the ozone problem in HG, (2) the model showed very high nitrogen oxides biases; that is, the model frequently overestimated nitrogen oxides concentrations by twice the amount observed, (3) the over prediction of volatile organic compounds (VOCs) concentrations was significant such that some were predicted ten times higher than observed, and (4) in spite of having available nitrogen oxides and ample VOCs, the model almost always underpredicted ozone especially for the highest observed times and locations. For the answer to the second question, additional modeling
sensitivity studies and other “weight of evidence” based on observations and the conceptual model must be considered along with the SIP model results. As an important model sensitivity study, here we consider the predicted effects of highly reactive VOC event emissions evaluate the efficacy of the short-term HRVOC rule in the Texas 2005 SIP.
4.1 Introduction

In 1990, the United States Environmental Protection Agency (US EPA) classified the 8-county Houston-Galveston-Brazoria area as a “severe” non-attainment area of the National Ambient Air Quality Standard (NAAQS) for 1-hr ozone (TCEQ, 2006a). For more than 15 years, the state of Texas has invested significant resources to develop a State Implementation Plan (SIP) for the Houston-Galveston (HG) to alleviate its ozone problems. The Clean Air Act Amendments 1990 (CAAA 1990) requires the state of Texas had to demonstrate future ozone attainment in a proposed SIP based on three-dimensional photochemical air quality models (PAQMs). Because these models can simulate complex interaction of various factors such as emissions and meteorological processes, they are considered the most effective tool to explore causality of ozone problems in a region (National Research Council, 1991; Russell and Dennis, 2000).

Applications of PAQMs to SIP development, however, often result in problems that can lead to unreliable model predictions (Fine and Owen, 2005; Roth et al., 2005). Ozone is a secondary pollutant formed by reactions of two major precursors, nitrogen oxides ($\text{NO}_X$) and volatile organic compounds (VOCs). Therefore, controlling ozone is essentially a matter of controlling these precursors. Flawed modeling results can mislead state modelers and policymakers to support ineffective (or, worse case, directionally wrong) control options, e.g. $\text{NO}_X$ control when VOCs controls are needed, or result in significantly delayed implementation of control strategies.

To meet the statutory requirement, the Texas Commission on Environmental Quality (TCEQ), formerly known as the Texas Natural Resource Conservation Commission, undertook a series of modeling efforts beginning in 1995 and extending through 2005.
During the course of SIP development, TCEQ has made several SIP revisions as requested by US EPA for improvement. Often, a SIP revision resulted in a change to a different modeling system and/or an alternative episode selection to improve the quality of modeling results.

In December 2000, TCEQ proposed a SIP revision based on the September 1993 episode using the Comprehensive Air quality Model with eXtensions (CAMx) version 2.0. TCEQ attempted to simulate five other episodes in 1993 that were within the period of a field study, the Coastal Oxidant Assessment for Southeast Texas (COAST) that was believed to provide ample observation data for testing and improving model performance. Unfortunately, the model performance for all of the episodes within the COAST study period failed EPA’s three statistical tests for acceptability. The last episode tried was the three day September 1993 episode that occurred just after the COAST study; this one could pass the three EPA statistical tests. For the SIP submission due in 2000, TCEQ claimed that the 1993 model passed the performance tests satisfactorily. Even though the US EPA agreed with TCEQ that the performance of 1993 model was acceptable, many questions were raised by an external peer-review sponsored by the Business Coalition for Clean Air Appeal Group (BCCAAG). The BCCAAG’s study showed that the model’s peak ozone performance was the result of a compensating error over a spatial difference of 55-km and showed that the model failed to reproduce the most important observed ozone characteristic, short duration (often one hour or less) very high ozone at mostly a single monitor at a time. This characteristic was present at every exceedance in the September 1993 observations but was never present in the model. Furthermore, the future case modeling predicted a peak ozone of 145 ppb, in spite of the use of every available NOX reduction that TCEQ could conceive to apply. TCEQ was eventually
lead by EPA Region VI to use a non-linear roll back procedure to estimate additional \(\text{NO}_X\) reductions based on curve fitting to arbitrarily chosen “across the board” \(\text{NO}_X\) category reductions. This additional 45 tons of needed \(\text{NO}_X\) reductions became known as the “\(\text{NO}_X\) gap” and was introduced into the 2000 SIP as a “commitment to be met by mid-course”, that is, by 2005.

Despite these issues of model performance, the TCEQ submitted a SIP that consisted of minimal \(\text{VOCs}\) controls proposed in the previous SIP revisions and added a new 90 % industrial \(\text{NO}_X\) reduction by 2006. Based on its own modeling studies, the BCCAAG challenged the proposed SIP in the Travis County District Court in 2001. After a five day bench trial, the TCEQ, by a mutual consent-decree, agreed (1) to insert into the SIP alternative \(\text{NO}_X\) emissions reductions tables based on 80% industrial point source controls and (2) to conduct by the mid-course correction date additional analysis and modeling to investigate the BCCAAG’s claim that highly reactive \(\text{VOCs}\) were responsible for the highest ozone observed and that selective VOC controls would be more effective than \(\text{NO}_X\) controls. Furthermore, the TCEQ agreed make the analysis and modeling activities more public and with greater stakeholder involvement. Specifically, they were to post the analyses and modeling files on a public web site and to supply model files to interested parties.

While these actions were dictated by the Texas Courts, the EPA was proposing to accept the TCEQ 2001 SIP as it was submitted. Several parties including the BCCAAG appealed EPA’s proposed approval to the 5th Circuit court by questioning EPA’s approval of TCEQ’s attainment demonstration based on the 1993 modeling. The 5th Circuit court denied all petitions by stating that TCEQ had followed EPA’s guidance (5th Circuit, 2003). As discussed above, TCEQ admitted that the 1993 model showed low reliability in its
predictions. In summary, EPA approved an attainment demonstration proposed by TCEQ with the 1993 model and Courts did not see reasons to revoke EPA’s approval because TCEQ followed the EPA’s modeling guidance including the model performance evaluation (MPE) protocol. Yet, TCEQ admitted that the 1993 model showed large uncertainties to be used in developing an effective SIP. Interestingly, all past HG SIP modeling studies, including the 1993 modeling, was subject to MPE that followed the MPE protocol approved by US EPA. At the end of the 18 month consent decree, the TCEQ chose to switch to the 80% point source NO$_X$ reduction and to undertake new VOC reductions based on the often presence of high concentrations of highly reactive VOCs (HRVOC, i.e., ethane, propene, butanes, and butadiene) in the Houston atmosphere.

The most recently adopted 2005 SIP by TCEQ is the result of the mid-course review (MCR) modeling effort that was based on another intensive field study, the Texas Air Quality Study 2000 (TexAQS 2000), which was conducted at the same time that the 2000 SIP modeling was being done. In this Houston-Galveston Mid-Course Review (HGMCR) modeling effort the TCEQ dropped the September 1993 modeling totally and undertook a new August-September 2000 episode using with CAMx version 4.0. The HGMCR modeling was also evaluated with a MPE method following the EPA guidance. In the newest SIP revision, TCEQ adopted a new rule that is composed of an 80% NO$_X$ reduction and highly reactive volatile organic compounds (HRVOCs) controls. Other control measures and rules in the 2001 SIP that were based on the 1993 model, such as the highway speed limit strategy were removed during this transition (TCEQ, 2006a).

As an independent modeling workgroup, we participated in several projects related to assessing and improving the HGMCR modeling. These projects provided us with
opportunities to examine the HGMCR modeling closely. Two primary weak aspects of the HGMCR modeling that we found were: (1) TCEQ did not take full advantages of PAQMs for the SIP development even though PAQMs can be operated to reflect area-specific conditions such as spatial and temporal variability of emissions, and (2) no significant efforts were made to detect and eliminate compensating errors. In other words, the HGMCR modeling was not evaluated rigorously with respect to the complexity of the ozone problems in HG. Recent studies on SIP MPE practices, however, showed that weak MPE practice was a common problem for essentially all past SIP modeling based on the EPA’s MPE protocol (Roth et al., 2005; Fine and Owen, 2005). That is, while TCEQ modelers struggled with the model performance by following EPA’s guidance, there was no good alternative MPE approach that they might follow to improve their MPE processes. We recently proposed an improved MPE methodology for SIP modeling (Kim and Jeffries, 2006b). The purpose of this study is to illustrate advantages of our MPE method by re-evaluating the performance of the Houston Galveston Mid-Course Review modeling as a case study.

4.2 Performance evaluation methodology

The MPE protocol used for this study is an instance of our new protocol class, the Protocol for Regulatory Ozone Model Performance Tests (PROMPT). PROMPT is a meta-protocol that assists modelers in constructing a particular MPE protocol for a specific SIP application, and we used it to construct the MPE protocol for this study. The MPE protocol used in this study was designed to be suitable for re-evaluating the HGMCR modeling case by external evaluators. Note that this protocol is a “snap-shot” protocol of evolving protocols that change with additions of new knowledge and new findings. Therefore, the MPE protocol used for this study only reflects our thinking and understanding of the
HGMCR modeling at present. Hereafter, we call the MPE protocol used for this study as the Protocol for Re-Evaluating HGMCR modeling (ProHGM).

The ultimate goal of the ProHGM application is to answer the following two questions: To what extent can we accept the predictions made by the HGMCR models at face value for the purpose of developing a Texas 2000 SIP for ozone problems at HG? (SIP-Q1), and if we cannot, then how should policy makers make judgments about the effectiveness of ozone control options that were proposed in the Texas 2000 SIP? (SIP-Q2)

We believe that SIP-Q1 and SIP-Q2 cannot be answered by using the EPA’s current MPE protocol. Essentially, these two questions can be summarized as a desirable question that should be asked by policymakers: ‘Why should I believe this modeling?’ (Fine and Owen, 2005). By following the instructions of PROMPT, ProHGM consists of four evaluation phases. Each phase aims at answering questions shown in the first column of Table 4.1. To answer the specific question posed for each evaluation phase, we developed several evaluation tasks. The summary of major evaluation tasks and subtasks for each evaluation phase is listed in Table 4.1. Note that some of the subtasks listed in the table may be beyond what can be done in practice or the content of the list may not be sufficient to achieve the MPE goals fully, depending on the results of each evaluation phase. Given that the purpose of this study is the demonstration of PROMPT implementation and applications, we focused on describing what we would conduct and how we would proceed based on our findings. Therefore, this article is not necessarily a “final report” of MPE for the HGMCR modeling.
Table 4.1. Summary of the Protocol for Re-Evaluating HGMCR modeling.

<table>
<thead>
<tr>
<th>Evaluation Phase #</th>
<th>Major task #</th>
<th>Subtasks (selected)</th>
</tr>
</thead>
</table>
| P1: Does the HGMCR modeling show or have all necessary components to produce the phenomena that we can expect from the current best conceptual model? | P1.1: Document the review of HGMCR modeling system setup | • Locate the information about how all the model input files were prepared  
• Identify possible sources of input uncertainties by modeling processes  
• Review the model configurations described in the TCEQ’s protocol including grid resolutions |
|                    | P1.2: Compile available observations and MPE tools for the rest of evaluation phases | • Create a table showing available observations at each ground monitor  
• Make detail notes about any specific aspects of monitors or measurements as needed  
• Select MPE tools and justify the rationale |
|                    | P1.3: Compare Base5b.Psito2n2 with the conceptual model with respect to the ozone behavior | • Review the characteristic of meteorology, ozone behavior, and emissions described in the conceptual model  
• Examine the model’s behavior by analyzing (1) morning and afternoon resultant winds, (2) daily peak ozone plots, (3) daily peak NO and NO₂ plots, and (4) ozone time series plots |
| P2: Can the HGMCR modeling distinguish which precursor(s) to control for ozone reduction? | P2.1: List areas that will be affected by proposed control policies | • Consult with TCEQ’s control option developers and/or look up the proposed control options in SIP  
• Discuss with policy developers what we find in P1  
• Clarify whether another episode selection or other alternative modeling can be worth |
|                    | P2.2: Examine surface winds at sites in the areas identified in P2.1 | • Examine surface winds  
• wind speed scatter plots  
• hodograms  
• wind error plots |
|                    | P2.3: Examine inorganic chemical signals at sites in the area identified in P2.1 | • Examine chemical signals at monitor sites by analyzing scatter plots and time series of NO, NO₂, and O₃ |
|                    | P2.4: Examine VOCs (plus CO, if available) at sites in the area identified in P2.1 | • Review the results of P2.3 at the sites available for P2.4  
• Examine chemical signals at monitor sites by analyzing scatter plots and time series of CO, ETH, OLE, ALD2, FORM, and ISOP |
|                    | P2.5: Assess model performance at each site on each day based on results of P2.2 through P2.4 | • Assess the usability of model explicitly by using one of following four categories: “None”, “NOₓ only”, “VOCs only”, or “NOₓ and VOCs” |
Table 4.1. Summary of the Protocol for Re-Evaluating HGMCR modeling (continued).

