SAB REDUCED FORM TOOLS (RFT) REVIEW PANEL
PRELIMINARY MEMBER COMMENTS*
MAY 26, 2020

Table of Contents

Turner, Jay ..... CHAIR OF THE RFT REVIEW PANEL

Belzer, Richard 2

Boylan, James 10

Cox, Jr., Louis Anthony (Tony) 19

Cullen, Alison C. 22

Lange, Sabine 26

Smith, Anne 39

Smith, Richard 48

*Please note the following are preliminary comments and may change as a result of further discussions and deliberations among panel members before final responses to the charge questions are developed.

Belzer, Richard

Question 1. Please comment on the evaluation approach developed by EPA to compare reduced-form models to full-form equivalents.

This approach provides interesting insights into the relative performance of reduced-form tools (RFTs) compared to full-form models (FFMs). We must remember at the outset that FFMs may be biased representations of potential regulatory outcomes. Thus, estimates of bias RFTs may instead be biased FFMs. Uncertainty about key assumptions not examined here, such as the choice of VSL and concentration-response relationship (which includes causality), is correctly identified as a study design limitation (Industrial Economics 2019, p. 4-5). A key question, which the present study did not address and cannot answer, is whether the presumptive biases observed in the RFTs are greater or less than uncertainties (or biases) in these other model parameters.

The remainder of my comments set aside these issues and assume that FFM outputs are always and everywhere unbiased, such that it is a reasonable and appropriate goal for RFTs to approximate FFM outputs as closely as possible.

Please comment on whether the emissions reduction scenarios used in the proposed evaluation approach provide enough diversity to adequately assess reduced-form performance over a range of possible applications (e.g., magnitude, type, and spatial variations of emissions reductions).

The extent to which these case studies are representative cannot be ascertained, and to IEc's credit, any such claim is expressly denied (Industrial Economics 2019, p. 4-4). As an analyst who spends more time on matters other that air quality, I was surprised by the variation in results across RFTs and point-source cases. As for Tier 3, generalization about changes in mobile-source emissions was never feasible given this study design. Tier 3 could only be illustrative.

At this time, the choice of RFT appears to have a material effect on benefit estimates, so if the IEc study's results are generalizable, the choice of RFT implies a choice of bias. Therefore, choosing any of these RFTs may exacerbate controversy, thereby diminishing or eliminating the intended gains from reduced analytic burden during regulatory development.

Please discuss whether the specific assumptions that EPA made to apply each tool as consistently as possible (e.g., emissions, meteorology, use of direct vs. BenMAP estimates, etc.) are appropriate and clearly explained. Please assess whether the report's description of its limitations is complete.

IEc appropriately fixed an array of assumptions to ensure that differences across RFTs were reasonably attributable to their respective architecture. What we cannot tell

from the report is the extent to which results across RFTs would vary if different arrays of fixed assumptions were used instead.[1]

The objectives of the charge to IEc appears to have been appropriately limited, but there are additional criteria that USEPA should consider. First, the Agency should establish a preference for RFTs that have substantial capacity for sensitivity analysis, including both model parameters and underlying assumptions that include substantial variability and uncertainty (e.g., causality and certain VSL parameters (discount rate, lag)). Other evaluative criteria not included by IEc include, transparency, reproducibility, robustness (stability in the face of changes in model specification), and validation.[2]

The IEC report does not (indeed, cannot) illuminate whether the choice of consistent assumptions, however reasonable, affects comparability. That is, does comparability depend on *which* consistent assumptions are imposed on the full-form models and reduced-form tools? Are results robust?

Question 2.

> *Please comment on the results of the reduced form tool evaluation in Section 3, considering both the quantitative and qualitative aspects of the model intercomparison.*

The exhibits in Section 3 would be more easily interpreted if the x-axis was logarithmic. The linear axis gives much greater visual attention to positive than negative biases.

> *Was the information clearly presented and informative?*

I found the discussion in Section 3 clearly presented, informative, and balanced.

> *Were EPA's conclusions reasonable?*

I am aware of no conclusions reached by *USEPA*. Although the IEc report does not include the customary disclaimer slapped on consultant reports, I interpreted it as reflecting the *authors'* analyses and views, not those of the Agency. The extent to which the report reflects *Agency* views should be made explicit to the Panel.

Certain caveats in the IEc report appear justified. For example, RFTs that cannot geographically identify where benefits are projected to be realized have limited utility for

---

[1] This hypothesis can be tested by substituting a different array of fixed assumptions and running the FFMs and RFTs again.

[2] On validation, see, e.g., National Research Council (2007) (which panel preferred the term *model evaluation* to *validation*).

federal regulations that are reasonably expected to have highly variable regional and local impacts.[3]

With respect to IEc's conclusions (Section 4.2), the analysis presented supports several inferences. First, CMAQ and CAMx appear to be interchangeable as FFMs to use as baselines for comparing RFT performance.[4] Second, SA Direct and EASIUR Direct performed relatively well on the Tier 3 case, whereas other RFTs did not.[5] (We do not know if this was real or artifactual.) Third, InMAP's performance was consistently poor, and based on IEc's discussion, it appears to be inferior on other key attributes that are not specific to the chosen case studies. High overhead in the form of time requirements, technical skills, proprietary software, and limited capability to audit results is generally undesirable for modeling that must comply with applicable information quality guidelines (IQGs).[6] The IQGs require transparency sufficient to ensure reproducibility by third parties.[7] This is unlikely to be achieved when analytic results depend on high-overhead RFTs. Before giving the go-ahead on complex models that are difficult to implement, reproduce, and audit, the Agency should devise an information quality compliance plan that absolutely ensures third-party reproducibility. The first step in implementing that plan is for the Agency itself to reproduce RFT-based

---

[3] Some advice is too vague to be useful. See, e.g., Industrial Economics (2019, p. 4-4): "[F]uture users should carefully consider the specific characteristics of a policy before deciding whether a specific reduced-form tool is or is not a good fit for estimating benefits."

[4] IEC goes too far when it says its analysis "validates" this inference; see Industrial Economics (2019, p. 4-1). Models that are highly non-independent cannot be used to make validation claims. IEc is on much firmer ground stating, "we can have confidence that the performance of reduced-form tools relative to CMAQ estimates would also hold if CAMx were the full-form model being used." And even this inference depends on the four point-source scenarios being representative. What would be helpful, but is missing in the IEc report, is comparisons of CMAQ and CAMx on other evaluative criteria such as ease of use, transparency, and capacity to perform robustness checks.

[5] I looked in vain for an explanation for this but drew a blank. It is important to know not just that SA Direct and EASIUR performed better on Tier 3; it's important to understand why. For example, was this result peculiar to the Tier 3 case or was it plausibly related to mobile source emissions generally?

[6] Office of Management and Budget (2002); U.S. Environmental Protection Agency (2002).

[7] Office of Management and Budget (2002, p. 8460): "'Reproducibility' means that the information is capable of being substantially reproduced, subject to an acceptable degree of imprecision. For information judged to have more (less) important impacts, the degree of imprecision that is tolerated is reduced (increased)… With respect to analytic results, 'capable of being substantially reproduced' means that independent analysis of the original or supporting data using identical methods would generate similar analytic results, subject to an acceptable degree of imprecision or error."

results prepared by its contractors; permit rigorous external and independent peer review, including review of information quality compliance; and show its work.

Some of IEc's inferences appear less supportable by the evidence provided. First, statements suggesting SA Direct and EASIUR Direct would perform well in other cases is not justified by the limited capacity to extrapolate from these case studies.[8] Second, generalization to mobile-source emission scenarios obviously is not justified based on the single case of Tier 3. Third, it is not clear how well any RFT should predict variation in FFM results. IEc may be correct that its reported 0.88 R^2 for EASIUR Direct is "reasonable," but that policy judgment is not IEc's to make.[9] Moreover, elsewhere IEc describes RFT outputs as "a quicker approach to generating *ballpark estimates* of the health-related benefits or costs associated with an air quality policy" (Industrial Economics 2019, pp. ES-7 and 5-1) – a considerably less demanding use than a Regulatory Impact Analysis. The Panel needs clarity about how USEPA intends to use RFTs, and in particular, the extent to which these estimates would be *influential information*[10] or *highly influential information.*[11]

IEc appears to have struggled to interpret differences in RFT results, suggesting that they thought the tools lacked transparency. USEPA can gain important lessons about the importance of transparency, proved by rigorous pre-dissemination review, by considering infamous missteps related to COVID-19 pandemic modeling. On March 16, 2020, the Imperial College London (ICL) COVID-19 Response Team predicted that an "unmitigated epidemic" "would result in 2.2 million U.S. fatalities and 510,000 U.K fatalities, "not accounting for the potential negative effects of health systems being overwhelmed on mortality" (Ferguson et al. 2020, p. 7 [hereafter "Report 9"]).[12] They also estimated 1.1–1.2 million U.S. fatalities "even if all patients were able to be treated" (p. 16) These predictions apparently led U.K. and U.S. decision-makers to panic and

---

[8] *See* Industrial Economics (2019, p. 4-2) (SA Direct and EASIUR Direct "would perform in a similarly reliable way for air quality policies beyond those considered in this analysis").

[9] *See* Industrial Economics (2019, p. 4-2). IEc reports EASIUR Direct "did a reasonable job capturing variation in benefits across large regions of the US," but does not report whether it did so *consistently* across regions, or it did so only on average. Regional differences often matter, and the lack of regional resolution in EASIUR limits its utility n analyzing polices with significant regional differences.

[10] Office of Management and Budget (2002, p. 8460 [definition of "influential"]).

[11] Office of Management and Budget (2005, pp. 2675-2676).

[12] These two qualifiers operate in opposite directions, the net effect of which is not clear. There is no evidence in Report 9 that the authors intended their estimates to be suggestive or illustrative.

impose Draconian economic shutdowns, which have had astoundingly large (but readily predictable) adverse effects on health and welfare.

Another predictable consequence of the ICL model's prominence and use was a demand for full disclosure. The ICL team declined to do so; elsewhere it was reported that their model was "a single 15,000 line file that had been worked on for a decade." Rather, the model code was "upgraded for over a month by a team from Microsoft and others," yet it "produce[d] very different results given identical inputs" (Denim 2020a). In a follow-up analysis, Denim (2020b) identified numerous additional errors, and further alleged that the ICL team engaged in unprofessional conduct, including deleting information and making false statements. Meanwhile, the updated code produced substantially different results than the original model used to produce the mortality estimates. The ICL team has since endured withering criticism, with likely reputational loss.[13]

> *Are there other results which would be useful to include in the comparison?*

Question 3. Exhibit ES-4 "Ratio of National Avoided Premature Mortality Benefits Estimates," shows how different reduced-form tools generated different estimates as compared to full-scale air quality models.

Exhibit ES-4 (and Exhibits 3-2, 3-3, and 3-4) are presented with linear x-axes. This makes upward biases more visually prominent than downward biases. More insight would be gleaned by using a log x-axis.

> *3a. Does the report provide a clear and thorough explanation for why some tools under- or over-estimated $PM_{2.5}$ health benefits as compared to the full-scale air quality modeling? Please add any additional explanations for the pattern of results observed.*

The reported explanation for why RFTs under- or overestimated CMAQ outputs focuses on tools' relative ability to predict the nitrate component of $PM_{2.5}$ and its effects (Industrial Economics 2019, p. 4-3). And Exhibit 3-4 shows extensive overestimation of benefits attributed to the nitrate component, though with the notable exception of AP3 BenMAP, which underestimated these benefits. Yet this does not explain systematic *under*-estimations for $prPM_{2.5}$ and sulfate by AP2 BenMAP, AP2 Direct, and EASIUR Direct. It appears that nitrate is a generalized problem across RFTs, but RFTs have their own model-specific biases, which are manifest across scenarios.

[13] A similar story can be told about the University of Washington Institute for Health Metrics and Evaluation (IHME) model.

3b. How do the results of this study inform our understanding of the suitability of these tools for regulatory economic analyses in their current form?

Reduced analytic burden is certainly a valuable goal, but it is not the only public policy goal. Nor does it capture the full array of costs the Agency bears when developing benefit estimates for air regulations. The reduced cost of analytic effort prior to issuing a proposal may be diminished or eliminated by increased workload responding to public comments and making revisions. A similar tradeoff arises at the final rule stage, where the choice of RFT could affect litigation risks and defense costs.[14]

The IEc report indicates that the single mobile-source case (Tier 3) is demonstrably different and most RFTs' performance is highly unsatisfactory. The best-case scenario is that Tier 3 is a one-off, and RFTs would do well in a range of other mobile-source applications. But we can't know this, and nothing in the IEc report indicates that any of the RFTs is ready for use for mobile-source regulations at this time.

Given the unknown representativeness of the case studies examined, combined with substantial biases and variability across RFTs, perhaps the best path forward is to collect more evidence by applying the method used by IEc to additional historic cases, and testing other hypotheses to identify additional structural differences. Each new application provides useful information to the Agency and RFT developers alike. This may enable developers to investigate sources for observed bias and modify their respective tools to remove it. Such iterative processes aren't always popular because they postpone the day USEPA can reliably substitute RFTs for FFMs. Nonetheless, iterative processes tend to lead to improvements that, at some point, would make this substitution a reasonable compromise.

Several RFTs performed especially poorly with respect to nitrate (Ex. 3-4).[15]

3c. Can any of the reduced-form tools explored in this report easily be modified to allow quantifying the extent to which the total health benefits accrue to specific geographic areas (e.g., by state, or where ambient concentrations are above or below the NAAQS)?

Given my primary experience is not in air quality modeling, I defer to others with respect to what specific modifications might be appropriate.

---

[14] Industrial Economics (2019, p. 3-4): "There is no apparent consistent difference between the performance of BPT reduced-form tools and the reduced-form air quality tools coupled with BenMAP." Thus, the choice of RPT could be construed as arbitrary and capricious.

[15] The scale in Exhibit 3-4 ranges from 0 to 10, not the range of 0-4.5 in Exhibits 3-2 and 3-3. This results in visual incomparability that may misled readers.

Question 4. Since 2008 EPA has used SA-BPT to estimate the health impacts of numerous regulations. Under the scenarios examined in this report, EPA's SA-BPT approach over-estimated $PM_{2.5}$-related health benefits by between 10 and 30 percent, depending on the sector. To ensure BPT estimates correspond to full-form results as closely as possible, the report recommends updating the underlying emissions inventories and air quality modeling used to inform the EPA SA-BPT approach over time.

4a. In the interim, how might EPA improve its characterization of results derived from the 2005 SA-BPT approach, specifically the potential degree of over- or underestimation in BPT-based results for a particular regulatory scenario?

The first step is to use this information in the context of retrospective analysis, required pursuant to EO 13563 (Obama 2011) and applicable Information Quality Guidelines (Office of Management and Budget 2002; U.S. Environmental Protection Agency 2002). If USEPA is confident that the FFMs are always "right" and the purpose of RFTs is cost-effectively get closer to "right," then the Agency should consider revising RFT-based benefit estimates for past rulemakings accordingly.

Where USEPA used a RFT to estimate regulatory benefits and the Agency has decided to repeal the regulation in question, it should use the same RFT for the deregulatory action. Whatever bias was imparted by the RFT in the original benefit estimate, using it again is likely to reduce or eliminate it.

