Comments on the ISA from Dr. Lianne Sheppard
(updated 08-04-15)

Chapter 3

General comments:
The reorganized text is a great improvement upon the previous version and I appreciate the efforts taken to be responsive to CASAC's concerns. I focused my attention on the long-term exposure and effects. I felt that while this chapter is a major advance, there are aspects that still need substantial improvement.

Organization

- Overall this chapter is much better organized and on target. Kudos to EPA staff! I focus most of the rest of my comments on suggested improvements rather than on the successes of this revision, of which there are many.
- There is text that appears under the wrong section heading. For instance, see page 3-19, the paragraphs starting lines 24 and 35.
- I'm surprised that in Section 3.4.5 "implication for epidemiologic studies" there is no attention to the new methods development for correcting for measurement error in cohort studies. The appropriate papers are (mostly) cited in the document, but there is absolutely no attention given to the methods development and the deeper understanding this brings to epidemiologic inference in cohort studies. Only the simulation studies and some of the definitions are reviewed. I suggest that under section 3.4.5.2 that a new subsection be added that covers the essence of the new methods developed by Spziro and others (specifically in papers published in 2011 that assume a parametric geostatistical exposure surface (note: the Epidemiology paper is cited but the Biostatistics one isn't)(Szpiro et al., 2011a; Szpiro et al., 2011b), and the Environmetrics paper published with Paciorek in 2013 that assumes a fixed exposure surface.)(Szpiro and Paciorek, 2013a, b) More generally, for all subsections of 3.4.5, I think the state of the methods for correcting for exposure measurement error should be the focus of the discussion of the subsection. Right now each subsection mostly focuses on comparisons of different estimate results, some are simulated and some are based on real datasets, though these distinctions aren't clear enough.

Clarity of writing: It will be essential that the text be edited for clarity, cohesion, and to ensure the appropriate points are being made. I have listed examples where corrections are needed in my detailed comments; this is not comprehensive. There are some confusing, unclear, overly general and/or misleading statements in this chapter. My written comments identify some of the sections that need attention. In particular Sections 3.2 and 3.4 need work.

Judgments and insights

- I'm finding it difficult to believe claims in the document that certain exposure estimation approaches are quite generally better than others, particularly when such statements are made without consideration of any additional information. For

instance, as mentioned repeatedly in Chapter 6 (section 6.2), this document judges IDW and dispersion modeling to be more uncertain than e.g. LUR in their ability to represent the spatial variation in NO2, and thus by implication the reader is led to believe that these exposure estimates produce poorer health effect estimates. While there may be many examples of pairs of studies where this conclusion is valid, I could easily describe a pair of studies as a counterexample: one which uses LUR and the other which uses IDW to estimate exposure but where I would trust the IDW estimate more than the LUR estimate of long-term average NO2. A few of the reasons would include the relative number of locations used in each study (more for the study using IDW), the representativeness of the monitoring locations (poorly aligned with study subjects in the study using LUR), the time period of measurement in each (much longer and better representing a full year in the study using IDW), the available covariates for the LUR (a limited or poorly chosen set would produce poorer estimates) and the authors' approach to model selection in the LUR (overfit LUR models will produce poor estimates of NO2, even when the resulting estimates are quite variable). It is not just the tools used to produce exposure estimates that matter, but the exposure study design and the application of the tools.

- The whole discussion of evaluation of models seems to be inadequately nuanced or informed by in-depth understanding of the prediction modeling methods. Early in the discussion of LUR models a footnote indicates "unless otherwise noted for the LUR studies, R2 refers to model fit." While not further defined, it appears that this means the R2 is equal to the square of the correlation coefficient from the regression of the predicted exposures on the monitored observations that are in the same dataset used to develop the LUR model. This kind of "in-sample" estimate is the last type of estimate of R2 I would wish to use to describe LUR model performance. It will tend to be too high, won't reflect overfitting of the LUR model, and won't actually inform us of the model performance we care about, namely predictive ability at subject locations. For this purpose, "out-of-sample" estimates of R2 are preferable; these are often computed using a technique called "cross-validation". Furthermore, out-of-sample R2 estimates can be obtained about the best-fit line (also called regression-based R2, computed as described above but using different measurement locations than the ones used to develop the model) or the 1:1 line (also called MSE-based R2). The latter R2 estimate gives a more complete picture of how the predictions at new locations compare to measurements. I suggest that the document warrants additional attention to how model evaluation is considered and discussed.