<table>
<thead>
<tr>
<th>Evaluation Phase #</th>
<th>Major task #</th>
<th>Subtasks (selected)</th>
</tr>
</thead>
</table>
| P3: How precisely can the HGMCR modeling estimate control requirements? | P3.1: Communicate with policy developer about their concern on model’s performance | • Provide tables for each day showing the status of model performance  
• Decide which days are worth for further analysis based on discussion with policy developer |
|                    | P3.2: Conduct comprehensive performance analysis at observed locations for selected days | • Answer the following questions for each day selected in P3.1  
• If predictions generally match history, is there any way this might be due to compensating errors?  
• If predictions do not match the history, what are the likely causes of the failure?  
• Divide monitor locations into three categories: “Reliable”, “Ambiguous”, or “Unreliable”  
• Visualize the model’s inputs for the important processes in the area of interest  
• Performing dispersion simulations for selected emissions without chemistry to determine how various sources are contributing to the focus area.  
• Perform process analysis of the focus region to visualize and understand the interaction among the physical and chemical processes and to explain the state of the chemical transformations.  
• Conduct selected sensitivity analyses  
• Divide locations into three categories: “Reliable”, “Ambiguous”, or “Unreliable”  
• Conduct traditional statistical evaluations for days qualified for P3 and compare them with those days not included in P3  
• Make comprehensive assessment by integrating the results of P3.2 and P3.3 with statistical tests  
• Document issues found in P3 and suggest further model performance improvement |
|                    | P3.4: Assess model performance analyses at locations examined in P3.2 and P3.3 | • Present overall assessment by developing GIS maps showing evaluator’s confidence on model performance  
• Create publicly accessible documents that contain our judgment and information (or location of information) used in MPE  
• Discuss with policy developers (1) if resources and statutory timeline may allow efforts to improve model performance and (2) whether these efforts are worth given that the precision demands of proposed policy options  
• Make recommendation such as pursuing alternative episode or changing modeling system |
One of the most important characteristics of the ProHGM implementation is the use of performance measures such as time series plots in a “progressive manner.” That is, some of measures are examined throughout evaluation phases but the level of inspection will be deeper as the evaluation processes advance. For example, time series plots will be initially inspected quickly in the early phase of evaluation but will be more thoroughly examined in the next evaluation phases.

4.3 Implementation and results

4.3.1 Evaluation Phase One (P1)

4.3.1.1 Phase One, Task 1 (P1.1)

The first subtask of P1.1 is to review sources of information about model setup. The primary material we used in this study was information that TCEQ provided to the public and other information we could find in the literature. The description of how TCEQ modelers prepared their input files can be found in the documents posted in TCEQ’s SIP narrative web site (TCEQ, 2006d). The site is titled “HGB Mid-Course Review SIP Narrative” adopted by TCEQ at December 1, 2004. TCEQ’s SIP narrative web site contains the rationale of HGMCR modeling setup such as episode selection and grid configuration. The model TCEQ used was the Comprehensive Air quality Model with eXtensions (CAMx) version 4.03 (ENVIRON, 2003). The CAMx model run scripts and input files used for the HGMCR modeling can be found in the TCEQ’s FTP site (TCEQ, 2006b). We made special notes for documents and data that were not part of TCEQ’s SIP narrative web site or from the TCEQ’s FTP site directly.

The second subtask of P1.1 is to review model configuration. Modeling domains adopted for the HGMCR modeling are shown in Figure 4.1. Initially, the HGMCR modeling domains consisted of five nested grids labeled (with horizontal grid resolution in parenthesis)
as Regional (36-km), East Texas (12-km), HGBPA (4-km), HG (1-km), and BPA (1-km). Later, TCEQ excluded the two 1-km domains, i.e. HG and BPA, for the HGMCR modeling for the proposed SIP. More detailed descriptions of the HGMCR modeling grid configuration can be found in the TCEQ’s HGMCR modeling page (TCEQ, 2006b). As highly recommended by PROMPT, we prepared a GIS map for HG (Figure 4.2). This map includes geographical features such as ground monitors. This map was used frequently for locating monitors, associating model behaviors at monitors near each other, and creating important graphical performance measures throughout this study.

Upon review, we considered that most of the original model configurations were reasonable except the horizontal grid resolution. A major study was conducted at the request of the TCEQ commissioners to address the impact of industrial VOC emission variability on peak ozone concentrations (Allen D. et al., 2004). The study already showed that 1-km resolution was necessary to simulate the impact of reactive VOC in HG properly with three-dimensional PAQMs. The results of this study formed a basis of the short-term HRVOC rule in the HGMCR SIP. In spite of this study, however, the TCEQ staff decided to only use the 4-km modeling results in the SIP.
Figure 4.1. Domains for the HGMCR modeling. BPA domain was considered in the early phase of modeling but excluded in the HGMCR modeling. More detailed information such as map projection parameters can be found on TCEQ’s HGMCR modeling web site (TCEQ, 2006c).
Figure 4.2. GIS map of HG 74 km by 74 km domain. Green squares depict ground or research monitors and provide four-letter labels. Triangles are major VOC point sources and they are colored depending on their annual VOC emissions in 2000 (red > yellow > green > blue > white). Gray filled polygons include major airports and other major area sources. Cyan areas represent water bodies. The Ship Channel runs from east of C35C to south of H07H and below highway 146.
The last subtask of P1.1 is to review input files. The TCEQ called their final base case “base5b.psito2n2” (called Base5b.Psito2n2 here). The Base5b part indicates the version of meteorological input files used for the SIP modeling and the Psito2n2 is a short name for emissions input files used for the base case modeling. The modeling period of the Base5b.Psito2n2 simulation is from 2000-08-22 to 2000-09-06. Here, we only discuss important aspects of Psito2n2 that are related to this study such as its historical background and emission adjustments made by TCEQ. For more details about adjustments made for Base5b and Psito2n2, see the Chapter 3 of TCEQ’s SIP narrative (TCEQ, 2006d).

Psito2n2 stands for “Point plus Special Inventory plus Terminal Olefin equal to NOX-2”. Originally, in 2001, TCEQ used standard procedures to develop a set of unadjusted emission inputs called “regular” (called Regular here) that was combination of the annual-based Ozone Seasonal Daily inventory and a “Special Inventory” (SI) based on daily self-reported hourly emissions at industrial point sources. The SI is the product of TCEQ’s significant efforts (1) to refine the speciation of various VOCs in HG because HG is very unique in the number of VOC species, and (2) to account for specific operating conditions such as upsets, start-ups, and shut-downs of point sources during the TexAQS 2000 study period. Unfortunately, this information was requested a year after the emissions had occurred. That is, the information was not concurrent with the field program and modeling period. In spite of the use of the SI inventory, the CAMx modeling with Regular emission inputs showed gross under-estimation of peak ozone concentrations. Consequently, it seemed that the SI enhanced the details of chemical speciation but added little to the ozone formation capabilities to the model.
In 2001, a study analyzing ambient measurements in HG showed that there was Transient High Ozone Events (THOEs) detected at monitors in 1993 (Blanchard, 2001). THOEs are observed high ozone concentrations accompanied with large hourly ozone concentration changes (> 40 ppb/hr) during short period of time (typically less than 3 hours). Additionally, there were very high concentrations of some reactive VOCs species in 1993 and it was hypothesized that these high VOCs streams could cause THOEs. As part of Texas Air Quality Study 2000 (TexAQS2000), researchers found that HG area had various large highly reactive volatile organic compounds (HRVOCs) emissions in HGMCR modeling period (Kleinman et al., 2002; Daum et al., 2003; Daum et al., 2004; Berkowitz et al., 2004; Berkowitz et al., 2005). In general, it was agreed among researchers that HG has unique large HRVOC emissions compared with other cities of similar sizes. Kleinman and Daum were able to show that very high ozone production (>100 ppb/h) were associated with a set of HRVOCs. On the other hand, such values only occurred in 17 of the 211 samples over 12 days. Subsequently, Allen et al. (2004) were able to show that episodic “event” emissions contributed about 4% to the total VOC emissions and about 12% to the HRVOC emissions.

The problem of the Regular emissions inventory was that it did not seem to have enough potential to create sufficiently high ozone concentrations due to the lack of enough HRVOC emissions. To fix these gaps of inadequate HRVOCs emissions, TCEQ scaled the emission of selected terminal olefins from some point sources with a factor.

Originally, given the time limit and need for making decisions in time, TCEQ took an approach to fix these gaps of HRVOCs by scaling all terminal olefin emissions from point sources with a factor derived from a single-pass measurement over the Ship channel by the
Baylor aircraft on October 19, 2001. From the Baylor aircraft measurements, TCEQ estimated that terminal olefins mixing ratio was equal to $\text{NO}_X$ mixing ratio. Later, the factor was re-calculated from the HRVOC-to-$\text{NO}_X$ emissions ratio, i.e. dividing total HRVOC emissions by total $\text{NO}_X$ emissions of some point sources that release more than 10 tons/year of some terminal olefins such as ethylene. The final adjustment factor was applied at 81 accounts that had >5% HRVOCs in emissions. The exact value of final scaling factor was, however, not available to public. According to TCEQ’s document, the total amounts of HRVOC additions were 318 ~ 358 tons/day of VOCs during the HGMCR modeling episode.

In the review of emission inputs, we found two major issues that might affect the overall model performance: (1) TCEQ modelers took an adjustment approach that was scientifically not defensible. As they stated in the SIP document, the device of Baylor aircraft was primarily for measuring isoprene. That is, it is not a well-established method for olefin species at that time. Moreover, these was no sound evidence that the actual amount of HRVOCs emissions were directly correlated with $\text{NO}_X$ emissions, and (2) the adjustment was made for 24 hours across the modeling domain, i.e. assuming the missing HRVOC emissions were relatively homogeneous in time and space. A study showed that this assumption is incorrect (Allen D. et al., 2004); there was large variability in industrial HRVOC emissions. Therefore, we could see the conceptual issues in emission inputs of the HGMCR modeling at the very first phase of our MPE procedures.
4.3.1.2 Phase One, Task 2 (P1.2)

The first subtask of P1.2 is to create a table for observational data base for MPE. In terms of observational database for MPE tasks, the HGMCR modeling had not only the ample routine ground monitor measurements but also special data from research and aircrafts for performance evaluations. The actual observational data we used are from TCEQ’s web page (TCEQ, 2005; TCEQ, 2006b). Table 4.2 shows the summary description of all monitoring sites used for this study. In HGBPA, the total number of sites (including WILT and LAPT) was 32, but we focused on 20 sites (18 routine ground monitors plus WILT and LAPT) in HG because the primary issue was the ozone problem in the HG domain.

The second subtask of P1.2 is to describe any particular aspect of observational data base for MPE. As part of the SIP modeling improvement projects, the State of Texas sponsored a large field campaign, Texas Air Quality Study 2000 (TexAQS 2000), and TexAQS produced observational data that are often not available in typical SIP modeling support databases (Daum et al., 2005). Most distinctive data were: (1) high resolution gas chromatography measurements at C35C and LAPT (TCEQ, 2005), (2) aloft measurements including high resolution CO and FORM observation at WILT (~ 250 m AGL) (Berkowitz et al., 2004), and (3) aircraft observations including 1-second measurements of gaseous species and 1-minute samples of VOCs (Daum et al., 2005). These non-routine observations could be used for enhancing performance evaluations.
Table 4.2. Summary of observational database used in this study. Both WILT and LAPT were non-AIRS research monitors and are included to enhance MPE. The geographical information about each site, such as neighborhood monitors, can be found in Figure 4.2.