4b. What criteria (e.g., geographical scale, regulated sector, pollutants/precursors) should EPA examine to determine the potential for divergence between SA-BPT results vs full-form air quality modeling results (resulting in over- or under-estimation)?

4c. Based on the results of this study, does the panel have any additional recommendations about BPT-based approaches?

Question 5. How do the results of this study inform the future development of reduced-form tools that are capable of providing reliable estimates of impacts associated with different sectors, across a variety of spatial scales, and for different portions of the air quality distribution? Are there other, less resource intensive approaches than full-scale air quality modeling for informing the public about the size and distribution of PM health benefits associated with alternative regulatory scenarios?

References

Denim, Sue. [Pseudonym]. 2020a. "Code Review of Ferguson's Model." In *Lockdown Sceptics*, edited by Toby Young. London UK: Lockdown Sceptics. Available: https://lockdownsceptics.org/code-review-of-fergusons-model/.

———. [Pseudonym]. 2020b. "Second Analysis of Ferguson's Model." In *Lockdown Sceptics*, edited by Toby Young. London UK: Lockdown Sceptics. Available: https://lockdownsceptics.org/second-analysis-of-fergusons-model/.

Ferguson, Neil M, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Ilaria Dorigatti, Han Fu, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Lucy C Okell, Sabine van Elsland, Hayley Thompson, Robert Verity, Erik Volz, Haowei Wang, Yuanrong Wang, Patrick GT Walker, Caroline Walters, Peter Winskill, Charles Whittaker, Christl A Donnelly, Steven Riley, and Azra C Ghani. 2020. "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand." London UK: Imperial College COVID-19 Response Team. Available: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/.

Industrial Economics. 2019. "Evaluating Reduced-Form Tools for Estimating Air Quality Benefits; Final Report." Cambridge MA: IEc.

National Research Council. 2007. *Models in Environmental Regulatory Decision Making* (National Academies Press: Washington DC).

Obama, Barack. 2011. "Executive Order 13563: Improving Regulation and Regulatory Review," *Federal Register*, 76: 3821-23.

Office of Management and Budget. 2002. "Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies; Notice; Republication," *Federal Register*, 67: 8452-60.

———. 2005. "Final Information Quality Bulletin for Peer Review," *Federal Register*, 70: 2664-67.

U.S. Environmental Protection Agency. 2002. "Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Environmental Protection Agency." *EPA/260R-02-008*. Washington DC: USEPA.

Boylan, James

Question 1. Please comment on the evaluation approach developed by EPA to compare reduced-form models to full-form equivalents.
Overall, the guiding principles of the evaluation approach seem reasonable. It was good that the report examined results from two full-form models (i.e., CMAQ and CAMx) so that any biases associated with using CMAQ alone as the full-form comparator could be identified. It is unfortunate that the CAMx results were not available for Tier 3. Considering multiple (four) reduced-form models (RFMs) with multiple (five) policy scenarios was good since it allowed for comparisons across different geographic distributions, magnitudes of emissions species, temporal patterns of emissions, and emission release heights.

The report should include more details describing how BenMAP-CE works, including the governing equations and the data sources. This would give the reader a better understand of how air quality concentration fields are converted into health benefits. In addition, the report should include more details related to the algorithms adopted for each reduced-form tool. For example, APX uses source-receptor matrices produced by the Climatological Regional Dispersion Model (CRDM); InMAP starts with source-receptor relationships from WRF-Chem and estimates air quality surfaces with variable grid resolutions using its own dispersion-reaction algorithm; and EASIUR relies on statistical regression for emissions and benefits based on air quality fields derived as "averaged plume" out of randomly selected full-form modeling (i.e., CAMx PSAT) grid cells. These details are not part of the report although this type of information is critically important to better understand the RFM formulation and the strengths and weaknesses of each RFM (e.g., meteorology for different years can be explicitly considered in InMAP, while it cannot in other models).

Reduced surface models (RSMs) estimate concentrations based on concentrations out of full-form models). Reduced form models (RFMs) use unique algorithms to estimate concentrations. It would be helpful to discuss the differences between RSMs and RFMs and identify which models in the report are RSMs and which are RFMs. This might provide insight when comparing the reduced form tools to the full-form models. Using air quality surface outputs from some RFMs as inputs to BenMAP was a good approach to examine additional RFM options. In addition, it would be helpful to see the analysis conducted in two separate steps with the first part comparing concentration fields generated by RFMs and full-form models and the second part comparing monetized benefits estimated by each RFM and BenMAP using a consistent concentration field. This would help readers better understand which component (concentration fields or estimated monetized benefits) were responsible for differences between the RFMs and full-form models.

Please comment on whether the emissions reduction scenarios used in the proposed evaluation approach provide enough diversity to adequately assess

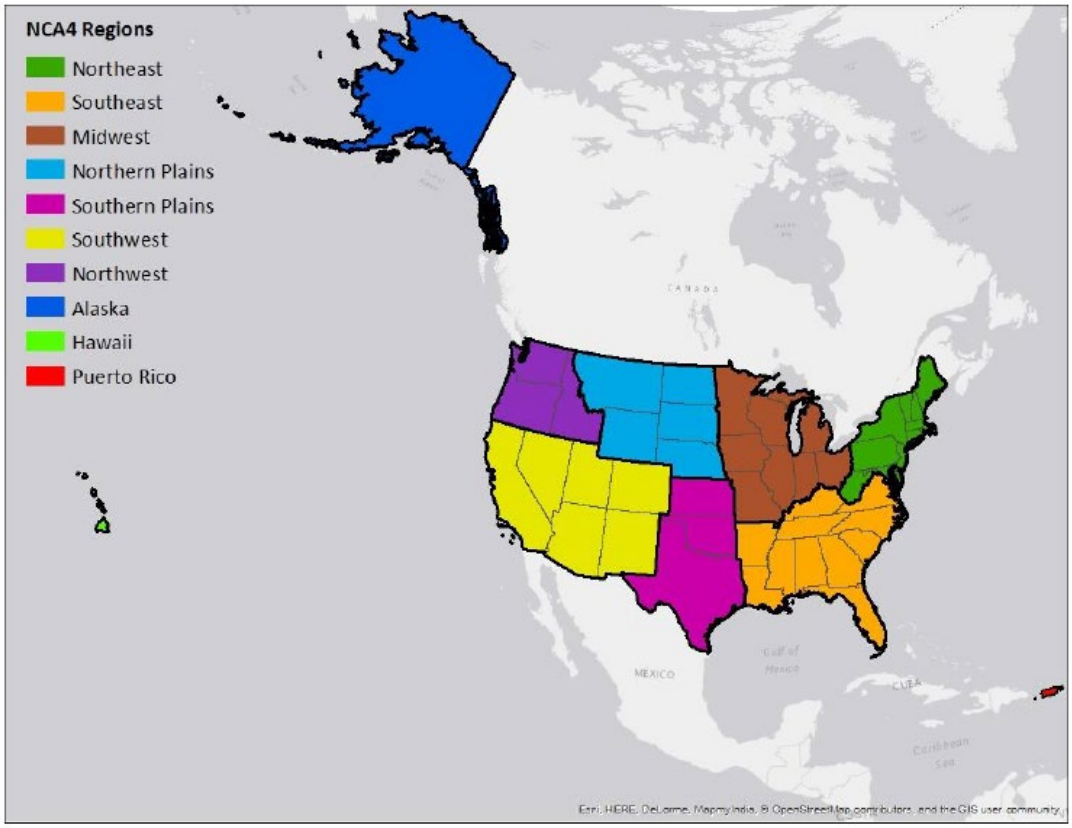
reduced-form performance over a range of possible applications (e.g., magnitude, type, and spatial variations of emissions reductions).
For this initial report, the five emission reduction scenarios provide a diverse range of possible applications. However, to fully evaluate the full range of possible applications, additional scenarios should be evaluated including area sources (e.g., residential wood combustion), MAR (marine/aircraft/rail) sources, additional industrial point sources, and on-road diesel emission reductions.

Please discuss whether the specific assumptions that EPA made to apply each tool as consistently as possible (e.g., emissions, meteorology, use of direct vs. BenMAP estimates, etc.) are appropriate and clearly explained.
It is good that EPA made an attempt to standardize key inputs such as all-cause mortality and the use of a consistent VSL value. However, some assumptions are not clearly described. For example, it was not clear why 36 km inputs were used for EASIUR only. In addition, the population and health incidence data for the EASIUR runs were not consistent with inputs to other RFMs. A more detailed explanation and potential implications of these differences should be discussed.

Please assess whether the report's description of its limitations is complete.
In general, most major limitations were adequately described in Section 4.4 (Limitations of the Analysis). However, one major limitation was mentioned multiple times in the report but was not included in Section 4.4. Specifically, the limitation that BPT approaches assign health impacts to the county in which the emissions change occurs rather than where the health impacts really occur should be added to Section 4.4.

Question 2. Please comment on the results of the reduced form tool evaluation in Section 3, considering both the quantitative and qualitative aspects of the model intercomparison. Was the information clearly presented and informative?
Overall, results were adequately documented for the intended evaluations. However, this report did not include any region-by-region results similar to the tables and graphics developed for the national-scale evaluations. Section 3.2 presents regional statistics, but they seem to be averaged across all regions and not broken down by individual regions. It is unclear if the regional average is a simple average of the regional statistics or a weighted average based on the regional health impacts. Exhibit C-2 is hard to digest and results might be better suited for presentation in graphical format. Exhibit C-2 includes "regional statistics", but they seem to be averaged across all regions and not broken down by individual regions. Exhibit B-1 contains the states in each NCA region. The report does not define the acronym "NCA". Also, it looks like multiple states are contained in multiple NCA regions. It is not clear if the whole state is included in these regions or just part of the state. If they are partial states, this should be indicated with "(partial)" next to the state name. The report gives a reference for the NCA regions on page 2-18 (https://www.epa.gov/cira); however, it does not clearly indicate which report contains the NCA regions. It was assumed that it is the report titled "2017 Report". In this report, the following map was found on page 17:

Figure 1.6. NCA4 Regional Aggregations[53]



This map is drastically different than the breakdown of states listed in Exhibit B-1. In addition to the list of states in each NCA region, the report should add a map to clearly show the groupings.

Finally, Exhibit 3-8 should include a breakdown of the time requirement for “Pre-Processing”, “Post-Processing”, and “Model Run”. The “High” time requirement definition should include an upper limit number of hours to help put these models into perspective compared to full-form models.

Were EPA’s conclusions reasonable?
EPA’s conclusions were generally reasonable based on the information presented in the report. However, more detailed and meaningful conclusions could be presented if some of the additional analyses presented in these comments were performed.

Are there other results which would be useful to include in the comparison?
As previously discussed, it would be helpful to see the analysis conducted in two separate steps with the first part comparing concentration fields generated by RFMs and full-form models and the second part comparing monetized benefits estimated by each RFM and BenMAP using a consistent concentration field. This would help readers better understand which component (concentration fields or estimated monetized benefits) were responsible for differences between the RFMs and full-form models.

Also, the report should include region-by-region results similar to the tables and graphics developed for the national-scale evaluations.

Question 3. Exhibit ES-4 "*Ratio of National Avoided Premature Mortality Benefits Estimates*," shows how different reduced-form tools generated different estimates as compared to full-scale air quality models.

3a. Does the report provide a clear and thorough explanation for why some tools under- or over-estimated $PM_{2.5}$ health benefits as compared to the full-scale air quality modeling? Please add any additional explanations for the pattern of results observed.

The report attempted to explain observed differences. However, the explanations are not as thorough as they could be and do not discuss the root cause of the differences. The report should start by discussing differences observed in $prPM_{2.5}$ concentration fields because $prPM_{2.5}$ results are driven more by transport (advection and diffusion) rather than chemistry. Then, the report should add a discussion of the added complexities associated with secondary $PM_{2.5}$ formation due to photochemistry and aerosol dynamics. For example, the amount of sulfate and nitrate production is related to the ozone formation and the presence of OH· radicals. When photochemistry is diminished (e.g., nighttime and wintertime) or under high NOx conditions (e.g., inside stack plumes), NOx can titrate ozone and slow the secondary formation of sulfate and nitrate. In addition, the amount of free ammonia in the atmosphere will have a significant impact on the formation of nitrate PM since the nitrate must be fully neutralized with ammonium (ammonium nitrate, NH_4NO_3). However, the amount of free ammonia will have a smaller impact on the formation of sulfate since sulfate can exist as ammonium sulfate ($(NH_4)_2SO_4$ which is fully neutralized), ammonium bisulfate (NH_4HSO_4 which is half neutralized), or sulfuric acid mist (H_2SO_4 which is not neutralized). If the report had performed the analysis in two separate steps with the first part comparing concentration fields generated by RFMs and full-form models and the second part comparing monetized benefits estimated by each RFM and BenMAP using a consistent concentration field, it would be much easier to distinguish estimated benefit differences due to air quality concentration fields compared to the estimated monetized benefits step.

In most regulatory applications, full-form model results are not used directly from the model. EPA's "Modeling Guidance for Demonstrating Air Quality Goals for Ozone, $PM_{2.5}$ and Regional Haze" (EPA, 2018) recommends that when air quality models are used for regulatory application to predict future year concentrations and future year control scenarios that the models be used in a "relative" sense rather than an "absolute" sense:

> *Air agencies should determine whether a control program scenario will provide sufficient emission reductions to demonstrate attainment of the NAAQS using the modeled attainment test. The modeled attainment test is a technical procedure in which an air quality model is used to simulate base year and future air pollutant concentrations for the purpose of demonstrating attainment of the relevant NAAQS.*

> *The recommended test uses model estimates in a "relative" rather than "absolute" sense to estimate future year design values.*
>
> *…this approach has the effect of anchoring the future concentrations to a "real" measured ambient value, which is important given model bias and error in the base year simulation(s). It is reasoned that factors causing bias (either under or over-predictions) in the base case will also affect the future case.*
>
> *The EPA has developed the Software for Modeled Attainment Test-Community Edition (SMAT-CE) tool to enable completion of the modeled attainment tests for $PM_{2.5}$ and ozone, as well as for calculating changes in visibility in Class I areas.*
>
> *The modeled attainment test is primarily a monitor-based test. As such, the focus of the attainment test is whether attainment can be reached at existing monitors. An additional "unmonitored area analysis" can also be performed to examine ozone and/or $PM_{2.5}$ concentrations in unmonitored areas.*

Many times, absolute modeled nitrate concentrations are significantly over predicted by the full-form photochemical models. The approach described above reduces biases in future year projections and policy scenarios by using the model in a "relative" sense rather than an "absolute" sense. Therefore, the future nitrate concentrations calculated with SMAT-CE can be significantly lower than the absolute nitrate concentrations directly from the model.

It is unclear if the CMAQ and CAMx full-form models used in this report are used in a "relative" or "absolute" sense. Also, it is not clear how the RFMs will be impacted by this choice. The report should clearly describe the approach used for each RFM. If the CMAQ and CAMx full-form models evaluated "relative" impacts and the RFMs evaluated "absolute" impacts, that could help explain the large nitrate overpredictions by the RFMs compared to the full-form models.

On page 3-2, the report states "Some reduced-form tools tend to consistently underestimate CMAQ benefits, while others tend to overestimate." The report should list the tools that fit into each category. It looks like SA Direct and InMAP consistently overestimate benefits; however, no tools seem to consistently underestimate benefits.

