ICF INTERNATIONAL

MEMORANDUM

To: Loren Hall, USEPA

From: Arlene Rosenbaum, Jonathan Cohen, and Ed Carr

Date: January 5, 2011

Re: **Response to Peer Reviewers Comments on Development of Ambient Exposure Models for Predicting Methyl Bromide Concentrations in California (DRAFT)**

BACKGROUND

EPA must ensure that recipients of financial assistance comply with the relevant non-discrimination requirements under federal law. These laws include Title VI of the Civil Rights Act of 1964, which encompasses a broad spectrum of impacts of discriminatory activity, including adverse health effects. Title VI prohibits discrimination on the basis of race, color, or national origin. The provisions of EPA's non-discrimination regulations implement Title VI and other non-discrimination laws. For EPA to make a finding of a Title VI violation, it must conclude that the harm alleged is both adverse and disproportionate to the members of a class of persons as defined above. A Title VI violation does not require finding that an action was taken with a discriminatory intent.

EPA's regulations allow persons to file administrative complaints alleging discrimination by recipients. The EPA Office of Civil Rights has the responsibility to process and review complaints that meet certain jurisdictional criteria. If these jurisdictional grounds are met, the complaint is generally accepted for investigation. Such acceptance for investigation does not imply any EPA finding of violation has or will be made.

The Title VI Administrative Complaint that resulted in the investigation that developed these models was received in 1999, and includes allegations concerning disproportionate adverse impact from methyl bromide (MeBr) exposure on California's Hispanic schoolchildren, including those at five schools named in the complaint. The alleged recipient named in the complaint is the California Department of Pesticide Regulations (CDPR), and the action that resulted in an alleged adverse disparate impact was the 1999 recertification of MeBr for use in California. EPA found that the complaint met jurisdictional requirements. The scope of the investigation for this complaint was determined after review of complaint allegations and recipient scope of authority.

The early stages of the investigation involved planning for an exposure assessment to use in evaluating the allegations. The geographic extent includes most schools in California, and the time period includes historical through then-current conditions (1995-2001). The exposures involved may be cumulative from multiple fumigant applications, and range over averaging periods of days to months. To evaluate the potential for disparate impacts, the exposure estimates need to be specific to a time and place to link with population characteristics. The approach used must be practical to apply for a wide geographic area with many receptors and usage events.

The Office of Civil Rights (OCR), with contract support by ICF International, developed a number of alternative models for estimating ambient air concentrations of the soil fumigant pesticide MeBr from multiple agricultural applications for receptor locations in California. The model development process entailed regression analysis of daily ambient monitoring data in multiple locations against proximate

MeBr usage amounts and meteorological conditions. The model formats involved calculating the daily atmospheric concentration from the surrounding area (for the same and previous days') usage of MeBr. These usage amounts were combined with alternative sets of "adjustment factors," in a model whose coefficients were fitted by a linear regression process. The adjustment factors included wind and other weather conditions as well as distance.

To evaluate the performance of the alternative formulations, a number of measures which compared the model concentration estimates with the monitored concentration data were developed, and a composite ranking was generated. After reviewing the alternative models' performance ranking data, along with physical plausibility of model components, and the statistical significance of component coefficient values, a single model was selected to use in conducting a statewide exposure assessment for California MeBr agricultural use between 1995 and 2001.

PEER REVIEW PROCESS

OCR began a contractor-supported peer review process for the exposure model development procedure and documentation in 2004, and completed it in 2005. A contractor provided support in identifying potential reviewers, providing compensation to reviewers as subcontractors, and managing the review process by distributing materials and receiving comments from reviewers.

After undergoing internal review by EPA scientists, the model development documentation "Overview of MeBr Ambient Exposure Model Development Process"(ICF 2005a) was provided to scientists outside of EPA for peer review, along with a set of charge questions to guide the review process. The results of the review are presented in a contractor report (Eastern Research Group, 2005).

This memorandum summarizes the comments received by the peer reviewers and provides responses. It is organized by charge question. The four charge questions provided by EPA (and a fifth provided by Eastern Research Group) were:

1.Does the document, "Model Development for Assessing California Methyl Bromide Ambient Concentrations", provide a clear and adequate description of the goals and methods EPA used to develop and review alternative exposure models? What additional information, if any, is critically needed to complete the documentation?

2.What are the overall strengths and weaknesses of the model development process as described?

3.What are the strengths and weaknesses of the data quality assurance activities conducted during the model development process?

4.What are the strengths and weaknesses of the model ranking elements, and the model ranking process? Can you identify alternative ranking measures that would be likely to present significantly different information about model performance that should be considered in model selection? Would these alternative measures be likely to change the selection process outcomes as described?

5.Provide any additional comments or recommendations you feel are important to improve the quality of this document.

When reviewers raised an issue in response to more than one question, the discussion and response were merged and presented together at the first instance an issue was mentioned.

SUMMARY COMMENTS AND RESPONSES

1. Does the document, "Model Development for Assessing California Methyl Bromide Ambient Concentrations", provide a clear and adequate description of the goals and methods EPA used to develop and review alternative exposure models? What additional information, if any, is critically needed to complete the documentation?

A. Reviewers suggested that the document provide more information about intended uses of the model, and indications of acceptable performance. Specific suggestions included adding more discussion about the range of exposure averaging periods expected to be used, performance criteria used for evaluating suitability of predictions, and considerations for estimates of long-term averages.

"The document should include discussion of the intended use of the predicted exposures. The nature of the effects of exposure over longer than acute time periods is dependent not only on the average exposure over the period, but also on shorter-term exposure peaks and their duration."

"The document outlined three types of exposure of interest: short term to long term. The model appears to target subchronic (7-8 weeks) for final application of results, but leaves it open to apply it to other periods of duration. It acknowledges in the text that short tem daily exposure is most difficult to model with the highest range of uncertainty. To what extent the model is intended for other periods than subchronic needs to be better expressed. It is not demonstrated that this model would be useful for longer periods."

"The performance criteria for the application of the model need to be more clearly outlined and distinguished so that the advantages and limitations of the model can be more readily evaluated for the desired application, and the potential results of model application more easily interpreted according to the demands of the application. . . .How low [are the concentration levels that models] are needed to be able to predict and at what degree of uncertainty is acceptable? As we get to lower levels of air concentration, most model performance goes down for a number of factors (accuracy of input data [meteorological data, use date, etc.] and model uncertainty).

"Describe the required accuracies of the exposure estimates. For example, is it satisfactory for the predictions to be within a factor of 10? Factor of 2? 10%?"

"What are the desired averaging times and distances for the concentration outputs?"

Response:

The models developed and reviewed for use were all designed to predict daily averages, with longer term averages estimated by averaging predicted daily averages. The longer term averages of interest for this study include 7 days, 30 days, 42 days (6 weeks), 182 days, and 365 days.

Using monitoring data collected on three or four days per week over periods of 7-9 weeks, models were ranked using several performance criteria for estimates of various averaging periods. While daily concentration estimates were weighted most highly in the ranking, other performance measures included ones for 2-week (i.e., 6-8 monitoring days), 4-week (14-16 days), and 7/8-week (22-30 days) averages, which were computed from the predicted daily averages.

In general, models that perform well over short periods will perform even better when the predictions are averaged over longer periods. The model performance results show that the performance as measured by R-squared values increases, and the mean square error decreases, as the averaging

period increases. Similar improvements in other performance measures also occur as averaging periods increase. Explicitly considering the longer-term measures in the ranking helps reduce the possibility of coincidentally predicting one or a few averaging periods better than others.

The model performance results show that the 8-week averages are approximately unbiased with some variance V. Because the correlation between consecutive predicted 8-week averages should be small, since they are primarily impacted by different MeBr applications and meteorology, predicted values for longer multiples of 8-week averages would also be approximately unbiased with a smaller variance V/s. Therefore we expect that a model successful in this measure would be even more useful for longer periods than 8 weeks.

After receiving peer review comments, the model evaluation process was also modified to explicitly consider performance for monitored concentrations below the median value for each averaging period, i.e., "low daily values". The updated ranking process included measures of model performance such as mean square error for this subset, in order to identify models that perform adequately for lower concentrations.

The document has been edited to expand the discussion of appropriate model uncertainty targets. In historic EPA exposure model development, acceptable performance has often been measured by the fraction of predicted values within a factor of 2 of measured values. (US EPA, 1992, Hanna et al., 2003) This performance statistic was explicitly included in the revision to the model ranking system used in final model selection, which was developed after the receipt of peer review comments.

The selected model predicts the observed daily average concentrations within a factor of 2 more than half the time for data sets of both all daily values and low daily values. For averaging times of 2 weeks or longer the selected model predicted values within a factor of for more than two-thirds of the time.

B. Reviewers also suggested that the document should be edited to more completely explain the rationale for the approach used in the analysis.

"Why are so many regression models being considered and discussed?"

"A better justification is needed for considering tens of thousands of "models" (regression equations)."

Response:

The reason for fitting so many regression models is that we did not know the specific form of the relationship between the concentrations and the usage. We used physical principles to determine the general form of the independent variables codified in the usage adjustment component factors. The concentration was predicted as a sum of the "adjusted" MeBr usage across the surrounding geographic area (defined by grid cells within various distance radii). Each of the many regression models used alternative factors and combinations of factors to "adjust" the usage according the physical principles (e.g., wind speed, wind direction, temperature).

As a result, we needed to compare the performance of alternative models with different numbers of terms to determine empirically the best overall performance using a combination of assessment criteria. The range of models included a balance between the most complex models with more terms and, therefore, more uncertainty for each parameter, and the simpler models with fewer terms and parameters that could each be estimated with more precision. This is a very common situation in

statistical modeling when there are a large number of "independent" regressor variables that could be included in a regression model.

C. Reviewers suggested that the document should more thoroughly discuss how conditions during model calibration relate to the conditions during expected uses.

"It's not clear how the wide variability in emissions characteristics of MeBr volatilization from soil are considered in the models."

"In order to evaluate the proposed model and methods of the model, we need to know how well the model is likely to perform to the target application receptors. For this evaluation, the target application receptors need to be characterized. This is missing from the document. The similarities and differences from the conditions of the monitored receptors used to construct the model need to be better known to evaluate the ability to apply the model outside the three counties used to develop the model."

"My understanding from the conference call is that the three counties monitored represent a wide range of soil conditions and meteorological conditions and that the differences tend to average out. It was also stated in the call that the range of conditions are inclusive of the counties to be applied. This needs to be stated and documented."

"Soil moisture is regarded to be a significant factor for emission rates. It was stated in the conference call that MeBr is generally applied to irrigated fields, so that soil moisture is not likely to be an issue for emission, as there is no significant variance. Since MeBr is applied prior to planting, this needs to be better established and documented."

"It would be useful for the application receptors to be mapped in conjunction with MeBr typical use patterns. Relevant information to be characterized includes the distribution of intensity of use, the variation in source to receptor distance, differences of conditions affecting emission and atmospheric transport (e.g. soil, soil moisture, irrigation, meteorological and topological conditions)."

Response:

MeBr volatilization from soil is known to depend upon a variety of factors, including the application method and sealing, soil type, moisture, organic content, and temperature. The application receptors for this study are public schools throughout the state of California. Thus, as noted by the commenters, the selected model should be applicable to all locations in California with appreciable agricultural applications of MeBr, with respect to application method, soil type, moisture, organic content, and temperature. Ideally, the model would incorporate these factors into the independent variables included in the adjustment factors.

Unfortunately, while California collects uniquely detailed pesticide usage data, it does not collect information on the application method, precise location or local environmental characteristics. And while sources of data for meteorological conditions were available, data on detailed soil characteristics were not.

As explained in the model documentation, models with alternative approaches to accounting for the effects of application rate and meteorological conditions were developed, and their performance compared. Factors which were developed to model the effects of volatilization and dispersion of the MeBr usage included several categories:

Application Method and Sealing The most common pre-plant application methods for MeBr described by CDPR include untarped, shallow tarped, deep untarped, and bed tarped, as described in the model documentation. A previous CDPR study had measured the fractional daily volatilization for each method, but the prevalence of particular application methods was unknown for either the monitoring period or the exposure evaluation period.

To account for the effect of application method on MeBr volatilization over multiple days, we used two alternative model formulations. The first approach used CDPR's empirical data on the percentages of emitted MeBr for the application day and each following day for selected application methods, and for an average across three of those methods. Thus the pounds of MeBr used are multiplied by the experimentally obtained percentages to estimate the proportional mass of emissions on each day. This first approach makes the approximation that all MeBr applications use the same method or an average method.