<table>
<thead>
<tr>
<th>Site</th>
<th>Easting</th>
<th>Northing</th>
<th>AIRS ID</th>
<th>O3</th>
<th>CO</th>
<th>NOₓ</th>
<th>VOCs</th>
<th>Wind</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAYP</td>
<td>435.02</td>
<td>-1114.33</td>
<td>482010055</td>
<td>X</td>
<td>-</td>
<td>NO, NO₂</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>HLAA</td>
<td>435.12</td>
<td>-1098.91</td>
<td>482010047</td>
<td>X</td>
<td>X</td>
<td>NO, NO₂</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>HCQA</td>
<td>437.89</td>
<td>-1122.20</td>
<td>482010051</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>WILT</td>
<td>438.33</td>
<td>-1109.58</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>NO, NO₂, NOY, NOZ, HNO₃, HONO, PAN</td>
<td>FORM</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ALB</td>
<td>432.33</td>
<td>482010067</td>
<td>X</td>
<td>X</td>
<td>NO, NO₂</td>
<td>PAR, ETH, OLE, TOL, XYL, ALD₂, ISOP</td>
<td>X</td>
<td>Most of NOₓ and VOC data available after 2000-08-31</td>
<td></td>
</tr>
<tr>
<td>HLAA</td>
<td>436.49</td>
<td>-1108.93</td>
<td>482010070</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>HCQA</td>
<td>434.92</td>
<td>-1098.45</td>
<td>482010046</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>WILT</td>
<td>437.88</td>
<td>-1120.85</td>
<td>482010062</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>H03C</td>
<td>459.11</td>
<td>-1108.76</td>
<td>482010035</td>
<td>X</td>
<td>X</td>
<td>NO, NO₂</td>
<td>PAR, ETH, OLE, TOL, XYL, ALD₂, ISOP</td>
<td>X</td>
<td>high resolution GCs</td>
</tr>
<tr>
<td>HOEA</td>
<td>460.44</td>
<td>-1104.75</td>
<td>482010070</td>
<td>X</td>
<td>-</td>
<td>NO, NO₂</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>H04H</td>
<td>465.27</td>
<td>-1104.83</td>
<td>482010083</td>
<td>X</td>
<td>-</td>
<td>NO, NO₂</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>H10H</td>
<td>469.54</td>
<td>-1096.83</td>
<td>482010084</td>
<td>X</td>
<td>-</td>
<td>NO, NO₂</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DRPK</td>
<td>470.97</td>
<td>-1115.14</td>
<td>482011039</td>
<td>X</td>
<td>X</td>
<td>NO, NO₂</td>
<td>PAR, ETH, OLE, TOL, XYL, ALD₂, ISOP</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>LAPT</td>
<td>477.60</td>
<td>-1114.70</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>NO, NO₂, NOY, HNO₃, PAN</td>
<td>PAR, ETH, OLE, TOL, XYL, ALD₂, ISOP</td>
<td>X</td>
<td>high resolution GCs</td>
</tr>
<tr>
<td>H08H</td>
<td>478.11</td>
<td>-1117.34</td>
<td>482010088</td>
<td>X</td>
<td>-</td>
<td>NO, NO₂</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>H07H</td>
<td>481.02</td>
<td>-1103.46</td>
<td>48201007</td>
<td>X</td>
<td>-</td>
<td>NO, NO₂</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>H11H</td>
<td>489.51</td>
<td>-1090.66</td>
<td>480710900</td>
<td>X</td>
<td>-</td>
<td>NO, NO₂</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TLMC</td>
<td>491.38</td>
<td>-1103.15</td>
<td>480710901</td>
<td>X</td>
<td>-</td>
<td>NO, NO₂</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Note that all VOCs listed above were speciated to match the chemical mechanism, CB-IV, used in HGMCR modeling.
The last subtask of P1.2 is to select a MPE tool. During a study in which we used an prototype of PROMPT (Jeffries et al., 2005), we found that traditional model analysis tools constrained our ability to accomplish our study goals fully in a timely manner. Thus, as a parallel to that study, we developed the Python-based Performance Analysis Support System (pyPASS) as a primary evaluation tool. For details about pyPASS, refer to Kim and Jeffries, 2006c). Since consistency of information and the replication by third-parties are also important requirements of PROMPT, we accomplished these two aspects by (1) pre-defining the formats and layouts of all graphical measures produced by pyPASS and (2) making scripts used for this study available to public. The guidance on how to read pyPASS graphical measures is also available in Kim and Jeffries, 2006c).

4.3.1.3 Phase One, Task 3 (P1.3)

The first subtask of P1.3 is to review the conceptual model. For this task, we examined the conceptual model described in Appendix A of the HGMCR Modeling Protocol (TCEQ, 2004). The conceptual model in the Appendix A was for the HGMCR Phase I modeling and no separate conceptual model was available for the Phase 2 modeling on TCEQ’s web site. Therefore, we assumed there was no update in the conceptual model. Even though this conceptual model was found to be incomplete, the construction of a newer conceptual model is beyond the scope of this study.

Conceptual modeling is a technique representing certain aspects of a system with objects (e.g. ozone and precursors) in the system, the attributes of objects (e.g. chemical concentrations), and the relationship between objects (e.g. chemical reactions) in the system (Based on Boman et al., 1997). It is important to note that these objects, their attributes, and their relationship are variable in space and time and depends on specific occurrence of events.
(e.g. meteorological conditions and emissions). That is, a conceptual model must be a representation of an understanding about how a system works. Therefore, we would expect a conceptual model for an ozone problem in an area should contain an explanation (more likely qualitative) about ozone formation based on description of temporally and spatially variable meteorology and precursor emissions.

Unfortunately, we found that the TCEQ’s conceptual model is somewhat ambiguous by losing particularly important ozone system behavior in HG. General points that TCEQ asserted with regard to ozone mechanism in HG domain can be summarized as following: (1) there were nighttime accumulation of high concentrations of precursors due to light winds and low mixing heights, (2) high concentrations of precursors were carried to the Southeast Houston by weak Northwesterly winds in the morning, (3) there would be formation of ozone clouds during the morning and movement of those clouds to the west toward Houston, (4) the entrainment of ozone clouds back over the Houston would occur by reverse flows (exact path depends on minor variations in winds.), (5) there were possibilities of large NOx and VOCs emissions from point sources near the Ship Channel, (6) there was significant day-by-day and place-to-place variation of mixing heights, and (7) large hourly ozone concentration changes (e.g. > 40 ppb/hr) at monitor sites, i.e. THOE, have been frequently observed.

We also reviewed daily specific conceptual models presented in the Appendix B of the HGMCR Modeling Protocol. Among the days of HGMCR modeling period, a few days at the beginning of model episode are considered as model “spin-up” days that are likely influenced by the model’s pre-specified initial conditions significantly. Subsequently, these days are not useful in the actual control strategy development. Further, according to TCEQ,
there were large wildfires at the beginning of September that could cause exceptional ozone exceedances during the end of the episode. Therefore, we focused our analysis here on the period from 2000-08-25 to 2000-08-31.

TCEQ states two common characteristics among the HGMCR episode days: (1) the possibility of monitoring network failures in detecting the majority of ozone clouds as they moved between monitors, i.e. the size of ozone clouds might be smaller than the distances between monitors, and (2) strong overnight winds that might flush previous day emissions out of the region on most days except 2000-08-31. Interestingly, the second characteristic is controversial to the first general point that we introduced above. In the following phase of evaluation, we examined at this issue.

The original title of the Appendix B was “Meteorological and ozone characteristics in the Houston area from August 23 through September 1, 2000.” That is, the Appendix B was not a complete conceptual model. Rather, it describes how meteorological conditions and ozone behaviors were. It missed two more major players in ozone formation: \( \text{NO}_X \) and \( \text{VOCs} \). Surprisingly, we found that the TCEQ’s conceptual modeling was a type of effort describing observations of phenomena rather than an attempt to provide casual explanation with the objects and events in the ozone formation system of HG. The most significant weakness was the lack of explanation about the role of specific sources of \( \text{NO}_X \) and \( \text{VOCs} \). For example, we would expect how downtown \( \text{NO}_X \) emissions would interact with the \( \text{VOCs} \) emissions near the Ship Channel on a certain day. This lack of adequate conceptual model limited our analysis here.

The second subtask of P1.3 is to compare model behaviors with the conceptual model that was reviewed in the first subtask of P1.3. To examine if the HGMCR model showed
overall ozone and meteorological characteristics comparably with the conceptual model, we
inspected four features of model outputs at monitor sites: (1) daily peak ozone concentrations
and daily maximum hourly ozone changes, (2) daily peak NO and NO$_2$ concentrations (3)
morning and afternoon resultant wind speeds (MRWS and ARWS) and (4) ozone time series.
Note that all performance measures at this stage were evaluated grossly and simultaneously.
An example “package” of graphical measures we examined for this screening evaluation task
is shown in Figure 4.3 and Figure 4.4. Figure 4.3 is necessary for day-by-day assessments
and Figure 4.4 was accompanied with Figure 4.3 for site-by-site and day-by-day analyses.
Figure 4.3. Example package needed for daily graphical performance analyses as part of P1.3 in which a model behavior is compared with the conceptual model. Each package consists of four graphical measures: a bar chart of unpaired peak ozone concentrations (first from top), a bar chart of unpaired maximum hourly ozone changes (second), a bar chart of unpaired peak NO (third) and NO₂ (last). The X-axis of each bar chart shows monitors sorted from east to west. Model data were depicted in red in bar charts.
Figure 4.4. Example package needed for site-by-site and day-by-day graphical performance analyses as part of P1.3. Each package consists of two graphical measures: surface wind speed scatter plots (left), and time series plots (right). For detailed guidance on how to use each graphical measure, refer to the pyPASS article (Kim and Jeffries, 2006c) and to a partial implementation of PROMPT for this case (Jeffries et al., 2005). In the wind speed scatter plots, yellow right-facing triangle and red left-facing triangle represent morning resultant wind speed (RWS) and afternoon RWS. Model data were depicted with dotted lines in time series plots.
Two factors we considered in our wind analyses: (1) the ratio of MRWS and ARWS of prediction to those of observation (RMRWS and RARWS) and (2) the pattern of daily wind speed change. Given that (1) the grid size of Base5b.Psito2n2 was 4-km and (3) most days showed about or faster than 4 km/hr MRWS (or ARWS whichever slower) in observations or predictions, we considered 0.5 ~ 2.0 as the initially acceptable value of RMRWS and RARWS. The pattern of daily wind speed changes were examined by tracking the spline curve in wind speed scatter plots that connects each pair of observation-prediction wind speeds. In the ideal case, the spline curve will be near 1:1 line, i.e. “diagonal” line pattern. If modeled (or observed) wind speeds were relatively invariant while observed (modeled) wind speeds were variant during a day, spline curves in our scatter plots will be “horizontal” (or “vertical”). If model-observation wind speed pairs are biased throughout the day but in opposite ways during morning and afternoon, “looping” spline curves can be found.

Daily peak ozone plots were inspected to check following features: (1) domain wide spatial biases of daily ozone peaks, and (2) the status of exceedances in predictions and observations. The daily peak hourly ozone concentration change plot was examined if monitor cells experienced at least one THOE when monitors in real world did. We considered the difference of peak ozone concentrations between model and observation was “significant” if the difference was larger than 20 ppb. Time series plots were examined to check the temporal discrepancies of ozone peak in predictions and observations. Daily peak NO and NO₂ plots were inspected to check following features: (1) domain wide spatial biases of daily peak NO and NO₂ concentrations and (2) the relative biases of NO and NO₂ at each monitor site.
We inspected four bar charts, 14 wind speed scatter plots, and 20 time series plots for P1.3 on each study day. The results of P1.3 on 2000-08-25 are summarized in Table 4.3. As shown in the table, the total number of sites providing surface wind data was 14 out of 20. On 2000-08-25, eight out of 14 sites showed looping pattern in wind speed scatter plots. Given that the ozone problem in 2000-08-25 was caused by ozone clouds moving across Houston from west to east, the quality of wind inputs might not be sufficient for accurate simulation. In the examination of peak ozone concentrations, we found that there were 12 out of 20 sites showed larger than 20 ppb ozone concentration differences between predictions and observations. Among those 12 cases, four cases showed significant maximum hourly ozone change differences between model and observation, i.e. > 40 ppb/h. Total six out of 20 sites showed False Negative, i.e. model predicted no exceedance while there were exceedances in the real world. Surprisingly, our nitrogen oxides analyses revealed that the model showed “high” NO and NO₂ biases at nine and eight sites out of 14 sites. The results implied that the model had serious performance issues that can influence precursor control options.
Table 4.3. Summary results of comparison of Base5b.psito2n2 with the conceptual model (P1.3) on 2000-08-25.