On page 3-4, the report states "AP2 BenMAP, AP2 Direct, and EASIUR Direct all underestimate CMAQ benefits except for Tier 3, while SA Direct, AP3 BenMAP, AP3 Direct, and InMAP BenMAP all overestimate CMAQ results to varying degrees." AP3 BenMAP and AP3 Direct do not overestimate CMAQ results for Pulp and Paper.

On page 3-4, the report states "Of all the models, AP3 BenMAP and AP3 Direct estimates of health benefits are within 10% of CMAQ benefits estimates for more

scenarios (3: CPP Proposal, Cement Kilns, and Pulp and Paper) than any of the other reduced form tools." AP3 Direct is within 10% for two scenarios, not three.

On page 3-8, the "Nitrate" chart shows a ratio of 0.0 for AP3 BenMAP with Pulp and Paper compared to a ratio of 1.8 for AP2 BenMAP with Pulp and Paper and a ratio of 2.4 for AP3 Direct with Pulp and Paper. This large discrepancy between similar models should be examined and explained in the report.

3b. How do the results of this study inform our understanding of the suitability of these tools for regulatory economic analyses in their current form?
In their current form, some RFM tools could be used for regulatory economic analysis, but only for specific policy scenarios at the national level after comparison to full-form models and acceptable model performance. Criteria for acceptable model performance should be developed and used to guide these applications. "Performance Goals" and "Performance Criteria" could be adopted from the recommendations published by Emery et. al. (2017) and Boylan and Russell (2006).

3c. Can any of the reduced-form tools explored in this report easily be modified to allow quantifying the extent to which the total health benefits accrue to specific geographic areas (e.g., by state, or where ambient concentrations are above or below the NAAQS)?
It appears that SA Direct can be modified to produce regional/state/county-level outputs easily. For the other tools, their native outputs are readily usable for specific geographic areas with minimal efforts. Once county level benefits are obtained, it is simple to aggregate by state or nonattainment areas.

Question 4. Since 2008 EPA has used SA-BPT to estimate the health impacts of numerous regulations. Under the scenarios examined in this report, EPA's SA-BPT approach over-estimated $PM_{2.5}$-related health benefits by between 10 and 30 percent, depending on the sector. To ensure BPT estimates correspond to full-form results as closely as possible, the report recommends updating the underlying emissions inventories and air quality modeling used to inform the EPA SA-BPT approach over time.

4a. In the interim, how might EPA improve its characterization of results derived from the 2005 SA-BPT approach, specifically the potential degree of over- or underestimation in BPT-based results for a particular regulatory scenario?
EPA could scale the benefit results up or down based on the degree of over- or underestimation in BPT-based results for a particular regulatory scenario. While this approach could use ratios based on total $PM_{2.5}$ at the national scale, it would be much better to use scaling ratios that are species-specific (sulfate, nitrate, $prPM_{2.5}$) and at the state or regional scale. For example, if the RFM has a +10% relative bias in EC predictions compared to the full-form model, benefits due to 1 TPY EC reduction can be reduced 10% to account for the model bias.

4b. What criteria (e.g., geographical scale, regulated sector, pollutants/precursors) should EPA examine to determine the potential for divergence between SA-BPT results vs full-form air quality modeling results (resulting in over- or under-estimation)?
In addition to examining $PM_{2.5}$ components ($prPM_{2.5}$, sulfate, nitrate, ammonium, SOA), EPA should examine model performance by season of the year. This might identify seasons where the models perform poorly (e.g., nitrate overprediction in wintertime) and help identify methods for reducing divergence between models on an annual basis.

4c. Based on the results of this study, does the panel have any additional recommendations about BPT-based approaches?
Any BPT-based approach will have an issue dealing with long-range transport of pollutants such as secondary $PM_{2.5}$. Therefore, it is the best to use a BPT-based approach at very large geographic scales (e.g., national scale). In addition, any BPT and/or RFM approach that utilizes a source apportionment technique to prepare underlying source-receptor relationship will suffer performance issues when direct/indirect NH_3 effects are involved in secondary inorganic $PM_{2.5}$ formation. For policy applications, it might be more appropriate to use a sensitivity approach (change in concentrations due to changes in specific emissions) such as the Brute Force method (Hwang *et al.*, 1997 and Clappier *et al.*, 2017) or High-Order Decoupled Direct Method in Three Dimensions (Zhang *et al.*, 2012 and Huang *et al.*, 2017) rather than a source apportionment approach (contribution to concentrations from specific emissions). Specifically, EPA should review Clappier et al., 2017 and discuss the potential impacts of not using brute-force runs to account for complex PM chemistry.

Question 5. How do the results of this study inform the future development of reduced-form tools that are capable of providing reliable estimates of impacts associated with different sectors, across a variety of spatial scales, and for different portions of the air quality distribution?
Future developments of reduced-form tools should:

- Evaluate concentrations fields and benefits separately
- Evaluate reduced surface models (RSMs) vs. reduced form models (RFMs)
- Focus on understanding the root-cause reasons for differences between models
    - Nitrate model performance
    - Surface level emissions
    - Seasonal impacts
    - Region specific analyses
    - Benefit algorithms
- Develop performance guidelines for acceptable model performance

Are there other, less resource intensive approaches than full-scale air quality modeling for informing the public about the size and distribution of PM health benefits associated with alternative regulatory scenarios?
On page 2-9, the report states "We conducted an extensive literature review to identify reduced-form approaches for predicting policy-related air quality changes and associated benefits.[10] Based on this review, we selected four reduced-form tools for this

analysis." The detailed literature review to identify all reduced-form models and the selection of the four reduced-form tools are a critical part of this report. Footnote "10" refers to a personal communication memorandum (November 17, 2017). The reference to a single personal communication memorandum seems far from "an extensive literature review." The report should include a copy of the personal communication memorandum in the Appendix. Also, the report should list all references that were reviewed and list all the reduced-form tools that were considered for selection. Finally, the report should discuss why the four reduced-form tools were ultimately selected for this report while others were not selected.

In terms of alternative RFM options, EPA should review ABaCAS (http://www.abacas-dss.com/abacas/Default.aspx). In fact, ABaCAS is a reduced-form model framework developed by EPA itself.

Additional Comments

On pages ES-3 and 1-2, the report refers to "nitrous oxides" rather than "nitrogen oxides". "Nitrous oxide" is N_2O while "nitrogen oxides" is NO_X ($NO + NO_2$).

On page 2-11, the report states "For this analysis, the original Fann et al. (2012) SA BPT values were updated in December 2017. The Fann et al. (2012) BPT values were used with a newer version of BenMAP-CE v. 1.3.7.1, which included updated population, baseline incidence rates, and income growth, in currency year 2015.[13]" It is not clear if these are same values reported in footnote #12 (US EPA. Technical Support Document: Estimating the benefit per ton of reducing $PM_{2.5}$ precursors from 17 sectors. Retrieved from: https://www.epa.gov/sites/production/files/2018-02/documents/sourceapportionmentbpttsd_2018.pdf). Also, additional details should be provided for the error that IEc discovered in the baseline mortality rates in BenMAP-CE and discuss how they concluded that the error would result in "the overestimation of benefits by less than three percent for aggregate benefits values."

Exhibits 2-3, 2-4, 2-5, 2-6, and 2-7 show total emission change maps for each scenario. It would be informative to see additional emission change maps for $prPM_{2.5}$, SO_2, and NO_X for each scenario.

The benefit per ton (BPT) values used in this report should be included in the Appendix.

The reference section did not include the reference cited on page 2-17 (Boylan and Russell, 2006). The report should add "Boylan, J.W. and Russell, A.G. (2006) PM and Light Extinction Model Performance Metrics, Goals, and Criteria for Three-Dimensional Air Quality Models; *Atmos. Environ.*, Volume 40, pp. 4946-4959".

References

Clappier, A., Belis, C.A., Pernigotti, D., Thunis, P., 2017. Source apportionment and sensitivity analysis: two methodologies with two different purposes. Geoscientific Model Development 10, 4245–4256. https://doi.org/10.5194/gmd-10-4245-2017

Emery, C., Liu, Z., Russell, A.G., Odman, M.T., Yarwood, G., Kumar, N., 2017. Recommendations on statistics and benchmarks to assess photochemical model performance. Journal of the Air & Waste Management Association 67, 582-598.

EPA, 2018. Modeling guidance for demonstrating air quality goals for ozone, $PM^{2.5}$, and regional haze. https://www3.epa.gov/ttn/scram/guidance/guide/O3-PM-RH-Modeling_Guidance-2018.pdf. Office of Air Quality Planning and Standards, United States Environmental Protection Agency.

Huang Z., Hu Y., Zheng J., Yuan Z., Russell A.G., Ou J., Zhong Z., 2017. A new combined stepwise-based high-order decoupled direct and reduced-form method to improve uncertainty Analysis in $PM_{2.5}$ simulations. Environ Sci Technol, 51(7):3852-3859. doi: 10.1021/acs.est.6b05479.

Hwang, D., Byun, D.W., Odman, M.T., 1997. An automatic differentiation technique for sensitivity analysis of numerical advection schemes in air quality models. Atmospheric Environment 31, 879-888.

Zhang W., Capps S. L., Hu Y., Nenes A., Napelenok S. L., and Russell A. G., 2012. Development of the high-order decoupled direct method in three dimensions for particulate matter: enabling advanced sensitivity analysis in air quality models. Geosci. Model Dev., 5, 355–368, www.geosci-model-dev.net/5/355/2012/, doi:10.5194/gmd-5-355-2012.

Cox, Jr., Louis Anthony (Tony)

My main thought on BPT is that "benefit per ton" (or "marginal benefit") is quite an ambiguous concept except in the rare case of linear, additively separable contributions (which are unrealistic for PM2.5 when interactions with income, co-exposures, and co-morbidities, among other factors, are important). If the intended meaning of the BPT concept can be clarified, then various technical methods (e.g., Shapley regression or partial dependence plots, depending on its intended meaning) might be useful for quantifying it. At the moment, different possible meanings of BPT seem to me to be comingled, which makes it difficult to give specific technical advice on the best way to improve its calculation. The reduced-form approximation issue does not address this ambiguity.

Response to Question 2 (Question 2. Please comment on **the results of the reduced form tool evaluation in Section 3, considering both the quantitative and qualitative aspects of the model intercomparison. Was the information clearly presented and informative? Were EPA's conclusions reasonable? Are there other results which would be useful to include in the comparison?)**

- VVUQ methods are not used (https://asmedigitalcollection.asme.org/verification). Why not?
- Mean squared error (MSE) is not reported (See Exhibit 2-11, p. 2-18). This could add useful information to the mean absolute error metrics.
- The coefficient of determination is insensitive to many types of errors. (Example: If each predicted value is 1000 times greater than the observed value, $r^2 = 1$, the same as for a perfect fit.)
- The mean bias and normalized mean bias likewise can have 0 values (the same as for a perfect model) even if all predicted values are extremely wrong (e.g., way too high for all small values and way too low for all high values).
- It would be more informative to show entire error distributions instead of just summary statistics.
- It would be more informative to test whether distributions of observed and model-predicted values (or full- and reduced-form analysis results) are significantly different from each other.
- It would be more informative to use visualizations such as regression diagnostics to understand when and how the different reduced-form model predictions differ significantly from each other and from full-form results.

- It would be more informative to use predictive analytics to see how accurately the model that generated a set of predictions can be identified from its prediction errors (or error residuals, taking the predictions of the full-form model as the gold standard for defining prediction errors). That is, does each reduced-form model have a characteristic error signatures?
- Something like a CART tree for prediction errors vs. reduced form model characteristics and policy scenario characteristics would be more informative about what characteristics drive errors.
- An optimization approach that seeks scenarios to maximize error metrics would help reveal how large errors from reduced form model could be. (This is a variation of extreme condition model testing, which is widely used in model testing and validation. It would probably be useful here.)
- The small sample size of reduced-form models (N = 8 at most, and fewer if the AP models are not counted as independent observations) makes it difficult to draw confident general conclusions.
- The small sample number of policies analyzed makes it difficult to draw confident general conclusions. It is not possible to get a sense of the error surface for different policies from this small sample.
- It would be helpful to use sensitivity analysis techniques (some of which have also been discussed in connection with Info-Gap robust design methods) to understand the types of scenarios that lead to relatively large or small prediction errors for some or all of the reduced form models. For what kinds of policies and scenarios do they work relatively well or badly?
- The concept of "avoided premature mortality" is problematic (e.g., Exhibit 3-2, p. 3-3), as noted by many people over many decades.
- All the reduced-form tools misapply the BenMAP-CE tool by treating its statistical regression equations for health impacts (representing statistical relationships with model specification errors, unmodeled errors in variables, omitted confounders, omitted interaction terms, etc.) as if they were valid causal models (Exhibit 2-10, note b, p. 2-16). As detailed in the CASAC's comments on the PM2.5 NAAQS review, this is a fundamental misinterpretation of the regression models. In general, regression equations (e.g., showing that reading ability in young children increases with shoe size after adjusting for grade level, age in years, sex, and parent's income and education) do not address the question of how the dependent variable would be changed by a change in a predictor (e.g., how, if at all, giving children larger shoes would affect reading ability. Presumably it would not, at least if the association is due to residual confounding by age, but a regression model does not reveal this. The key methodological point is that regression equations such as those in BenMap-CE do not in general give correct answers to causal questions,

such as how changing a predictor (e.g., pollutant levels) would change health effects (Pearl 2009, https://projecteuclid.org/euclid.ssu/1255440554).

Response to Question 3a ("Does the report provide a **clear and thorough explanation** for why some tools under- or over-estimated PM2.5 health benefits as compared to the full-scale air quality modeling? Please add any additional explanations for the pattern of results observed.")

I did not find a clear and thorough explanation. This report is more descriptive than explanatory: it reports here's what we ran, here's what it produced, here are some patterns we observe. The question of what would be produced if different examples were chosen has not been addressed in a principled or systematic way, or illuminated by large samples and visualization of error surfaces and their key drivers and sensitivities. Maybe designed experiments could be useful here, as in much of simulation modeling and evaluation.

Response to Question 4c (Based on the results of this study, does the panel have any additional recommendations about BPT-based approaches?)

The concept of BPT does not seem to be well-defined when the causal C-R functions (which should be assessed and distinguished from the statistical ones) are nonlinear and involve high-order interactions among factors. Is the BPT intended to reflect benefits attributed to a marginal ton of pollutant (e.g., using Shapley regression), or a causal relationship between changes in pollutant and changes in benefits when certain other factors are held fixed? If the latter, which factors are assumed to be held fixed at what levels, and why? Is BPT intended to reflect the direct effect on health of the pollutant, or total effects (including effects of any co-pollutants or other factors that mediate the effect of the pollutant on health)? In general, BPT should probably be replaced with a multivariate response surface or a partial dependence plot (with a clearly defined adjustment set of other factors and the levels at which they are assumed to be fixed) showing how changes in a pollutant are related to changes in health effects when other causally relevant factors are held fixed at specified levels.

Cullen, Alison C.

Question 3. Exhibit ES-4 "Ratio of National Avoided Premature Mortality Benefits Estimates," shows how different reduced-form tools generated different estimates as compared to full-scale air quality models.