The second approach treats the daily percentages of mass release as unknown parameters. Thus the emissions on the *r*'th day after the application are multiplied by an unknown factor *f(r)* to estimate the emissions on each day. This second approach can be regarded as making the same approximation that all MeBr applications use the same method, which has an unknown set of daily mass release percentages. Alternatively, the second approach can be regarded as being based on the assumption that there are some number of different daily emission rate profiles for each possible application method and soil type, each with some probability of being used, so that *f(r)* is a weighted average of the emission percentage for day *r* and profile *p*, weighted by the probability for profile *pr*. Since the daily emissions are also multiplied by an unknown dispersion factor to estimate the concentration, the regression coefficients do not provide estimates of *f(r)*, but the ratio of the regression coefficients for day *r* to day *s* provides an estimate of *f(r)*/*f(s)*.

Temperature To account for the effects of ambient air or soil temperature on emissions, we used an empirical approach, since the volatilization flux equations in the literature are highly non-linear and include several parameters that depend upon the soil type, tarp type (which affects MeBr diffusion), and other physical parameters that are also not available from the pesticide usage database. We assumed that the temperature effect could be approximated by a simple equation

$$\textit{Temperature adjustment factor} = a + b * \text{Average } [\max(\text{temp} - 4, 0)]$$

where *a* and *b* are unknown parameters (that are included in the derived regression coefficients), temp is the hourly temperature in ºC, and the average is taken over the daytime, nighttime, or entire 24 hours, depending upon the model formulation. Thus the usage is multiplied by one factor *f(r)* to estimate the emissions on day *r* after the application under some average temperature conditions, and then is multiplied by the *temperature adjustment factor* to account for the temperature effect. The temperature above 4 ºC was used, since that is the boiling point of MeBr, and volatilization at temperatures below this was assumed to be negligible. Hourly temperatures below 4 degrees were counted as 0 values in the average.

We used either ambient air temperature or soil temperature in this manner, depending on the specific model formulation. We did not attempt to estimate an adjustment for soil temperature and ambient temperature together, due to project limitations as well as the consideration that these effects would be expected to be similar (due to an expected correlation between the soil and ambient temperatures) and combining the two adjustments would increase the model complexity and therefore tend to reduce the precision.

Soil Moisture Soil moisture is not expected to vary as widely as other soil conditions, since it can be changed by irrigation, and is usually controlled to improve the effects of fumigation (Cohen et al. 2011, Section III D). Prior to planting in a typical strawberry field, the field is typically sprinkler irrigated for 12 hours. Then (for bedded fields) two drip-lines are installed per bed, using a tape layer machine, which buries the pipe in the beds prior to fumigation. (University of California Cooperative Extension, 2004a, 2004b). There is also evidence that soil fumigation with MeBr is more effective in moist but unsaturated fields because of higher soil retention (Gan et al., 1996), which provides an incentive for applications to occur within a relatively narrow range of soil moisture. Because of these issues, and the scarcity of relevant data, soil moisture is not explicitly considered in the models.

Soil Type and Climate The overall variability in emissions characterization due to soil type differences were considered in the regression model by calibrating the model with ambient monitoring data from several locations in three different California counties, i.e., Kern, Santa Cruz and Monterey Counties. These locations are expected to cover much of the range of soil types found in agricultural areas of California where MeBr is used.

Principal areas throughout the state where frequent applications of MeBr have historically occurred include the Monterey and Santa Cruz Counties, Ventura and Santa Barbara County and the San Joaquin Valley (Sacramento, San Joaquin, Stanislaus, Merced, Madera, Fresno, Tulare, Kings, and Kern). In general, the climate and soils of the San Joaquin Valley counties are similar, and so should be reasonably well represented by Kern County. The coastal counties (Santa Cruz, Monterey, Santa Barbara, and Ventura) experience similar climates, although Monterey and Santa Cruz experience somewhat cooler and wetter conditions than Santa Barbara and Ventura. We anticipate that the ranges of conditions used in the model development are broadly representative for these counties as well other major agricultural areas of the state.

Distance between Use and Receptor Locations: CDPR collects pesticide use data which is geographically coded to a 1 mile square area, called a Meridian-Township-Range-Section (MTRS) location. Prior CDPR model development studies used MTRS center points to represent these locations out to 7 sections distant from a receptor, but found limited benefit to explicitly considering the distance in estimating concentrations. The CDPR reports summarizing the monitoring data also provided some information regarding the proximity of nearby MeBr usage by MTRS section. Detailed usage data from the Pesticide Use Reporting (PUR) database for the monitoring periods were used in the EPA calibration process.

In order to specify MeBr application locations more precisely for the purpose of calibrating these regression models, a denser grid of quarter-mile square cells was used in calibrating the models and in the exposure assessment, with usage allocated to the cells from the larger MTRS sections based on the proportion of the farmland present. The authors also obtained paper copies of usage reports and identified specific field locations for several of the closest application sites to the monitoring locations (predominantly within 2 miles). Applications in these field polygons were then assigned to the more detailed grid cell array.

By including monitoring data in several high-use counties, the model calibration process incorporated a fairly wide range of climate, soil and application characteristics for California agricultural areas. The monitoring locations included sites in predominantly rural and suburban areas, including several schools, similar to the characteristics of many receptor locations. Several of these characteristics such as proximity, wind speed and direction, and temperature were explicitly included as independent variables in the selected model, which will help account for the effect of variations in them when modeling receptor concentrations. The model development report has been updated to describe these features more completely (Cohen et al. 2011).

D. Reviewers suggested that additional documentation of several important assumptions be provided, including maximum distance to MeBr uses included in the model calibration analysis, and the choice to use a regression-based approach instead of an existing conventional air dispersion model such as ISC, or developing a model based on power law coefficients. Reviewers suggested some established models that might be as or more useful than a newly-developed regression model, and requested better explanation of the physical basis of the regression model components.

"It is presumed in the project framework … that receptor air concentration is limited to a few miles. This has not been demonstrated. The results of the model underestimating lower concentrations suggest that it is not correctly characterizing longer distant sources. Longer distant sources are more important as nearby source applications are more infrequent or lacking altogether."

Response:

The issue of bias for the low concentrations is addressed in the response to comments in section 2.D. below.

"Other precedents of the use of regression modeling have been for conventional pollutants such as ozone. Numerical dispersion modeling and regression modeling have been used to predict air concentration data. Regression modeling works well where extensive air concentration data with meteorological data exist. It may be helpful to reference these approaches and their relative successes."

Response:

The report added several references to published examples of numerical dispersion models and regression models for predicting fumigant air concentrations that were most relevant to this study.

"This approach may ultimately prove to be useful, but at the moment, the justifications are weak for using a regression equation in place of a traditional model. As an example of the lack of an adequate review, the well-known and widely used regression model known as the OBDG (Ocean Breeze - Dry Gulch) regression model described by Nou (1963) is not mentioned.

Nou, V, 1963: Development of the diffusion prediction equation - The Ocean Breeze Dry Gulch Diffusion Program. Vol. II (Eds. D.A. Haugen and J.H. Taylor) AFCRL-63-791 (II) DOC Technical Services, Washington DC, pp 1-21."

"The [Ocean Breeze Dry Gulch] OBDG model is an example of how an equation should be formulated using basic physics assumptions and using unknown power law coefficients which are then best-fit by the regression process. This method would have been more appropriate than the current EPA method where the powers are prescribed before hand rather than being solved."

"Explain why the linear equation format (Y = a + bX) was chosen instead of the multiplicative format (which is linear in logarithms):

$$C - Cb = aZ1^{p1}Z2^{p2}Z3^{p3}\ldots Zi^{pi}\ldots Zn^{pn},$$

such as used in the Nou (1963) formula and in many other similar formulas for atmospheric boundary layer processes. For the OBDG equation, the Zi include basic scientific measures such as

wind speed, downwind distance, and vertical temperature gradient. This can be written in logarithmic form by taking log of both sides:

$$log(C\text{-}Cb) = log(a) + p1log(Z1) + p2log(Z2) + \ldots + pnlog(Zn)$$

By first calculating correlations between (C-Cb) and the various proposed Z1, the less important independent variables can be eliminated."

"*I do not follow the rationale for eliminating this [logarithmic] equation[4]. The statement 'However, this is a very different formulation to the physically-based equations 1, 2, and 3' is incorrect, since all that has been done is the logarithm taken of the same equation. And if there is "zero reported usage", the correct result is obtained, since log (C-Cb) would be minus infinity, implying correctly that C = Cb."*

Response:

The regression model formulations evaluated in this project were all based on physical assumptions, with some simplifications necessary to make the problem tractable. In general, the effects of wind speed or distance were modeled as either linear or quadratic functions of the reciprocal of the wind speed or distance, and the temperature effects were modeled as linear functions of the average number of degrees above 4 ºC. The coefficients of these linear and quadratic functions were solved for, using regression, rather than being prescribed beforehand.

It is certainly possible that the true relationship may involve non-integer powers instead of powers of 1 or 2. However, making the powers an unknown parameter would have created a set of non-linear regression models, which would have required even more computational effort than the selected approach. Furthermore, we believe that a linear or quadratic approximation with two or three terms is often an accurate approximation of a single term raised to an unknown power.

The OBDG regression model approach is based on an equation of the form

(Eqn 1) $$C - Cb = aZ1^{p1}Z2^{p2}Z3^{p3}\ldots Zi^{pi}\ldots Zn^{pn},$$

where C = concentration, Cb = background concentration, the Zi are scientific measures such as wind speed, downwind distance, and vertical temperature gradient, and a and pi are unknown parameters. Cb could also be treated as an unknown parameter. If this equation applies, then we can rewrite the equation in logarithmic form by taking logarithms of both sides:

(Eqn 2) $$\log(C\text{-}Cb) = \log(a) + p1\log(Z1) + p2\log(Z2) + \ldots + pn\log(Zn),$$

and the unknown parameters a and pi can be easily estimated by a linear regression.

Unfortunately, the OBDG equation (1), can only be applied to situations with a single emission source. For this assessment, we have multiple sources impacting each receptor site, where each source is characterized as a combination of usage grid cell and application date. Each of the multiple sources can have a different usage amount, distance from the receptor, and direction from the receptor. Each source also releases a proportion of its total application over several days, which changes day by day.

To apply the OBDG equation, we would have to replace the right hand side of Equation 1 with a sum of products, one product for each source, where each product had the form of the right hand side of equation 1. The sum would have a large number of terms, since potentially each of the

approximately 3000 quarter-mile grid cells within an 8-mile diameter could be included. And each day of application within a grid cell would also be a separate source, since the daily emission rate depends upon the number of days since the application. As a result, the regression equation would have far more parameters than the number of measurements, so that the unknown parameters could not be uniquely estimated by regression.

Moreover, since the OBDG equation must relate the concentration to the usage, one of the *Zi* terms would have to be the MeBr usage. Thus, most of the product terms would be zero, since in practice a given measurement is impacted by emissions from relatively few of the grid cells within 8 miles of each receptor point. Thus, the relevant data available to estimate the parameters for a given grid cell and emissions day are quite limited, making the model fit very poor. An approach to address the sparse matrix problem might be to combine grid cells and/or emissions days by reducing the geographical and/or temporal resolution of the usage data. However, such an approach would likely lead to reduced model performance since the directions and distances to the emissions sources would be much less precise. In summary, application of the OBDG approach to this situation would require either using a high spatial and/or temporal resolution that would result in a poor model fit, or reducing the geographic and/or temporal resolution, leading to poorer model performance.

Furthermore, for this application equation 1 cannot be simplified to the logarithmic version in equation 2, as described below. Therefore, the non-linear regression model given by equation 1 would have to be fitted instead of the simpler linear form in equation 2.

There are two reasons for this. First, since the OBDG equation must relate the concentration to the usage, one of the *Zi* terms would have to be the MeBr usage. Since the observed usage is very frequently zero, equation 2 cannot be used because the logarithm of zero is undefined. If some minimum possible non-zero usage were substituted for the zero usage variables in order to avoid this problem, the results would be highly sensitive to the choice of minimum value. Second, because there are multiple sources, the logarithmic transformation would create a term of the form

$$\log(aZ1^{p1}Z2^{p2}Z3^{p3}\ldots Zi^{pi}\ldots Zn^{pn} + bY1^{q1}Y2^{q2}Y3^{q3}\ldots Yi^{q}i\ldots Ym^{qm})$$

which cannot be simplified to the form of equation 2.

In the modeling documentation, Equation 3

$$\text{Log(Conc)} = \log(\text{Usage}) + \log(a(\text{distance})) + \log(b(k)) + \log(c(\text{wind direction})) + \log(d(w))$$

is equivalent to Equation 4

$$\text{Log(Conc)} = \text{Intercept} + A \times \log(\text{Usage}) + B \times \log(\text{distance}) + \log(b(k)) + C \times \log(\max(\cos(\text{angle}, 0)) + D \times \log(w)$$

only if the Intercept = 0 and the values of *A*, *B*, *C* and *D* are each equal to 1. For example, taking anti-logarithms of Equation 4 implies that the concentration is proportional to the Usage raised to some power *A*, instead of the Usage, so that doubling the usage multiplies the concentration by 2^A instead of doubling the concentration. This is a very different physical formulation when *A* is not equal to 1.