<table>
<thead>
<tr>
<th>Site</th>
<th>Surface winds</th>
<th>O3</th>
<th>NOx</th>
<th>NO bias</th>
<th>NO2 bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAYP</td>
<td>No Yes Looping</td>
<td>ΔO3</td>
<td>Δ(dO3/dt)</td>
<td>Skill score</td>
<td>ΔTmax</td>
</tr>
<tr>
<td>HLAA</td>
<td>- - -</td>
<td>12</td>
<td>6</td>
<td>TP 1</td>
<td>High</td>
</tr>
<tr>
<td>HCQA</td>
<td>No Yes Looping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WILT</td>
<td>- - -</td>
<td>23</td>
<td>16</td>
<td>TP 0</td>
<td>Low</td>
</tr>
<tr>
<td>HCFA</td>
<td>- - -</td>
<td>53</td>
<td>51</td>
<td>TP 1</td>
<td>High</td>
</tr>
<tr>
<td>HALC</td>
<td>Yes Yes Looping</td>
<td>45</td>
<td>19</td>
<td>FN 1</td>
<td>No</td>
</tr>
<tr>
<td>HROC</td>
<td>Yes Yes Looping</td>
<td>52</td>
<td>40</td>
<td>TP 0</td>
<td>-</td>
</tr>
<tr>
<td>HWAA</td>
<td>- - -</td>
<td>23</td>
<td>27</td>
<td>TP 1</td>
<td>-</td>
</tr>
<tr>
<td>HSMA</td>
<td>- - -</td>
<td>38</td>
<td>38</td>
<td>FN 1</td>
<td>-</td>
</tr>
<tr>
<td>C35C</td>
<td>Yes Yes Diagonal</td>
<td>39</td>
<td>16</td>
<td>FN 1</td>
<td>High</td>
</tr>
<tr>
<td>HOEA</td>
<td>Yes Yes Looping</td>
<td>38</td>
<td>27</td>
<td>FN 0</td>
<td>High</td>
</tr>
<tr>
<td>H03H</td>
<td>Yes Yes Diagonal</td>
<td>15</td>
<td>6</td>
<td>TN 0</td>
<td>High</td>
</tr>
<tr>
<td>H04H</td>
<td>No No Looping</td>
<td>41</td>
<td>1</td>
<td>TN 1</td>
<td>High</td>
</tr>
<tr>
<td>DRPK</td>
<td>Yes Yes Looping</td>
<td>45</td>
<td>24</td>
<td>FN 2</td>
<td>-</td>
</tr>
<tr>
<td>LAPT</td>
<td>- - -</td>
<td>10</td>
<td>21</td>
<td>TN 3</td>
<td>No</td>
</tr>
<tr>
<td>H08H</td>
<td>Yes Yes Looping</td>
<td>7</td>
<td>3</td>
<td>TN 1</td>
<td>No</td>
</tr>
<tr>
<td>H07H</td>
<td>No Yes Diagonal</td>
<td>29</td>
<td>7</td>
<td>FN 1</td>
<td>High</td>
</tr>
<tr>
<td>H10H</td>
<td>No Yes Diagonal</td>
<td>3</td>
<td>4</td>
<td>TN 1</td>
<td>High</td>
</tr>
<tr>
<td>H11H</td>
<td>No Yes Vertical</td>
<td>5</td>
<td>6</td>
<td>TN 2</td>
<td>High</td>
</tr>
<tr>
<td>TLMC</td>
<td>Yes Yes Vertical</td>
<td>15</td>
<td>10</td>
<td>TN 1</td>
<td>-</td>
</tr>
</tbody>
</table>

RMRWS: Ratio of modeled morning resultant wind speeds (MRWS) to observed MRWS
RARWS: Ratio of modeled afternoon resultant wind speeds (ARWS) to observed ARWS
ΔO3: the difference between predicted peak ozone and observed peak ozone (unpaired)
Δ(dO3/dt): the difference between predicted peak ozone change per hour and observed peak ozone change per hour (unpaired in time and paired in space)
Skill score: FP (False Positive, predicted ozone ≥ 125 ppb and observed ozone < 125), FN (False Negative, predicted ozone < 125 ppb and observed ozone ≥ 125), TP (True Positive, ozone ≥ 125 ppb and observed ozone ≥ 125), and TN (True Negative, ozone < 125 ppb and observed ozone < 125)
ΔTmax: the difference in hours between predicted peak ozone and observed peak ozone
NO and NO2 biases are denoted “high” if the predicted concentrations are higher than twice of observation. In opposite case, it is marked “Low”. Otherwise, it is noted “No.” If there is no observation available, “-” mark is used.
From 2000-08-25 to 2000-08-29, MRWS were faster than ARWS. From 2000-08-30 to 2000-08-31, the pattern became opposite. In most of days, the RMRWS was within 0.5~2.0 except that six out of 14 sites on 2000-08-25 showed poor performance on RMRWS. In general, large ozone discrepancies appeared at the western Houston on 2000-08-25. On the contrary, significant ozone discrepancies on 2000-08-30 and 2000-08-31 were on the eastern Houston near the Ship Channel. From 2000-08-26 to 2000-08-29, there were few exceedances in observation and the model agreed with observation well with respect to the status of exceedances with a few exceptions on 2000-08-26 and 2000-08-29. Hours of days when ozone peaks occurred in the model and observation were matched within two hours in most cases. High or close-to-high NO and NO\(_2\) biases were found in our peak nitrogen oxides analyses throughout the modeling period.

For the question for P1 in Table 4.1, we concluded that the HGMCR modeling showed, in general, what we needed to see: (1) the model replicated some important meteorological phenomena such as relatively high wind speeds during nighttime even though the magnitude was often over-estimated by large margin, (2) the model was able to produce high ozone concentrations and some THOEs even though the magnitude and timing needed significant improvement, and (3) the hour of day when peak ozone observed was closely estimated by the model. Yet, the model performance was not considered adequate, especially because of significant NO\(_X\) biases. This issue was investigated further in the following phase of evaluation when we inspected time series plots. In addition, we decided to examine two model structural issues in the following evaluation phase: (1) the horizontal grid resolution of HGMCR modeling may not be sufficient to reproduce THOEs and (2) the rationale of TCEQ’s domain wide HRVOC imputation in Psito2n2 was ambiguous.
4.3.2 Evaluation Phase Two (P2)

4.3.2.1 Phase Two, Task 1 (P2.1)

One of the requirements by PROMPT is to evaluate a model performance with respect to policy questions. Thus, we consulted the documents in TCEQ’s SIP narrative web site as material representing concerns by TCEQ policy makers. One of the control options in the revised SIP was the short-term HRVOC emissions rule; each site in Harris County should comply with not-to-exceed limit of 1200 lbs/hr HRVOC release. Since the area influenced by the short-term cap rule is primarily Harris County, especially those areas near the Ship Channel, and the surrounding areas included in 1-km domain, we conducted follow-up analyses for all 20 monitor sites listed in Table 4.2. Our focus was to evaluate how potentially useful the Base5b.Psito2n2 is for testing the short-term HRVOC rule in the HG area. We also investigated the spatio-temporal behavior of Base5b.Psito2n2 to seek solutions to two model setup issues identified in P1: the adequacy of 4-km grid resolution for HG ozone in 2000 and the HRVOC emission imputation across domain for the entire modeling period.

4.3.2.2 Phase Two, Task 2 (P2.2)

For every day from 2000-08-25 to 2000-08-31 at every monitor sites, we examined three graphical performance measures for surface winds: (1) wind speed scatter plots, (2) hodograms, and (3) wind error plots. The focus was to find temporal information of surface winds at each site and to identify large wind errors. Also, the overall surface wind pattern at monitor sites during a day was inspected. Our criteria for relatively reliable surface wind input data were (1) 0.5~2.0 for RMRWS and RARWS, and (2) less than 30 degree differences between modeled wind directions and observed wind directions. Our final judgment, however, was not mechanically made from these two numeric criteria. We
considered two more factors: (1) the wind speed scatter plot pattern and (2) tendency of wind
directions and wind directional errors, especially the existence any prevalent directional
biases. We divided the hours of day into four groups: (1) Midnight-to-Morning (0000-0600),
(2) Morning-to-Noon (0700-1200), (3) Noon-to-Afternoon (1300-1800), and (4) Afternoon-
to-Midnight (1900-2400). Then, we graded the quality of surface winds into three
categories: (1) “Reliable”, (2) “Ambiguous”, and (3) “Unreliable.” Model performance at a
site can be Unreliable due to various reasons: model performance at site might show
significant wind errors near critical hours such as noon to 1400 when ozone formation
potential was high or the model showed persistent wind direction errors.

An example package of graphical measures we used for P2.2 is shown in Figure 4.5.
From Figure 4.5, we could find that the model showed very poor wind performance at HOEA
on 2000-08-26. If the peak ozone concentrations at HOEA on 2000-08-26 showed good
matching, it is doubtful that it could be compensating errors between ozone formation and
advection. Given that all sites on this day were considered to show very poor morning wind
predictions, we would recommend further investigation to use the model prediction on this
day. The total number of plots we examined was 48 plots (16 sets of three types of plots) for
each model day.
Figure 4.5. Example package needed for graphical performance analyses required by P2.2. Each package consists of a set of graphical measures: wind speed scatter plots (top), hodogram (bottom left), and wind error plots (bottom right). This example represents one of the worst cases. For example, observed wind speeds were invariant and close to 4 km/h from 1100 to 1300 while predicted wind speeds changed from 4 km/h to 12 km/h. The afternoon wind speeds were mostly twice as fast as observed wind speeds.
Further, we found that the model showed unreliable nighttime surface winds for most of the days at almost all sites. For most days, afternoon winds are at least ambiguous, i.e. it is very hard to tell whether the surface wind errors are clearly making the wind inputs unusable for modeling. The most unreliable day was 2000-08-26 and 2000-08-29 even though the other two days between these two days did not show good performance necessarily. The best day we could identify in the P2.2 was 2000-08-30.

4.3.2.3 Phase Two, Task 3 (P2.3)

For every day from 2000-08-25 to 2000-08-31 at every monitor sites, we examined two types of graphical performance measures for $O_3$, NO, and NO$_2$: scatter plots and time series plots. The focus was to find general behavior of chemical signals: (1) when NO concentrations became higher than NO$_2$ concentration during daytime, (2) how ozone changed during a day, (3) if the model was able to capture THOEs if observational data showed it, and (4) whether any significant biases existed in nitrogen oxides predictions. An example package of graphical measures we inspected for the P2.3 is shown in Figure 4.6. H04H site is the only place that the model showed False Positive with large (> 40 ppb) margin on 2000-08-29. In general, the model over-predicted more than 20 % of ozone concentrations and there were clearly gross over-estimation of NO$_2$ throughout the day except some hours in the afternoon. For each site on each day, we inspected 4 plots and the total number of plots we used for P2.3 was 80.
Figure 4.6. Example package needed for graphical performance analyses required by P2.3. Each package consists of a set of graphical measures: a time series (top left) and scatter plots for NO, NO₂, and O₃.
After examining all sites for all days, we found several points in the model performance. First, there were large over-predictions of NO₂ at many monitor locations, which were worse during nighttime. Second, the model often produced ozone signal similar to the observation, but it missed THOEIs in several cases. Third, at the same time, the model often grossly over-predicted ozone as shown in Figure 4.6. Fourth, there were many cases that the model showed large morning over-prediction of NO, especially around 0700. On the contrary, the model showed significant morning under-prediction of NO and it often coincidentally occurred with large ozone over-prediction. This did not seem to be associated with NO₂ over-prediction necessarily.

4.3.2.4 Phase Two, Task 4 (P2.4)

As shown in Table 4.2, we had VOCs data available at five sites. We looked at the model predictions of CO, ETH, OLE, ALD2, FORM, and ISOP at these sites. The graphical measure we used was seven-day time series of VOCs as shown in Figure 4.7. Unfortunately, the quality of CO data was not considered good enough for fine analyses except at the WILT site. Apparently, the precision of CO measurement in most sites was 100 ppb. These poor-resolution CO data, however, could still be used to find that there were large CO over-prediction (e.g. more or equal to the double of observed CO concentrations) at some sites on 2000-08-30 and 2000-08-31.
Figure 4.7. Example graphical measures used for P2.4: CO at WILT (top), ETH and OLE at C35C (middle and bottom).
The most interesting observation for CO performance at WILT (the top plot of Figure 4.7) was that the model predicted diurnal variations of CO concentrations while the observation did not show these evident cycles. Additionally, the WILT site often showed CO spikes in the morning while the model could not reproduce them. The WILT site is in the 4th layer of model grid and these data give important insights into the aloft conditions in the model. The likely cause of the predicted CO concentrations decreasing down to the specified background CO concentrations is that the predicted nighttime wind speeds are too high and the higher daytime CO is being blown away. Other more important aloft pollutants would also be blown away at night resulting in little entrainment of important radical sources on the next morning when the mixing height rises.

Our examination of ALD2, ETH, OLE, and ISOP showed that there was gross over-estimation of these reactive species at virtually all monitor sites. At the same time, the model missed “spikes” of these reactive VOCs. On the contrary, FORM was predicted relatively well except some occasions such as LAPT site on 2000-08-25. Given that FORM could result from photochemical reactions as well as direct emissions, the performance judgment by this species is hard to interpret without additional information. In general, the model performed very poorly on reactive VOC species.

4.3.2.5 Phase Two, Task 5 (P2.5)

For the question for P2, we concluded that the HGMCR modeling showed significant biases in its predictions. First, wind data used for the HGMCR modeling showed significant discrepancies during nighttime. Often, afternoon wind data for simulations showed shortcomings by much slower wind speeds than observed winds. Additionally, the surface wind inputs showed large directional biases. Second, the model showed gross over-
predictions of NO\textsubscript{2} at many locations in many days. Most of over-predictions were during nighttime. Given that the nighttime wind speeds in the model were much faster than the observation, the NO\textsubscript{2} over-prediction issues could be worse with more correct wind speeds. Third, the model often missed important THOEs. Yet, the model showed some False Positive predictions. This aspect should be examined more thoroughly in the next phase of evaluation. Last, the model showed persistent over-estimations of reactive VOCs such as ETH and OLE during the modeling period. At the same time, the model missed reactive VOCs “spikes”.