3b. How do the results of this study inform our understanding of the suitability of these tools for regulatory economic analyses in their current form?

This analysis of how the reduced form models compare to the full form CMAQ and CAMx models is interesting. The potential to save resources and effort when generating model predictions of benefits is important, as long as there is a clear understanding of the tradeoffs that are inherent.

It would be helpful to see a bit more critical review in this report of CMAQ and/or CAMx themselves – as representations of reality. Is aligning with these full form models the right goal for the reduced-form models? The report appears at present to represent more a reflection of how closely one model reproduces another, and less an analysis of how each model would be expected to align with real world phenomena. This is not to say however that any observations or measurements selected as a comparison set would not also reflect assumptions and decisions about how, when and where to measure - and that these characteristics would not in turn influence the alignment of model intermediates and/or results - but that is a separate issue.

The fact that comparisons are restricted to point estimates of various metrics associated with each model is of concern. Without generating a distribution for a metric of interest associated with each model it is not possible to gauge whether the two are actually statistically different predictions. More detailed uncertainty analysis, as well as additional analysis of the sensitivity of model alignment across more policy scenarios is needed. For example, there is only one policy scenario for EGU, only one for mobile sources.

Regardless of alignment between full form and reduced-form models, it is important to note that if there are large changes in precursor emissions, and/or non-linear responses, the reduced-form models generally may not be suitable for application.

On the positive side, being able to consistently anticipate whether a reduced-form model is likely to over- or under-estimate full form results allows the reduced form model to provide valuable information even if the magnitude of the divergence may vary across policy scenarios. For example, if there is confidence that a reduced form model consistently under- or over-estimates the results of a full form model for a given compound, the results could be used to bound or bracket expectations. This information may be valuable for decision-making even though uncertainty would persist about discrepancies between the reduced-form model estimate and the expected full form model estimate.

Several of the reduced-form models consistently over- or under-estimate results for each of the compounds with only one exception (with the exception being different for each model). This includes the SA-Direct, AP2-BenMAP, AP2-Direct, InMAP BenMAP, and EASIUR models. Additional analysis of why the particular exceptions arise is warranted.

The greatest deviations from full form model results are seen for nitrate. The SA-Direct and EASIUR models show less extreme outlying ratio values across policy scenarios for sulfate and nitrate than any of the other reduced-form models. As such, those two models seem to be good candidates for approximating full form models. For the others, it is likely important to understand what is causing the extreme overestimates for sulfate and nitrate in different policy scenarios. Without understanding what is causing variability in the ratio values for a given model across policy scenarios, there is no way to compensate for it. If the causes of that variability can be pinpointed, they can be addressed by adjusting the methods/inputs of the reduced-form models or by adjusting how the results are interpreted.

Looking at Table ES-4 it further appears that EAISUR Direct always understates relative to full form models (except for Tier 3) while SA Direct always overstates. For Tier 3, they were the two best performers. Is there an integration possibility where these two are combined in some way, (perhaps weighted if information warrants), to generate a combined form? Since these two models are identified as being among the easiest and least time consuming to use, more exploration of the integration of the two may be warranted. It is also important to understand more about the reason for the consistent under/over performance of the two. It was noted that these models are not the best suited for regional estimates as they attribute benefits to where emissions are reduced and not where air quality is necessarily improved, but for national level the two taken together could be very informative.

Question 4. Since 2008 EPA has used SA-BPT to estimate the health impacts of numerous regulations. Under the scenarios examined in this report, EPA's SA-BPT approach over-estimated $PM_{2.5}$-related health benefits by between 10 and 30 percent, depending on the sector. To ensure BPT estimates correspond to full-form results as closely as possible, the report recommends updating the underlying emissions inventories and air quality modeling used to inform the EPA SA-BPT approach over time.

4a. In the interim, how might EPA improve its characterization of results derived from the 2005 SA-BPT approach, specifically the potential degree of over- or underestimation in BPT-based results for a particular regulatory scenario?

EPA could consider running more policy scenarios for one or more select reduced-form models and compare the results against the full form models to refine understanding

about the degree of over- or under-estimation. This would allow for better comparison of expected projections of the reduced scale versus full form models.

The agency should also carry out sensitivity analyses to discern which inputs are playing the largest roles in driving the divergence of the reduced-form model results from the full form models. The report states that because SA-BPT values exist only for $prPM_{2.5}$, NOx, and SO2, the benefits of those three precursors are summed to get a SA Direct benefits estimate for Total $PM_{2.5}$. The report does disaggregate this and looks at how the national benefits ratios differ for each of those three compounds. Looking at the results of that disaggregation (Exhibit 3-4), it looks like all models, SA-BPT included, diverge most from the full form models in their prediction of national benefits when considering nitrate. This raises a question as to whether nitrate is playing an outsized role in creating consistent overestimates when considering Total $PM_{2.5}$. Additional investigation into consistency of significant contributors to deviations under different policy scenarios would be valuable for identifying overall patterns relative to the full form models.

Overall, as raised above under question 3b, it would be helpful for this report to reference more earlier work constituting critical reviews of CMAQ and/or CAMx themselves – as representations of reality. Is aligning with these full form models the right goal for the reduced-form models? The report appears at present to represent more a reflection of how closely one model reproduces another, and less an analysis of how each model would be expected to align with real world phenomena.

4b. What criteria (e.g., geographical scale, regulated sector, pollutants/precursors) should EPA examine to determine the potential for divergence between SA-BPT results vs full-form air quality modeling results (resulting in over- or under-estimation)?

There is only one set of BPT values for each sector, originally derived from full form BenMAP results, applicable to emissions changes within the contiguous US. The agency should consider additional sensitivity analysis - generating a separate set of BPT values, for several geographic area subsets, and for each regulated sector.

The report states that “the SA BPT values we applied...do not contain different values for different emission heights.” This is acknowledged as a potential contributing factor as to why reduced form model results deviate for the Tier 3 scenario. What are the opportunities for accounting for this as a step towards improving the reduced form modeling? The agency could consider assessing performance for a more recent mobile source scenario with more contemporary aspects.

4c. Based on the results of this study, does the panel have any additional recommendations about BPT-based approaches?

More exploration could be targeted to whether selection of a reduced form tool should depend more on the specific characteristics of the policy scenario of interest. Point

source emissions are generally reported to be better handled by reduced form models than mobile source emissions are for example – although this is based on a very limited set of policy scenarios assessed thus far.

Benefits per ton has been cast as a key metric but the absolute scale of an impact (given that the absolute magnitude of benefits can vary so much by policy scenario) has a major influence on the real significance of a given percentage difference.

Treatment of uncertainty should be augmented. Although uncertainty in some factors is fairly consistent across the board for all models, there are other uncertainties that can be expected to make highly variable contributions to overall results depending on both model particulars and policy scenario.

The absence of CAMx based full form results for Tier 3 for comparison with CMAQ raises further questions about how to interpret the comparisons between full form models and reduced-form models. Whether CMAQ and CAMx align well with each other for point source scenarios does not lend insight into how comparisons for mobile source scenarios might align.

Question 5. How do the results of this study inform the future development of reduced-form tools that are capable of providing reliable estimates of impacts associated with different sectors, across a variety of spatial scales, and for different portions of the air quality distribution? Are there other, less resource intensive approaches than full-scale air quality modeling for informing the public about the size and distribution of PM health benefits associated with alternative regulatory scenarios?

There is a role for reduced-form models, especially for national scale assessment. The performance of these models does vary with policy scenario, and more work to understand the differences related to characteristics of each relevant scenario is warranted. That understanding, in turn, can help provide guidance about when to apply different reduced-form models.

The report illuminates several sources of observed deviation between the reduced-form and full form models. For instance, accounting for ground-level emissions leads to challenges for reduced-form models. Also, the percentage of emissions changes attributable to nitrate affects the alignment of predictions of reduced form models with full form models.

The study suggests that the SA-Direct and EASIUR reduced-form models, which require less time and technical expertise than the rest, can produce results that are competitive with more involved reduced-form models and are accurate enough to meaningfully inform decisions – for national scale assessments. Therefore, pursuing ways to update and improve those models, as discussed in the report, should be prioritized to explore their limits.

Lange, Sabine

Question 1. Please comment on the evaluation approach developed by EPA to compare reduced-form models to full-form equivalents. Please comment on whether the emissions reduction scenarios used in the proposed evaluation approach provide enough diversity to adequately assess reduced-form performance over a range of possible applications (e.g., magnitude, type, and spatial variations of emissions reductions). Please discuss whether the specific assumptions that EPA made to apply each tool as consistently as possible (e.g., emissions, meteorology, use of direct vs. BenMAP estimates, etc.) are appropriate and clearly explained. Please assess whether the report's description of its limitations is complete.

Evaluation Approach

Scaled Benefits Methods

On page ES-4 the authors stated that "We multiplied the prPM2.5 benefit per ton based on EC by the total amount of primary PM2.5 emissions to estimate benefits related to all primary PM2.5 emissions." It is not clear here or elsewhere in the document what the total amount of primary PM2.5 emissions are. This also seems to contradict Exhibit 2-2, which says that "For all scenarios Primary PM2.5 is represented by EC emission only". Appendix A, which discusses each of the RFTs in more detail, does not mention scaling at all (and in several places states that EC emissions are the only emissions used for PrPM2.5). The methods and purpose of the primary PM2.5 scaling needs to be clarified. On page A-14 – in the "apportioning total benefits" footnote, the authors state: "[32] Total PM2.5 benefits included contributions from OC and Crustal. Therefore, total benefits were also allocated to those precursors and then later subtracted out of total PM2.5 benefits. Such that total PM2.5 benefits ultimately only reflected the contribution of EC." The authors should specify when and how these OC and crustal benefits were apportioned - was a delta characterized like it was for EC? Or were the apportionments for SO4, NO3, and EC calculated in step three, and then any leftover percentage of total PM2.5 benefits were removed? If the final calculations for PrPM2.5 actually do include OC and crustal, shouldn't that have just been left in? (This concern also applies for APX BenMAP calculations.)

In addition, the idea of multiplying the BPT for EC prPM2.5 seems problematic, because the CPP RIA for example (Table 4-8) values crustal PM2.5 at about 10-times less than OC and EC PM2.5, so using the EC PM2.5 BPT for crustal PM2.5 would over-estimate the PM2.5 total benefits.

Additional Methods Information

Exhibit 2-2: the authors should specify how the changes in emissions tonnage were calculated. If the full-form model was used to calculate the emissions changes, then it

just seems like the full-form model air quality surfaces could be input into BenMAP and the full benefits calculated.
Pg 2-12 states that "AP2 and AP3 estimate the marginal cost of emissions by quantifying the total health burden and monetized costs associated with a baseline emissions scenario, systematically increasing the baseline emissions by one ton, recalculating the total health burden and monetized costs, and taking the difference between the two estimates." The authors should include more information about the methods used by the APX RFTs to calculate changes in air quality with changes in emissions (to generate the air quality surfaces that are used by BenMAP). For example, does APX use a photochemical model like CMAQ and CAMx, or a regression model like EASIUR, or a chemical transport model like INMAP?

In section A.4 when discussing the BenMAP model methods, the authors state ""The precursors of interest for this analysis were EC, SO4, and NO3." However, Exhibit 2-9 says that the outputs were NH3 and SOA as well as those emissions. The methods in Appendix A should specify where and how NH3, SOA, and VOCs were accounted for.

Assumptions in Application of RFTs

Exhibit ES-3 on page ES-4 states in footnote A that: "These population and incidence datasets are also reflected in the SA Direct and APX Direct BPT values. The only model that does not reflect these inputs is the EASIUR Direct reduced-form tool." It is not clear from this table or from the rest of the document what population and health incidence data are used for EASIUR Direct – this information should be provided in this table and in the Methods sections of the document.

Pg 2-1: "These policies were projected to impact PM2.5 emissions from sources that have varying geographic distributions within the US (and consequently proximity to population centers);" Did the different RFT models in fact take into account the proximity of the source to population centers?

The full-form model for the cement kiln, refinery, and pulp & paper scenarios are described as hypothetical in Chapter 2. Are there any references of EPA modeling for these scenarios, or were they run specifically for this exercise? This should be clarified.
Exhibit A-3 shows the adjustment factors used to convert from mortality + morbidity health effects to mortality effects only. The source of these factors should be provided. It would be helpful if the authors noted in Appendix A (when describing the RFT models), what the averaging time and form of the concentration metrics were (e.g. annual average 24-hour concentrations of PrPM2.5?).

The benefits per ton estimates from USEPA (2018) seem to be for EC and OC together (based on subscript a in Table 1 of that document), but it isn't made entirely clear. The basis for the BPT estimates for PrPM2.5 as it relates to OC, EC, and crustal PM2.5 from USEPA (2018) should be described.

*Question 2. Please comment on the results of the reduced form tool evaluation in Section 3, considering both the quantitative and qualitative aspects of the model intercomparison. Was the information clearly presented and informative? Were EPA's conclusions reasonable? Are there other results which would be useful to include in the comparison?

Rough Estimate of SA Direct Results

I conducted a rough calculation of the benefits estimated using the SA Direct method, to help myself better understand the inputs and outputs. I used the emissions changes for policy scenarios in tons from Exhibit 2-2, and the BPT for each PM2.5 species from USEPA (2018), using the Krewski et al. estimates with a 3% discount rate for the year 2025 for the different sections (Tables 69 (cement kilns), 71 (pulp & paper), 73 (refineries), 100 (EGUs), 131 (2030, on-road mobile)). While calculating the estimates using the national level emissions changes won't perfectly recapitulate the results from county-level calculations, it should provide a roughly comparable estimate.

I multiplied the emissions change for SO2, NOx, or PrPM2.5 by the matched BPT (and by the appropriate mortality-only adjustment factor in Exhibit A-3) and compared the result to the data provided in Exhibit C-1. The full tables of data for my calculations for the CPP are shown here (Table 1), and for the other scenarios are included at the end of this document (Tables 4 to 7). In general I found that while this rough estimate did a very good job of recreating the SO2 and NOx benefits estimates, the PrPM2.5 estimates using the rough estimate were lower by a factor of 4 to 14, depending on the scenario.

Table 1. Calculation of benefits for CPP Rule via SA Direct method (using EGUs Benefits per Ton estimates and a mortality-only adjustment factor of 0.973 from Exhibit A-3), calculations rounded to 2 significant digits.