Thus the conversion is not simply a matter of taking logarithms of the same equation. Introducing the Intercept term into equation 4 implies that the concentration is represented by exp(Intercept) multiplied by the exponent of the remaining terms, and exp(Intercept) has no counterpart in equations 1 to 3.

The last part of the comment suggests that the reviewer is considering an equation of the form

$$\text{Log(Conc – background)} = A \times \log(\text{Usage}) + B \times \log(\text{distance}) + \log(b(k)) + C \times \log(\max(\cos(\text{angle}, 0)) + D \times \log(w)$$

instead of equation 4. If so, the issue about exp(Intercept) is not a problem but the issue about the non-unity powers *A*, *B*, *C*, and *D* still applies. For this equation, setting Usage = 0 would indeed lead to an equality if Conc = background, but this equation could not be used to fit the regression to all the data since the log(Usage) value is undefined for days with zero usage. Furthermore, the equation would be physically incorrect on days with zero usage because the observed concentrations are not uniform on those days.

In summary, there were several major reasons why the OBDG model approach is not feasible for this analysis. These include: 1) the difficulties in defining and applying non-linear regression models; 2) the difficulties in applying this approach to the numerous possible sources impacting each receptor; and 3) the reduction in geographic specificity, and likely model predictive accuracy, that would be required to adapt to these constraints.

"[T]he rationale for eliminating the ISC model from consideration is weak, and the authors did not survey the literature to identify simpler but scientifically correct area source models. For example, the ATDL area-source dispersion model was specifically designed for adjacent grids of area sources, which is the problem of interest in the current document. That reference is: Hanna, Briggs, and Hosker, 1982: Handbook of Atmospheric Diffusion. DOE/TIC-11223, pp 59-60."

"[The section entitled] 'Evaluation of the Possible Use of ISC Dispersion Model' is very weak, since most of the rationale is faulty. For example, it is stated that "Application of the ISC model requires specification of the pollutant emission rate from each source …. , the actual emission rates are uncertain." Yet the same problem occurs for the regression model, and the authors went ahead with the model anyway and calibrated the area source strength. A similar calibration procedure could have been used for the ISC model, which is indeed appropriate for the multiple source emissions scenarios and receptors. Also, other science-based simplified area source dispersion models exist (e.g., the ATDL area-source model) and would have been revealed by a thorough literature review."

"One reason stated for not using ISC is that the source terms are not known. Yet the authors then proceed to develop measures of the source terms from available usage data for use in the regression equations. This same procedure could have been used in ISC, and then the source magnitude calibrated using the observed concentrations."

Response:

The emission rate for a particular time period (e.g., daily, hourly) is a function of (1) the fraction of the total applied MeBr that is volatilized as opposed to remaining or decaying in the soil, and (2) the temporal profile of the volatilization over a period of several days (i.e., the *relative* daily emissions generated by a MeBr application). Specification of an emission rate is not required for the regression model. Instead, the MeBr daily application rates are specified. For some of the candidate regression models we also specified the daily proportion of emissions. For other candidate regression models the regression process itself was used to estimate the proportion of usage emitted that day. In neither case did we have to specify the actual (absolute) emission rates.

In contrast, the ISC, ATDL, and other air dispersion models require specification of actual emission rates, usually in mass per time per unit area, i.e., they require an estimate of both the fraction of applied MeBr that was volatilized and the temporal profile. Furthermore, for ISC the temporal profile

must be specified on an hourly, rather than a daily, basis. Estimating the emission rates by calibration of ISC runs would be extremely challenging, since each receptor concentration is influenced by volatilization from multiple fumigant application sites at various proximities and directions from the receptor; and the concentration for a given hour is influenced by volatilization from nearby applications which occur on both on the current day and from several prior days' usage.

"The multiple source-receptor scenario described is not a reason for not using ISC. This scenario is exactly what ISC is intended for and many examples exist of this type of application (e.g., the current Houston modeling exercise being carried out by EPA/OAQPS)."

"The difficulties of using the alternative method of numerical dispersion models are stated, but are not altogether convincing, since they many of the difficulties can be overcome with greater effort, albeit with some limitations. Unfortunately, the best way to evaluate the two approaches is to do an inter-comparison study of the construction and application of the two approaches, but this is probably beyond the scope of this study."

"The document states the numerical models requirement of computer time to be a disadvantage, especially from multiple sources. This however should not be a criterion to reject this approach, as it is a difficulty that can be addressed by running in parallel by banks of PCs or on a workstation or supercomputer. There are many models out there that use more and less computer time, depending upon complications. Another objection to the use of numerical models is the uncertainty of precise location and emission profile. This is indeed a difficulty for all modeling attempts, especially to derive parameters from monitoring data. Numerical models however can be employed using emission profiles derived from experimental data, or emissions can be modeled separately as in input to a dispersion model "

"Advances in PC processor speed and the "upcoming release of 64 bit processors" limit the computational advantage of regression models compared to "process-based" simulation models such as ISC. A more acceptable justification for using regression models would involve developing a case study comparison of the performance of a regression model and a process-based model in predicting air concentrations."

"This reviewer is not convinced by the series of weak justifications for not using ISC. If the authors are going to argue "too many sources and too many receptors", then they should back up their qualitative statements with specific quantitative numbers concerning ISC model run times. For example, they should show that it takes 11 minutes to carry out an ISC run for a single source and n receptors, etc. If the total run time for all sources and receptors exceeds, say, one week, then I might buy their argument. But I have been involved in many ISC projects where there have been thousands of sources and receptors and ISC run times of 1 or 2 weeks are acceptable"

"It is stated that 'After reviewing … models such as ISC, …we determined to pursue development of an enhanced version of a regression model …'. However, the rationale given to back up this statement is relatively weak. A much stronger case is needed. If a permit application came to the EPA from an industry, the applicant would not be allowed to proceed without providing extensive justifications and back-up data."

"It is possible that one or more of [these models] may be useful. However, the data used to develop the regression models are still quite limited and one is not sure that a complete range of conditions has been sampled. For example, are there scenarios anticipated that are outside of the range of the data used for development? A critical test would be to compare the regression model concentration outputs with the standard (e.g., ISC) model outputs for some well-defined standard scenarios."

Response:

The issue of the range of applicability of the regression models is addressed in the response to comments in section 1.C.

We agree that the multiple source-receptor scenario is not a reason per se for not using ISC, but the size and scope of this study (much of the state of California) limits the feasibility of this type of model for this application.

To test the initial rationale for using a regression approach rather than ISC, which was based on professional judgment, the authors conducted two investigations. For the first, a limited study of one receptor location was performed using the ISC3 dispersion model, and the results compared to those from one of the candidate regression models. Findings from that comparison are presented in the revised model development report (Cohen et al. 2011, Attachment 2). They suggest that the regression model is likely to provide concentration predictions that are at least as accurate as air dispersion modeling for this application, given the uncertainty of the emission profile for soil fumigation with a variety of possible methods.

We also investigated the computational resources that would be required to apply the ISC model for this study, and presented the findings in the revised report: (Cohen et al. 2011, Attachment 1). The findings suggest that on a 2.2 GHz Intel Celeron system, it would take anywhere from 7.5 CPU-months to 19 CPU-months, depending on the average density of surrounding grid-cell emission sources, to simulate MeBr concentrations at each of the approximately 8,400 public schools in California over a 7-year period. There would also be considerable labor resource requirements for input preparation, file handling, and output processing. Use of 64-bit and multi-core processors will not appreciably improve such times, since the model code must be optimized for such features, and used with a 64-bit operating system. Since ISCST3 was created as 16-bit code, it requires recompilation as a 32-bit application even to successfully execute within the Windows 64-bit operating system, but is clearly not likely to benefit from it without code modification. Similarly, use of multiple cores would require parallel execution of model calculations, which would require significant updating of the original source code.

Further, the intended application for this model is not to assess a facility permit approval. While such model applications are subject to the provisions of EPA modeling guidance and model applicability determinations, there is no approved model for this application. The only comparable modeling precedent for this problem is the set of CDPR-developed models (described in detail in this study's model development report), which were developed using a similar regression approach.

E. Reviewers suggested that including additional information concerning the air monitoring data collection procedures and methods would be useful.

"The document needs revision to clarify its descriptions, and to better distinguish activities conducted under this project from previous work. More information is needed about the physical characteristics of monitoring sites. More detailed description of the air monitoring data collection and analysis methods, and to what extent they are consistent across sites and monitoring periods."

"List the ambient monitoring concentration thresholds and the accuracy of the concentration monitors. Discuss backgrounds."

"The information on detection limits and other details is very important and should receive more discussion here and in previous sections. A summary table is needed."

"A weakness was the limited description of monitoring data measurement methods to assure consistency."

Response:

In the updated report (Cohen et al. 2011) we have clarified the discussion to distinguish between previous CDPR model development work and this project. The updated report also includes additional discussion of background concentrations observed in several monitoring situations in section III.E. In addition to the expanded discussion, the California Air Resources Board (CARB) and CDPR reports cited in the document provide a detailed discussion of monitoring methods, including detection limits and quantitation limits, which are also noted in the revised report section III.A.1.

In general, the three counties in which the ambient MeBr measurements were conducted represent locations with a relatively high number of MeBr applications, and thus relatively high ambient concentrations. However, the daily ambient concentrations were low even in these locations during substantial portions of the monitoring studies used in the development of the regression model. For example, at stations in Monterey and Santa Cruz counties, daily average concentration ranged from 0.06 to 30.77 ppb in 2000 and from 0.07 to 36.65 ppb in 2001, with the majority of measured concentrations below 1 ppb in both years. Daily concentrations in Kern County were somewhat less in 2000 and 2001, ranging between 0.001 and 14.17, and 0.02 and 25.33, respectively.

Statewide monitoring conducted by CARB since 2002, which is located primarily in urban population centers, has shown much lower MeBr concentrations with a maximum observed daily value of slightly less than 1 ppb for Simi Valley, CA (Ventura County), while the statewide average was only 0.015 ppb (which is the value substituted when measurements were below detection limits of that project's monitoring methodology). These locations are expected to be characteristic of many target application receptors that receive minimal exposure. As noted, separate performance measures for very low concentrations were also included in the regression model calibration ranking.

However, because of the difficulty of specifying the varying background concentration of MeBr, background concentrations are not included in the regression model, and thus it has a tendency to underpredict these very low concentrations. But since these very low concentrations (less than 0.01 ppb) are not much different than zero, and are not of health concern, underpredicting them is not likely to affect the outcome of the analysis.

More generally fractional bias is discussed in section V.B.10 and V B.11 of the revised report (Cohen and Rosenbaum (2011). The results suggest that for the portion of the daily average monitoring data below the median (0.3080 ppb daily average) fractional bias shows a moderate overestimate (~ 0.50) for the top performing models, including the selected model. The fractional bias for the 7/8 week average concentrations below the median (0.6844 ppb) shows a moderate underestimate (~ -0.50).

2. What are the overall strengths and weaknesses of the model development process as described?

Strengths

"Overall strength of the modeling approach is its relative simplicity. "

"The base regression approach was informed by the experience with the CDPR regression model. Improvements were attempted to be made by the inclusion of parameters in the regression equation relevant to dispersion and emission processes. Difficulties in creating separate regression terms

representative of the processes with interpretable coefficients were acknowledged. A creative approach for developing regression terms was devised. An attempt to mirror the Gaussian dispersion equation was made. . . .The longer period average air concentrations were made from daily predictions. There was acknowledgment that all models have difficulties on smaller time scales."

"Tens of thousands of alternative formulations were explored. In many respects, this was done somewhat objectively from a statistical standpoint, with all permutations formulated and run. This approach has the potential of discovering a formulation that one might not have thought would be effective. At its best, it could discover the relative importance of physical processes and conditions that determine pollutant air concentration, highs and lows."

Weaknesses

A. Reviewers identified some general limitations of the regression approach and model evaluation process.

"The main weakness is the empirical nature of the regression models, which limit their utility at locations other than the monitoring sites."

Response:

We agree that a general weakness of regression models is their empirical nature. However, for this study there are sufficient monitoring data for reasonably reliable regression model development. Further, the monitoring dataset includes most of the situations to which we foresee applying the model, such as days with several, few or no nearby applications, and a variety of wind conditions.

"The main weaknesses are that the regression model formulation does not allow for solution of the power law coefficients associated with the independent inputs, does not employ knowledge of standard scientific formulations, does not adequately justify eliminating the logarithmic formulation, and the use of thousands of "models" is irrational. In this reviewer's experience, the regression equation can be solved so that a best fit is obtained for multiple inputs and parameters, without having to independently list and test each combination of inputs."

"The sheer amount of regression model formulations makes the evaluation and ranking the models very difficult to comprehend, manage and evaluate. Although this approach has the potential to discover relationships of physical factors and their relative importance to determine the distribution of air concentration, this was not fully realized."

