9

4.8 Evaluating Regional-Scale Air Quality Models

Alice B. Gilliland, James M. Godowitch, Christian Hogrefe, and S.T. Rao

Abstract Numerical air quality models are being used to understand the complex interplay among emission loading, meteorology, and atmospheric chemistry leading to the formation and accumulation of pollutants in the atmosphere. A model evaluation framework is presented here that considers several types of approaches, referred to here as the operational evaluation, diagnostic evaluation, dynamic evaluation, and probabilistic evaluation. The operational evaluation is used to reveal the overall performance of the model, and diagnostic evaluation approaches are then used to identify what processes and/or inputs significantly influence the predictted concentrations and whether they are simulated correctly. Dynamic evaluation entails assessing a model's ability to reproduce observed changes in pollutant concentrations stemming from changes in weather and emissions. Probabilistic evaluation approaches will provide the confidence that can be placed on model results for air quality management or forecasting applications. Here, we present example results from several different model evaluation studies that consider questions related to the operational, diagnostic, and dynamic evaluation of a model, and discuss their complementary goals toward model improvements and characterization of model performance.

Keywords Air quality modelling, evaluation methods, regional scale

1. Introduction

Photochemical air quality models are being used to simulate ozone (O_3), particulate matter $\leq$2.5 μg m^{-3} ($PM_{2.5}$), and other pollutants across regional domains. Performance evaluations play a critical role in both regulatory and research applications of the models. For example, air quality model simulations must be evaluated against observational data prior to using the model to make decisions about emission control strategies. In research, improvements to process-level model algorithms or inputs are in part judged based on whether these changes improved model performance. In model applications that have either or both regulatory and research purposes, models can further be used to infer relationships between atmospheric pollutant concentrations and relevant processes, meteorology, and emissions. Given the influence that model evaluation results can have on regulatory decisions and scientific

C. Borrego and A.I. Miranda (eds.), *Air Pollution Modeling and Its Application XIX.*

nclusions about air pollution, it is critical that model evaluation studies are mprehensive and characterize model performance in insightful ways that not ly reveal how well model predicted pollutant levels compare to observed data, ıt also increase confidence in the inputs (e.g., meteorology and emissions) and the odelled processes. Here, a model evaluation framework is presented that organizes evaluation approaches to represent how they differ and complement one ıother, and a few examples are discussed.

. Proposed Air Quality Model Evaluation Framework

ı Figure 1, we present a framework for model evaluation approaches, which is ased on the purpose and specific questions being asked as part of an analysis.

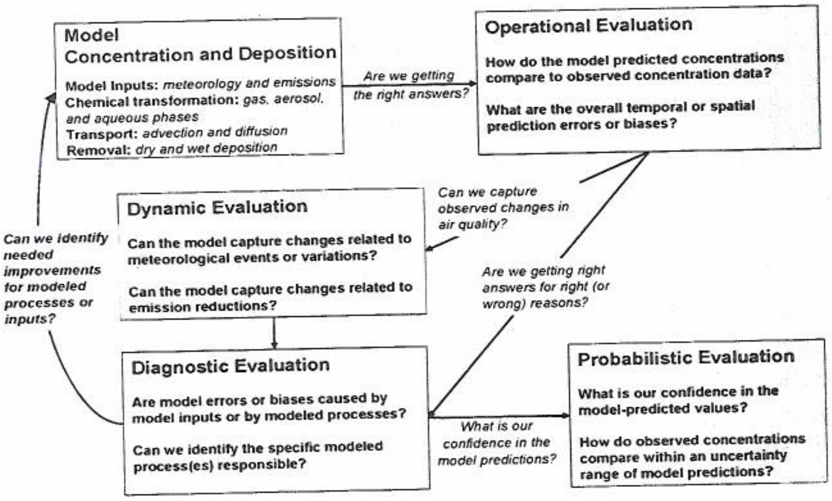
Model
Concentration and Deposition
Model Inputs: meteorology and emissions
Chemical transformation: gas, aerosol, and aqueous phases
Transport: advection and diffusion
Removal: dry and wet deposition
Are we getting the right answers?
Operational Evaluation
How do the model predicted concentrations compare to observed concentration data?
What are the overall temporal or spatial prediction errors or biases?
Dynamic Evaluation
Can the model capture changes related to meteorological events or variations?
Can the model capture changes related to emission reductions?
Can we capture observed changes in air quality?
Are we getting right answers for right (or wrong) reasons?
Can we identify needed improvements for modeled processes or inputs?
Diagnostic Evaluation
Are model errors or biases caused by model inputs or by modeled processes?
Can we identify the specific modeled process(es) responsible?
What is our confidence in the model predictions?
Probabilistic Evaluation
What is our confidence in the model-predicted values?
How do observed concentrations compare within an uncertainty range of model predictions?