Data Source:	Exhibit 2-2	USEPA 2018 Table 100	Calculated	Exhibit C-1
Pollutant	**Ton Reductions**	**Benefits per ton ($)**	**Total Benefit ($ Mill)** [1]	**Total Benefit ($ Mill)** [2]
Pri-PM2.5	2,481	$170,000	$410	$5,800
NOx	414, 479	$6,700	$2,700	$2,700
SO2	422,670	$46,000	$19,000	$19,000
Total PM2.5			$22,000	$28,000

Note: estimates marked in red show substantial differences between calculated and presented total benefits

[1] Calculated Total Benefit ($ Millions) = Ton Reductions x Benefits per ton x 0.973 mortality-only factor

[2] Exhibit C-1 Total Benefits taken directly from the appropriate row of Exhibit C-1 in the SA Direct column

I investigated whether this discrepancy was caused by the "scaling" of EC-only PrPM2.5 to include OC & crustal PrPM2.5. To do this I looked at the PM2.5 emissions reductions estimates from the USEPA Proposed CPP RIA (2014), which was the basis of the CPP emissions reductions (as stated on page 2-3), specifically Option 1 State estimates. Table 4-11 from the CPP RIA shows that nationally for 2025, 49,000 tons of crustal PM2.5 and 6,000 tons of EC+OC PM2.5 were projected to be reduced. So, as per the

calculation specified on pp 2-16 to 2-17 ("We scaled the results by multiplying the prPM2.5 benefit-per-ton based on EC only by the total amount of primary PM2.5 emissions to generate an estimate of impacts for total primary PM2.5 emissions."), I multiplied the PrPM2.5 BPT ($170,000 x 0.973 mortality-only factor) by 55,000 tons (49,000+6,000) = $9,097 M, which still does not match the value in Exhibit C-1 for SA Direct, PrPM2.5, CPP ($5,800 M, also shown in Table 1 of this document).

I also compared the CPP scenario SA Direct results in Exhibit C-1 to the PM2.5 benefits provided in the 2014 CPP RIA (because the CPP RIA used the same BPT approach). Table 4-14 of the CPP RIA presents the Summary of Estimated Monetized Health Co-Benefits for the Proposed EGU GHG Existing Source Guidelines in 2025 (millions of 2011$). Using Option 1 – State, 3% Discount Rate, the lower end of the range provided (which represents the results from Krewski et al. 2009), and multiplying by 0.973 for mortality-only and by 1.05 to roughly convert to $2015 I generated the results in Table 2. I could largely recreate the benefits presented in the CPP RIA using the inputs from that document (emissions reductions tons and BPT) and I converted it to a comparable number for the current analysis (conversion to $2015, mortality-only benefits). This generally produced estimates that were similar to my SA Direct calculations of SO2 and NOx benefits, and would be comparable for the PrPM2.5 if the same PM2.5 source were used (EC for SA Direct, EC+OC and crustal separately for CPP RIA). Table 3 shows the presented results (from the CPP RIA or from this RFT Report) versus calculated benefits for the CPP RIA and for the SA Direct model. The estimates presented in the CPP RIA compared to Exhibit C-1 were very different for the PrPM2.5 estimate and therefore for the total PM2.5 estimate. For comparison, the CMAQ-BenMAP estimate for total PM2.5 was quite similar to the CPP RIA estimate, but this was generated in the CPP RIA by higher estimates of SO2 and NOx benefits and lower estimates of PrPM2.5.

Table 2. Estimates of benefits for CPP emissions changes based on data from the CPP RIA (2014) and the SA Direct benefits calculated based on the RFT (2020) analysis.

Pollutant	CPP RIA (2014) (2015$)			SA Direct Benefits (2020) (2015$)		
	BPT (2011$)	Tons	Calculated Benefits (millions) [1]	BPT (2015$)	Tons	Calculated Benefits (millions) [2]
SO2	$41,000	425,000	$17,800	$46,000	422,670	$18,900
NOx	(for NOx as PM2.5) – $6,000	436,000	$2,670	$6,700	414,479	$2,700
PrPM2.5 (EC+OC)	$150,000	6,000	$920	$170,000		
PrPM2.5 (Crustal)	$17,000	49,000	$850	Not provided		
PrPM2.5 (EC)	Not provided	Not provided		Not provided	2,481	$410 [3]
Total PM2.5			$22,500			$22,200

[1] Benefits = BPT x tons x 0.973 (mortality-only adjustment factor) x 1.05 ($2011 to $2015 adjustment)
[2] Benefits = BPT x tons x 0.973 (mortality-only adjustment factor)
[3] Benefits for PrPM2.5 (EC) calculated using the BPT estimate for EC+OC

Table 3. Estimates of benefits for CPP emissions changes based on the presented benefits from the CPP RIA (2014) and calculated from the inputs of the CPP RIA, and the SA Direct benefits presented in the RFT (2020) analysis and calculated based on the RFT (2020) inputs, and the CMAQ-BenMAP benefits presented in the RFT (2020) analysis.

Pollutant	CPP RIA (2014) (Millions 2015$)		SA Direct Benefits (2020) (Millions 2015$)		CMAQ-BenMAP Benefits (2020) (Millions 2015$)
	Presented (Table 4-14) [1]	Calculated in Table 2	Presented (Exhibit C-1)	Calculated in Table 2	Presented (Exhibit C-1)
SO2	$18,400	$17,800	$19,000	$18,920	$15,000
NOx	$3,000	$2,670	$2,700	$2,700	$1,700
PrPM2.5 (EC+OC)	$920	$920			
PrPM2.5 (Crustal)	$850	$850			
PrPM2.5 (EC)				$410	
PrPM2.5 (EC+OC+crustal)			$5,800 [3]		$3,500 [3]
Total PM2.5	22,500 [2]	$22,200	$28,000	$22,000	$21,000

Note: estimates marked in red show substantial differences between calculated and presented total benefits

[1] Benefits = Benefits value for CPP RIA 2014 Table 4-14 Option 1-State, 3% Discount Rate, lower end of presented range x 0.973 (mortality-only adjustment factor) x 1.05 ($2011 to $2015 adjustment)

[2] Total PM2.5 Benefits = Total – NOx (as Ozone) from Table 4-14, then calculated as in footnote 1

[3] Assumed to be the benefits from total primary PM2.5 (EC+OC+crustal) based on language about scaling on pages 2-16 to 2-17

Therefore, a rough calculation of the SA Direct benefits cannot reproduce the PrPM2.5 or total PM2.5 estimates presented in this report, likely because of an unpresented calculation of OC & crustal PrPM2.5. In addition, the presented PrPM2.5 and total PM2.5 benefits don't match the benefits presented in the Proposed CPP RIA (2014) upon which the scenario is based. The authors need to resolve these problems, and/or explain the discrepancies in the estimates.

Scaled PrPM2.5 Benefits Calculations

As discussed in my responses to charge question 1, I have trouble following the scaled conversion of PrPM2.5 (EC) to PrM2.5 (EC+OC+crustal). One way to better explain to readers about the PrPM2.5 benefits would be to include the raw and scaled versions of the PrPM2.5 and total PM2.5 benefits, as well as a better explanation for how and why they were generated.

Exhibit 3-4 shows the ratio of benefits from PrPM2.5 (labeled as EC only) for the RFTs compared to CMAQ BenMAP. The ratios are the ones that would be expected based on the values shown in Exhibit C-1, but Exhibit C-1 presents the scaled estimates (defined

on page ES-4 as prPM2.5 BPT based on EC multiplied by the total amount of primary PM2.5 emissions EC scaled to OC + crustal) - would these ratios be expected to be the same for EC only? This is another example of a lack of clarity in the estimates due to PrPM2.5 scaling.

Benefits Calculations from NH3, VOC, SOA

Some of the RFTs take into account changes in NH3 and VOC emissions (noted in Exhibit 2-9). The authors should discuss how much these emissions contribute to the total PM2.5 benefits (they are not included in Appendix C Exhibits or discussed in the Appendix A methods) and how those may impact the relative outputs of the RFTs versus the full form tools. For example, as shown in Exhibit 2-2 a substantial amount of the reductions from the Tier 3 rule were from VOCs (33% of the change) - so does that mean that those models that don't capture VOCs (SA Direct, AP2 BenMAP, AP3 BenMAP, EASIUR Direct) won't capture these estimates? That could explain the variety of estimates, although it doesn't explain the magnitude of over-estimates.

Regional Results

Page 2-16 2-16 "We generated county-level results for each full-form model and reduced-form tool and expressed these in terms of monetized benefits of avoided premature mortality ($2015)." This statement makes it seem like county-level results were derived for all the RFTs, but the results from regional analysis were only presented in Exhibit C-2 and Section 3-2 for the BenMAP tools. The authors should present regional analyses from all the tools. I understand that there are concerns about health benefits and emissions reductions being ascribed to different counties, but that doesn't seem like it should be a problem with the large scale of the regions. Just in general for the regional analyses the authors should present more information, including the model results for each region (perhaps in a map?), as well as the correlation, error, and bias metrics for all the models.

In addition, it seems that some of these bias, correlation, and error analyses were in fact completed for the other RFTs, but not presented. For example, on page 4-2 the authors state, "EASIUR Direct also did a reasonable job capturing variation in benefits across large regions of the US (0.88 r2 value on average)." This information is not present in the Results chapter or in the appendices. Similarly, on page 5-2 the authors state "In our analysis we saw differences in how the tools performed at different geographical scales and locations". The authors should provide more information about what data this conclusion was based on.

The metrics generated from the regional analysis (correlation, error, and bias metrics) are based on only 7 datapoints – from the 7 US regions. This does not seem like a lot of datapoints for conducting this analysis. Could the authors use perhaps state-level regions? At the least the authors should note when discussing these results that they are based on 7 datapoints. In addition, the regional results and their implications should be considered to a greater extent in the Discussion chapter (4).

For section 3.2 - Regional Results – result interpretation would be aided by knowing how much bias would be considered large, similar for error and R2 (there are some general metrics for R2 - higher than 0.75 is high correlation, etc). E.g. Simon et al. 2012 state ""For these metrics to be meaningful, the evaluator must be familiar with typical observed magnitudes in order to understand whether bias and error are "large"."

Other Comments

Premature mortality results are presented in Exhibit 3-1, but it would be good if the authors could also present those in a tabular form in the appendix (similar to the benefits dollar estimates in Exhibit C-1).

Page 3-4: "Of all the models, AP3 BenMAP and AP3 Direct estimates of health benefits are within 10% of CMAQ benefits estimates for more scenarios (3: CPP Proposal, Cement Kilns, and Pulp and Paper) than any of the other reduced form tools." I'm not sure that noting that these tools are within 10% is really helpful in characterizing this tool, because it is so far off for the other two scenarios.

For Exhibit 3-6 it would be helpful if the authors could scale the NMB to be equal above and below zero. Because the positive goes to infinity and the negative only goes to -100%, it is hard to judge the magnitude of positive versus negative bias.
For section 3.3 - Reduced-Form Tool Complexity and Level of Effort - it would be helpful if the authors also presented the complexity and level of effort required for the full-form tools. It is also a bit misleading to say that SA Direct took no tools and little time, because the emissions information had to come from somewhere - there should be a discussion of what is required on the front end to plug into these tools.

Page 4-3 states "The primary drawback of the AP3 BenMAP model is its complexity." Additional primary drawbacks are that AP3 BenMAP does a very poor job estimating the Tier 3 benefits, and it is completely inconsistent for NOx estimates - for pulp and paper the ratio with CMAQ was 0 (actually about 50-times lower based on information in Exhibit C-1). In addition, it was not consistent with AP3 Direct, suggesting a problem in translation to BenMAP, instead of just in the air quality surface estimation. The next paragraph on this page attributes the problem of inconsistency between AP3 BenMAP and AP3 Direct to AP3 Direct, but I don't think that attribution is correct because the problem seems to be with NOx estimates in AP3 BenMAP.

Question 3. Exhibit ES-4 "*Ratio of National Avoided Premature Mortality Benefits Estimates*," shows how different reduced-form tools generated different estimates as compared to full-scale air quality models.

> *3a. Does the report provide a clear and thorough explanation for why some tools under- or over-estimated $PM_{2.5}$ health benefits as compared to the full-scale air quality modeling? Please add any additional explanations for the pattern of results observed.

Page 3-4 discusses the possibility that the amount of NOx reductions in the scenario are in part responsible for the variability in the RFT performance for different scenarios, particularly the poor performance for Tier 3. However, this explanation doesn't seem to actually explain the variability, either by volume of emissions (e.g. cement kilns have the same percent of NOx reductions as Tier 3), or the types of emissions changes. In addition, page 4-4 states that "Regardless, the reduced-form tools considered in this analysis should be applied with caution to policies with large changes in NOX emissions." This does not seem to be a generalization that is well-supported by the results.

I think that the way to really determine the source of the differences between the model results (particularly all of the BenMAP models) is to investigate the air quality surfaces being produced and comparing those directly, grid cell by grid cell. That is a more direct way of answering the questions about the models, rather than running many different scenarios and trying to back-calculate.

For the Tier 3 rule, the CMAQ-BenMAP estimates are being treated as "truth" for the PM2.5 concentration changes, but what if the RFTs are actually better at predicting PM2.5 concentration changes from mobile sources than CMAQ is? For the other scenarios CAMx could be used as a kind of control, but that isn't available for Tier 3. The PM2.5 concentration changes projected to occur for Tier 3 needs to be investigated for each of these models, and ground-truthed in some way to figure out which may be correct.

While the explanation of the variability caused by the RFTs is one of the goals of this project, I'm not sure that it was attained. It is not clear to me why some tools were better than others and why some scenarios were closer to the CMAQ-BenMAP result than others. I think that looking at air quality surfaces will help to answer these questions.

3b. How do the results of this study inform our understanding of the suitability of these tools for regulatory economic analyses in their current form?

One aspect that is relevant to application of these tools to regulatory economic analyses are the underlying assumptions about the BenMAP tool itself. In this report the CMAQ-BenMAP results are being assumed to be the "true" benefits, with the other model's results being more or less similar to BenMAP's "truth". I think that there needs to be some language or caveat noting that even if the RFT is perfectly aligned with BenMAP, there are still uncertainties and assumptions being made if the resulting benefit estimates are applied to a regulatory economic analysis. (For examples of concerns about applying benefits estimates to regulatory assessments, the EPA can refer to the Clean Air Scientific Advisory Committee's comments[16] on the risk assessment section of the *Policy Assessment for the Review of the National Ambient Air Quality Standards for Particulate Matter, External Review Draft*, *Sept 2019*.)

[16] https://yosemite.epa.gov/sab/sabproduct.nsf/LookupWebReportsLastMonthCASAC/4713D217BC0710348525851500635 9BA/$File/EPA-CASAC-20-003.pdf

3c. Can any of the reduced-form tools explored in this report easily be modified to allow quantifying the extent to which the total health benefits accrue to specific geographic areas (e.g., by state, or where ambient concentrations are above or below the NAAQS)?

Question 4. Since 2008 EPA has used SA-BPT to estimate the health impacts of numerous regulations. Under the scenarios examined in this report, EPA's SA-BPT approach over-estimated $PM_{2.5}$-related health benefits by between 10 and 30 percent, depending on the sector. To ensure BPT estimates correspond to full-form results as closely as possible, the report recommends updating the underlying emissions inventories and air quality modeling used to inform the EPA SA-BPT approach over time.

4a. In the interim, how might EPA improve its characterization of results derived from the 2005 SA-BPT approach, specifically the potential degree of over- or underestimation in BPT-based results for a particular regulatory scenario?

In response to charge question 2 I have outlined my concerns about the methods used to calculate benefits via SA Direct in this document, which should be addressed before judging the appropriateness of this tool.

On page 4-4 the authors note that "Thus, the fact that the Tier 3 scenario is exclusively comprised of ground-level emissions may be a secondary contributing factor, as may the use of a different base year emissions inventory (2005) than the other policies." The SA Direct tool uses NEI from 2005 as well – perhaps this is why SA Direct does a better job with Tier 3 than the other models?