Major suggestions for improvement

- I suggest making a table summarizing all the studies discussed in Section 3.2. Easily being able to compare the models, evaluations, input data (including number of locations, number of time points (if relevant, i.e. more than 1), time scale), and results would help readers better understand this section. It is clear from the text describing the papers I am familiar with that the write-up is somewhat misleading or worse. Hopefully the addition of the table will help address this concern. The table should also indicate the time period and spatial domain of interest for each study.

- I suggest completely revamping the section on spatio-temporal modeling in Section 3.2 since this is becoming much more common in epidemiology (and it is the major approach used in EPA's MESA Air study). (This should replace/update the "spatiotemporal interpolation modeling" discussion starting on p 3-11.) Cite Lindstrom (2013)(Lindström et al., 2013), Szpiro (2010 Env)(Szpiro et al., 2010), Sampson (2011 AtEnv)(Sampson et al., 2011), Keller (2015 EHP)(Keller et al., 2015) and Li (2013 AtEnv; explicitly recognizing that Li is a fairly minor extension of the MESA Air spatio-temporal model)(Li et al., 2013). The spatio-temporal models discussed above use LUR, universal kriging (UK), and temporal trend function estimation using a singular value decomposition (SVD) of basis functions to capture both temporal and spatial variability. Many of their evaluations focused on spatial variation since that is the source of variation if interest for long-term epi studies. There are other papers that also report spatio-temporal models that might be cited in the spatio-temporal modeling section, even though they don't focus on NOx/NO2 applications. These include Paciorek (2009 AnnAppStat)(Paciorek et al., 2009), Yanosky, (2008 AtEnv; 2009 EHP)(Yanosky et al., 2008, 2009).
- Clarify the nuances of the R2's being reported in the document. (See my comments above.) It is not particularly helpful for readers to be comparing in-sample and out-of-sample R2 estimates across studies as though they are the same quantity. Furthermore, for out-of-sample estimates, there are additional distinctions to consider (see above). Precisely defining the R2's being reported throughout the document would help readers make fair "apples to apples" comparisons.

<u>Response to charge question</u>

1. The exposure discussion is re-organized to clarify: a) the connection between particular exposure assessment methods and epidemiologic study designs, and b) the influence of exposure error on health effect associations from epidemiologic studies of specific designs. How explicitly and accurately is epidemiologic study design considered in the discussion of the utility and uncertainties of various exposure assessment methods, the nature of exposure measurement error, and the impact of exposure measurement error on NO2-health effect associations? How effective is the discussion in facilitating the evaluation of the strength of inference from epidemiologic studies in Chapters 5 and 6?

The reorganized text is a great improvement upon the previous version and I appreciate the efforts taken to be responsive to CASAC's concerns. However, there are aspects that still need improvement. The measurement error discussion still needs to be refined and improved to actually focus on the methods advances as do some of the comments and generalizations about various exposure assessment strategies for application to epidemiology. (see my detailed comments above) EPA's goal of facilitating evaluation of the strength of inference from epidemiologic studies is not yet fully met. The text includes many details that make this chapter's material less useful than ideal for making judgments about the epidemiologic studies. Yet in other ways important details are missing (e.g., a synopsis of the statistical methods advancements on handling

measurement error in cohort studies). The existing reviews of exposure assessment studies don't give the reader the deep insight needed to really understand their utility in epidemiologic study applications.

There are broad claims in the document that aren't always well supported. While I would agree that certain exposure estimation approaches should be better, claims are often made without good appreciation of essential additional information needed to evaluate any particular application. I'm finding it difficult to believe claims in the document that certain exposure estimation approaches are quite generally better than others, particularly when such statements are made without consideration of any additional information.