Fig. 1 A suggested framework for organizing and identifying the purpose and questions addressed in various evaluation analyses

As the first step in model evaluation, model predictions are compared to observed data and statistical metrics are computed, which is referred to here as "operational evaluation." Typically, most of the observational data is focused on the endpoint pollutants that are monitored for air quality, such as O_3 or $PM_{2.5}$ and component species of $PM_{2.5}$. However, the ability of a model to predict the endpoint pollutant of interest does not address whether the predicted concentrations result from correct or incorrect processes, which is commonly referred to as diagnostic evaluation. For secondary pollutant species that are not directly emitted, diagnostic evaluation methods are critical for insuring confidence in a model as a tool and for identifying model improvements. Figure 1 also includes a new evaluation approach referred to as "dynamic evaluation" that focuses on the model predicted change in air quality concentrations in response to either emission or meteorological changes. This requires historical case studies where known emission changes or meteorological changes occurred that could be confidently estimated, and dynamic evaluation also requires that these changes had an observed impact on air quality. Operational, diagnostic, and dynamic evaluation approaches complement one another by not only characterizing how well the model captured the air quality levels at that time, but how well the model captures the role and contributions of individual inputs and processes and the air quality responses to changes in these factors. For the remainder of this discussion, examples will be shown of how these three approaches in concert capture a more comprehensive evaluation of model performance for specific model applications and support the priority of further model improvement.

A fourth aspect of model evaluation in Figure 1, referred to as probabilistic evaluation, attempts to capture the level of confidence in model results for regulatory or forecasting applications, and a classic example would be ensemble modelling for meteorology forecasting. With computer efficiencies improving exponentially, methods such as ensemble modeling that introduce a range of uncertainties into air quality model predictions become increasingly realistic for decision-making or forecasting. This topic of model evaluation is only included here in a very limited extent, but additional research and advancements are needed to develop more innovative and creative approaches that consider the confidence in air quality models for various applications (see Gégo et al., 2003).

The following examples illustrate how these evaluation approaches can help provide increased confidence that model performance is well characterized and suitable for air quality regulatory and forecast application. Example results are shown using the Community Multiscale Air Quality (CMAQ) model version 4.5 (Byun and Schere, 2006) For the purpose of illustration, only scatterplot illustrations are shown, but it is of course critically important to examine the full range of spatial and temporal scales.

3. Operational and Diagnostic Evaluation Methodologies: Complementary Roles

Previous studies have provided operational model evaluation results for O_3 for both retrospective and forecasting cases (e.g., Eder et al., 2006; Tesche et al., 2006). While the results on average show quite good performance in most studies, the results are often based on more than 500 observational sites and extremely large subcontinental regions. An example of typical operational evaluation results for O_3 are shown in Figure 2, where results from a summer 2002 CMAQ model simulation are compared against observational data. If one looks only at the scatterplot and statistical metrics, it gives the impression that the model performance is very good.