Just a thought: If SA Direct tends to over-estimate, and EASIUR Direct tends to under-estimate, then perhaps EPA can run both models and provide the range. (This would also address some model uncertainty questions).

*4b. What criteria (e.g., geographical scale, regulated sector, pollutants/precursors) should EPA examine to determine the potential for divergence between SA-BPT results vs full-form air quality modeling results (resulting in over- or under-estimation)?

4c. Based on the results of this study, does the panel have any additional recommendations about BPT-based approaches?

*Question 5. How do the results of this study inform the future development of reduced-form tools that are capable of providing reliable estimates of impacts associated with different sectors, across a variety of spatial scales, and for different portions of the air quality distribution? Are there other, less resource intensive approaches than full-scale air quality modeling for informing the public about the size and distribution of PM health benefits associated with alternative regulatory scenarios?

Future Development of RFTs

An additional question that has to be addressed in the evaluation of these RFTs is: what happens if changes are made to BenMAP - will these models be expected to change concomitantly? Will this analysis still be accurate?

Other Tools for Regulatory Benefits Calculations

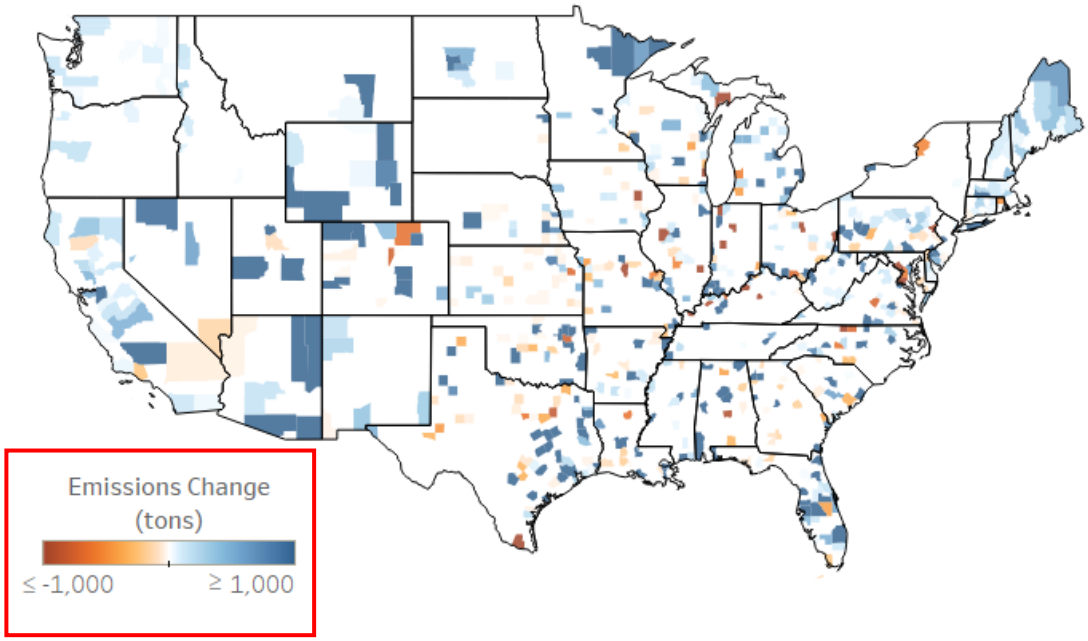
As is implied in the charge question, the absolute numbers from the benefits analysis may not be the output that is used in a regulatory analysis, but rather the differences between modeled alternative scenarios. The ability of these tools to differentiate between differential alternatives for the same policy type isn't really covered here. For example, a rule would not normally be comparing benefits from regulating cement kilns versus refineries, but they would be comparing a 20% reduction in emissions versus a 40% reduction in emissions. Therefore, it may be more useful for the authors to determine the relative "accuracy" of modeling different levels of emissions changes within the same type of scenario. The current evaluations show the magnitude of effect of no change versus some change, but not among different levels of changes.

General Notes About the Report (Presentation-Related)

Pg ES-1, "The study did not evaluate the ability of each approach to characterize the distribution of PM2.5-related premature deaths according to the annual mean concentration at which they occurred." This statement was the first mention in the report about PM2.5 (the acronym should be defined), premature death, and annual mean concentrations. This should be moved to later in the executive summary (maybe the discussion section) after the authors have explained what the report and analyses do address.

Exhibit ES-4 and similar exhibits present the main full-form model as "CMAQ estimates" in the title. However, CMAQ is not the only difference between the models and therefore I suggest labeling as "CMAQ BenMAP estimates".

Exhibits 2-3 to 2-7 show maps of emissions decreases as blue and increases as orange, but the way the figure legends are labeled makes the emissions changes appear to go in the opposite direction.



Pp 2-5 to 2-6 “This hypothetical scenario assumed uniform emissions reductions from the 2025 emissions baseline that was developed as part of the analysis for the CPP proposal: 40% reduction in baseline NOX emissions, 50% reduction in baseline SO2 emissions, and 40% reduction in baseline prPM2.5 emissions.” A later sentence on page 2-6 reads, “Two thirds of the emissions reductions are attributed to NOX and one third of the emissions reductions are attributed to SO2.” These seem to be conflicting statements about the amount of reductions of NOx and SO2 from the CPP rule. (From Exhibit 2-2 the reductions seem to be close to 50-50 for NOx and SO2).

Page 2-6 “Emissions reductions are focused in industrialized areas of the continental US, particularly the rust belt region, Texas, and the desert Southwest.” The map doesn’t make this case – emissions reductions seems to be spread out more. Also, the “rust belt” should be defined. Earlier in the paragraph the reductions are described as “diffuse” – this is probably a better descriptor.

Page 2-13 “2-13: “The InMAP model estimates the annual average primary and secondary PM2.5 related to changes in emissions.” Are the authors referring to the annual average concentration of PM2.5 in this sentence? If so, this should be specified. Page 2-13 to 2-14 “3D annual average meteorology, air quality, and deposition information”. The authors should specify what 3D inputs are, where they come from, and how they are applied.

It would be easier to understand the RFTs and how they compare to the full form models if there were perhaps a figure or figures of the general method by which the tools generated their estimates.

In general, it would help if the models were always presented in the same order (in the tables in the appendices too).

Exhibit 3-8 would be useful in the Methods section, because it outlines the inputs and outputs of the tools in a more precise way.

Page 4-5 states "our observations are only accurate with respect to the versions of the reduced-form tools we tested." This is also true for CMAQ and BenMAP, and should be presented as such.

In Appendix A the air quality scenario with emissions reductions is referred to as the "control" scenario, whereas it is called the "policy" scenario in the main body of the text. For clarity these scenarios should be discussed using a consistent label.

All the referenced documents that are in the footnotes should also be in the reference list.

Additional Tables for Estimates of SA Direct Benefits

Table 4. Calculation of benefits for the Tier 3 Rule via SA Direct method (using on-road vehicles Benefits per Ton estimates and a mortality-only adjustment factor of 0.972 from Exhibit A-3), calculations rounded to 2 significant digits.

Data Source:	Exhibit 2-2	USEPA 2018 Table 131 (year 2030)	Calculated	Exhibit C-1
Pollutant	**Ton Reductions**	**Benefits per ton ($)**	**Total Benefit ($ Mill)**	**Total Benefit ($ Mill)**
Pri-PM2.5	1,322	$500,000	$640	$3,000
NOx	345,333	$10,000	$3,400	$3,500
SO2	13,002	$28,000	$350	$360
Total PM2.5			$4,400	$6,800

Note: estimates marked in red show substantial differences between calculated and presented total benefits

Table 5. Calculation of benefits for the cement kilns scenario via SA Direct method (using cement kiln Benefits per Ton estimates and a mortality-only adjustment factor of 0.977 from Exhibit A-3), calculations rounded to 2 significant digits.

Data Source:	Exhibit 2-2	USEPA 2018 Table 69	Calculated	Exhibit C-1
Pollutant	**Ton Reductions**	**Benefits per ton ($)**	**Total Benefit ($ Mill)**	**Total Benefit ($ Mill)**
Pri-PM2.5	557	$460,000	$250	$2,600
NOx	96,468	$7,100	$670	$670
SO2	55,398	$55,000	$3,000	$3,000
Total PM2.5			$3,900	$6,300

Note: estimates marked in red show substantial differences between calculated and presented total benefits

Table 6. Calculation of benefits for the refineries scenario via SA Direct method (using refineries Benefits per Ton estimates and a mortality-only adjustment factor of 0.971 from Exhibit A-3), calculations rounded to 2 significant digits.

Data Source:	Exhibit 2-2	USEPA 2018 Table 73	Calculated	Exhibit C-1
Pollutant	**Ton Reductions**	**Benefits per ton ($)**	**Total Benefit ($ Mill)**	**Total Benefit ($ Mill)**
Pri-PM2.5	424	$400,000	$160	$610
NOx	34,967	$8,400	$290	$290
SO2	16,421	$85,000	$1,400	$1,400
Total PM2.5			$1,850	$2,300

Note: estimates marked in red show substantial differences between calculated and presented total benefits

Table 7. Calculation of benefits for the pulp and paper scenario via SA Direct method (using pulp and paper Benefits per Ton estimates and a mortality-only adjustment factor of 0.973 from Exhibit A-3), calculations rounded to 2 significant digits.

Data Source:	Exhibit 2-2	USEPA 2018 Table 71	Calculated	Exhibit C-1
Pollutant	**Ton Reductions**	**Benefits per ton ($)**	**Total Benefit ($ Mill)**	**Total Benefit ($ Mill)**
Pri-PM2.5	278	$190,000	$50	$520
NOx	34,616	$4,700	$160	$160
SO2	36,464	$58,000	$2,100	$2,100
Total PM2.5			$2300	$2,800

Note: estimates marked in red show substantial differences between calculated and presented total benefits

References

US EPA (2014a). Regulatory Impact Analysis for the Proposed Carbon Pollution Guidelines for Existing Power Plants and Emission Standards for Modified and Reconstructed Power Plants. https://www.epa.gov/sites/production/files/2014-06/documents/20140602ria-clean-power-plan.pdf.

US EPA. (2018). Environmental Benefits Mapping and Analysis Program – Community Edition (BenMAP) User's Manual. Research Triangle Park, NC, Office of Air Quality Planning and Standards. US Environmental Protection Agency. July 2018. https://www.epa.gov/sites/production/files/2015-04/documents/benmap-ce_user_manual_march_2015.pdf

Smith, Anne

Question 1. Please comment on the evaluation approach developed by EPA to compare reduced-form models to full-form equivalents. Please comment on whether the emissions reduction scenarios used in the proposed evaluation approach provide enough diversity to adequately assess reduced-form performance over a range of possible applications (e.g., magnitude, type, and spatial variations of emissions reductions). Please discuss whether the specific assumptions that EPA made to apply each tool as consistently as possible (e.g., emissions, meteorology, use of direct vs. BenMAP estimates, etc.) are appropriate and clearly explained. Please assess whether the report's description of its limitations is complete.

The evaluation approach itself is clearly organized. It follows the structure used in many other model comparison exercises, such as those of Stanford University's Energy Modeling Forums. That is, it establishes a set of shared scenarios to be run with each model and establishes several shared assumptions. The shared assumptions usually focus on defining the baseline scenario against which the scenarios are to be run, but can also include fixing other key input parameters to be the same across all the models. The process then allows remaining differences in model structures and assumptions to drive differences in their results, i.e., in their predicted responses to the scenarios. Results of each model are then compared to each other, with effort to understand what structural aspects of the models cause their results to differ.

Below I respond to the charge question elements individually. I also comment on an issue that is not called out specifically by the charge question, but which I believe raises more significant issues. This is regarding how the results are compared.

(a) The emissions reduction scenarios used in the proposed evaluation approach
This model comparison covers five scenarios, intended to compare the reduced form models (RFMs) under a variety of policy change conditions. The five scenarios do reflect diversity in the ways that EPA regulations may affect ambient PM2.5 precursor emissions, and by having this range of alternative types of controls some interesting insights do emerge. In particular, it becomes clear that estimates of benefits from NOx emissions reductions are subject to the greatest inconsistencies, and that use of RFMs alone to value benefits of policies with such changes may be most questionable.

Although the five scenarios analyzed are reasonable, I feel that more efficient design could be applied (e.g., in future evaluations). The scenarios would still start by identifying a variety of types of sectors (the current five are not unreasonable choices). However, when deciding how much to change the emissions in each sector, it could be more straightforward to apply across-the-board percentage emissions reductions to all types of emissions within each sector. This is in contrast to the current set of scenarios that include two specific policies and three hypothetical policies (but which have differing percentage reductions for each pollutant emitted in those sectors). A more structured set of proportional reductions across different sectors might have helped this

evaluation process to more readily identify causes of differences in results from RFM to RFM. It would also have avoided a suggestion that this analysis provide insights on the relative benefits of specific policies (which I discuss further below).

(b) The specific assumptions that EPA made to apply each tool.

It is clear that the following assumptions were standardized across all the RFM runs: (a) to report benefits for chronic mortality only, using specifically the Krewski et al. (2009) C-R relationship that BenMAP treats as its default value from that study, (b) to apply the same value of statistical life (VSL) also relied on in BenMAP, and (3) that the emissions reductions (quantity and geographic location) associated with each scenario are the same for each alternative model. The report also provided a clear explanation of how the BenMAP model was substituted for the original RFM's internal ("direct") benefits calculations in some cases. I found the description of meteorology assumptions more difficult to follow. In particular, it is unclear to me why some RFMs were applied using different meterological-year assumptions than the other RFMs, and only in some scenarios.

Although I could follow the explanation of what was done, modelling-wise, to create the "RFM-BenMAP" cases in addition to the "RFM-Direct" cases, I think the report needs to do more to explain why these results should differ at all, given that almost all of the important assumptions determining the benefits portion of the calculations were standardized. Is this because the population and mortality rate assumptions still differ from those of BenMAP in the "RFM-Direct" calculations? Is it because the geographical detail differs? The differences between the RFM-direct and RFM-BenMAP results are not large, but why would they be expected to differ at all? If large differences were not expected, why were these extra types of runs important to conduct, given the complexity they have added to the study and report?

(c) The study's method of comparing results from the different RFMs

In model comparison exercises, the usual intent is to understand what structural aspects of each model may cause its results to differ from those of others. This process can help modelers decide whether there is some aspect of their model that they may wish to modify. It can also help the modeling community develop a shared understanding of important structural uncertainties of the system being modeled that cannot be reduced, and which make it useful to rely on results from multiple different models rather than any single model.

The report does discuss hypotheses on what structural differences cause some of the most significant inter-model differences – particularly in regards to the nitrate-related benefits calculations. However, two other possible goals are implicitly suggested in the way that the results are compared to each other in the report that lie outside of the usual model comparison exercise.

One implicitly suggested goal arises at the start of the Results chapter, (i.e., with Exhibit 3-1). Here, the total benefits of the 5 scenarios are compared to each other. This

leaves readers with the first "insight" from the results being that some policy scenarios would, if implemented, have higher benefits than others. This inappropriately focuses the report away from comparisons *of the RFMs* and onto comparisons *of the scenarios*. The use of scenarios that reflect (in some cases) actual policies adds to the apparent policy-evaluation intent. Use of only highly abstract scenarios (as I suggested in my response section 1a above would have helped avoid this suggestion, but the simplest fix of all would be to not compare benefits across scenarios at all.