"The ranking methodology is the most confusing aspect of the report. As mentioned above, this reviewer thought that one of the purposes of a multiple regression approach was to have the statistical procedure and software identify the best performing set of parameters and coefficients. Instead, the EPA authors seem to have included a convoluted intermediate step where thousands of possible combinations of parameters and coefficients are selected and denoted as "models", which are subsequently ranked. I can see possibly picking five to ten alternate basic model groupings that fall within a rationale set of criteria, but not thousands of them."

Response:

The issue of the number of model formulations considered was previously addressed in the response to comments in section 1.B. Given the history of model development of this type by CDPR (described in the report), in which simple models that did not incorporate widely used physical input parameters were evaluated to perform relatively well for the longer averaging periods examined, the

authors were not in a position to determine *a priori* which combinations of model features would be the best predictors. During the model development process, several sets of alternative formulations were considered and discarded if they did not provide identifiable improvements in performance.

The infeasibility of using a power law approach for this application is discussed in the response to comments in section 1.D.

Due to the huge variety of potentially plausible models, the large number of possible input variables makes it infeasible to apply a standard stepwise regression approach whereby a single regression model is fitted with some of the terms, and we either add or subtract candidate terms until the best model is found. Most of the resulting models would not be consistent with known physical principles, e.g., including terms for only selected days after the MeBr application.

Furthermore, a stepwise regression approach requires application of a single model performance measure. We judged that there are multiple important measures of model performance for each regression model, so that several criteria were used to rank the models, as described in the documentation. These include evaluations of performance for various exposure averaging periods, low as well as complete sets of ambient concentration ranges, subsets of the data by year and location, and several statistical measures of both mean and near-maximum over- and under-estimates. This combination of ranking factors provides better assurance that the most successful models are likely to provide adequate predictive power under a range of conditions.

The final model selected, as well as almost all of the best performing models, includes a number of important independent variables which are commonly included in air dispersion models. We consider this formulation an appreciable improvement over the situation prior to this development, in which the relatively successful regression models identified for predicting MeBr concentrations with these monitoring data varied significantly in formulation between the two years of monitoring data and did not adequately predict values in both years combined, did not include weather conditions and did not incorporate any discounting of usage contributions by distance from the receptor.

B. Reviewers identified several issues concerning how physical principles of atmospheric dispersion were incorporated in model formulations that were felt to be questionable or which needed additional explanation.

"Regression equation formulation - Despite the statement after the bullets that "the basic model underlying the many alternative models examined is formulated based on known physical principles of air dispersion", the authors have missed several important physical principles. For example, it is well known that the power on the distance (x) term varies with distance from an area source (the power is less near the edge of the area source and approaches about 1.5 at large distances). Also, the wind direction term should be related to the expected plume width."

"The wind direction coefficient ignores the physics-based knowledge of lateral dispersion from area sources. The term should be a function of (distance/area source size) and (plume width/distance)."

Response:

This model development process evaluated a number of model formulations which featured the inverse of the distance between source and receptor raised to a power of either 1 or 2, and compared the performance among these and other formulations. In the study by Nou (1963) (Ocean Breeze and Dry Gulch Diffusion Programs), cited by the reviewer, a regression approach was developed for predicting concentrations downwind of a source in rural conditions. The findings of the study

suggested that concentration varied inversely with distance raised to the 1.96 power, i.e., closer to 2 than to 1.5.

Additionally, an analysis of the Gaussian model formulation, as used in ISC3, shows that for rural conditions and D-stability the downwind distance has an exponent of 1.81 in the range of 0.3 to 1 km downwind, and becomes slightly lower at farther distances (1.6 at 3 to10 km downwind). For stabilities A, B and C the exponent ranges from 1.9 to 3, and for E and F stability (1.5 to 1.75). Overall, we feel that consideration of the distance formulations included should adequately capture the range of likely alternatives, and avoid further expanding the number of alternative formulations to be compared and ranked.

For these regression models, the wind direction term (actually the wind vector) approximates the crosswind exponential function term from the Gaussian formulation. In the regression model this is represented as Cosine(θ), where θ is the angle between the wind vector and the source-to-receptor vector.

In the Gaussian formulation the crosswind term is

$$\exp\left[-0.5\left(\frac{y}{\sigma_y}\right)^2\right]$$

where y is the crosswind distance from source to receptor.

When the wind vector is unaligned with the source-to-receptor vector,θ is large and cosine(θ) is small. Similarly, y is large and the crosswind distance term is small. Conversely, when the wind vector and the source-to-receptor vector are aligned, θ is small and cosine(θ) is large (e.g. cosine(0)=1), while y is small and the crosswind term is large (e.g., exp(0)=1).

Also, note that the "adjusted" usage, which is effectively a surrogate for emissions, is allocated to uniformly-sized grid squares. Since the size of the "area sources" is therefore uniform, the area size need not be considered in the regression model. This usage is allocated to each grid cell's center point, which is the location used to calculate the proximity distance.

"Attachment 7 on Gaussian Equation - This equation is for a point source. The area source equation should be used for the current application."

Response:

The Gaussian point source formulation is being used only as a guide for what factors should be considered as candidates for inclusion in the regression model. These factors are the same for both the point source and area source equations.

"Attachment 7: The u_s term cannot be the "wind speed at release height", since this is a ground level area source. Actually u_s is usually taken to be the speed at the NWS standard anemometer height (sometimes 3 m and sometimes 10 m). These σ_y and σ_z formulas do not appear to be from the ISC equation. What is the reference?"

Response:

The u_s term as applied here is either the 2-m CIMIS anemometer height or the typical 10-m NWS anemometer height, depending upon the meteorological station used. In most cases priority is given to the CIMIS data.

The σ_y and σ_z formulas are equations that approximately fit the Pasquill-Gifford curves for the rural mode. These are given as equations 1-32, 1-33 and 1-34 in the ISC3 User's Guide, Volume II (EPA, 1995). This reference information has been included in the modified model development documentation.

"A weakness of any regression approach is that it implicitly assumes that the basic scientific phenomena are linear. However, most atmospheric phenomena are non linear and sometimes even switch signs as the independent variable increases."

Response:

The regression approach used here is based on the assumption that the output is a linear function of the unknown parameters, but the unknown parameters can be constructed to include non-linear effects. For example, in some of our models we assume that the effect of distance is a quadratic function of the reciprocal of the form

$$a + b*(\text{distance}^{-1}) + c*(\text{distance}^{-1})^2.$$

This is a non-linear relationship, but the regression approach applies since the expression is linear in the coefficients *a*, *b*, and *c*. More generally, our approach is based on the assumption that the concentration is a linear combination of various adjusted usage variables, but the adjusted usage variables are often non-linear functions of the data. The primary linearity assumption made throughout our modeling is that for a given set of meteorological and usage conditions, the concentration is linear in the usage. This is the same linearity assumption incorporated into EPA's ISC regulatory dispersion model and generally into Gaussian dispersion models. It implies that if all the reported usage amounts were multiplied by *X* and the meteorology was unchanged, then the concentration would also be multiplied by *X*.

"Many of the variables could have been weeded out by doing a correlation analysis or a sensitivity analysis initially. Also, how do the authors know that a different formulation (e.g., the Nou approach with power laws) would not perform better?"

Response:

A correlation analysis can only examine the relationship between a single independent variable and the concentration and will therefore ignore the potential for synergies between independent variables. In such cases, a combination of two or more independent variables provides a better estimate of the concentration than might have been expected from the individual correlations. Moreover, for this application, the evaluation of single terms in isolation may not be helpful. For example, it is inconsistent with physical principles to have a term for the usage only on a single day *r* prior to the concentration measurement, since consideration of the usage on later days up to the measurement day will also contribute to the measured concentration. Ranking of the initial unconstrained regression modeling in rounds 1 and 2, as described in the documentation, was designed to weed out unimportant variables and combinations of variables.

Of course, we cannot know for certain that our set of formulations include the best possible ones based on our selection criteria. However, the fact that we fitted a large number and large variety of physically reasonable model formulations and did not find large differences in overall performance

among the best-performing models makes it less likely that there is another formulation that would perform substantially better than these models.

We have discussed the Nou approach in our response to section 1.D. above, and concluded that the approach is infeasible for this application because of the problem of multiple emissions sources.

"The statistical model performance package was intended for application to field data sets that are independent. If a regression equation has been best fit to a data set, then it is hardly an independent test. For example, I once fit a similarity formula to the Prairie Grass field data, and it then agreed much better with those same data than a Gaussian plume model such as the ISC model, which had not been tuned to those data."

"It should be stated that a true regression model test would involve independent data. The degree of independence of the cross-validation data should be discussed. Also, the effects of the "enhancements" to the emissions data should be discussed."

Our analyses addressed this important issue using several types of cross-validation. The cross-validation approaches were a vital part of our model performance evaluation and are an effective way to address the performance of a regression model on a different dataset to the one used to calibrate the regression. In this approach, we fitted the regression model formulation to a subset of the data based on both monitoring site and collection year, and then applied the fitted model to predict the concentrations for the remaining "independent" data. We then calculated the model formulation's performance statistics on the independent data set, and included these statistics in the model rankings.

A very similar cross-validation approach was used in the cited original Nou (1963) paper with the Prairie Grass field data. In their case, they randomly divided the data into two halves (stratifying by the delta T value) and used one half to predict the other half. In our analyses, we divided the data into the two year groups and used one half to predict the other half for the cross-validation by year. We also divided the data into the two groups with and without monitoring site X and used the group without site X to predict the group with site X. Rather than doing a separate analysis for each of the sites, we combined all the 14 sets of site X predictions to come up with goodness-of-fit summary statistics. Although the methods are similar, our approach with fixed rather than random subdivisions of the data was a more severe test of the model since if there is large variation across sites or years, then it will be hard for a model to predict the values for the excluded site or year from the remaining data, but it would be much easier to predict one random half of the data from the other random half. Our cross-validation test was designed to account for our intended use of the model at different sites and times.

The response to the comment concerning "enhancements" to the usage data is included in our response to comments in section 3.A.below.

"Using meteorological data averaging periods of one day has some disadvantages. The Gaussian dispersion equation is meant for one averaging period congruent with the time period of meteorological and measurement data. One to two hours averaging period are preferred. Greater averaging periods distort the realities of wind direction and variation of wind speed."

Response:

A daily vector average wind was calculated to correspond to the 24-hour measurement period for MeBr concentrations. This approach is somewhat analogous to the approach used in the ISC3LT model, which uses a joint frequency distribution to estimate concentrations for long-term averaging

periods. Many models, including the final selected one, split days into separate day and night averaging periods, which further reduced this time.

C. One reviewer identified several issues involving statistical significance testing of model coefficients for the adjusted usage variables.

"The models generated many insignificant regression terms, many with negative coefficients that defy physical reality. When a regression term coefficient has a distribution above and below zero, that is, it is statistically insignificant, that term should not generally be used. Forcing negative coefficients to be positive, was a worthy exercise, but does not give confidence to the results."

"Evaluation of the statistical significance of regression terms might be better used to construct a set of candidate models."

"The presence of statistically insignificant variables should be a factor in ranking. In the document, these variables were not identified. The problem was identified when it was stated that some of the coefficients were very low, and even negative. A comprehensive analysis of this problem needs to be done. EPA provided, in separate correspondence, statistical significance of the coefficients for some of the models. This should be in the report, along with similar data for a larger set of models. Generally, statistically insignificant variables are dropped. The document should provide a better rationale for their inclusion."

"Models that did not have insignificant variables need to be identified as candidates. They are likely to be improvements over the simple CDPR model, and more applicable to regions outside the three monitored counties."

"The use of a nonstandard regression approach that does not break down the independent variables with their own coefficients and the presence of insignificant variables in the models, which are not well documented, makes the evaluation of the models presented here more difficult to evaluate at this stage of development. "

Response:

The use of a nonstandard regression approach that does not break down the independent variables with their own coefficients was a consequence of the problem of having a very large number of intermittent emissions "sources" from different directions and distances. The standard regression approach could not be applied.

In Round 1 of the regression model set development, the coefficients of the regression models were unconstrained and therefore could have positive and negative values. In some cases a negative coefficient can make physical sense, provided that the predicted concentrations are not negative for the expected usage and meteorological patterns. Often a negative coefficient would occur for a daytime coefficient when the corresponding nighttime coefficient is positive, or vice versa, so that the physical model can be interpreted as a cancellation of opposing effects.

However, in order to make the resulting models more physically plausible, and hence be more likely to apply in different locations and time periods, the Round 2 model development constrained the coefficients to be non-negative so that all the predictions would be positive or zero. More precisely, we estimated the regression coefficients by minimizing the sum of square errors subject to various

constraints, including the constraint that the values are not negative.[1] We agree that these constraints by themselves do not guarantee good results. However, the selection of the best models based on various performance measures, including the cross-validation analyses, provides confidence that the selected models provide an adequate approximation to reality.