By PROMPT, after reviewing the model performance at each monitor sites, we had to classify all monitors into four groups depending on the model’s potential usefulness to support precursor controls: “None”, “NO\textsubscript{X} only”, “VOCs only”, or “NO\textsubscript{X} and VOCs”. Surprisingly, we had to conclude that Base5b.Psito2n2 did not show reliable performance for testing any types of precursor controls. Three major reasons were (1) unreliable surface wind inputs, (2) gross-overestimation of NO\textsubscript{2}, (3) over-predictions for important reactive VOCs that the short-term HRVOCs control rule aiming at, and (4) missing the important ozone characteristic in HG, i.e. THOEs and possibly associated high concentration short-duration HRVOC signals at monitor sites.

Nevertheless, we could see possible solutions because what we found in P1 were also connected to problems identified in P2. First, HRVOC over-predictions could stem from the TCEQ’s HRVOC imputation methodology that was based on the current NO\textsubscript{X} emission modeling. Given that the NO\textsubscript{X} was over-esitmated, these two problems can be related to each other. Second, missing THOEs could result from insufficient chemical activity due to the grid size, i.e. 4 km.
TCEQ had Regular emission inventory in which no special imputation was made with the TCEQ’s NO\textsubscript{X}-to-VOC scale factor. Therefore, Regular inventory could alleviate HRVOC over-estimation issues. Also, 1-km simulation had potential to enhance the model’s ability to predict rapid ozone formation. Therefore, we acquired 1-km emission files of Psito2n2 and Regular from TCEQ modelers. To ensure one of our hypotheses on TCEQ’s HRVOC imputation, we compared low-level ETH and OLE emission intensities of Psito2n2 and Regular in scatter plots of values in 1-km domain cells paired temporally and spatially. In P1, we identified the fact that TCEQ imputed the HRVOC emissions when they created Psito2n2 emission files but we could not find the actual scaling factor. By comparing the low-level emissions of ETH and OLE in Psito2n2 and Regular, we could find that it was close to eight.

4.3.3 Evaluation Phase Three (P3)

4.3.3.1 Phase Three, Task 1 (P3.1)

After reviewing the results of P1 and P2, we concluded that TCEQ’s Base5b.Psito2n2 was inadequate for testing the HRVOC short-term rule. Nevertheless, we found that Base5b.Regular could be useful with some improvements because Regular emission files did not have large adjustments of HRVOC emissions. That is, at least, Regular emission files might not cause domain-wide VOC overestimation so that it may be useful for VOC controls in a limited way. Therefore, we developed an alternative modeling case based on Regular emissions.

Four days showed significant issues: 2000-08-25 (six FNs), 2000-08-30 (seven FNs), 2000-08-29 (two FPs and two FNs), and 2000-08-31 (six FNs and one FPs). There was also performance issues in 2000-08-26 (two FPs). The policy question of HRVOC short-term
rule, however, primarily concerns if the HGMCR modeling can reproduce the high ozone concentration by HRVOC events. Therefore, we did not include 2000-08-26 as a day for P3 in this study. For our alternative modeling case, we requested TCEQ to develop 1-km emission input files and ran Psito2n2 at 1-km grids with 4-km meteorological inputs by activating the Flexi-Nesting option of CAMx. The Flexi-Nesting implemented in CAMx allows modeler to run CAMx at finer grid scale than the resolution of input files supplied to the model (ENVIRON, 2003).

The most challenging problem in developing the alternative modeling was that we simply did not have the actual HRVOC emission event data for the year 2000. To resolve the issue, we consulted a study investigating the variability of HRVOC emissions in HG (Allen D. et al., 2004). The study showed that there was high spatial and temporal variability in HRVOC emissions in 2003 and the amount of HRVOC event varied significantly. By assuming the nature of HRVOC events in 2000 would have been similar to that in 2003, we formulated a set of emission inputs by adding HRVOC event emissions to selected cells to Regular emission files for four days we focused. We called this new emission inputs as RegEvnt1 (i.e. Regular + Event version 1). While 163 ~ 203 tons/day of HRVOCs were added to variant emissions to build Psito2n2 from Regular, the total amount of HRVOCs needed to create RegEvnt1 from Regular was only 66.9 tons.

For the comparison purpose, we ran RegEvnt1 at 1-km and 4-km. We also conducted a preliminary Process Analysis. In this analysis, we found that the relative effect of horizontal advection process on ozone formation with HRVOC events was linear to the size of grids while the effect of chemical reaction process was by square. Given that THOEs were often characterized with its short retention time at monitor locations, i.e. possibly narrow plume
widths, we considered that 1-km simulation represented the effects of HRVOC events on ozone formation more properly than 4-km runs (Figure 4.9). For the detailed description of how to prepare RegEvnt1, refer to a study utilized partial implementation of PROMPT (Jeffries et al., 2005).

4.3.3.2 Phase Three, Task 2 (P3.2)

For P3, we focused on Psito2n2 and RegEnvt1 at 1-km. Results of RegEvnt1 and the new Psito2n2 were undergone P1 and P2 that we conducted for Psito2n2 at 4-km. Several things were identified between the modeling results of Psito2n2 and RegEvnt1 at 1-km: (1) there was virtually no differences in NOX predictions by two models, (2) RegEvnt1 showed better HRVOC predictions than Psito2n2, and (3) ozone performance was similar to each other. In general, 1-km simulations showed more spatially resolved results while Psito2n2 at 1-km often showed more over-predictions of HRVOC predictions than Psito2n2 at 4-km. An example of performance improvement of HRVOC at C35C by RegEvnt1 is shown in Figure 4.8. Except 2000-08-25, both Psito2n2 and RegEvnt1 at C35C showed almost identical ozone behaviors from 2000-08-25 to 2000-08-31.

After finishing P3.2, we found (1) all the extra HRVOC emissions in Psito2n2 on most of modeling days did not necessarily contribute high levels of ozone concentrations, and (2) RegEvnt1 showed almost the same amount of NOX over-estimation compared with Psito2n2 except WILT. RegEvnt1 showed much less over-predictions than Psito2n2 at WILT throughout the modeling period. At the same time, the shapes of NO and NO2 time series were similar between Psito2n2 and RegEvnt1. RegEvent1 made significant differences in the level of NOX concentrations in the 4th layer at WILT while those differences were not
apparent at the surface monitors nearby WILT. More analyses with Process Analysis are under way.
Figure 4.8. Comparison of Psito2n2 and RegEvnt1 for ETH (top) as well as O₃, NO, and NO₂ (bottom) at C35C. The fine dotted lines are for Pisto2n2. The dashed lines are for Regevent1. Observational data are depicted with solid lines with filled squares.
Figure 4.9. Effects of horizontal grid resolution on event simulation. X-axis is for hours of August 30, 2000 and Y-axis shows domain-wide peak ozone concentration for each hour of day. The amount of event was 1,450 lb ETH and 10,188 lb OLE for an hour and the event duration was two hours as denoted with the purple box. Same amount of event emission was added to each cell in 4 km grids and 1 km grids. In other words, 4 km grid cell where HRVOC event was added is the cell contains 1 km grid cell. As shown in this figure, fine grid simulation is necessary to simulate transient high ozone events (THOEs) that are the most unique ozone behavior in the HG area.
For each site in four focused days, i.e. 2000-08-25 and 2000-08-29 ~ 2000-08-31, we conducted more in-depth performance analyses for P3.2 listed in Table 4.1. One of these analyses required examination of winds in context of emission sources. Therefore, we synthesized a map showing important geographical features such as emission sources and monitors combined with hodograms we inspected in P2. In addition, we investigated graphical measures used in P1 and P2 more closely.

Graphical measures for evaluating performance at LAPT in P3.2 are shown in Figure 4.10. For 1100, H08H site missed the wind observation. However, H08H site likely had northerly winds and the prediction was probably close to the observed wind unless there was a large wind change in the real world. We did not have hourly wind data for LAPT. Therefore, H03H was examined for the wind pattern. For 1200-1300, the predicted winds had more northerly wind component in wind vectors of H03H while the observed winds were almost westerly. A rough estimation could be made; the $O_3$ plume might be pushed to east further in the real world than while the $O_3$ plume was pushed to south further in the model. What can be expected under this situation is that H08H, LAPT and DRPK would likely have a skewed THOE peak in the model relative to the observation, if it had a THOE at each site.
Figure 4.10. Example of graphical analyses for P3.2. Wind time series plot for H08H (top left) and for H03H (top right) overlaid on GIS map. The observed winds are in magenta and the predicted winds are in cyan. The O3, NO, and NO2 time series at LAPT are shown at the bottom.
The aircraft measurement around 1200-1400 would help to confirm this hypothesis. Unfortunately, there were no aircraft measurements available for this time period. Instead, the aircraft observation was made around 1450 near by DRPK. Figure 4.11 shows ozone tile plots of RegEvnt1 at the 9th layer for 1400 and 1500 overlaid on part of Figure 4.2 with an aircraft track for 1451~1453. In our aircraft analysis, we found that the peak ozone concentration predicted by Psito2n2 at 1-km for cells intersected by the aircraft track was lower than 110 ppb for 1451~1453 while the observed peak ozone was 206 ppb. As shown in Figure 4.11, however, RegEvnt1 clearly predicted ozone concentrations over 150 ppb. Nevertheless, it is hard to tell if RegEvnt1 could predict ozone concentrations close to 200 ppb because of temporal resolution of model outputs, i.e. hourly average concentrations. In summary, H08H was considered at least as an Ambiguous site to test the effectiveness of the short-term HRVOC rule.
Figure 4.11. Ozone concentrations of RegEvnt1 from 1400 to 1500 on 2000-08-30 at the 9th layer. The dots flowing diagonally at lower left corner of tile plots are an aircraft track from 1451 to 1453. Black squares are cells of very high concentrations (> 180 ppb) of ozone. Each tile represents 1 km by 1 km grid cells. The estimated ozone plume width is approximately 6 km.
The best performance by Psito2n2 and RegEvt1 could be found on 2000-08-30. We could identify one site (DRPK) as a Reliable site and three sites as possibly Reliable sites (H03H, H08H, and TLMC) on 2000-08-30. There were four sites classified as being Ambiguous such as LAPT due to the absence of winds or NO\textsubscript{x} measurements while the ozone signal at each site seemed to be reasonable. We examined the spatial pattern of chemical concentrations with predicted winds as well as observed winds and chemical concentrations by inspecting tile plots as shown in Figure 4.12. Our focus was how models described chemical signals and winds around DRPK. The black areas in Figure 4.12 indicate cells of very high (> 180 ppb) ozone concentrations. Compared with Psito2n2, RegEvt1 did not make high ozone concentrations over the Galveston Bay. While the spatial ozone distributions in Psito2n2 runs were broad, RegEvt1 clearly reproduced a very narrow (less than 6 km for width) ozone plume nearby DRPK. We created and examined tile plots of selected species such as ozone and ETH for our study period in animation and snap-shots. The total number of tile plots for a species was 216 and we inspected seven chemical species besides ozone: NO, NO\textsubscript{2}, ALD2, ETH, OLE, ISOP, and CO. Some species such as ISOP did not show significant different, which we could expect because ISOP emissions in Psito2n2 and RegEvt1 were identical. Eventually, we could conclude that predictions made by RegEvt1 on 2000-08-30 could be used within a limited area of modeling domain, i.e. nearby DRPK.
Figure 4.12. Ozone concentration predicted by three models at 1500 on 2000-08-30: Psito2n2 ran at 4 km (left), Psito2n2 ran at 1 km (middle), RegEvnt1 at 1 km (right).
4.3.3.3 Phase Three, Task 3 (P3.3)

In P3.3, we focused those areas showing high ozone concentrations in the future case for those days considered in P3.2. In TCEQ’s FTP site, the most recent future case simulation made by TCEQ is “fy07o”. This future case is the result of many control strategies such as Cap-and-Trade of OLE in Harris County and the seven counties nearby Harris County. Among four days we focused on, 2000-08-31 still showed 145 ppb of ozone over the Galveston Bay (large water body on the left of Figure 4.2). The peak ozone concentration on 2000-08-30 and 2000-08-29 was 122 ppb over the Galveston Bay and 113 ppb near H04H. On 2000-08-25, the daily maximum ozone concentration was 121 ppb outside of HG. The most important questions posed in P3.3 were (1) if a model performance over the Galveston Bay was reliable for 2000-08-29 – 2000-08-31, and (2) whether the predicted peak ozone concentration of 121 ppb on 2000-08-25 was acceptable as its face value, given that there was persistent wind errors during the day and the peak occurred outside of HG.