The other implicitly-suggested intent stems from the decision to compare every RFM's results to the what is predicted by the full-form model CMAQ. This sets a strong assumption that the "right" benefits estimate comes only from CMAQ. The EPA does use CMAQ as a routine choice for its "full-form" benefits analyses, and so this may seem like an appropriate point of comparison. However, one of the structural differences among the alternative RFMs appears to be use of different full-form models as their starting points. The differences in air quality changes predicted by different full-form models may be an important uncertainty that is lost sight of when deciding that the gold standard for performance is to match exactly the benefits that the CMAQ model would produce. Thus, in this report, any understatement or overstatement compared to the full-form CMAQ result is treated as a negative in an RFM's "performance."

Given that the SA BenMAP estimates have been calculated starting from CAMx, which produces air quality predictions very similar to those of CMAQ, this choice of a "sole comparator" makes the SA BenMAP estimates seem to have the best performance (based on regional comparisons). I have no opinion on the merits of CMAQ and CAMx compared to other full-form models that other RFMs may have been based on, but the strong assumption that CMAQ provides the "best benefits estimates of all" should be made clearer in the report.

Inter-RFM comparisons could have been made without reference to CMAQ or any other full-form model outputs by recognizing that the most fundamental output of an RFM is its benefit-*per-ton* (BPT) estimate. Policy benefits of any particular set of reduced form estimates are simply the BPT value multiplied by the change in emitted tons (summed over all emissions). Since the emissions changes will not differ from RFM to RFM, a simpler evaluation approach that would not involve the assumptions I've mentioned above would be to compare the RFS's BPT estimates to each other and to seek patterns in how their relative values differ from scenario to scenario. This would provide a more neutral way of comparing the options without making any assumption that one model's estimates are assumed to be "better" than others, and that would eliminate the potential to suggest that some policies have higher benefits than others.

(c) The report's discussion of analysis limitations

The section that outlines limitations of the analysis is clear but could be more complete. For example, it should include a discussion of the assumptions/limitations I have outlined above with respect to the way the comparisons have been structured. It should also provide more discussion of the fact that none of the comparisons have addressed uncertainties in the underlying concentration-response (C-R) relationship, including the

question of how accounting for non-linearities or non-equal toxicities of the PM2.5 constituents might affect the inter-RFM differences.

Question 2. Please comment on the results of the reduced form tool evaluation in Section 3, considering both the quantitative and qualitative aspects of the model intercomparison. Was the information clearly presented and informative? Were EPA's conclusions reasonable? Are there other results which would be useful to include in the comparison?

My comments under response (1c) above is relevant to this charge question as well, and I copy it below:
In model comparison exercises, the usual intent is to understand what structural aspects of each model may cause its results to differ from those of others. This process can help modelers decide whether there is some aspect of their model that they may wish to modify. It can also help the modeling community develop a shared understanding of important structural uncertainties of the system being modeled that cannot be reduced, and which make it useful to rely on results from multiple different models rather than any single model.

The report does discuss hypotheses on what structural differences cause some of the most significant inter-model differences – particularly in regards to the nitrate-related benefits calculations. However, two other possible goals are implicitly suggested in the way that the results are compared to each other in the report that lie outside of the usual model comparison exercise.

One implicitly suggested goal arises at the start of the Results chapter, (i.e., with Exhibit 3-1). Here, the total benefits of the 5 scenarios are compared to each other. This leaves readers with the first "insight" from the results being that some policy scenarios would, if implemented, have higher benefits than others. This inappropriately focuses the report away from comparisons *of the RFMs* and onto comparisons *of the scenarios*. The use of scenarios that reflect (in some cases) actual policies adds to the apparent policy-evaluation intent. Use of only highly abstract scenarios (as I suggested in my response section 1a above would have helped avoid this suggestion, but the simplest fix of all would be to not compare benefits across scenarios at all.

The other implicitly-suggested intent stems from the decision to compare every RFM's results to the what is predicted by the full-form model CMAQ. This sets a strong assumption that the "right" benefits estimate comes only from CMAQ. The EPA does use CMAQ as a routine choice for its "full-form" benefits analyses, and so this may seem like an appropriate point of comparison. However, one of the structural differences among the alternative RFMs appears to be use of different full-form models as their starting points. The differences in air quality changes predicted by different full-form models may be an important uncertainty that is lost sight of when deciding that the gold standard for performance is to match exactly the benefits that the CMAQ model would produce. Thus, in this report, any understatement or overstatement compared to the full-form CMAQ result is treated as a negative in an RFM's "performance."

Given that the SA BenMAP estimates have been calculated starting from CAMx, which produces air quality predictions very similar to those of CMAQ, this choice of a "sole comparator" makes the SA BenMAP estimates seem to have the best performance (based on regional comparisons). I have no opinion on the merits of CMAQ and CAMx compared to other full-form models that other RFMs may have been based on, but the strong assumption that CMAQ provides the "best benefits estimates of all" should be made clearer in the report.

Inter-RFM comparisons could have been made without reference to CMAQ or any other full-form model outputs by recognizing that the most fundamental output of an RFM is its benefit-*per-ton* (BPT) estimate. Policy benefits of any particular set of reduced form estimates are simply the BPT value multiplied by the change in emitted tons (summed over all emissions). Since the emissions changes will not differ from RFM to RFM, a simpler evaluation approach that would not involve the assumptions I've mentioned above would be to compare the RFS's BPT estimates to each other and to seek patterns in how their relative values differ from scenario to scenario. This would provide a more neutral way of comparing the options without making any assumption that one model's estimates are assumed to be "better" than others, and that would eliminate the potential to suggest that some policies have higher benefits than others.

Question 3. Exhibit ES-4 "Ratio of National Avoided Premature Mortality Benefits Estimates," shows how different reduced-form tools generated different estimates as compared to full-scale air quality models.

My first comment is that this exhibit does not show "how different RFMs compared to full-scale air quality model<u>s</u>" (emphasis added). Rather, it shows how different RFMs, some of which were originally estimated from full-scale models other than CMAQ, compare to estimates from a full-scale application of the single model, CMAQ.

> *3a. Does the report provide a clear and thorough explanation for why some tools under- or over-estimated $PM_{2.5}$ health benefits as compared to the full-scale air quality modeling? Please add any additional explanations for the pattern of results observed.*
>
> Some reasons for these differences are explained, but I sense that more could be done. In particular, there is insufficient explanation of why results of the "RFM-BenMAP" runs differ from those of their respective "RFM-Direct" runs.
>
> *3b. How do the results of this study inform our understanding of the suitability of these tools for regulatory economic analyses in their current form?*
>
> I do not feel that this report has not been structured to inform our understanding about the "suitability" of these tools for regulatory economic analyses. It shows that they differ from each other, as well as from one full-form model; some of these differences are quite large. This result of the study primarily suggests that

there is a very wide range of uncertainty revealed across multiple alternative models. Unless the differences from certain of the models are shown to be due to a faulty model structure or feature,[17] these differences do not suggest that any one RFM is more suitable than another. Rather, these differences suggest that this wide range of uncertainty across all the models should be acknowledged/reported if and when a benefits analysis is to be based on reduced form estimates. However, even using this study's results for that purpose would be problematic because there are many additional uncertainties associated with adopting one or another RFM that this study has not explored at all, such as how potential non-linearities and differences of the relative potency of the PM2.5 constituents may alter the range of model-to-model differences (see my comment on "Limitations" in response (1d) above).

Overall, the question of "suitability" must be tied to the question of "for what use." The charge question appears to mean for use in final regulatory impact analyses, but there are many other ways in which rough cut, ballpark estimates of benefits may be appropriate. While they can have a useful role in scoping analyses, for example, the results of this study are not presented in the context of usefulness in different roles in which benefits estimates are needed and with different degrees of accuracy.

3c. Can any of the reduced-form tools explored in this report easily be modified to allow quantifying the extent to which the total health benefits accrue to specific geographic areas (e.g., by state, or where ambient concentrations are above or below the NAAQS)?

One of the major limitations of relying on benefit-per-ton estimates from an RFM in a benefits analysis is the inability *after the fact* to evaluate the sensitivity of those benefit estimates to potential non-linearities in the assumed C-R relationship. Such alternative benefit-per-ton estimates can certainly be computed on a reduced-form basis by any of the models considered, but doing so requires that the analysts conduct multiple sensitivity runs of the RFM once the air quality baseline and changes have been estimated and report out the alternative BPT estimates. If *any* RFM is used to produce a *single* set of BPT estimates, even if detailed by geographic location, then these scientific uncertainties regarding the C-R relationship will no longer be possible to quantify. This issue is not sufficiently explored in this report, which has considered only a linear C-R relationship.

On a related note, the report makes an interesting point that some benefits estimates are computed for the locations in which the emissions reductions occur while others are computed for the locations where the air quality changes occur. This may be a key to understanding differences in the benefits estimates from

[17] Faulty model features have not been identified, with the possible exception of how NO3 formation is predicted for mobile source NOx emissions in some of the RFMs. Even this question requires more in-depth study.

different tools, but how this may be affecting the comparisons among the models is insufficiently explored in the report. I would have liked to have had more discussion of this point.

Question 4. Since 2008 EPA has used SA-BPT to estimate the health impacts of numerous regulations. Under the scenarios examined in this report, EPA's SA-BPT approach over-estimated $PM_{2.5}$-related health benefits by between 10 and 30 percent, depending on the sector. To ensure BPT estimates correspond to full-form results as closely as possible, the report recommends updating the underlying emissions inventories and air quality modeling used to inform the EPA SA-BPT approach over time.

4a. In the interim, how might EPA improve its characterization of results derived from the 2005 SA-BPT approach, specifically the potential degree of over- or underestimation in BPT-based results for a particular regulatory scenario?

Although SA-BPT has been found to overstate benefits relative to the "sole comparator" of a full-form CMAQ run, it has also been shown to produce even larger estimates relative to several other RFMs. This charge question is posed from the perspective that the gold standard of a correct benefit estimate does come from CMAQ. In better characterizing the potential degree of error in its estimates, it would be appropriate to recognize the range of all BPT estimates from multiple RFMs within which the SA-BPT estimates lie. However, even this range of uncertainty will be misleading because it would lack representation of the additional uncertainties not explored in this study of potential C-R non-linearities and potential for PM2.5 constituents to have non-equal toxicities. If a range of uncertainty is to be characterized, it should be inclusive of all the key forms of uncertainty in the estimate.

4b. What criteria (e.g., geographical scale, regulated sector, pollutants/precursors) should EPA examine to determine the potential for divergence between SA-BPT results vs full-form air quality modeling results (resulting in over- or under-estimation)?

N/a

4c. Based on the results of this study, does the panel have any additional recommendations about BPT-based approaches?

My responses to questions (4a), (3b) and (3c) are relevant here too. I copy those responses below for ease of reference:

4a: Although SA-BPT has been found to overstate benefits relative to the "sole comparator" of a full-form CMAQ run, it has also been shown to produce even larger estimates relative to several other RFMs. This charge question is posed from the perspective that the gold standard of a correct benefit estimate does come from CMAQ. In better characterizing the potential degree of error in its

estimates, it would be appropriate to recognize the range of all BPT estimates from multiple RFMs within which the SA-BPT estimates lie. However, even this range of uncertainty will be misleading because it would lack representation of the additional uncertainties not explored in this study of potential C-R non-linearities and potential for PM2.5 constituents to have non-equal toxicities. If a range of uncertainty is to be characterized, it should be inclusive of all the key forms of uncertainty in the estimate.

3b: I do not feel that this report has not been structured to inform our understanding about the "suitability" of these tools for regulatory economic analyses. It shows that they differ from each other, as well as from one full-form model; some of these differences are quite large. This result of the study primarily suggests that there is a very wide range of uncertainty revealed across multiple alternative models. Unless the differences from certain of the models are shown to be due to a faulty model structure or feature,[18] these differences do not suggest that any one RFM is more suitable than another. Rather, these differences suggest that this wide range of uncertainty across all the models should be acknowledged/reported if and when a benefits analysis is to be based on reduced form estimates. However, even using this study's results for that purpose would be problematic because there are many additional uncertainties associated with adopting one or another RFM that this study has not explored at all, such as how potential non-linearities and differences of the relative potency of the PM2.5 constituents may alter the range of model-to-model differences (see my comment on "Limitations" in response (1d) above).

Overall, the question of "suitability" must be tied to the question of "for what use." The charge question appears to mean for use in final regulatory impact analyses, but there are many other ways in which rough cut, ballpark estimates of benefits may be appropriate. While they can have a useful role in scoping analyses, for example, the results of this study are not presented in the context of usefulness in different roles in which benefits estimates are needed and with different degrees of accuracy.

3c: One of the major limitations of relying on benefit-per-ton estimates from an RFM in a benefits analysis is the inability *after the fact* to evaluate the sensitivity of those benefit estimates to potential non-linearities in the assumed C-R relationship. Such alternative benefit-per-ton estimates can certainly be computed on a reduced-form basis by any of the models considered, but doing so requires that the analysts conduct multiple sensitivity runs of the RFM once the air quality baseline and changes have been estimated and report out the alternative BPT estimates. If *any* RFM is used to produce a *single* set of BPT estimates, even if detailed by geographic location, then these scientific

[18] Faulty model features have not been identified, with the possible exception of how NO3 formation is predicted for mobile source NOx emissions in some of the RFMs. Even this question requires more in-depth study.

uncertainties regarding the C-R relationship will no longer be possible to quantify. This issue is not sufficiently explored in this report, which has considered only a linear C-R relationship.

On a related note, the report makes an interesting point that some benefits estimates are computed for the locations in which the emissions reductions occur while others are computed for the locations where the air quality changes occur. This may be a key to understanding differences in the benefits estimates from different tools, but how this may be affecting the comparisons among the models is insufficiently explored in the report. I would have liked to have had more discussion of this point.

Question 5. How do the results of this study inform the future development of reduced-form tools that are capable of providing reliable estimates of impacts associated with different sectors, across a variety of spatial scales, and for different portions of the air quality distribution? Are there other, less resource intensive approaches than full-scale air quality modeling for informing the public about the size and distribution of PM health benefits associated with alternative regulatory scenarios?

The report makes it obvious that the very act of trying to summarize a concept (i.e., "benefits") that is highly complex, location-specific, scenario-specific, and subject to scientific uncertainties affecting the underlying formulas with a single set of fixed "per ton" benefit values cannot produce "reliable" estimates of impacts. So many full-scale model runs would have to be run to obtain a more detailed set of "more reliable" reduced-form values that are tailored to particular analytic circumstances that it would probably be more efficient to just perform the full-scale models runs as each new and different policy arises. Even doing that leaves many important sources of uncertainty in the benefits estimates uncharacterized. The more important line of further research is not to attempt to improve on the reduced form approach, but instead to develop a thoughtful evaluation of when this rough and inaccurate estimation approach can be helpful to guide policy decisions -- and when is inherently too unreliable to be used to inform a policy decision. BPT estimates may be useful for screening out or refining potential regulatory options before reaching the proposed rule stage, even if they are deemed too unreliable to be used to inform the public about the benefits of a proposed or final regulatory option. The report could have done more to discuss the usefulness of such reduced form tools in different parts of the regulatory decision process. In thinking about this broader question in the future, the concept of data quality objectives may be useful to consider.