To meet the objective of facilitating the evaluation of the strength of inference from epidemiologic studies, I suggest EPA consider classifying exposure assessments for each epidemiologic study according to appropriateness for use in inference, using a system similar in spirit to those used for other judgments. For instance, each exposure assessment could be classified as strong, acceptable, weak, or inappropriate for the intended epidemiologic study. This would allow better judgment of epidemiologic studies based on the appropriateness of their exposure assessment. (A particular exposure assessment might get one judgment for one epidemiologic study and a different judgment for a different study.) The reasons behind the judgment should be provided as well.

Ultimately, given the current state of knowledge and the resources available, I think it will be difficult to successfully make all the changes needed to address this charge question and meet the objectives of this chapter.

Detailed comments for Chapter 3

- P 3-2 l 9: What does "research-grade" mean? How is this linked to central sites?
- P 3-7 paragraph starting line 4: This discussion is problematic. Please revisit.
- P 3-7 paragraph starting line 19: Ditto
- P 3-8 paragraph starting line 32: Isn't the point of the SA-LUR model to incorporate temporality into the model (through variables such as wind speed, etc)? So why does this paragraph open in a way that implies that the previous discussion wasn't about temporality?
- P 3-9 paragraph starting line 17: There are some misleading statements that should be corrected.
- P 3-10 line 23-5: While the statement is fine, I think the more important point is that for informing inference for epidemiological studies the comparison of the modeled estimates to measured values should be at locations that are relevant to the intended epidemiologic study.
- P 3-10: It is misleading to say Lindstrom 2013 "applied LUR" since the spatio-temporal model fit in that paper was much more complex than a LUR. Furthermore there were only two averaging times for model evaluation in that study: 2-week and long-term. There were no daily data; all the input data were on a 2-week time scale. The different model evaluation summaries (homes, snapshot, long-term averages at

monitoring sites) each had different strengths and weaknesses and gave different insights into the spatial performance of the model.

- P 3-11 paragraph starting line 3: I suggest this paragraph discussing multiple linear regression as an "emerging exposure assessment method" should be dropped.
- P 3-11 discussion on spatiotemporal interpolation modeling needs to be updated and merged with the suggested new section I described above. Spatio-temporal modeling methods are no longer "emerging" for application to epidemiologic studies.
- P3-13 line 30-1: I don't understand the relevance to the ISA of CA DOT's lack of support for CALINE. Omit or clarify.
- P 3-16: Is Fuentes and Raftery the right reference for BME? And what about all the BME work in air pollution by Serre and his group?
- P 3-18 line 12-3: I wonder how many epidemiologic studies would care about model performance averaged over multiple locations?
- P 3-18 l 14: First these models should be defined.
- P 3-19 l 1: A stronger statement than "are not typically used" is necessary here. All probabilistic exposure models are inappropriate for use as exposures in epidemiologic studies because the probabilistic component induces measurement error in the health effect estimate. Probabilistic exposure models are very useful for risk assessment.
- Table 3-1 is a useful addition. Some details need to be corrected:
    - Revise the title: Most of the methods listed in the table are not "sampling methods".
    - Consider redesigning to have a set of columns for epidemiologic applications that rely on short-term exposure variation and another for those that rely on long-term averages
    - I don't understand the phrase "if the monitors are sited at fixed locations" under passive monitors. How else are passive monitors sited?
    - Kriging is omitted (ordinary and universal). Ordinary kriging could be considered with IDW.
    - The summary for spatiotemporal modeling reveals a lack of careful reading of the papers cited in the document. Please update. (See also comments above)
    - Parameterization modeling was never mentioned in the text
- P 3-23 l 13: The statement "exposure is likely to be underestimated" is too general to be true. The key challenge with IDW and other spatial smoothing methods is that with too few monitors the surface will be too smooth to capture the spatial variation of interest.
- P 3-56 Section 3.4.3.4: The focus of this section should be on the impact of instrument accuracy on epidemiologic study inference. How many studies differentiate exposure over a day? I would mention the broad classes of short-term time series and long-term cohort studies and talk about the role of instrument error in inference for each.
- P 3-57 Section 3.4.4: The overview of the confounding section is nice and could be a model for other sections (e.g. instrument error).
- The section heading for "3.4.4.3 Personal and Indoor Relationships between Nitrogen Dioxide and Copollutant Exposures" focuses on personal-ambient relationships.
- Section 3.4.5: This section needs work. Specific comments