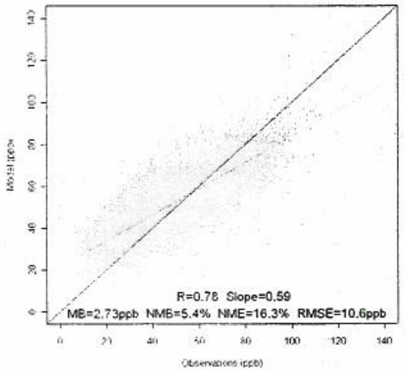

Fig. 2 Example scatterplot for daily 8-hour maximum O_3 from Summer 2005 comparing observations from the Air Quality System (AQS) network and the Community Multiscale Air Quality (CMAQ) model along with mean bias (MB). Normalized Mean Bias and Error (NMB and NME). and root mean square error (RMSE) from the same daily maximum 8-hour O_3 concentrations

However, further analysis of the operational model evaluation results for O_3 elucidates that model performance for O_3 is not equally good across all conditions. For example, Appel et al. (2007) compared model performance at different ranges of O_3 levels as well as evaluation under different synoptic meteorological regimes and demonstrated that the model's underpredictions are associated with high pressure, stagnant conditions typical of high O_3 events in the U.S and overpredictions are associated with frontal passages. Hogrefe et al. (2001) have also shown that the model predictions of O_3 are challenged most for the high-frequency variations that occur below the diurnal time scales. These types of evaluation results are needed to identify specific conditions associated with meteorological forcing that need further diagnostic evaluation for model improvements.

Modeling $PM_{2.5}$ introduces many additional challenges since it is comprised of a number of aerosol chemical species such as sulphate, nitrate, ammonium, organic and elemental carbonaceous materials and because the emission inputs are largely uncertain for many agricultural and diffuse sources. Continued research is needed to refine the modelled representation of the chemical transformation processes as well as the influences of emissions and meteorology. Operational evaluations of $PM_{2.5}$ components such as sulphate aerosol concentrations compare reasonably on the seasonal time scale compared to other aerosol species such as nitrate and carbonaceous aerosols where scientific advancements and model improvements are needed (e.g., Morris et al., 2006).

For model improvement of nitrate, as an example, diagnostic evaluations are needed to identify the factors that contribute to model deficiencies. Bhave et al. (2006) provide a summary of recent diagnostic work to understand and improve nitrate predictions related to chemical transformation processes, specifically the heterogeneous N_2O_5 pathway for HNO_3 production. Gilliland et al. (2003, 2006) and Pinder et al. (2006) demonstrate how critical NH_3 emissions as well as the heterogeneous N_2O_5 pathway can be to nitrate aerosol predictions. Here, an example is shown of additional diagnostic evaluation work that is ongoing to look more carefully at the role of meteorological forcing to wintertime nitrate predictions. Figure 3 illustrates that meteorological model inputs can have a substantial impact on model's predictions of total nitrate, and demonstrates the need for improving the estimated removal via wet and dry deposition.

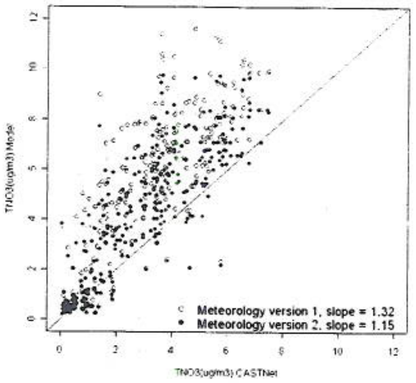

Fig. 3 Predictions versus observations of total nitrate from the Clean Air Status and Trends Network (CASTNet) from January 2002 using two separate sets of meteorological model inputs. "Meteorology version 1" simulation used a non-graupel microphysics scheme and had a large surface temperature cold bias that affect wet and dry deposition. "Meteorology version 2" used the same meteorological model (MM5) but used a microphysics scheme with graupel and had improved surface temperatures

4. Dynamic Evaluations: Challenges and Relevance

The previous examples provide illustrations of how operational and diagnostic evaluation studies can provide initial characterization of model performance issues and direction for model improvement. More uncommon are dynamic evaluation studies that explicitly focus on the model-predicted pollutant responses stemming from changes in emissions or meteorology.