Although it is often statistical practice to remove non-significant terms from a statistical model, such an approach is not required. If a coefficient is not significant, this simply means that we are confident (typically at the 95 % level) that the true value is within an interval that includes zero. This does not imply that the true value equals zero, only that the true value could be zero. The regression estimate of the coefficient is the best estimate of that true value, in the sense of least squares.[2]

Excluding the non-significant terms and refitting the model will increase the mean square error and can reduce the predictive power of the model in a more general sense. Furthermore, excluding certain terms can lead to physically unrealistic models. For example, it is quite possible that the term for the usage on day *r* after application is not statistically significant, but the terms for other days are significant, which would produce a model that ignores the impact of usage *r* days after the application if the non-significant terms were replaced by zeroes.[3]

But we agree that all other factors being equal, models with only significant coefficients (i.e., those whose distributions were entirely above zero) would be preferable as long as they are physically plausible. Thus, the final model selection was based, in part, on consideration of which models included only significant coefficients. The final model selected was one with all significant coefficients, since the performance of this model, as measured by multiple criteria, was not appreciably inferior to other highly-ranked models that included insignificant coefficients (i.e., those with confidence intervals that included zero).

D. One reviewer expressed concerns about regression model intercepts

"Forcing the intercept to zero, necessarily introduces a downward bias for low concentration predictions (as the slope pivots to reach zero), and neglects background sources, hence mis-specifies the model. Ordinarily the intercept is forced to zero when that is what we expect from our knowledge

[1] In Round 2 of the model development, some of the coefficients actually turned out to be extremely small, e.g. less than $1x10^{-10}$, but negative, and insignificant, due to the numerical precision of the computations used to apply the non-negativity constraints. Due to computer rounding, coefficient estimates with absolute value < 1.0E-10 can be treated as being equivalent to zero

[2] Subsequent analysis of the calculated residuals for the selected model (which are the best estimate of the errors) revealed that that were not normally distributed and were heteroscedastic. Statistical theory suggests that

- If the errors are heteroscedastic, the standard method for calculating the uncertainty of model coefficients assumed to have homoscedastic errors will in most cases yield an underestimate
- If the errors are heteroscedastic, less uncertain model coefficients can be estimated by accounting for the heteroscedasticity of the errors in the regression analysis, if we know its form.

Note, however, that even if the errors are heteroscedastic, the model coefficients estimated by assuming homoscledastic errors are still approximately unbiased. i.e., the coefficients are as accurate, even if more uncertain, as would be heteroscedastic estimates, given a large sample.These issues are discussed and addressed in Section V.B.11 of the updated document.

[3] A positive non-significant coefficient for a same-day usage term, paired with a positive significant coefficient for a related previous-day usage term would suggest that there is more uncertainty about the influence on concentrations of same day usage than about the influence of previous days' usage. For an individual measurement this could be the case due to differences in usage levels and meteorological conditions from day to day, but it seems unlikely to be the case for the full set of 659 daily measurements used for the regression analysis. It is likely that such results are an anomaly of regression analysis applied to a formulation with explanatory variables that are somewhat correlated.

of the phenomenon. In this case, there is a rationale, because there is an attempt to model background. However, background comes from a greater distance than modeled, so we would not expect a zero background."

"There is the problem that the intercept represents to some extent the unique proximity of the modeled data points to other background sources, or the level of intensity of use in the immediate area, not captured by other regression terms. In any case it is problematic to devise a regression model in this situation that can have an intercept that will be accurate in regions other than what was used to construct its coefficients."

Response:

As noted by the commenter, since the intercept represents the unique proximity of the monitor locations to other background sources, it is very likely to be spatially variable. If we included an intercept in our modeling of the monitoring data, this intercept would represent the typical background concentration in the monitoring region, but could be a very poor estimate of the background concentration in other regions.

Because of the difficulty of specifying the varying background concentration of MeBr, background concentrations are not included in the regression model, and thus it has a tendency to underpredict these very low concentrations. But since these very low concentrations (less than 0.01 ppb) are not much different than zero, and are not of health concern, underpredicting them is not likely to affect the outcome of the analysis

More generally, fractional bias is discussed in section V.B.10 and V.B.11 of the revised report (Cohen et al. 2011). The results suggest that for the portion of the daily average monitoring data below the median (0.3080 ppb daily average) fractional bias shows a moderate overestimate (~ 0.50) for the top performing models, including the selected model. The fractional bias for the 7/8 week average concentrations below the median (0.6844 ppb) shows a moderate underestimate (~ -0.50).

"An additional weakness is that there is no indication that the developers discussed the "model" with the air modeling experts at RTP, either in the Air Modeling Division of NERL or in OAQPS."

Response:

The model development was based in the work of California DPR (see for example, Li, Johnson and Segawa, 2005). In addition, we consulted with EPA's OPP technical staff. The experts at both of these organizations are familiar with the key issues surrounding soil fumigant volatilization and dispersion.

The model developers at ICF have extensive experience in developing air dispersion models for EPA and other government agencies, including ASPEN, REMSAD, HYROAD and the UAM.

Moreover, the regression model development approach is not typically used by air model experts at NERL or OAQPS, so that their expertise for the issues involved in this study may be limited.

3. What are the strengths and weaknesses of the data quality assurance activities conducted during the model development process?

Strengths

"Specific efforts were made to review and assess the relevance of monitoring data, and to account for the impact of one nearby commodity fumigation source."

"Dramatic improvements were made to data inputs of MeBr usage from PUR data by labor intensive means. Meteorological data was analyzed and missing data was replaced by appropriate methods."

"Outliers that most likely arose from data errors were discovered and taken out with convincing analysis. Errors (differences between predicted and observed values) were presented extensively."

"Questions of non-linearity, and normal distribution were addressed by presentation of plots of observed vs predicted observation. A predominance of underestimating low values was presented."

"Data sets were run separately and combined by years, counties, and high vs low, to determine if these factors resulted in biases. This provides more confidence to apply the models to other years and regions."

"The observed concentrations and source usage term are fairly well-defined and have been measured over many weeks during two years."

Weaknesses

A. Reviewers suggested that further exploration of factors that may correlate with model performance would be helpful.

"Physical process and data factors that resulted in the tendency to underestimate, and have difficulty estimating low predicted values, and estimate zero values, when observed values resulted, should have been further explored."

Response:

This is discussed in the responses to comments in section 2.D.

"It would help to have error bars on these observations clearly explained in the text. For example, what is the uncertainty in the magnitude, the location, and the timing of the source information? Also, the rationale for air sampler site placement could be more clearly given (e.g., are the samplers placed in expected "hot spots", or near sensitive locations such as schools, etc.?)."

Response:

Our methodology for estimating usage by grid cell is explained in the database development document (ICF, 2011) and provides qualitative information about the sources of uncertainty in the usage distribution. It would be extremely difficult to provide quantitative error bars for those grid cell usage estimates, since we don't know how accurate the reported usage amounts are in the PUR and our analyses have shown that the accuracy of the reported MTRS sections can also vary. Estimating these uncertainties to calculate "error bars" would be a major effort, outside the resource limitations of this project. Also, it is unlikely to provide useful information without the additional effort of propagating these error estimates throughout the entire modeling process.

The CARB and CDPR reports cited in the document provide a detailed discussion of monitoring methods, and note that sites were generally placed in high-use areas, and in some cases were placed at public buildings, including schools.

"How can the regression equation be used in the future if it is not based on basic input data but instead is based on a data set made up of the basic data plus "enhancements"? These enhanced data will not be available at other sites and times."

Response:

To clarify the discussion, note that there were two types of "enhancements" made to the usage data. The raw usage data from PUR is reported as pounds of usage in a specified MTRS location, which is typically about 1 mile × 1 mile square in size. The first enhancement was to reallocate the usage spatially to ¼ mile × ¼ mile square grid cells containing the farmland portion of each MTRS, and to reallocate unrealistically large usages listed on a single day temporally over two or more days. Both these temporal and spatial enhancements were made to the raw data for all sites and times where the regression model was calibrated and where it was applied.

The second type of enhancement used additional detailed information on usage (a) to add some usage data records from filed reports in Kern County which were not included in the PUR database, and (b) to improve the spatial allocations near monitoring sites by linking them to specific fields and application days where possible. The second type of enhancement was only feasible for a limited amount of data, namely the usage data collected near the monitoring sites during the monitoring period, and in the vicinity of the complaint schools during the modeling study period. The field-specific usage data near the monitors were used to improve the calibration of the candidate regression models, in order to obtain a more accurate estimate of derived model coefficients. The field-specific data near complaint schools were collected to provide enhanced accuracy for modeled exposures there. [4]

To examine this comment, we can simplify the problem by assuming that the usage data used to calibrate the regression model, which has had both types of enhancements, represents the true usage for the monitoring period and regions. When applying the fitted regression model to other sites and times, the independent adjusted usage values, which have only the first type of enhancement (additional spatial precision), will be subject to bias and imprecision. Clearly the predicted concentrations will have a corresponding additional imprecision beyond the uncertainties of the fitted regression model, and possibly bias. The fact that some MeBr applications may be left out of the PUR usage database could introduce bias in the disparate impact estimates, which are the focus of this analysis, only if the omissions are biased with respect to the proximity and demographics of the surrounding receptor populations. Similarly, the lack of secondary enhancements of type (b) (i.e., enhanced spatial reallocation) for the non-complaint schools should contribute only negligibly to bias.

Nevertheless, by using the enhanced data for the regression model calibration, the uncertainty of the regression-derived model coefficients due to the uncertainty of the inputs is reduced, so that the predictions are more accurate than they would otherwise be, even if they are applied using less precise usage data.

B. Reviewers suggested that more information be included concerning procedures used to adjust MeBr usage data, to the monitoring data for one site, and to fill gaps in input meteorological data.

[4] Field-specific maps were collected for methyl bromide usage from Santa Cruz, Monterey, and Ventura Counties. Due to the effort involved in requesting, finding and then digitizing field maps, only maps in high priority areas were requested. The areas that were within 3 miles of a complaint school, or reported greater than 4000 lbs of usage per MTRS and had less than 2% farmland acreage, were submitted to the counties in an effort to collect associated field maps for the usage reports.

"More than the current ten lines should be devoted to this discussion of the 'Correction for Commodity Fumigation Effect.'"

"I hope that there will be more discussion of the "monitoring data that were adjusted to account for the influence of a nearby commodity fumigation chamber".

Response:

In addition to a citation to the report describing this modeling in detail, the following information has been added to the report.

> Emissions from the Crowley fumigation chamber may occur from three different chambers each separately vented to the atmosphere. The flow rate from each stack is 12,900 cfm with a stack temperature of 303 K. and each stack is 40 feet high. The distance from the Crowley fumigation chamber stack to the Parajo Middle School monitoring site ranged from 280 to 284 meters. Emission rates ranged from zero up to a maximum of 22.7 lbs over the standard two-hour venting time period. At most the three stacks vented the maximum amount for two hours each day. During the monitoring period 25 days in 2000 and 24 days in 2001 had concurrent commodity fumigation. However, on only 9 of these 49 days was the wind direction such that the monitored concentration was affected by the commodity fumigation.

The report also includes a table which provides the modeled commodity fumigation daily values which were deducted from the corresponding monitoring daily values near that site.

"Several methods of "improving" the pesticide usage data are discussed. Some are more justified than others. However, more importantly, there is a need for a quantitative estimate of the final uncertainty in the usage data magnitudes and spatial location. Also, the arbitrary corrections compromise the statistical independence assumptions needed for the regression analysis."

Response:

The infeasibility of an uncertainty assessment of the usage data is discussed in the response to comments in section 3.A.

Usage data was corrected and spatially and temporally reallocated using the available information in an objective, scientific manner, not arbitrarily. More details on the procedures used were presented in section 3.A, and the updated report text has been revised to clarify the descriptions of these procedures (Cohen et al. 2011).

We believe the issue of the statistical independence of the data is not an appreciable concern, for several reasons. For the most part, the regression analysis required no statistical assumptions, since most of the analyses did not apply any statistical distribution theory. The regression modeling was based on the principle of least squares, i.e., choose the coefficients to minimize the sum of squared errors. This principle does not require any assumptions to be made about the underlying distributions.

However, if certain distributional assumptions are made, then the least squares principle follows from another statistical principle, the maximum likelihood principle, which chooses the model to maximize the estimated joint probability density. Those distributional assumptions are that the error terms are independent and have normal distributions with a mean of zero and the same variance for every observed value (i.e., the errors are homoscedastic). Those distribution assumptions were used to estimate the uncertainty of the estimated coefficients, and hence their significance. The uncertainty estimates were also used as part of the outlier analyses.

Furthermore, the improvements to the usage data do not necessarily compromise the usual independence assumptions if they are correctly stated as requiring that the error terms for each day are statistically independent, conditionally on the complete set of adjusted usage variables for every day. This statement of the regression assumptions is plausible but not easy to justify. Even without any adjustments to the usage data, the daily values of the "adjusted usage" variables must be statistically correlated since most of the adjusted usage variables are based on the sum of usage across multiple days, and usually across multiple, geographically proximate, grid cells. Nevertheless, the assumptions about the conditional distributions of the errors are quite possible.