Smith, Richard

Comment about the structure of the Charge Questions

Specifically, I don't understand why questions 2 and 3 were separated. Charge Question 2 asks us to review Chapter 3, and Charge Question 3 is about Exhibit ES-4 from the Executive Summary. But Exhibit ES-4 is also Exhibit 3-3 of Chapter 3, and to answer the questions about it, it is necessary to read the whole of Chapter 3. The only difference is that Charge Question 3 is asking some more specific questions about the report that may or may not get covered under Charge Question 2. It seems to me we should merge Charge Questions 2 and 3, though with particular attention to the specific issues raised under parts a, b and c of Charge Question 3.

Response to Charge Question 1

This question asks us to comment on the evaluation approach and whether the choice of five emission reduction scenarios was adequate.

To answer the second question first, I am not an expert on emission scenarios and had never heard of four of the five scenarios considered (the exception being the CPP). However, I feel they should also have included the Mercury and Air Toxics Standard (MATS), which is topical since the SAB produced a report on the proposed revisions of MATS in February. Like the CPP, this was also focused on electricity generating units, but as I understand, the detailed pollution reduction measures were quite different, with the CPP focusing on greenhouse gas emissions while MATS was directed at mercury, but EPA also claimed substantial "co-benefits" in PM2.5 reductions. Therefore, it would be interesting to see how robust EPA's evaluation of these co-benefits would be under the proposed alternative modeling approaches.

In commenting on the overall approach, I appreciate of course that for an exercise like this to deliver useful information, it's necessary to limit its scope, but there are a number of places where I wish the scope had been drawn somewhat more widely. In particular, the report does not make any comparisons based on varying either the value of statistical life (VSL) or the concentration-response relationship (CRR), so I presume these values were provided to the contractor by EPA and do not represent the contractor's own judgment.

With respect to either VSL or CRR, one could make the comment that simply substituting one number for another will result in all the estimates being scaled up or down by the appropriate scaling factor, and will therefore not affect the relative values of estimated costs which are the main focus of the report. In the case of VSL, which the report fixes at $8.7 million in 2015 dollars, I think that assumption could very well be true, though even here, the question of whether income levels (one component of the VSL) vary systematically across spatial regions, which are affected differently by changes in air pollution, could be an issue. Here, I am more focused in the CRR, which

numerous studies have shown varies both spatially and with the level of PM2.5 (i.e. a non-linear CRR).

The authors say they are using the CRR from Krewski et al. (2009), but they don't even include that reference in the bibliography of their report. I am fairly sure they are referring to this:

Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y, Turner MC, Pope CA III, Thurston G, Calle EE, Thun MJ. 2009. Extended Follow-Up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality. HEI Research Report 140. Health Effects Institute, Boston, MA.

So I would start by adding that reference to the bibliography of the report.

However, I don't think matters are quite as simple as just citing the Krewski report. That report is full of tables estimating hazard ratios (HRs) for PM2.5 under various statistical modeling assumptions, and nowhere does the report under review indicate which specific estimate they were using. I downloaded the user manual for EASIUR (https://barney.ce.cmu.edu/~jinhyok/easiur/EASIUR-Users-Guide-200505-Jinhyok.pdf) and they quote a HR of 1.06, associated with a 10 $\mu g/m^3$ rise in PM2.5, with a 95% confidence interval (1.04,1.08). They attribute this to the above Krewski report, but they don't say specifically which estimate from that report they are using. I searched through the Krewski report to see if I could find an estimate that was consistent with their reported value, and I conclude that the EASIUR authors were taking the "All Causes – Standard Cox Model" estimate from Table 10, page 27 of the Krewski report. So my first question to EPA is: was this in fact the same estimate as was used for the report under review?

My second comment about the CRR is that even if that estimate and its confidence interval are accepted at face value, that still implies a 2:1 ratio between the upper and lower confidence bounds of the mortality differential associated with a specific adjustment of PM2.5, which puts in perspective some of the discussion in the report under review (whether a 2:1 ratio between different modeling approaches is an acceptable level of agreement). In fact the true uncertainty about the CRR is almost certainly much larger than that confidence interval would imply, since this is just one of a large number of estimates produced by Krewski et al. under different statistical modeling assumptions without any clear-cut criterion which of those (if any) is really the best estimate.

My third comment relates to the whole question about whether this is just a debate about a multiplicative factor that doesn't affect the inter-model comparisons: I don't think it is. There is substantial evidence that the CRR varies both with the level of PM2.5 (i.e., is non-linear) and spatially, and either of these would invalidate a simple scaling relationship among estimates of total benefit. I don't think any alternative estimate is definitively established as best, but my feeling is that the report should have somehow

reflected that there are different ways of going about estimating the CRR and they don't necessarily lead to the same thing in an exercise like this.

Response to Charge Questions 2 and 3

I will try to focus specifically on the issues raised under 3a, b, c, but I think they are subsumed in a general review of what the report achieves.

National results

The same data are presented a number of different ways: for example Exhibits 3-2 and 3-3 are the same data in two different forms of display, and I presume both are derived from Exhibit C-1 in Appendix C. However, there seem to be some minor inconsistencies in the way the data in Exhibit C-1 were reduced to the two figures: for example, if we look at the results for AP2-Direct v. CMAQ-BenMAP under total PM2.5 for the Tier3 scenario, Exhibit C-1 shows a benefit of $4,100 (millions) under CMAQ-BenMAP and $11,000 under AP2-Direct, a ratio of 2.68, not 2.8 as reported in Exhibits 3-2 and 3-3. There are numerous minor inconsistencies like that, and while none of them is large enough to affect the broad comparisons being made, I do feel that the report's authors should have taken care to ensure self-consistency of results from different parts of the report.

Following Exhibits 3-2 and 3-3, the report tabulates the separate calculations made for the three major components of PM2.5: primary PM2.5, which the authors abbreviate to prPM2.5 (as noted in the report, the estimate is actually based only on the EC component of prPM2.5, and scaled up to reflect the OC and crustal components), sulfates and nitrates. These three are plotted together in Exhibit 3-4, which is formatted to look like Exhibit 3-2 and not Exhibit 3-3. The following comment does not affect the overall evaluation of the report, but I would have preferred that they had used Exhibit 3-3 as the main format for their plots rather than Exhibit 3-2. I have two reasons for saying that. First, ultimately this report is trying to compare the different modeling approaches, not the different emission reduction scenarios. Therefore, it makes more sense to compare models within scenarios, rather than scenarios within models. My second reason for preferring Exhibit 3-3 over 3-2 is that Exhibit 3-3 is the one extracted as Exhibit ES-4 which is the subject of Charge Question 3. Therefore, it would have been better to use the format of Exhibit 3-3 for all the model comparison plots.

As for the results themselves, I learned a great deal from Exhibit 3-4, which I think goes a long way towards explaining the differences in overall PM2.5 results. There doesn't seem to be too much of a discrepancy in the way the different modeling techniques treat prPM2.5, but this component is also of relatively minor importance compared with sulfates and nitrates, at least for the policy scenarios being considered in this report. The results for sulfates also seem reasonably consistent overall, though with underestimation in some approaches (notably AP2 and EASIUR), but it seems that all versions of the APX class of models overestimate the sulfate component of the benefits

for the Tier3 scenario. This may be saying something specific about how APX handles the Tier3 scenario but it is difficult to draw any broader conclusions than that.

However, the nitrate results are all over the place and this may be the single most significant take-away from the whole report. All versions of the APX model and (even more so) the InMAP model seem to grossly overestimate the benefits due to nitrate reductions, and moreover, this is not just for one scenario but seems to happen across the board. However, there are also some substantial underestimates, most notably for AP3-BenMAP and the Pulp and Paper scenario. The results for SA-Direct and EASIUR also show a lot of variability but not nearly as much as the APX and InMAP results.

I think one question that should be raised here is whether the problem might be with CMAQ rather than the other modeling approaches (CMAQ may be the best-established approach but one shouldn't assume it works uniformly well under all circumstances!). However, I don't really think that is the explanation, for two reasons, (a) there is still good agreement between CMAQ-BenMAP and CAMx-BenMAP, (b) recalibrating the CMAQ results wouldn't do anything to help the variability among the other four modeling approaches – for example, the "Refineries" and "Pulp and Paper" scenarios are fairly similar in terms of CMAQ or CAMx estimated benefits but differ wildly in their estimates under the APX and InMAP models, and to a lesser extent, the SA-Direct and EASIUR approaches (Exhibit C-1).

I would be wary of using these results to conclude too quickly that the SA-Direct and EASIUR approaches are therefore better than the APX and InMAP approaches, though the relative failure of InMAP is certainly a disappointment. I'm not really familiar with the specifics of any of the four models, but it seems to me that an approach that attempts to model the transport of different pollutants through the atmosphere, even if imperfectly, should be better suited than a more statistical approach, which is what I understand SA and EASIUR to be, when it comes to evaluating a completely new scenario. The disadvantage of all statistical approaches is that they are only as good as the data used to calibrate them, and they may not perform well when applied to scenarios outside the range of past experience. Therefore, as a matter of very broad general principle, I would prefer something that seems closer to the full modeling approach of CMAQ or CMAx, which I understand the APX models and InMAP to be. Maybe there is some way these models can be retuned to better estimate the nitrate component.

Regional Results

In contrast to the national results, which I found overall well explained, I don't think the authors ever quite got to grips with the regional parts of the analysis. According to Appendix B, the continental US was divided into 7 regions defined by states. County-level results for each modeling approach were aggregated into the 7 regions, but instead of presenting separate results for all 7 regions, the authors calculated summary statistics (principally R^2, normalized mean bias and normalized mean error). I found these hard to interpret. Why can't we see individual results for the 7 regions? Exhibits 2-

3 through 2-7 show substantial variability among the policy scenarios in terms of which parts of the country they affect – it's therefore entirely plausible that the results for different scenarios will be quite different in different parts of the country. Also, given the different results that were obtained nationally for total PM2.5 and for different components, I feel those differences should be shown in the regional results as well.

The authors of the report may have felt it would be too much to produce 28 different figures (7 regions multiplied by 4 air pollution components) looking like Exhibit 3-2 or 3-3, but I think the information could have been shown more compactly in tabular form (a picture isn't always worth a thousand words…). In particular, I should point out that Exhibit C-2 only shows in tabular form what is in the figures Exhibits 3-5 through 3-7; I would have preferred to see the raw numbers for the seven regions being calculated.

Another part of the presentation that I don't understand is why the authors didn't include any of the "Direct" model results in the regional summary. This is especially important if the overall tenor of the report seems to be favoring the SA-Direct and EASIUR models – if either of these doesn't work well in regions, we need to know about it! I thought the explanation might have been because of the spatial mismatch between where the emission reductions occur and where the health benefits are felt (this point is first made on Page ES-6), but in fact, the discussion on page 2-18 implies that regional results should still be valid ("By aggregating the results to the regional scale, we minimized the distinction between emissions locations and receptor locations caused by emissions transport."). In fact, there may still be some discrepancy here, because transport obviously does occur across regional boundaries, but that is precisely why I would like to see the results.

Complexity and Level of Effort

Exhibit 3-8 is informative. This table describes the format of each of the modeling tools and qualitatively evaluates them according to their pre- and post-processing requirements, time requirements, and level of skill and software required. The APX tools require MATLAB, which (unlike R) is not a free package, but my feeling is this should not deter an agency responsible for national policy. MATLAB is a very well established package and its mathematical routines are widely applied across a large number of scientific disciplines.

Response to Charge Question 3(a)

In terms of providing a "clear and thorough explanation" (why some tools over- or under-estimated health benefits), I would say the report is partially successful, but overall, it is more successful at documenting the results themselves than it is at providing an explanation for them. The one clear theme to emerge is that there seems to be a problem with the nitrate component of several of the models, but especially APX and InMAP. I don't think the report really explains exactly what went wrong with these models, but it's not clear to me what else the contractors were expected to do. I do feel the report could have done a better job of documenting the regional results, which could

provide further insights into what is going wrong with some of the models in their national projections.

Response to Charge Question 3(b)

This question addresses the suitability of the reduced-form models for regulatory analysis "in their current form", which I assumed means without making any modification to the code. I think the report clearly shows that there are problems with all versions of the APX model and also with the InMAP-BenMAP model in that these models especially seem to miss the nitrate component and perform badly on the Tier3 scenario. In contrast, SA-Direct and EASIUR have more consistent results across the full range of scenarios, though it appears that SA-Direct systematically overestimates the benefits and EASIUR systematically underestimates them (except for Tier3). However, I would be extremely cautious about making blanket generalizations from these results. I believe these are primarily statistical models and, although I am a statistician myself, I am inclined to be cautious about using statistical approaches to complex environmental modeling problems. Statistical approaches such as multiple regression generally work well when applied to input configurations that are similar to those on which the model was originally trained, but may perform very badly in configurations that are a long way from that space.

Response to Charge Question 3(c)

The question asks whether "any of the reduced-form tools explored in this report [can] easily be modified…" to achieve more detailed geographic results; that is a hard question to answer, because I don't know what would be involved in modifying the tool itself, rather than giving more detailed regional results for the tool as it currently exists. Given the stated intention of the report, that the tools would be used in their current form rather than adapted to the scenarios being assessed, it seems to me that the safest response to this question would be no, we can't answer that question because the study wasn't designed to. However, if EPA is willing to consider extending the present study, to consider how various modifications of the tools could improve the results, it seems to me that could be a very useful exercise.

Response to Charge Question 4

Since I'm not assigned to discuss this, I'll leave it to others to provide more detailed review, but I'd like to contribute some short responses.

Charge Question 4(a): I support the recommendation that the SA model should be updated to reflect more recent emissions inventories and air quality modeling. In the meantime, however, I would be very cautious about any implication that EPA could just scale up or down the SA results to match the CMAQ results more closely. The national results are still based on just five test scenarios and there is really no basis for assuming they will apply to others. However, the case for doing this would be stronger if

it could be demonstrated that similar scaling factors also apply to all the regional results, something it seems the authors of the report could have included.

Charge Question 4(b): The question is hinting that other analyses based on geographical scale, regulated sector and different pollutants/precursors could also be evaluated to test further the agreement between the SA and CMAQ models. In principle, I agree. As noted in response to 4(a), it seems to me that the authors of the report already have the data they need to report on the geographical scale component. It would be worth extending the analysis to include different regulated sectors and pollutants/precursors.

Charge Question 4(c): Nothing at the moment. I may have more comments during the meeting.

Response to Charge Question 5

Although I have been critical of a number of aspects of this report, I strongly support making the kinds of comparisons that the report does: if we don't look at these issues, we'll never know how the models compare. Although I don't necessarily accept that CMAQ (or CAMx) should always be the gold standard of modeling approaches, I see nothing in this report to challenge that assumption. Therefore, it seems to me that CMAQ and CAMx will remain the preferred approaches for the foreseeable future.

One final comment: it seems to me that (in addition to Krewski), the reference for Boylan and Russell (2006) is missing (see page 2-17). I believe the paper being cited is this:

PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models; Boylan, James W. and Russell, Armistead G., Atmospheric Environment, Volume 40, Issue 26, p. 4946-4959, DOI: 10.1016/j.atmosenv.2005.09.087