The difficulty about the model performance over the Galveston Bay was that no measurement was available over the Bay. Therefore, it was hard to assess if the model predicted for the right reason. Prominent issues were relative contribution of transport processes and mixing depth described in the model to the future case ozone concentrations. Additionally, as shown in Figure 4.12, another root of the problem was that the prevalent ozone over the Bay in Psito2n2 was not shown in the results of RegEvnt1. We examined FORM tile plots at 1500 on 2000-08-30 and found that FORM concentrations over the Bay were 8–12 ppb in Psito2n2 and 4–8 ppb in RegEvnt1. One possible study to resolve this issue is to compare the CAMx with other model such as Community Multiscale Air Quality model (CMAQ) because these two models have different representation of mixing processes.
such as vertical diffusion. The performance comparison between CAMx and CMAQ is under investigation and it is beyond the scope of this study.

The performance issue on 2000-08-25 was investigated by examining tile plots at the 1st layer. Both of Psito2n2 and RegEvnt1 were able to create ozone clouds and both models could move those clouds across the modeling domain. The problem was that wind inputs seemed to be biased in direction persistently about 30 degree counterclockwise during the period that ozone clouds moved across HG domain; the observed winds tended to be more north than the predicted winds. In other words, the predicted winds would make air mass stay in the downtown of HG if wind speeds were same as observed wind speeds. In general, both of Psito2n2 and RegEvnt1 might not be able to see ozone clouds on right places and times. Nevertheless, there were some differences: Psito2n2 showed moderate (c.a. 80~100 ppb) ozone concentrations around ozone clouds and these surrounding ozone concentrations resulted in errors at the monitors did not see ozone clouds. These surrounding ozone concentrations were not formed in RegEvnt1 simulation. Given that the ozone clouds in RegEvnt1 on 2000-08-25 were disappeared much faster than Psito2n2, the role of these moderate concentration ozone surrounding high concentration ozone in Psito2n2 needs to be explained.

4.3.3.4 Phase Three, Task 4 (P3.4)

For P3.4, we estimated five types of statistical performance measures at each site. Three of them are traditional measures (US EPA, 1991): mean normalized bias (MNB), mean normalized gross error (MNGE), unpaired peak accuracy (UPPA). Two of them are found in a recent literature to improve shortcomings of correlation-based measures (Legates and McCabe Jr., 1999): modified index of agreement ($d_1$), and modified coefficient of efficacy.
The results of traditional measure estimation of Psito2n2 and RegEvnt1 for 2000-08-30 are shown in Figure 4.13. In general, Psito2n2 and RegEvnt1 did not show significant differences except at WILT and LAPT. There were cases that two models showed directionally different errors such as MNB at LAPT. In P3.2, we identified one site as a Reliable site: DRPK. The magnitude of UPPA by RegEvnt1 was smaller than by Psito2n2 while the trend of MNB and MNGE was opposite. The values of d₁ and E₁ by two models at DRPK were almost identical. Interestingly, statistical measures alone were not enough to distinguish the performance of Psito2n2 and RegEvnt1, which confirmed the general concern about heavy dependence on statistical performance measures for MPE of environmental modeling including air quality modeling (Willmott, 1984; Legates and McCabe Jr., 1999; Russell and Dennis, 2000; Beck, 2002).

We found that Psito2n2 and RegEvnt1 showed very similar statistical performance to each other. Yet, two models showed very different performance in our graphical analyses. Psito2n2 showed gross over-estimation of \( \text{NO}_X \) and VOCs while RegEvnt1 significantly alleviated VOC biases. The poorer ozone performance with better \( \text{NO}_X \) performance at WILT by RegEnvt1 implied an important issue; \( \text{NO}_X \) biases at WILT by RegEvnt1 were much lower than Psito2n2, which was unique because other sites showed almost identical \( \text{NO}_X \) biases. Therefore, we suspected that poor ozone performance at WILT with improved \( \text{NO}_X \) estimation by RegEvnt1 indicated that conditions near WILT in Psito2n2 could be \( \text{NO}_X \) inhibition aloft while the real world condition was not in \( \text{NO}_X \) inhibition status.
Figure 4.13. Results of statistical performance tests of Psito2n2 and RegEvnt1 for 2000-08-30. All of statistical measures were estimated for each monitoring site using the equations proposed by EPA (US EPA, 1991).
In P3, we identified several points in Psito2n2 and RegEvnt1 with respect to the short-term HRVOC rule: (1) four days in the period of 2000-08-25 ~ 2000-08-31 showed significant ozone performance issues, (2) wind inputs were in poor quality for most days, (3) 1-km grid simulations replicated THOEs better than 4-km grid simulations, (4) the most useful predictions were made by RegEvnt1 near DRPK on 2000-08-30, (5) NOX overpredictions were commonly made by both of models except at WILT by RegEvnt1, (6) the model predictions over the Galveston Bay should be examined further because the future ozone exceedances were predicted there by a future case based on Psito2n2 while eliminating HRVOC events from RegEvnt1 itself would remove the ozone exceedances in the future at the same place, (7) there was possible wind direction biases that could be the root of poor performance on 2000-08-25, (8) Psito2n2 showed much higher HRVOC biases than RegEvnt1, and (9) statistical performance measures for both of models indicated that there was no significant differences between both models.

4.3.4 Evaluation Phase Four (P4)

The Task 1 of Phase Four (i.e., P4.1) requires a GIS map showing our confidence on the performance of RegEvnt1. In P2.1, we considered the short-term HRVOC rule as the policy question that modelers need to answer. From the results of P2 and P3, we concluded that Psito2n2 was not reliable to test the efficacy of short-term HRVOC rule. Instead, we found a possible reliability in Regular and developed an alternative modeling case, RegEvnt1. Finally, we could identify partially useful predictions by RegEvnt1 in limited area and time for testing the short-term HRVOC emissions.

With respect to the posed policy question, we graded several places in the model based on our confidence about the reliability of model performance at sites or in areas. Based on
our performance assessment of model predictions on 2000-08-30, we created Figure 4.14 that shows our confidence on the reliability of RegEvnt1’s performance with regard to the short-term HRVOC rule.

Since this study aimed at demonstrating characteristics and advantages of implementation and application of PROMPT compared with traditional SIP MPE practices, we skipped the Task 2 and Task 3 of Phase Four (i.e. P4.2 and P4.3.) In general, evaluators need to create publicly accessible documents that contain all information used in MPE to accomplish the goal of P4.2. For P4.3, evaluators will consult with policy makers to recommend possible improvements of model performance including an alternative episode selection, etc.
Figure 4.14. Suggested reliability of RegEvnt1 on evaluating the efficacy of short-term HRVOC rule based on the results of P4.1 conducted for 2000-08-30. Red circles depict the daily peak ozone concentrations observed at each site. Green, gray, and purple areas represent Reliable, Ambiguous, and Unreliable model performance. The response of RegEvent1 to HRVOC short-term rule application can be properly assessed at most two sites based on the performance of model on 2000-08-30.
4.4 Conclusions

In this study, we re-evaluated the HGMCR modeling by applying an improved MPE methodology for evaluating the reliability of ozone SIP modeling, PROMPT. With PROMPT, we could assess the model performance with respect to control strategies considered by policy makers. We showed that the results of PROMPT application can help modelers answer the question would be asked by prominent policy makers more directly than traditional MPE approach: “Why should I believe this model?”

To answer this question, we considered three types of characteristics of observed and modeled systems: numeric characteristics, enumerated characteristics, and non-numeric characteristics. Table 4.4 summarized the system characteristics we considered in this study. Numeric characteristics analysis includes the use of statistical measures. Numbers that we used as guideline criteria for characteristics are presented in parentheses. Note that we did not use these values as “bright line” standards. Some characteristics are enumerated to classify certain aspects of observed and modeled system behaviors. While we used these numeric/enumerated values for our analysis, PROMPT requires examining non-numeric characteristics such as “spatial pattern of ozone and other chemicals.” Inevitably, proper analyses of these non-numeric performance characteristics require training to make proper judgments. By combining various measures from very quantitative criteria (e.g. resultant morning wind speeds) to highly qualitative measures (e.g. spline curve patterns in wind speed scatter plots), we could incorporate all available information in a comprehensive way. It is, however, clear that these characteristics alone are not sufficient to judge if a model will “fulfill” its task. Therefore, we considered regulatory characteristics that led us to conduct necessary sensitivity studies.
Table 4.4. Characteristics of observed and modeled systems considered in re-evaluating Houston-Galveston Mid-Course Review modeling for day-by-day and site-by-site analyses

<table>
<thead>
<tr>
<th>1. Numeric characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Number of monitors for important chemical species and meteorological variables such as surface winds</td>
</tr>
<tr>
<td>b. Observed and predicted resultant morning wind speeds and resultant afternoon wind speeds at each monitor site (0.5–2.0)</td>
</tr>
<tr>
<td>c. Number of Transient High Ozone Events in observed ozone time series and predicted ozone time series at each monitor site</td>
</tr>
<tr>
<td>d. Differences of daily maximum hourly ozone concentration change between observation and prediction at each monitor site (40 ppb/hr difference)</td>
</tr>
<tr>
<td>e. Differences of daily maximum ozone concentration between observation and prediction at each monitor site (20 ppb difference)</td>
</tr>
<tr>
<td>f. Temporal differences of observed daily peak ozone and predicted daily peak ozone at each monitor site</td>
</tr>
<tr>
<td>g. Wind direction errors</td>
</tr>
<tr>
<td>h. Aloft ozone difference between observation and prediction</td>
</tr>
<tr>
<td>i. Predicted ozone concentration locations in the future case</td>
</tr>
<tr>
<td>j. Width of high concentration (&gt;180ppb) ozone plumes (expecting about 8-12km)</td>
</tr>
<tr>
<td>k. Statistical measures: unpaired-peak accuracy, mean normalized bias, mean normalized gross error, modified index of agreement, and modified coefficient of efficacy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Enumerated characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Nitrogen oxide biases: High if prediction is greater than double of observation</td>
</tr>
<tr>
<td>b. Model’s skill score (e.g. True Positive) for ozone</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. Non-numeric characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Pattern of spline curves in wind speed scatter plots for observation and predictions: Diagonal, Vertical, Horizontal, and Looping</td>
</tr>
<tr>
<td>b. Shape of O₃, NO, and NO₂ time series for observation and prediction</td>
</tr>
<tr>
<td>c. Persistency of wind directional errors at each monitor site</td>
</tr>
<tr>
<td>d. Spatial distribution of chemicals with wind field patterns</td>
</tr>
</tbody>
</table>
In the re-evaluation of HGMCR modeling, we posed the short-term HRVOC emission rule as an example policy question that needed scientifically defensible answers. Our conclusion was that Psito2n2 itself has low reliability to test the short-term HRVOC rule. The two major reasons were (1) wind inputs were in low quality to meet the necessary spatial and temporal resolutions needed to replicate ozone formation in HG, and (2) the model showed too large NOX and VOC biases. By P1 and P2, it became evident that the cause of VOC biases stemmed from TCEQ’s HRVOC imputation approach, i.e. increasing HRVOC emissions by factor of 8.0 from its predecessor model emission inventory, Regular. In consecutive analyses, we found a possible performance improvement in Regular because it showed much lower VOC biases than Psito2n2. Therefore, we developed an alternative modeling case, RegEvnt1, based on Regular. In summary, we concluded that RegEvnt1 can be used with more confidence than Psito2n2 for testing the short-term HRVOC rule because we found that RegEvnt1 showed better performance than Psito2n2 even thought the usability is still limited in space and time.

The required HRVOC additions to Regular for RegEvnt1 was 66.7 tons for the entire modeling period while the requirements for Psito2n2 was approximately 318 tons/day ~ 358 tons/day. The differences in the amount of HRVOC imputed in each model pose important issues in policy developments; the future case will carry the imputed VOCs in its emission inventory from base case models. That is, we may end up controlling HRVOCs that might not be present in 2000. In other words, we may control some artificial HRVOCs. This is especially important if policy makers want to use Psito2n2 for their decision.

Compared with past MPE practice following the EPA’s current modeling guidance, our application of PROMPT showed promising results. With PROMPT, we could formulate
detail answers to the policy question; can the HGMCR modeling be used for testing the short-term HRVOC emission rule? We could also recommend and suggest possible ways to improve model performance during the course of our MPE. By emphasizing the use of graphical performance measures and utilizing all available information such as aircraft measurements and GIS maps, we could distinguish the performance of RegEvnt1 from that of Psito2n2. We expect that the implementation and application of PROMPT to the 8-hour ozone SIP modeling will improve the quality of MPE by guiding modelers to conduct systematic and consistent MPE and by assisting modelers to assess model performance sufficiently to formulate scientifically sound answers to policy questions.