Statistical theory suggests that:

- If the errors are not independent, the standard method for calculating the uncertainty of model coefficients assumed to have independent errors will in most applications yield an underestimate because the errors will be positively correlated
- If the errors are not independent, less uncertain model coefficients can be estimated by accounting for the dependence (i.e., covariance matrix) of the errors in the regression analysis, if we know its form and/or have sufficiently precise estimates of the covariances.

Note, however, that even if the errors are not independent, the model coefficients estimated by assuming independent errors are still approximately unbiased. i.e., the coefficients are as accurate, even if more uncertain, as would be estimates that account for dependent errors, given a large sample such as that used here. In addition, the model coefficients estimated by assuming homoscedasticity are also approximately unbiased, even if the errors are actually heteroscedastic. Therefore the regression model estimates based on the principle of least squares provides approximately unbiased estimates of the coefficients even if the assumptions of independence and homoscedasticity are violated.

"Present a justification for the "75% completeness criterion" [for meteorological station data]."

Response:

The documentation has been updated to provide additional information on this issue. Since the usage and concentrations were measured over 24-hour periods, the hourly meteorological data was also averaged over a 24-hour period, but we also averaged the hourly meteorological data over daytime and nighttime periods to allow for separate terms to model the impacts of daytime and nighttime emissions. In each period, we required 75 % completeness for the hourly meteorological data, and combined primary and secondary meteorological station data if necessary in order to meet the completeness criteria. The details are in Attachment 3 of the revised document (Cohen et al. 2011).

The choice of a 75 % completeness criterion was made based on our "engineering judgment" that the errors in using an 18-hour average to estimate a 24-hour average are negligible. The same 75 % completeness requirement is frequently used by EPA in their analyses of air quality data, such as determining compliance with ozone and particulate matter national air quality standards (US EPA, 2001).

"Provide the justification for the arbitrary choice of 0.5 m/s as the 'minimum daily average wind speed'".

Response:

The documentation has been updated to provide additional information on this issue (Cohen et al. 2011, footnote 30). The choice of 0.5 m/s as the minimum wind speed comes from a study done for the California Air Resources Board (CARB, 1998). This study collected continuous 10 Hz and 1-minute wind measurements with conventional and sonic anemometers at four agricultural locations in the Central San Joaquin Valley over a one month time period between December 1995 and January 1996 at a height of 2-meters. The period from 28 December 1995 through 11 January 1996 was an intense stagnation period with no synoptic scale forcing. During this time the minimum hourly vector average wind speed at any of the four locations was 0.5 ms^{-1} and the minimum daily average wind speed was around 1.0 ms^{-1}.

For this study we conservatively used a minimum daily average wind speed of 0.5 ms^{-1}, instead of the 1.0 ms^{-1} used for conventional NWS weather data and measured in the CARB-sponsored study, to allow for the possibility that stronger stagnation periods may occur than the CARB-sponsored study period.

4. What are the strengths and weaknesses of the model ranking elements, and the model ranking process? Can you identify alternative ranking measures that would be likely to present significantly different information about model performance that should be considered in model selection? Would these alternative measures be likely to change the selection process outcomes as described?

Strengths

"The choice of major ranking criteria appears appropriate."

"Regression model R2 and MSE are appropriate statistical measures of model performance and are essential to a ranking process of regression models, but cannot be used alone. Known physical processes of emission and dispersion and interpretability of the model in those terms, is an appropriate guidance to evaluate the application of the model to the real world (conditions in other areas and times). Analysis of the tails of distribution, the 95th percentile, is also appropriate, since the distribution and variance of the predictions, given the inputs, is important to the applicability of the model."

Weaknesses

A. Reviewers identified several areas in the model ranking process that they suggested should be modified or explained better.

.

"Too much focus and importance was given to R2 and MSE in evaluating the models. . .Model specification is the first hurdle to cross. It is important that there is an adequately high R2 and low MSE, but model specification is ultimately more important. . . . More attention to the reasonableness of the model specification should be done. Parameters that will capture differences between regions to be applied, are most important for the ability to generalize and apply the model outside the three counties that were used to generate the models."

Response:

To address the issue of the reasonableness of model specification, we considered a broad set of models, each of which is a potentially plausible approximation to the underlying physical reality. The

candidate models differ in the set of physical factors included and in the level of complexity of the estimated functional form. Given unlimited data, (e.g., soil type, precipitation, cloud cover, in addition to the data used) it may have been reasonable to only use the most detailed model specifications. However, the available data were limited. Furthermore, the best-fitting model balances the need for detailed modeling of the underlying physical process with the increased uncertainty of the parameters in highly parameterized models.

We carefully considered the parameters to be modeled and the possibility of including additional parameters to explain differences between regions that are not adequately captured by differences in the meteorological and MeBr usage distributions. During earlier stages of the model development process, we defined and compared several alternatives for key input parameters, and identified and focused on the ones which consistently performed better as measured by a smaller MSE, overwhelming presence among the combined top-performing models, and other evaluation factors.

Initially in the model development process, we considered the county as an explanatory variable and fitted separate models to each county group. Later, we decided against such an approach, since the statewide application of the regression model would require each location, and possibly each time period / location combination, to be allocated to one of the county groups, and we did not have sufficient data to justify any particular choice. Many variables describe differences between counties (e.g., inland versus coastal, soil type, types of agriculture) and it is not possible from the available three county monitoring data to determine which variable to use and which levels of that variable to group the locations.

Later in the model development process we evaluated the possibility of stratifying the data by the seasonal mean concentrations for each site**,** for example, sites with averages above or below 1 ppb, under the hypothesis that relative usage distributions are different between locations with high and low concentrations. We decided against that approach because the statewide application would then require an additional modeling step to allocate time period / location combinations to these two groups, introducing additional uncertainty. The development of hybrid distance models, with separate combined terms for "near" and "far" usage, also provided more consistently good performance for the two groups of sites, and substantially reduced the differences in best-performing models for each site category. However, we ultimately used some ranking measures for prediction of the subset of concentration values (from any site) below the median for each exposure averaging period, with the intent of insuring better performance for other sites with consistently lower concentration values.

The final model selection process also relied partially on the evaluation of model specifications. Models were examined to insure that their formulations incorporated expected and consistent input variables similar to conventional air dispersion models, such as consideration of wind speed and direction for at least the "near" terms of the hybrid distance models.

"An alternative ranking process incorporating fractional bias might be considered, but it's not clear that such an approach would change the model ranking or selection process appreciably."

Response:

Although we did not incorporate the fractional bias in the overall ranking scheme, we reviewed these data in the final part of the revised model selection process to compare the detailed performance of selected candidate models. The fractional bias statistics for those models are presented in section V.B.10. of the revised report (Cohen et al. 2011). We noted that for all the final candidate models the degree of bias was quite small for the total data set for both daily averages and 7/8 week averages.

The degree of bias was moderate for the portion of the data set below the median for both averaging times, showing overestimates for daily averages and underestimates for 7/8 week averages.

"Adjusted R2 was not reported. Adjusted R2 corrects for the increasing R2 due to the mere addition of another variable (degrees of freedom relative to sample size). When comparing models of different amount of variables with unadjusted R2 it should be documented that the differences are not due to this statistical effect. In the correspondence from EPA it was stated that there were a large enough sample data set, that the difference from R2 and unadjusted R2 was not significant. This needs to be documented in the report."

Response:

As described in the updated report (Cohen et al. 2011), and as discussed in detail below, the usual adjusted *R* squared approach is not appropriate for averaging periods greater than one day, constrained regressions, or cross-validated regressions. For daily values from unconstrained, non-cross-validated regressions, some statisticians might prefer using the adjusted *R* squared statistics, but for many comparisons the adjustment will make very little difference.

Also, the report notes that the *R* squared values were not included in the overall ranking statistics, which instead used the similar rankings based on the mean squared error. We included the *R* squared values primarily as a convenience, to provide readers with a common method to evaluate more easily one major type of the performance measures used.

The *R* squared values reported in the tables were calculated as the squared Pearson correlation coefficient between the observed and predicted values. This approach was used for all *R* squared values, including unconstrained and constrained regressions, regressions with or without an intercept, cross-validated and non-cross-validated analyses, and averaging periods of one day, 2, 4 or 8 weeks. The formula used is as follows:

$$R2 = \{\Sigma\ (\text{Observed} - \text{Mean Observed})\ (\text{Predicted} - \text{Mean Predicted})\}^2 / \{\Sigma\ (\text{Observed} - \text{Mean Observed})^2\ \Sigma\ (\text{Predicted} - \text{Mean Predicted})^2\}$$

Using standard regression calculations, for unconstrained, non-cross-validated regressions with an intercept, it can be shown that:

$$R2 = 1 - SSE/SST,$$

$SSE = \Sigma\ (\text{Predicted} - \text{Observed})^2$, *SSE* = Sum of Squared Errors

$SST = \Sigma\ (\text{Observed} - \text{Mean Observed})^2$, *SST* = Sum of Squares Total.

One possible problem in using *R* squared to compare models is that the *R* squared value cannot decrease if you add terms to the regression model. For this reason, some statisticians have advocated the use of an adjusted *R* squared value that includes a penalty effect for the number of fitted parameters, and therefore will not necessarily increase with added regression terms. For models with an intercept the adjusted *R* squared value is calculated using the formula

$$\text{Adjusted } R2 = 1 - \{SSE/(n-p)\}/\ \{SST/(n-1)\}$$

where n = total number of measurements and p = total number of fitted parameters.

For models without an intercept, these formulae are usually redefined as follows:

$R2\_NOINT = 1 - SSE/SSTU$,

$SSE = \Sigma\ (\text{Predicted} - \text{Observed})^2$,

$SSTU = \Sigma\ (\text{Observed})^2$ = Sum of Squares Total Unadjusted,

Adjusted $R2\_NOINT = 1 - \{SSE/(n\text{-}p)\}/\ \{SSTU/(n)\}$

The adjusted *R* squared value tends to be a less biased estimate of the corresponding population value than the unadjusted *R* squared, but on the other hand it is still biased. Another potential difficulty with the adjusted *R* squared is the possibility of negative values, which are hard to interpret.

For the daily average concentrations from the non-cross-validated, unconstrained regressions with an intercept, the difference between the adjusted and unadjusted *R* squared depends upon *R2*, $n = 659$ days, and the number of parameters, *p*. The difference equals $(1\text{-}R2)\ (p\text{-}1)\ /\ (n\text{-}p)$ and so cannot exceed $(p\text{-}1)/(n\text{-}p)$. For the best-performing top models, *p* is generally no more than 20, giving a difference of no more than 0.03. If the *R* squared values for two models are sufficiently close and the numbers of parameters are sufficiently different, then the relative ranking based on *R* squared could be different from the relative ranking using the adjusted *R* squared. In summary, for this case of daily values from the non-cross-validated, unconstrained regressions with an intercept, there are some reasons to prefer the adjusted *R* squared statistic, but it will very often make very little difference.

For the regressions without an intercept, the bigger difference is between *R2* and *R2_NOINT*, rather than between *R2_NOINT* and adjusted *R2_NOINT*. To allow comparisons between models with and without an intercept, we chose the same *R* squared definition for both cases. Using *R2_NOINT* instead of *R2* does not change the relative ranks of models that both have an intercept or both do not have an intercept, but provides a meaningless comparison between one model with an intercept and another model without an intercept. Arguing as above, for the case of daily average concentrations from the non-cross-validated, unconstrained regressions without an intercept, there are some reasons to prefer the adjusted *R* squared statistic, but it will very often make very little difference.

For the majority of the analyses using an *R* squared statistic, the above rationale for preferring instead the adjusted *R* squared statistic does not apply: For the analyses of 2-, 4-, and 8-week averages, the formulae for the adjusted *R* squared values are not applicable. The justification for the degrees of freedom adjustment factors relies on the fact that each of *n* predicted values is regressed against *p* "independent" regressor variables, but for these longer-term averages, the model predictions were instead derived by averaging the daily predictions.

Similarly, for the cross-validated regressions, it is not at all obvious how to adjust the *R* squared, because the total number of fitted parameters is the number of terms in the model multiplied by the number of subsets cross-validated (e.g., 2 for cross-validation by year since the regression is separately fitted to each year of data). Finally, for the constrained regressions, the true number of fitted parameters depends upon the numbers of constraints as well as the number of terms, and the adjusted *R* squared formulae will not apply.

" There is a precise mathematical equation relating R2 and MSE and it should be given instead of making this vague subjective conclusion. Also, since R2 is dimensionless and MSE has dimensions and is affected by a mean bias, it is not true that R2 and MSE "track closely". The equation would allow this to be seen."

Response:

The relevant mathematical equations for unconstrained, non-cross-validated regressions with an intercept, are

$R2 = 1 - SSE/SST$,

$SSE = \Sigma\ (\text{Predicted} - \text{Observed})^2$, SSE = Sum of Squared Errors

$SST = \Sigma\ (\text{Observed} - \text{Mean Observed})^2$, SST = Sum of Squares Total,

MSE = SSE / n, where n = number of observed values.