  - Preface to Eq 3-12: Is this a model for cohort studies and a continuous outcome? This statement is so general to not be helpful. (We do use a model like this in our measurement error methods work where we make it clear the above two aspects are assumed.)
  - Lines 4-6: OK but I think this generality gets this statement into trouble. Usually we want exposure on the native scale for inference about health effects. The logit is a transformation of the outcome (for a glm) not a normalization; also there are details omitted here that are relevant to glms.
  - Line 7 "most …": This is much too strong a statement. What about survival outcomes or binary our count outcomes?
  - Paragraph starting line 9: Make sure to clearly distinguish pure vs. "-like" error and to define the latter. The definitions of "-like" error were developed for modeled exposures from e.g. a LUR. Also recent research has shown that there can be bias in either direction from Berkson-like error (see Szpiro & Paciorek 2013).
- P 3-83 l 27: Be precise. Were these pure or "-like" errors? I expect the former. Similarly, address statements on the following page on lines 2, 12-13, 13-14, 16-17. Some may be incorrect, or at least misleading as written.
- P 3-88 l 8-10: Strike this sentence. This work was based on simulation studies. The data were made up so the work could certainly be repeated for cohort studies. However the recent measurement error work for cohort studies makes important progress in a different way.
- P 3-38 l 4: If the central site monitor is truly systematically higher or lower, with no other missing features, then the slope (beta coefficient of interest) won't be affected.
- P 3-89 l 23-26: In this section these results deserve more comment since they are highly counter-intuitive to me. I'd like to know the details of what was done in the IDW vs. LUR to understand why this is true. For instance, if the IDW had the right time period, but the LUR didn't, this could affect the epi findings.
- P 3-90 paragraph starting line 33
  - l 33 "spatial errors": Be clear with terminology. Paciorek focused on confounding not error that is uncorrelated with exposure and outcome.
  - L 38: This reference to "effect of specification of spatial conditions" is unclear and misleading. Szpiro (2011)(Szpiro et al., 2011b) should not be reviewed in the same paragraph with Paciorek (2010 Stat Sci 25: 107-125) since the foci of the papers were entirely different!
  - The summary of Szpiro (2011) has numerous confusing, poorly worded, or erroneous statements. There was absolutely no confounding in the simulation study in that paper! (continues on following page)
- P 3-91:
  - Please also carefully review the discussions of Basagna and Szpiro & Paciorek for clarity and correctness. I had difficulties with both. In particular, with respect to S&P: This entire discussion completely ignores the new methods for correction that were developed and the assumptions that were made to accomplish this. One important role of this review would be to note that this approach is the wave of the future and that future epidemiologic

cohort studies should be using measurement error correction methods since studies that don't do the correction get the wrong variance estimate for the health effect and may also need to correct for bias.
  - L 3: "predicting the true concentration" is not what is meant by the true exposure model in Szpiro (2011). The correctly specified model had all 3 covariates included while the misspecified model only included two of them.
  - L 5: Statement incorrect
  - L 6: Poorly worded
  - L 8-10: I don't understand this sentence.
  - L 11: Replace "of the correctly specified exposure model" with "in simulating the monitoring network data".
  - L 13: The bias in Szpiro (2011) was trivial and probably not worth mentioning in this discussion.
  - L 16: Insert "in the monitoring data" at the end of this sentence.
  - Paragraph starting line 36: This discussion needs to be fully revamped.
- P 3-92 l 17: In place of "criticized" it would be more correct to say "was pointed out to be a version of". Suggested new wording: "Spiegelman noted that the new measurement error correction methods developed by Szpiro & Paciorek (2013) are a version of regression calibration. This study ..."
- P 3-93 l 5: Will overestimating exposure *always* drive health effects towards the null? I can show a simple counterexample.
- P 3-93 paragraph starting line 14: Make sure the generalizations stated are correct and correctly qualified.
- P 3-95 l 25-7: This is too broad-brush of a statement. It needs to be qualified
- Section 3.5: There are some confusing, unclear, overly general and/or misleading statements in this section.

Chapters 5 & 6

I prioritized Chapter 6 and linkages with long-term exposures for my review.