Acknowledgement

This work was supported in part by the Houston Advanced Research Center under Contracts H-12.8HRB and H-13. Partial support was also made by the Houston-based stakeholder Mid-Course Correction Group and by the 8-Hour Coalition Group. We wish to thank Dr. Jim Smith at Texas Commission on Environmental Quality for providing modeling inputs and outputs for the Houston Galveston Mid-Course Review modeling.
5. SUMMARY AND RECOMMENDATIONS FOR FUTURE STUDIES

5.1 Summary of this study

Upon the review of past MPE practices and the US EPA’s MPE protocol that the past MPE practices followed, I found significant shortcomings in the US EPA’s MPE protocol under the shed of the recent advances in MPE theories for environmental modeling. These shortcomings include allowing model users (1) to accept modeling results that may lead to directionally incorrect emission controls or (2) to reject, as a whole, partially useful modeling results for policy decisions. As a solution to this problem, I developed the Protocol for Regulatory Ozone Modeling Performance Tests (PROMPT), a meta-protocol to improve the MPE for regulatory ozone modeling. PROMPT was formulated based on the principles derived from discussions appearing in the recent MPE literature. Two major characteristics of PROMPT are (1) the improved utilization of graphical evaluations and (2) the assessment of model performance explicitly with regard to ozone control policy questions. The detailed structure of PROMPT was designed and constructed based on the integration of (1) the principles upon the review of recent theoretical advances of MPE and (2) my practical experiences in real-world SIP modeling analyses. I concluded that PROMPT can serve the regulatory photochemical ozone modeling community better by supporting regulatory modelers to develop a case-specific protocol with explicit guidelines for more systematic and comprehensive performance evaluation than traditional approaches.

To implement PROMPT-like MPE protocols, however, I found that the existing tools designed to assist traditional MPE practices fell short of meeting my MPE needs. That is,
existing tools are inadequate for permitting a more comprehensive MPE. Thus, I developed a computerized MPE tool, the Python-based Performance Analysis Support System (pyPASS) that facilitates the implementation of the new MPE approach for regulatory photochemical modeling. I demonstrated that pyPASS can provide more focused information for comprehensive model performance evaluation with less resources than can traditional tools.

As a way to demonstrate the advantages of PROMPT, I re-evaluated the performance of the Houston Galveston Mid-Course Review (HGMCR) modeling as a case study. I attempted to answer two questions: To what extent can I accept the predictions made by the HGMCR models at face value for the purpose of developing a Texas 2000 State Implementation Plan (SIP) for ozone problems at HG?, and if I cannot, then how should policy makers make judgments about the effectiveness of ozone control options that were proposed in the Texas 2000 SIP? To answer these questions, I developed a specific MPE protocol for the HGMCR modeling and utilized pyPASS to implement the MPE protocol.

For the first question, I concluded that the HGMCR modeling showed significantly low reliability for developing and testing ozone control options. Three major reasons are: (1) the precision of meteorological inputs used for the model simulation was inadequate with respect to the resolution anticipated by the ozone problems in Houston, (2) the model showed too high nitrogen oxides biases; that is, the model frequently overestimated nitrogen oxides concentrations twice as high as observed, and (3) the over-prediction of volatile organic compounds (VOCs) concentrations was significant such as some VOCs concentrations predicted ten times as high as observed.

For the answer to the second question, I suggest to consider the predicted effects of highly reactive VOCs (HRVOC) event emissions and the consequence of controlling those
emissions estimated with the model simulations based on the emission inputs used for the precedent version of HGMCR model as the judgmental basis for the efficacy of short-term HRVOC rule in the Texas 2000 SIP.

In this study, I could show that a MPE following the PROMPT with suitable tools such as pyPASS can help modelers answer the critical policy question with regard to the application of regulatory ozone modeling for ozone policy decisions: “Why should I believe this model for the ozone air quality decisions?”

The answer to this question is obtained when modelers assess the reliability of model in light of the posed policy concerns that the model is supposed to provide scientific information. The PROMPT and the pyPASS can assist modelers formulate the answer by providing a systematic guidance on how to exercise MPE with properly crafted performance measures.

5.2 Recommendations for future studies

I could identify several future research needs upon the completion of this study. First, I found that it is very critical to develop a conceptual modeling methodology for ozone air quality management. The lack of a good conceptual modeling framework for ozone problems resulted in that modelers depend on “descriptions” of the ozone problems in an area rather than a “causal explanation” (or “a hypothesis”) for the ozone problems. Moreover, those descriptions in the past MPE often missed the role of precursors. One of the most important evaluation tasks must be the examination of model’s ability to realize the causes of ozone problems identified in a conceptual model. That is, a flawed conceptual model will likely result in defected numerical modeling. Thus, I argue that the regulatory air quality
modeling community needs to develop a framework to build the conceptual modeling systematically.

Second, an improvement of PROMPT is necessary to resolve the issues related to MPE questions for 8-hour SIP modeling. That is, how modelers need to interweave performance analyses and to interpret analyses results in the context of 8-hour ozone NAAQS. I can see this task will be a particularly difficult research subject when a state utilizes “baseline” (not “basecase”) emissions in their attainment demonstration for the 8-hour SIP modeling. This challenging issue should be investigated to ensure the reliability and effect of adopting “baseline” emissions before any states begin serious MPE processes and attainment demonstration modeling for the 8-hour ozone NAAQS.

Third, more improvements of performance analysis support tools and performance measures are necessary. My suggestions are (1) incorporation of performance “probes” for diagnostic model performance investigation, (2) utilization of high resolution measurements such as profiler observation and Lidar, (3) visualization of emissions in a meaningful way, (4) proper treatment of spatial and temporal resolution of aircraft VOCs measurements, (5) improvement in presenting statistical measures, and (6) methods to improve how to carry meta-information related to MPE in graphical measures.

Last, PROMPT needs be applied by state modelers in a real SIP modeling case. One of the most important aspects of PROMPT application is to accommodate the participation of policy makers. The re-evaluation of HGMCR modeling lacks this “real” interaction with policy makers. This is due to the fact that the case study was performed in a retrospective way. That is, the case study was intended to be a third-party evaluation. An MPE study that
carries the traditional approach and the PROMPT-based approach in a real SIP modeling in parallel would clarify the further research needs.
APPENDIX B. Model Performance Evaluation Protocol for Houston-Galveston Mid-Course Review Phase 2 modeling

Introduction

The model performance evaluation (MPE) protocol described here is an instance protocol of the Protocol for Regulatory Ozone Performance Tests (PROMPT) (Kim and Jeffries, 2006b). PROMPT is a meta-protocol that was developed as an alternative guidance protocol for MPE to the EPA’s guideline MPE protocol. PROMPT was designed to guide state modelers for their modeling as well as third party evaluators for MPE of modeling done by state modelers. For the purpose of HGMCR modeling evaluation, we used PROMPT to build a protocol for the third-party evaluation. In general, MPE protocols evolve as new information and new findings are added. Thus, this protocol should be considered as an evolving protocol that reflects most recent (i.e. by the time this protocol was written) scientific understanding, engineering knowledge, and outcomes of previous HGMCR modeling performance assessments.

This protocol frequently refers to HGMCR modeling related information described in ‘Protocol for Ozone Modeling of the Houston/Galveston/Brazoria Area: Combined 1- and 8-hour Ozone Modeling Analysis’ prepared by Texas Commission on Environmental Quality (TCEQ). Hereafter, we will call this modeling document as “TCEQ’s protocol”. The version of TCEQ’s protocol we used was published in February, 2004 and is publicly available on the TCEQ’s HGMCR modeling site (TCEQ, 2006b). Note that TCEQ’s protocol is a modeling protocol and includes its own MPE protocol. What we attempt is to replace the MPE protocol with our protocol for our own analyses.
We consider the ultimate goal of protocol application is to answer following two questions: To what extent can we accept the predictions made by HGMCR modeling at face value for the Texas 2000 SIP development? And if we cannot, then how should we make judgments about the effectiveness of ozone control options that may be proposed in Texas 2000 SIP?

To answer these two questions, four evaluation phases are implemented by following PROMPT. Guidelines for detail implementation and the rationale of PROMPT can be found in the PROMPT paper (Kim and Jeffries, 2006b). Following elements will form each evaluation phase:

- the goal of each phase
- the information required for each phase
- the list of evaluation tasks
- the evaluation material for each evaluation task
- the suggestion for follow-up analyses
- the implication of analysis result for other evaluation phases or tasks
- the documentation requirement

Unless we find significant sources of biases, our assessment for each target of evaluation tasks will be one of statements such as “We agree with the rationale described in corresponding sections of TCEQ’s protocol” or “We conclude that the rationale is the best we can achieve with given resources and knowledge”. The possible other assessment can be “We did not find any significant sources of biases.” We will, however, make more detailed statements delivering assessment results, if necessary.
As a support for following evaluations, we will also create high-resolution GIS maps with ArcGIS. The GIS maps will have, as a minimum, following features: major roads including interstate highways with proper labels, important borders such as county borders, airports, ground monitors, water, and major emission sources.

All plots should reflect the resolution of data. That is any hourly averaged data should be expressed properly. For example, all ground monitors report ozone values as hourly averaged value then the plot should be stair-step type plots to show that the data are not point values. For our evaluation and to make graph production efficient, we will use pyPASS (Kim and Jeffries, 2006c) for the main analysis tool. This protocol is the Version 2 (last modified 2006-03-10) prepared by Byeong-Uk Kim and Harvey E. Jeffries.

Evaluation Phase One (P1)

EP1 aims at answering the following question: Does the HGMCR modeling show or have all necessary components to produce the phenomena that we can expect from the current best conceptual model? The goal of this phase is to confirm that HGMCR modeling is potentially capable of describing a specific ozone problem dealt with in the 2000 SIP.

The goal of EP1 will be achieved by conducting two major tasks:

- Task 1 (P1.1): documenting and reviewing HGMCR modeling system setup
- Task 2 (P1.2): comparing the ozone and nitrogen oxides behavior of the current (‘base5b.psito2n2’) HGMCR modeling with the observed ozone and nitrogen oxides pattern at ground monitor locations that was described in the conceptual model, if any.

Two pieces of information are mandatory for this phase of evaluation.

- an operational model
- a conceptual model
The modeling case to be examined with this protocol is ‘base5b.psito2n2’. We may, however, examine precedent base cases and perform modeling with possible minor variations of ‘base5b.psito2n2’. We will examine the future case to ensure that the model can make good predictions in the base case where the future ozone is high. The conceptual model that we will look is the one described in the Appendix A of TCEQ’s protocol even though we may incorporate additional information, if we can find.

For P1.1, we will conduct following sub-tasks. We will locate the information about how all the model input files were prepared and identify possible sources of input uncertainties by modeling processes. Note that we do not attempt identifying the epistemic root of input uncertainties and the impact of these uncertainties on ozone prediction because it is impossible just by examining inputs only. Rather, we will examine and track how model inputs were prepared, i.e. procedures of preparing input files.

We will review the model configurations described in the TCEQ’s protocol including grid resolutions. However, at this evaluation phase, the review will not be comprehensive because it is hard to do detail analyses without examining the effects of model configurations on model predictions. Thus, more in-depth analyses should be done in subsequent phases of evaluations. We will also check if there is a newer modeling system and their possible impacts on the model predictions.

For P1.2, We will conduct following sub-tasks. We will review the characteristic of meteorology, ozone behavior, and emissions described in the conceptual model proposed in HGMCR modeling protocol. We may look into additional information by incorporating recent literatures. After compilation of the conceptual model, we will examine the model’s behavior by analyzing (1) morning and afternoon resultant winds, (2) daily peak ozone plots
and daily maximum hourly ozone change plots, (4) peak nitrogen oxides plots, and (4) ozone
time series plots. We may add and examine more graphical measures if necessary such as
daily wind speed scatter plots for each monitor and hourly ozone change for a monitor on
each day. All time series plots may span more than a day to provide an “overview”.

For resultant winds analyses, we will draw wind vectors of observation and prediction at
each monitor sites on a GIS map. Morning resultant winds and afternoon results winds
represent the vector sum of hourly wind vectors from 0700 to 1200 and from 1300 to 1800,
respectively. Note that each hour notation indicates the starting of an hour, i.e. 0700 wind
indicates an hourly average wind from 0700~0759.

We will make plots of daily peak ozone concentrations for the modeling period after
sorting monitors in order of their locations and importance. We will divide monitors into
four groups for each day by the model’s predictability of ozone exceedances: false positive
(FP), false negative (FN), true positive (TP), and true negative (TN). Depending on the
availability of resources for MPE, we will examine monitors for subsequent evaluation tasks
in order of the model’s status of predictability as describe above, i.e. from FP to TN. Peak
nitrogen oxides plots will be examined if there are significant discrepancies in predictions.