As noted in the report (section IV.B.1.), we used the above *MSE* formula rather than dividing by the degrees of freedom, *n-p*. Using these equations, the *R2* and *MSE* relationship is given by

$R2 = 1 - n \times MSE/SST$.

This formula shows that for the daily average concentrations from unconstrained, non-cross-validated regressions with an intercept, the *R* squared value is a simple linear function of the *MSE*, with constant coefficients, since *n* and *SST* are constant across all regression models for the given dataset. In this case, the *MSE* and *R* squared values will track exactly. Although *MSE* has the dimensions of ppb-squared for this model, the multiplicative factor *n* / *SST* in the *R2* equation makes the *R2* dimensionless.

The above derivation of the relationship between *R2* and *MSE* does not apply for averages over a longer period, constrained regressions, cross-validated regressions, or regressions without an intercept. Nevertheless, if the true intercept is small, then these formulae will apply approximately in these more general cases, so that the *MSE* and *R2* values and their ranks will track closely but not exactly.

B. Reviewers also suggested additional documentation and discussion of the monitoring data used for model calibration.

"*The arbitrary 1 ppb line drawn between 'High' and 'Low' [concentration data categories] should be justified. Relate 1 ppb to background."*
"Justify the arbitrary choice of 1 ppb for separating 'high' and 'low'"

Response:

In the model calibration activities described in the report, the "low" values were used in a model ranking criterion for one exposure averaging period. As previously discussed in this section, the subsequently revised ranking process includes criteria for "all" and "low" values in each of the four evaluated exposure averaging periods. The revised report (Cohen et al. 2011) notes that in the final model calibration step, the observed and predicted data for each averaging period were divided into two categories defined as above and below the median value for that averaging period. Tests showed that the performance for "all" and "high" categories were substantially the same, while the included performance measure for "low" values provided an additional test of performance across a range of concentration values. These "low" concentration value performance measures were included in the model ranking methodology along with the "all" results ranking factors.

"The models generally have greater difficulty in the low ambient concentrations, below 1 ppb, and the most, below 0.1 ppb MeBr. Most models will have the greatest difficulty accurately predicting low

concentration levels, since longer distant sources, and a more complex path of atmospheric transport and fate is at work. If it is desired to have greater accuracy in the distribution of concentrations at the low end, numerical dispersions models are likely to provide better results."

"This approach, at this stage, will be less certain estimating air concentrations at receptors arising from more distant source applications. This situation is likely to be important for receptors that are not nearby heavy applications, or nearby applications are infrequent."

Response:

The authors agree with the remarks about the difficulty of modeling very low ambient concentrations, but we believe that the consequences for this application are not serious. This issue is also discussed further in the response to comments in section 2.D.

It is not clear that the available numerical dispersion models would be more accurate for the low concentrations given the limitations of the usage data and other input data, and the uncertainty and variability of background concentrations.

"It's impossible to "conclusively determine which (if any of these) models are actually capable of accurately characterizing the ambient exposure from multiple fumigant sources" because of the empirical nature of regression models, the lack of information on monitoring data collection and analysis, the 'ambiguity on the treatment of emission flux patterns' and other issues"

Response:

The issue of the empirical nature of regression models is discussed in the response to comments in section 2.A. As mentioned in the response to comments in section 1.E , the report references others with detailed information on monitoring data collection and analysis, and now includes an expanded description of key monitoring data characteristics. The issue of the 'ambiguity on the treatment of emission flux patterns' is discussed in the response to comments in section 1.C. ("Application Method").

"There seems to be too much "fiddling" with data (e.g., correcting some data, tossing out other data, eliminating data due to "flow rate deviations" (whatever that means). A statistical regression process relies on independent data. Subsequent estimates of R2, MSE, and evaluations also rely on independence. The authors should discuss the effects of this "data fiddling" on the conclusions."

In the air modeling and monitoring studies that I have done, the rule is used that "all data is innocent until proven guilty". That is, you should not throw out data just because they are outliers. A specific reason based on facts and reviewed by an advisory team is needed.

Response:

Other than the deletion of a few regression model prediction outliers (concentration days that no or virtually no models successfully predicted), and a site location correction, the corrections and modifications made to the air monitoring data were all taken from the CARB reports, since the CARB was responsible for collecting and analyzing the monitoring data.

CARB corrections: The corrections to (*i.e.*, omission of) nine daily measurements from Kern County in 2001 were CARB's corrections of tabulated values in their Monitoring Report cited in the model development documentation. The monitoring process uses initially empty vacuum canisters which continually draw in air during the collection period. In order to provide a reliable average values, the

flow rates should be as constant as possible. The CARB flagged some data as having excessive flow rate deviations if the ending flow rate measured at the end of sampling differed from the starting flow rate (3.0 sccpm) by more than 25 %. This criterion was recommended by reviewers in the CARB Quality Management Branch to ensure representativeness of the integrated 24-hour samples. These cases were treated as invalid samples by the CARB and so we excluded them from our analyses. These corrections to the monitoring data have no impact on the statistical independence of the data.

Regression outliers: The four reported concentrations that were deemed to be outliers were very high (from 14.6 to 36.6 ppb). The great majority of the best-fitting regression models significantly underestimated these concentrations. This suggests that either (a) the monitored values were somehow erroneous, (b) some nearby usage was unreported or under-reported, or (c) the relationship between concentrations and the usage and meteorology for these four values is very different from the relationship on other days, for example, much higher volatilization or atypical local weather. If either (a) or (b) were the case, excluding these concentrations would be justified on the basis of erroneous data, either monitoring or usage data. If (c) is the case, excluding these concentrations would be justified because we did not have sufficient information to model a different relationship on those four days and, even if we did, we would not know how to find similar "outlier" days for the statewide application. Excluding them also prevented the coefficient regression development process, which took into account the magnitude of such errors, from over-compensating for what were expected to be atypical conditions.

One could speculate that, although CDPR's 2010 modeling used seasonal averages rather than daily averages, CDPR might have obtained different seasonal averages had they also excluded these high daily concentration days from their analysis, and these revised seasonal averages might have produced different regression models. If daily values with high concentrations but relatively low usage values had been excluded, the resulting seasonal averages might have had more similar patterns to the Ventura county data, which generally had high usage and lower concentrations. This might have led CDPR to include some of the Ventura county data in their regression model calibration instead of excluding all those data from their final township cap regression model calibration.

Monitoring site location correction: The location coordinates published in the CARB and CDPR reports for all monitoring sites were reviewed using maps and aerial photographs. They were also compared with other data where available, such as CARB locations for the same monitoring sites in other reports, and with information collected during site visits by EPA staff. The specific location coordinates for one site (Pajaro Middle School) were updated as a result. The report continues to include a description of that process along with a copy of the original and corrected location displayed on an aerial photograph, which clearly depicts the school at the corrected monitoring location.

5. Provide any additional comments or recommendations you feel are important to improve the quality of this document.

A. One reviewer commented on the structure and organization of the report.

"The document's writing needs to be made more structured and organized to be easier to follow. The presentation of data sources, model development and model evaluation should be described in a more logical sequence.

Response:

The updated document (Cohen et al. 2011) has been restructured and clarifying text addressing reviewer comments has been added.

B. Two reviewers commented on a description of the models' tendency to underestimate concentrations.

"Why does the document indicate that it's desirable to avoid over-predicting modeled concentrations which could lead to a mistaken finding?"
"It seems inconsistent that the EPA would endorse a "model" that underestimates concentrations. Couldn't the regression equation be constrained to provide a better fit to the higher concentrations?"

Response:

The updated report (Cohen et al. 2011) has been revised to reflect simply that "neither under- nor over-estimation of concentrations is desirable". As previously mentioned, our examination of the fractional bias statistics for the best-performing models indicated that the degree of underestimation for all of them was quite small when the full data set is included for both daily and 7/8 week averages. The degree of bias was moderate for the portion of the data set below the median for both averaging times, showing overestimates for daily averages and underestimates for 7/8 week averages.

"The statement about the accuracy of air model predictions refers to deterministic, scientific models such as AERMOD, not to regression equations fit to data. It is well known that if a regression equation is fit so the best-fit line passes through the middle of the data, then the agreement should be much better."

Response:

We agree that air dispersion model predictions have a target accuracy rate of a factor of two or better. We also agree the mean predictions equal the mean observed concentrations for unconstrained regression models with an intercept, but not necessarily in other cases.

However, it does not follow that a regression model should necessarily provide better agreement with observations than would a dispersion model. The agreement depends upon the uncertainty in the representation of physical phenomena by the regression formulation, as well as uncertainties in the input and monitoring data, and upon the importance of unknown or unmeasured factors that would be expected to affect the concentrations.

If the regression model and dispersion model use exactly the same mathematical representation for the relationship between concentrations and the independent variables (*e.g.*, usage, meteorology), but the regression model uses empirically estimated parameters rather than independently-specified values, then the regression model would be expected to perform at least as well, and probably better, since the parameter values are calibrated to match the observed concentrations. However, although we used the standard Gaussian formulation variables as candidates for inclusion, these regression model formulations did not use exactly the same mathematical representation. For example, because the monitoring and usage data were collected on a 24-hour basis, the finest temporal resolution possible for the regression models was 24 hours (although for some model formulations we disaggregated the meteorological data into day and night subsets). Also, because the available concentration data were limited, the regression model had to balance the need for detailed representation of the underlying physical process with the increased uncertainty of the parameters in highly parameterized models.

Given the mix of "advantages" and "disadvantages" for the regression models, it is not clear *a priori* whether they should perform better than a dispersion model. We believe that target and observed model accuracies are sufficiently similar here to warrant applying similar performance criteria.

"It is stated that met data was obtained from either CIMIS sites or NWS sites. There could be a large difference in wind speeds since the NWS sites are usually at airports with small surface roughness and hence larger wind speeds. Were the two types of data compared for nearby sites?"

"Summarize the similarities and differences between the CIMIS met monitoring stations and the NWS stations. This is important because often NWS airport winds are much larger than winds at more sheltered sites or near trees and agricultural fields"

"Discuss possibly biases between the NWS and CIMIS Met data. For example, the NWS wind speeds are likely to be higher."

Response:

CIMIS and NWS data characteristics were compared in the report of the data bases developed for this study (ICF 2011, section 6.1). As noted therein, CIMIS data was judged to be better suited for this application based on the criteria of proximity to the study area, elevation of the measurement, and time period of the measurement.

However, because of incomplete data, it was necessary to use NWS data for some locations and time periods. This was done both in the model calibration procedure described in the report, and later in applying to selected regression model to all the public schools in the state.

As explained in the report of the data bases developed for this study (ICF 2011, section 6.2) schools were paired with meteorological stations based on the following primary criteria, whether a CIMIS or a NWS station.

- Proximity: the closest site within 20 miles which meets other criteria
- Elevation: within 500 feet
- Match water-body influence: coastal or inland

For receptor locations where one or more of these primary criteria could not be met, the following "professional judgment" criteria were applied to select a matching meteorological station.

- Surface terrain features in the vicinity of the school and candidate meteorological stations
  - avoid pairing school with a meteorological station in a different valley;
  - minimize differences in valley axis orientation (e.g., up-valley station versus down-valley station).
  - avoid differences in water-body influences;
- Surface roughness features in the vicinity of the school and candidate meteorological stations
  - minimize differences in roughness values (e.g., farmland versus forested or built-up area);
- Proximity of the school and meteorological stations to urban areas
  - minimize differences in urban heat island influences;

In the case of the model calibration sites, meteorological data were taken from CIMIS for all 14 monitoring site when complete data were available (i.e., primary site). But for 3 sites NWS data were used if complete proximate CIMIS data was not available for a particular time period (*i.e.*, additional data was taken from a secondary site). For the statewide application of the selected model, out of

8,493 schools, NWS was used as the primary meteorological data for 1,026 schools and as the secondary data for 1,385 schools.

"If a physics-based power law had been proposed such as 1/dp or 1/wq, and then the powers p and q best fit by a regression package, then the agreement would be better. Also, I thought that the sigma-theta method for estimating stability had been discarded long ago, since many groups have shown that large sigma-thetas can occur during both light wind unstable or light wind stable conditions, and hence the relation is not monotonic."

Response:

The infeasibility of using a power law approach for this application is discussed in the response to comments in section 1.D.

The sigma theta approach is still a suggested method for estimating stability as shown in EPA's *Guideline on Air Quality Models* Appendix W Section 9.3.3.2 (US EPA, 2003, with detailed procedures presented in US EPA, 2000). As noted the method is non-monotonic with respect to large sigma-thetas, but this issue was addresses in this analysis by varying the treatment for daytime and nighttime conditions. For daytime conditions large sigma thetas were assigned to "A" stability, while for nighttime periods large sigma thetas were assigned to "F" stability. Also, note that the final selected model does not include an adjustment for stability.