I found the discussion of exposure assessment brought forward from Chapter 3 to be still in need of further refinement for clarity, accuracy, and utility for the purpose of judging the inferences that should be made from the epidemiologic studies. For example, text on pages 6-19 and 6-20 should be refined.

<u>Response to Charge question</u>

3. Drawing from Chapter 3, the health effect evaluations more critically evaluate the exposure assessment methods used in epidemiologic studies. Please comment on the adequacy and consistency with which exposure assessment, including the utility and uncertainties of the methods used and potential impact of exposure measurement error, is considered in describing the strength of inference from epidemiologic results. To what extent is available information on health effects related to personal and indoor NO2

adequately considered in conclusions?

I appreciate the inclusion of the exposure modeling approach in the summary figures and the exposure assessment details in the tables. The discussion of the utility and uncertainties of the methods used and the potential impact of exposure measurement error is less successful. For instance, the discussion in section 6.2.2.1 (starting p 6-19) needs some refinement.

The material in Chapter 3 did not focus on indoor and personal NO2 other than to look at their relationships with ambient NO2. Indoor and personal sources provide important information about health effects separate from ambient source NO2/NOx.

Detailed comments:

Please verify all the numbers in Table 6-1 reported for the McConnell (2010) study. I had a hard time finding all the HRs in the paper.

References

Keller, J. P., C. Olives, S. Y. Kim, L. Sheppard, P. D. Sampson, A. A. Szpiro, A. P. Oron, J. Lindstrom, S. Vedal, and J. D. Kaufman, 2015, A Unified Spatiotemporal Modeling Approach for Predicting Concentrations of Multiple Air Pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution: Environmental Health Perspectives, v. 123, p. 301-309.

Li, L. F., J. Wu, J. K. Ghosh, and B. Ritz, 2013, Estimating spatiotemporal variability of ambient air pollutant concentrations with a hierarchical model: Atmospheric Environment, v. 71, p. 54-63.

Lindström, J., A. Szpiro, P. Sampson, A. Oron, M. Richards, T. Larson, and L. Sheppard, 2013, A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates: Environmental and Ecological Statistics, p. 1-23.

Paciorek, C. J., J. D. Yanosky, R. C. Puett, F. Laden, and H. H. Suh, 2009, PRACTICAL LARGE-SCALE SPATIO-TEMPORAL MODELING OF PARTICULATE MATTER CONCENTRATIONS: Annals of Applied Statistics, v. 3, p. 370-397.

Sampson, P. D., A. A. Szpiro, L. Sheppard, J. Lindström, and J. D. Kaufman, 2011, Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data: Atmospheric Environment, v. 45, p. 6593-6606.

Szpiro, A. A., and C. J. Paciorek, 2013a, Measurement error in two-stage analyses, with application to air pollution epidemiology: Environmetrics, v. 24, p. 501-517.

Szpiro, A. A., and C. J. Paciorek, 2013b, Rejoinder: Environmetrics, v. 24, p. 531-536.

Szpiro, A. A., C. J. Paciorek, and L. Sheppard, 2011a, Does More Accurate Exposure Prediction Necessarily Improve Health Effect Estimates?: Epidemiology, v. 22, p. 680-685.

Szpiro, A. A., P. D. Sampson, L. Sheppard, T. Lumley, S. D. Adar, and J. D. Kaufman, 2010, Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies: Environmetrics, v. 21, p. 606-631.

Szpiro, A. A., L. Sheppard, and T. Lumley, 2011b, Efficient measurement error correction with spatially misaligned data: Biostatistics, v. 12, p. 610-623.

Yanosky, J. D., C. J. Paciorek, and H. H. Suh, 2008, Predicting Chronic Fine and Coarse Particulate Exposures Using Spatio-Temporal Models for the Northeastern and Midwestern US: Epidemiology, v. 19, p. S211-S212.

Yanosky, J. D., C. J. Paciorek, and H. H. Suh, 2009, Predicting Chronic Fine and Coarse Particulate Exposures Using Spatiotemporal Models for the Northeastern and Midwestern United States: Environmental Health Perspectives, v. 117, p. 522-529.