Gilliland et al. (2008) provide the most direct example of a dynamic evaluation study, where air quality model simulations were evaluated before and after major reduction in the NOx emissions. The U.S. Environmental Protection Agency's NOx SIP Call required substantial reductions in NOx emissions from power plants in the Eastern U.S. during summer O_3 seasons beginning in June 2004. Gégo et al. (2007) and USEPA (2006) offer examples of how observed O_3 levels have decreased noticeably after the NOx SIP Call was implemented. Since air quality models are used to estimate how air quality will change due to various emission control strategies, the NOx SIP Call is an excellent opportunity to evaluate a model's ability to simulate the response of O_3 to known and quantifiable O_3 changes. Figure 4 provides an example from this study where changes in O_3 are compared from before (summer 2002) and after (summers 2004 and 2005) the NOx emission reductions occurred. Meteorological differences were much greater between 2002 and 2004 than 2002 and 2005, and, hence, larger O_3 decreases in 2004 were also due to the cooler/wetter conditions in 2004. Figure 4 also illustrates model underestimation of O_3 decreases as compared to observations, which could be due to either the underestimation of NOx emission reductions or a dampened chemical response in the model to those emission changes, or other factors. Analysis methods such as the e-folding distances (Godowitch et al., 2007; Gilliland et al., 2008) have been used show that NOx emissions in these simulations are not impacting O_3 levels as far ownwind as observations suggest, which could be a factor here.

Dynamic evaluation approaches introduce several new challenges. First, retro-ective case studies are needed that offer observed changes in air quality that can e closely related to known changes in emissions or meteorology. The NOx SIP all has offered a very strong case study to test model responses via dynamic eva-iation, but next steps must include further diagnostic evaluation to identify what hemical, physical, or emission estimation uncertainties are contributing to the urrent model results. Findings from additional analysis of this case study can ulti-iately lead to model improvements that are directly relevant to the way air quality iodels are used for regulatory decisions.

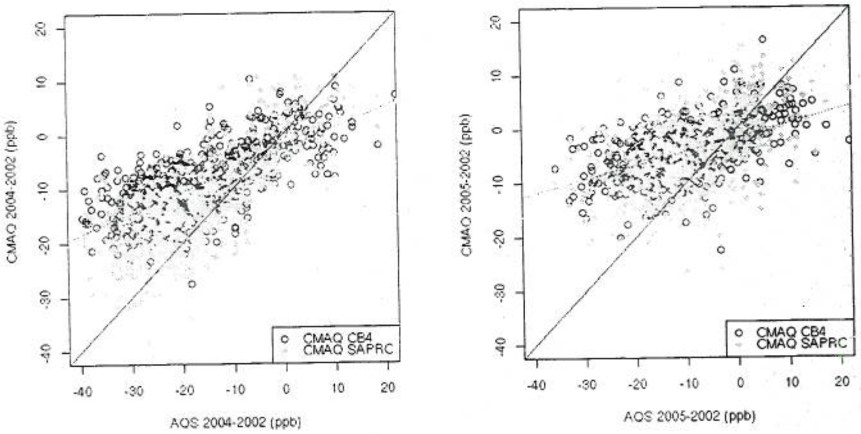

Fig. 4 Summer (2004–2002) and (2005–2002) comparison of the average of upper 95th% of naximum daily 8-hour average O_3 values at the Air Quality System (AQS) network sites in the Eastern U.S. Results are shown using both the CMAQ CB4 and SAPRC99 chemical mechanisms. See Gilliland et al. (2007) for further description

5. Summary

The topic of this paper, evaluation of regional air quality models, is indeed challenging and broad. The intention here is to present a perspective about how many different studies all contribute to a multi-faceted area of research referred to as regional photochemical air quality model evaluation. It can be challenging to characterize model performance for a number of air pollutants via operational methods, but we encourage analyzing model results in ways that characterize model performance across a range of scales and dis-aggregation. Diagnostic evaluation perspectives are needed to extend operational results to the next stage of identifying processes or model inputs that have an influential role on model predictions and how they compare to observations. The model's sensitivity to meteorological and emission uncertainties should also be addressed within a diagnostic evaluation context, as well as the more traditional diagnostic studies such as chemical indicators that consider the chemical state within the model simulation. As a next challenge to traditional evaluation studies, we introduce dynamic evaluation to stress-test the model's ability to reproduce known changes in air quality "forcings" such as meteorological and emission changes that can directly impact the way that air quality models are used in regulatory decision making.