"Fig 9 containing observed vs "predicted" concentrations - A threshold (say 0.01 ppb) should be applied to both observed and predicted concentrations to avoid the large collection of points at predicted C = 0.0. I made a rough calculation on the unpaired "highest 5" observed and predicted concentrations and find that the predicted highest concentrations are about 50 to 60 % of the observed"

Response:

The issue of background concentrations is discussed in the response to comments in section 2.D.

The updated report (Cohen et al. 2011) contains a revised Figure 9 presenting the results for models selected based on revised ranking criteria in response to the comments received here. The authors believe it is still helpful to display the full range of values which particular models predict to be equal to zero, since the range can vary somewhat from model, and include observed values near 1.0 ppb.

"Page 9 last paragraph, questionable assumptions: (1) all usage distributed somewhere throughout the MTRS is located at the section center, (2) the receptor is located at the section center where the monitoring site is located, and (3) 0.5 mile minimum distance. There is a conflict because assumptions (1) and (2) would put the receptor right on top of the fumigant source while assumption (3) separates the two by > 0.5 miles. Why?"

Response:

The assumptions noted here were those used in previous work by CDPR, not those used for the current study. This is stated more clearly in the updated report (Cohen et al. 2011).

CDPR applied assumptions 1 and 2 to calculate distances between the source and receptor, except in the cases where the source and receptor were in the same section. If the source and receptor were in the same MTRS section location, a separation distance of 0.5 miles was assumed.

"Page 15 last paragraph, how 24-hour methyl bromide concentration is averaged? Or is it a cumulative measurement of methyl bromide over 24-hr period"?

Response:

The CARB used a canister sampler to make a continuous and cumulative measurement of MeBr over each 24-hr period. Using this approach provides the equivalent of a 24-hour average value, which is clarified in the updated report (Cohen et al. 2011).

"Page 34 middle section, the temperature effect is accounted for using degree-hours. How is time entered in the formulations for ambient and soil temperatures? Why subtracting 4 from the temperature?"

Response:

The temperature factors were calculated as the sum of degree-hours divided by the number of hourly measurements in the time period. A degree-hour for each measured hourly value equals the number of degrees Celsius minus 4 when that number is positive, and equals zero for each measured hourly temperature at or below 4 degrees Celsius. The time period is either the 24-hour monitoring period, the daytime part of that 24-hour period, or the nighttime part of that 24-hour period, depending upon the regression model formulation. The subtraction of 4 was made to better approximate volatilization/temperature curves found in the analysis by Yates, Gan and Papiernik (2003), who examined and modeled the impacts of ambient temperature and other factors on emissions. Because 4°C is the boiling point of MeBr, volatilization is expected to be minimal at temperatures of 4°C or less.

This is clarified in the updated report (Cohen et al. 2011).

"Page 36 Eq. [3], based on Eq. [2], the log (Conc) on the left side of Eq. [3] is actually log (Conc-Intercept)?"

Response:

For equation 3, we specified that we were ignoring the background term, represented by the Intercept from equation 2, so that equation 3 represents the case where there is a single source and no background term. If log(Conc) is replaced by log (Conc – Intercept), then we would get a different version of equation 1, which is a non-linear model because of the presence of the "Intercept" term on the left hand side. (This "Intercept" is put in quotes to emphasize that it is not the usual intercept for a regression model, defined as the predicted value when the independent variables are all zero). Since you cannot take the logarithm of a negative number, that would force the "Intercept" to be smaller than all the daily values, i.e., less than 0.001 ppb (since one measured daily value was a non-detect, estimated as half the MDL). In any case, the primary issue of the discussion is that none of the equations 1 to 4 are applicable for modeling the MeBr concentrations because there are multiple emissions sources.

"It is stated that a "simpler" method is preferred. But simpler than what?"

Response:

The updated report (Cohen et al. 2011) has been revised to read "a relatively simple, and computationally tractable prediction method is preferred."

"Explain why the AMBI monitoring data were not used."

Response:

The documentation has been edited to provide additional information on this issue. CDPR's review identified some issues with the quality and consistency of the AMBI data (Segawa 2002). For example, the 2001 AMBI data included a number of contaminated trip blanks. The criterion for identification of unacceptable uncertainties in particular daily values also used a different flow rate variability threshold than CDPR, and monitoring procedures in 2001 included turning off a monitor when MeBr applications were known to occur nearby.

The AMBI 2002 data were expected to be more consistent with CDPR's approach, but CDPR's review still identified some uncertainties and apparent inconsistencies. In addition, dramatic differences in CDPR regression model calibration results when these data were included (Johnson and Li 2003), also suggests that there were fundamental collection method discrepancies.

"The rationale concerning the uncertainties does not make sense. Are you saying that if there are uncertainties in input data, then there is no need to run ISC and you might as well run a regression model?"

Response:

The updated report (Cohen et al. 2011) has been revised to read as follows.

> "Application of the ISC model requires specification of a pollutant emission rate for each source. The emission rate for a particular time period is a function of (1) the fraction of the total applied MeBr that is volatilized as opposed to remaining or decaying in the soil, and (2) the temporal profile of the volatilization over a period of several days (i.e., the relative daily emissions generated by a MeBr application). As discussed elsewhere in this document, these are both a function of several variables, including the application method. Although we have some information about the date and volume of MeBr usage, the method used for any particular application is unknown"

As noted in response to comments in section 1.D., specification of an emission rate is not required for the regression model. Instead, the MeBr daily application rates are specified.

"Again to help the reader, three tables should be inserted summarizing the ambient monitoring data, the pesticide usage data, and the weather data."

Response:

Ambient monitoring values are presented in Table 2 of the report. It is not clear to the authors what summary information is being requested by the reviewer.

"Several options for determining background concentration are described. Which method was eventually chosen and why?"

Response:

The documentation has been edited to provide additional information on this issue. We omitted the specification of a background concentration by constraining the regression intercept to a zero value. This was done because the intercept represents the unique proximity of the monitor locations to background sources, and thus it is very likely to be spatially variable. If we included an intercept in our modeling of the monitoring data, this intercept would represent the typical background concentration in the monitoring region, but could be a very poor estimate of the background concentration in other regions.

Please see the response to comments in section 2.D. for additional discussion.

"Several arbitrary and unjustified 'adjusted usage' factors are listed. For example, why should the usage be multiplied by the inverse of the wind direction standard deviation? What is the effect of these assumptions on the regression equations?"

Response:

The adjustment of usage by the inverse of the wind direction standard deviation is based on the relationship identified by Draxler (1976) in which

$$\sigma_y = \sigma_\theta x f_y$$

where σ_θ is the standard deviation in wind direction, x the downwind distance and f_y is a function with unity near the source but a decreasing function of travel time. Thus, to first order the σ_y is approximately proportional to σ_θ. According to the Gaussian plume formulation the concentration is inversely proportional to σ_y, and thus approximately inversely proportional to σ_θ.

As noted above, the relationship between σ_θ and stability is different during nighttime than during daytime, and therefore separate daytime and nighttime sigma-theta values were tested as candidate factors for emission adjustment. However, the adjustment of usage by the inverse of the wind direction standard deviation did not add much explanatory value to the regression model development, and was not included in the final model selected for the statewide application.

"Specific rationale should be listed rather than saying 'makes sense'."

Response:

The updated report (Cohen et al. 2011) has been revised to read:

> "[Eqn 9] is consistent with physical principles, properly addresses the problem of multiple sources, and can be fitted to the data using regression."

"Arbitrary constants are chosen (757.53 feet and 459.57 feet) with no justification. Please show the calculations that led to these numbers. Why are five significant figures needed for an arbitrary constant?"

Response:

The calculations leading to these numbers are presented in Attachment 11 of the revised report (Cohen et al. 2011). The distance 459.57 feet was the average radius for a sample of 20 California school grounds. The distance 757.53 feet was calculated as the inverse-distance-weighted average distance between a receptor and the usage, based on the geometry of a circular school, a square grid

cell of side 1320 feet, a uniform random distribution of the school centroid, and a uniform distribution of usage at all points outside the circular school boundary.

"This reviewer finds the notation impenetrable. It would be better to simply use the mathematical format."

Response:

The updated report (Cohen et al. 2011) includes a mathematical equation for the final selected model.

"Where is the line drawn between an "error that is so extreme …" and an error that is not? What is the justification?"

Response:

The excerpt at issue occurs in section IV.A.1 of the report. The remainder of that section presents a precise definition of how (and why) we defined monitoring data outliers for the model calibration process, which then defines the line we drew to distinguish extreme errors (i.e., outliers) from other values and justifies our procedure. We believe that our approach is reasonably conservative (not too many outliers) and consistent with statistical principles.

The 1 1/5 page Observations and Conclusions section is too brief and appears to have been written as an afterthought.

Response:

In the updated report (Cohen et al. 2011) this section has been expanded to more thoroughly present conclusions.

References

California Air Resources Board (1998). "San Joaquin Valley Regional PM_{10} study Characterizing Micrometeorological Phenomena: Mixing and Diffusion in Low Wind Speed Conditions Phase III: Monitoring and Data Analysis" by Systems Applications International, July 1998.

Draxler, R.R. (1976) "Determination of Atmospheric Diffusion Parameters." *Atmospheric Environment*, 10, pp 99-105.

Eastern Research Group, Inc. (2005) *External Peer Review of the Draft Report on Model Development for Assessing California Methyl Bromide Ambient Air Concentrations*. July 12, 2005.

Gan, J., Yates, S., Wang, D., Spencer, W. (1996). "Effect of Soil Factors on Methyl Bromide Volatilization after Soil Application." *Environ. Sci. Technol.* 1996, 30, 1629-1636

Hanna, S., Britter, R., Franzese. P. (2003)"A baseline urban dispersion model evaluated with Salt Lake City and Los Angeles tracer data." *Atmospheric Environment* 37 (2003) 5069–5

ICF (2005a). "Model Development for Assessing California Methyl Bromide Ambient Concentrations." Draft Report for Peer Review. May 26, 2005.

Cohen JP, Rosenbaum AS (2011) "Model Development for Assessing California Methyl Bromide Ambient Concentrations" Draft report prepared by ICF International for Loren Hall, US EPA Office of Civil Rights, Washington DC

ICF (2011)."Database Development for Model Inputs to Assess California Methyl Bromide Ambient Concentrations" Draft report prepared by ICF International for Loren Hall, US EPA Office of Civil Rights, Washington DC

Johnson, B and Li, LY, (2003). "Calculation of a tolerance interval for a township limit on methyl bromide use to control subchronic exposure". Department of Pesticide Regulation, Environmental Monitoring and Pest Management Branch.
http://www.cdpr.ca.gov/docs/dprdocs/methbrom/docs/mebr_tolerance_cap.pdf

Li, L., Johnson, B., Segawa, R. (2005)."Empirical Relationship between Use, Area, and Ambient Air Concentration of Methyl Bromide." J. Environ. Qual. 34:420–428 (2005).

Nou, V, (1963): *Development of the diffusion prediction equation – The Ocean Breeze Dry Gulch Diffusion Program. Vol. II* (eds., D.A. Haugen and J.H. Taylor) AFCRL-63-791 (II) DOC Technical Services, Washington DC, pp 1-21

Segawa, R (2002). Memorandum to John Sanders dated May 29, 2002 on "Review of registrant revised report on methyl bromide air monitoring".

University of California Cooperative Extension (2004a). "Sample Costs to Produce Strawberries – South Coast Region –Santa Barbra County," 2004

University of California Cooperative Extension (2004b). "Sample Costs to Produce Strawberries – South Coast Region –Ventura County," 2004

US EPA (1992). Protocol for Determining the Best Performing Model. Office of Air Quality Planning and Standards, Research Triangle Park, NC. EPA-454/R-92-025. December, 1992.

US EPA (1995). ISC3 User's Guide, Volume II. EPA-454/B-95-003b

US EPA. (2000) Meteorological Monitoring Guidance for Regulatory Modeling Applications. EPA Publication No. EPA–454/R–99–005. Office of Air Quality Planning & Standards, Research Triangle Park, NC. (PB 2001–103606). http://www.epa.gov/scram001/guidance/met/mmgrma.pdf

US EPA (2001). "Part 50: National Primary and Secondary Ambient Air Quality Standards." Revised July, 2001. http://www.epa.gov/oar/oaqps/greenbk/40cfr50.html

US EPA (2003). 40 CFR Part 51 Revision to the Guideline on Air Quality Models: Adoption of a Preferred Long Range Transport Model and Other Revisions; Final Rule. *Federal Register*. Vol. 68, No. 72. Tuesday, April 15, 2003. http://a257.g.akamaitech.net/7/257/2422/14mar20010800/edocket.access.gpo.gov/2003/pdf/03-8542.pdf

Yates, Gan and Papiernick (2003). "Environmental Fate of Methyl Bromide as a Soil Fumigant." *Reviews of Environmental Contamination and Toxicology*. Volume 177. pp. 45-122.