For time series analyses, we will make ozone time series plots of ozone at each monitor
location. We will use step line plots to clarify the temporal resolution of observation and
model outputs. We may classify the modeling domain into sub-regions, if necessary. Each
time series will be examined in terms of the hour of peak ozone and the ozone concentration
within 3~4 hours of peak ozone observed and modeled. The diurnal patterns between
observation and prediction will be compared and any apparent discrepancies will be marked
for each site on each day for further analyses.
Evaluation Phase Two (P2)

EP2 aims at answering the following question: Can the HGMCR modeling distinguish which precursor(s) to control for ozone reduction? The goal of this phase is to clarify the possible issues in model’s ability to estimate ozone response to precursor control.

Information required for EP2 is as following.

- description of ground observation including location of monitors, types of measurements, and data resolution for each site
- scatter plots of surface winds and hodograms for each site on each day
- scatter plots and time series plots of $O_3$, $NO$, and $NO_2$ (plus $CO$, if available) for each site on each day
- time series of VOCs (continuous measurements preferred) for each site on each day
- proposed control policies (requires communication with policy developers)

The goal of P2 will be achieved by conducting following tasks:

- Task 1 (P2.1): Compiling available observations for the rest of evaluation phases
- Task 2 (P2.2): Listing areas in modeling domain that will be affected by proposed control policies
- Task 3 (P2.3): Examining surface winds with emphasis on those sites in the areas identified in P2.2
- Task 4 (P2.4): Examining $O_3$, $NO$, and $NO_2$ with emphasis on those sites in the area identified in P2.2
- Task 5 (P2.5): Examining VOCs (plus CO, if available) with emphasis on those sites in the area identified in P2.2
• Task 6 (P2.6): Assessing model performance at each site on each day based on results of P2.3 through P2.5

For P2.1, we will conduct following sub-tasks. We will create a table showing available observations at each ground monitor. Following items will be included for each ground monitor:

• Location of monitor
• AIRS ID
• Availability of O₃, CO, and surface winds
• List of species of nitrogen oxides, VOCs, and other measurements (if any)
• Remarks

We will make detail notes about any specific aspects of monitors or measurements as needed.

For P2.2, we will conduct following sub-tasks. We will consult with TCEQ’s control option developers and/or look up the proposed control options in SIP. Following information about the contents of control options will be obtained: location, time, and types of emissions (in terms of VOCs control vs NOₓ control).

During P2.2, we may discuss with policy developers what we find in P1 and P2.1. We will clarify whether another episode selection or other alternative modeling can be worth if the results of P1 and P2.1 do not satisfy the need of policy developer. This process should be iterative and may involve discussion about the possible model precision improvement and design of control options. How to do this is, however, beyond the scope of this protocol.

For P2.3, we will conduct following sub-tasks. We will examine surface winds with wind speed scatter plots and hodograms for each monitor on each day. Hourly averaged
winds will be used as data for these plots. Proper data preparation should be done and justified. The main focus of analysis with wind speed scatter plots is to test if there is any significant wind speed bias for some period. We will divide 24 hours into 4 groups: midnight to morning (0000-0659), morning to noon, (0700-1259), noon to evening (1300-1859), and evening to midnight (1900-2359). Based on our observation of this analysis, we will examine wind error plots and wind plots (i.e. hodograms). With wind error plots, we will identify the relative error of wind speeds and the overall wind direction biases at each monitor. The results of this analysis will guide us to focus data points in some monitors that we need to pay more attention. What we will focus on in wind analysis is that if wind prediction at a site is in general usable for more meaning evaluations. We will make rough estimation of wind direction errors wind speed error that is prevalent. We will use the estimation as a permissible error at the beginning. Note that the actual acceptance of errors, however, will depend on other evaluation results such as the precision anticipated by proposed control options.

For P2.4, We will conduct following sub-tasks. We will examine chemical signals at monitor sites by analyzing scatter plots and time series of NO, NO₂, and O₃. We will pay special attention to NO₂ biases. Scatter plots will make clear how model is biased and the results of scatter plots will help us to identify to locate when those biases occur when we do time series analysis. Similar analysis will be done for O₃ and NO. For O₃ analysis, we will make clear whether model’s O₃ biases show apparent coincidence with NOₓ biases. For example, we will analyze if O₃ underprediction is associated with overprediction of NO by looking at relative strength of NO and NO₂. Also hour of day will be considered because the
presence of solar radiation can change the interpretation of prediction biases of these three species.

For P3.5, we will conduct following sub-tasks. We will review the results of P3.4 at the sites available for P3.5. Because severe NO\textsubscript{x} biases make other analysis hard, we will judge whether we can make meaningful assessment on sites where VOCs (and/or CO) measurements are available. Once we consider a site worth for P3.5, we will first focus on following model species depending on their signal strength in predictions and observations: CO, ETH, OLE, ALD2, FORM, and ISOP. Because CO is much less reactive than other VOCs, we can use CO signal comparison to examine the possible physical process biases in the model. ETH and ISOP are explicitly described by CB4 mechanism that is used in current HGMCR modeling, which is most meaningful when we do predictions-observations comparison of VOCs.

For P3.6, we will conduct following sub-tasks. For each monitor on each modeling day, we will make explicit assessment about the usability of model. Our recommendation for each site on each day will be one of following four categories:

- None
- NO\textsubscript{x} only
- VOCs only
- NO\textsubscript{x} and VOCs

We will conclude ‘None’ when the modeling results show severe flaws at monitors. ‘NOx only’ recommendation will be made when (1) the model shows reasonable NO\textsubscript{x} agreements and surface wind performance (2) but it does not have any VOCs measurements while the site is located distant from monitors have VOCs measurements. Note that this is
crude assessments and more refined assessment will be made in the next phase of evaluation. When \( \text{NO}_x \) concentrations are not highly biased and surface wind performance is reasonable, we may mark sites as ‘\text{VOCs only}’. We will present our assessments in form of maps showing distribution of the quality of model performance over modeling domain to policy developers and clarify any concern related to model performance.

**Evaluation Phase Three (P3)**

P3 aims at answering the following question: How precisely can the model estimate control requirements? The goal of P3 is to test if HGMCR modeling estimates the necessary precision for the control requirement estimation, depending on the precision demanded by policies.

Information required for P3 is as following:

- Policy developer’s feedback for the assessment made in P2.6
- Model’s precision anticipated by policy makers (other basic information should be available through P2)
- Graphical measures made for P2
- Graphical measures with more refined observational data (e.g. aircraft measurements)
- Future case modeling, if available

The goal of P3 will be achieved by conducting following tasks:

- Task 1 (P3.1): Communicating with policy developer about their concern on model’s performance assessed in P2 and selection of days for further analysis
- Task 2 (P3.2): Conducting comprehensive performance analysis at observed locations for selected days
• Task 3 (P3.3): Conducting comprehensive performance analysis at non-observed locations for selected days

• Task 4 (P3.4): Assessing model performance analyses at locations examined in P3.2 and P3.3

For P3.1, we will conduct following sub-tasks. We will provide tables for each day showing the status of model performance. Information included in these tables is: results of P1 (i.e. the predictability of ozone exceedance), the quality of surface wind prediction, any remarks we need to inform policy developers. We will decide which days are worth for further analysis based on discussion with policy developer.

For P3.2, we will conduct following sub-tasks. We will answer the following two sets of questions for each day selected in P3.1:

• If predictions generally match history, is there any way this might be due to compensating errors among processes such that the apparently good match occurs for the wrong reasons? Are the process rates from the modeling system used in the study consistent with those from modeling systems that are apparently working well?

• If predictions do not match the history, what are the likely causes of the failure? Are the physical conditions correctly simulated by the model? Is the wind speed and direction approximately correct? Is the volume of the mixed layer approximately correct? Is the vertical mixing process too slow or too fast? Are emission and deposition processes or magnitudes atypical? Are the chemical rates as expected?

After reviewing all the monitors for each day, the results can be divided into three categories:
• Category MH-R: Those monitor sites where the model matches history reasonably well and there are no indications of compensating error.

• Category MH-A: Those monitor sites where the history matching is ambiguous and there is little evidence as to the cause.

• Category MH-U: Those monitor sites where the physical conditions simulated by the modeling system preclude a good history match for chemical concentrations, especially for secondary products like ozone.


For P3.3, We will conduct following sub-tasks. We will set some focus areas after identifying locations show high ozone concentration in the base case. Then, we will conduct several tasks including but not limited to the following:

• Visualize the model’s inputs for the important processes in the area of interest. This might include plotting the vertical diffusivities over land and water; visualizing the low level and high level emission inputs of the area; visualizing the model’s predicted wind field; and performing dispersion simulations for selected emissions without chemistry to determine how various sources are contributing to the focus area.

• Perform process analysis of the focus region to visualize and understand the interaction among the physical and chemical processes and to explain the state of the chemical transformations.

• Conduct selected sensitivity analyses by varying important inputs or process representations and determine the effects these have on the model. Each sensitivity
analysis may require the performance analysis component and the prediction analysis component, i.e. a recursive application of the model evaluation.

- Review the state of the science and the alternative representations available and assess if the current representation in the model is adequate. It is desirable to acquire auxiliary tools for this procedure such as a modified version of the PAQM, if possible.

After reviewing all such areas for each modeled day, the results can be divided into three categories:

- Category MP-R: Those model locations where the model’s performance is more likely than not adequate and one should accept the predictions as reliable.
- Category MP-A: Those model locations where the model’s performance is ambiguous and there is little evidence as to the cause.
- Category MP-U: Those model locations where the model’s performance is either the physical conditions simulated or chemical conditions simulated by the modeling system are more likely than not resulting in biased results.


For P3.4, we will conduct following sub-tasks. We will conduct traditional statistical evaluations for days qualified for P3 and compare them with those days not included in P3. We will make comprehensive assessment by integrating the results of P3.2 and P3.3 with statistical tests. We will also document issues that we found in P3 and suggest further model performance improvement.

**Evaluation Phase 4 (EP4)**
EP4 aims at answering the following question: What are the possible biases in the prediction and the impact of biases on the policy choice? The goal of EP4 is to assess the potential effects of model biases found in P2 and P3 on the proposed policy options for ozone control.

Information required for P4 is as following:

- Assessment of model prediction’s consistency with conceptual model
- Assessment of model’s support for various precursor controls
- Assessment of anticipated precision

The goal of P4 will be achieved by conducting following tasks:

- Task 1 (P4.1): Performing science-policy bias assessment
- Task 2 (P4.2): Documenting compiling the whole evaluation processes
- Task 3 (P4.3): Making decisions on next steps

For P4.1, we will conduct following sub-tasks. We will present our overall assessment by developing GIS maps showing our confidence on model performance. This map will succinctly show which areas HGMCR modeling shows good performance in modeling domains. Depending on their classification made in P3. We will encode locations in the modeling domain with numerical values representing our confidence. Locations between encoded areas will be interpolated with inverse distance weighting function. Determination of necessary parameters for the function will be documented.

- 1 (Reliable): MH-R or MP-R
- 0 (Uncertain): MH-A or MP-A
- -1 (Unreliable): MH-U or MP-U
For P4.2, we will conduct following sub-tasks. During the course of P1 through P3, we will create publicly accessible documents that contain our judgment and information (or location of information) that we used.

For P4.3, we will conduct following sub-tasks. We will discuss with policy developers (1) if resources and statutory timeline may allow efforts to improve model performance and (2) whether these efforts are worth given that the precision demands of proposed policy options. We will also make recommendation such as pursuing alternative episode or changing modeling system.
REFERENCES


5th Circuit, 2003. BCCA vs. US EPA

66 FR 63972, 2001. Potential Clean Air Reclassification and Notice of Potential Eligibility for Attainment Date Extension and Approval of Attainment Demonstration, Georgia: Atlanta Nonattainment Area; Ozone.


69 FR 8126, 2004. Clean Air Act Reclassification, San Joaquin Valley Nonattainment Area; California; Ozone.


National Research Council., 1991. Rethinking the Ozone Problem in Urban and Regional Air Pollution, National Academy Press, D.C.


Roache, P. J., 1998. Verification and Validation in Computational Science and Engineering, Hermosa, NM.


Space Telescope Science Institute, 2005.
http://www.stsci.edu/resources/software_hardware/numarray


TCEQ, 2004. Protocol for ozone modeling of the Houston/Galveston/Brazoria Area: 
Combined 1- and 8-hour ozone modeling analysis, TX.


TCEQ, 2006c. 
http://www.tceq.state.tx.us/implementation/air/airmod/data/hgb1_camx_domain.html


Tesche, T. W., Georgopoulos, P., Seinfeld, J. H., Cass G., Lurmann, F. W., and Roth, P. M, 
1990. Improvement of Procedures for Evaluating Photochemical Models, California Air 
Resources Board, Sacramento, CA.


Photochemical Oxidant Models, PhD Dissertation, University of North Carolina at Chapel 
Hill.

Tonnesen, S., Jeffries, H. E., 1994. Inhibition of Odd Oxygen Production in the Carbon Bond 


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

ambient air quality standards and proposed regional haze rule.

additional emission reductions, not modeled.


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