Acknowledgments The research presented here was performed under the Memorandum of Understanding between the U.S. Environmental Protection Agency (EPA) and the U.S. Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) and under agreement number DW13921548. This work constitutes a contribution to the NOAA Air Quality Program. Although it has been reviewed by EPA and NOAA and approved for publication, it does not necessarily reflect their policies or views.

References

Appel KW, Gilliland AB, Sarwar G, Gilliam R (2007) Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impactting model performance; Part I – ozone, Atmos. Environ., 41, 9603–9615.

Bhave et al. (2006) 6th Annual CMAS Conference, October 1–3, 2007, Chapel Hill, NC, http://www.cmascenter.org/conference/2006/ppt/session1/bhave.ppt

Byun D, Schere KL (2006) Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. Appl. Mech. Rev., 59, 51–77.

Eder B, Kang D, Mathur R, Yu S, Schere K (2006) An operational evaluation of the Eta–CMAQ air quality forecast model, Atmos. Environ., 40, 4894–4905.

Gégo et al. Probabilistic assessment of regional scale ozone pollution in the eastern United States (2003) In Air Pollution in Regional Scale. Proceedings of the NATO Advanced Research Workshop, Kallithea, Halkidiki, Greece, June 13–15, 2003. NATO Science Series: IV. Earth and Environmental Sciences. D. Melas, and D. Syrakov (Eds.). Kluwer, Dordrecht, 87–96.

Gégo E, Porter PS, Gilliland A, Rao ST (2007) Observation-based assessment of the impact of nitrogen oxides emissions reductions on ozone air quality over the eastern United States, J. Appl. Met. Climatol., 46, 994–1008.

Gilliland AB, Hogrefe C, Pinder RW, Godowitch JM, Rao ST (2008) Dynamic evaluation of regional air quality models: assessing changes in O_3 stemming from changes in emissions and meteorology, Atmos. Environ. doi:10.1016/j.atmosenv.2008.02.018.

Gilliland AB, Appel KW, Pinder R, Roselle SJ, Dennis RL (2006) Atmospheric environment, seasonal NH_3 emissions for an annual 2001 CMAQ simulation: inverse model estimation and evaluation, Atmos. Environ, 40, 4986–4998.

liland AB, Dennis RL, Roselle SJ, Pierce TE (2003) Seasonal NH_3 emission estimates for the Eastern Unites States using ammonium wet concentrations and an inverse modeling method, J. Geophys. Res.-Atmos., 108, 10.1029/2002JD003063.

ˌdowitch JM, Hogrefe C, Rao ST (2007) Influence of point source NOx emission reductions on modeled processes governing ozone concentrations and chemical/transport indicators, in review with J. Geophys. Res.-Atmos.

ɔgrefe C, Rao ST, Kasibhatla P, Hao W, Sistla G, Mathur R, McHenry J (2001) Evaluating the performance of regional-scale photochemical modeling systems: Part II - O_3 predictions, Atmos. Environ., 35, 4175–4188.

orris RE, Koo B, Guenther A, Yarwood G, McNally D, Tesche TW, Tonnesen G, Boylan J, Brewer P (2006) Model sensitivity evaluation for organic carbon using two multi-pollutant air quality models that simulate regional haze in the southeastern United States, Atmos. Environ., 40, 4960–4972.

nder RW, Adams PJ, Pandis SN, Gilliland AB (2006) Temporally resolved ammonia emission inventories: Current estimates, evaluation tools, and measurement needs, J. Geophys. Res.-Atmos., 111, doi:10.1029/2005JD006603

esche TW, Morris R, Tonnesen G, McNally D, Boylan J, Brewer P (2006) CMAQ/CAMx annual 2002 performance evaluation over the eastern US, Atmos. Environ., 40, 4906–4919.

SEPA (2006) NOx Budget Trading Program, EPA-430-R-07-009. http://www.epa.gov/airtmarkets