Commenter A

Nitrogen Management Project Protocol
Minimum Data Standard Public Comment Period

General Comments:

- Please explain a little bit better in the document what the Reserve is seeking: (1) developing an in-house database of N_2O data combining results from many different studies for in-house or contracted development of new quantification approaches or (2) inviting people to submit datasets that they deem are extensive enough to develop a quantification approach which can then be considered by the Reserve for adoption. In some sections of the minimum data standard, (1) and (2) are a little bit convoluted, especially when it comes to applicability and extrapolation. You want to make sure not to discourage people that have data from one field with one side-by-side comparison not to submit their data, just because they don't think it is applicable to e.g. the whole central valley of California. I would focus the minimum data standard around (1),and really focus on the rigidity of the cumulative N_2O value for each treatment-field-year combination. Then, you could consider developing another standard with requirements for quantification approaches, including calibration, validation, uncertainty assessment, extrapolation criteria and applicability. But that's of course fully up to your ambitions and intentions.

- You could look at the outreach page of the former lab of Johan Six at UC Davis. http://www.plantsciences.ucdavis.edu/Agroecology/outreach.html There are a lot of data and photos of fields, that give an idea of the type of spatial stratification you can expect, the duration and occurance of N_2O peaks, etc.

- Is the data submittal sheet available yet? Will there be guidelines with respect to the units in which the data should be delivered, other input variables, etc.? Will there be another standard for proposed quantification approaches (e.g. rigidity of uncertainty analysis, etc. number of data (field-year-treatment observations) that need to be included in the calibration and validation dataset, etc.?)

- If the Reserve develops a database, will you only include the cumulative annual N_2O emissions or the flux emission time series as well? Having the time series might be useful for more rigorously calibrating and validating biogeochemical models. For Tier 2 type approaches, the annual fluxes would be enough I guess.

Reserve MDS Package Public Comment 1:

Recommended recent studies and resources for the Reserve to explore further:

- Decock, Charlotte, "Mitigating Nitrous Oxide Emissions from Corn Cropping Systems in the Midwestern U.S.: Potential and Data Gaps," *Environmental Science & Technology*, **2014**, *48* (8), pp 4247–4256. http://pubs.acs.org/doi/abs/10.1021/es4055324 (Recently published meta-analysis on N_2O emissions in corn-cropping systems in the Midwestern USA and Southeast Canada. IPNI has indicated that the full database underlying this work will be freely available, upon publication.

- Final report on the project "N_2O emissions from the application of fertilizers in agricultural soils" by the California Energy Commission. This report is finalized and should be coming out soon.

Reserve MDS Package Public Comment 3:

I think the submittal of 'non-published' data-sets should be allowed. Sometimes, it takes a while before a dataset gets published because of turnover of students and staff working on the project, the duration of the peer-review process, etc. It would be a shame to miss out on such data.

Reserve MDS Package Public Comment 4:

I am not really familiar with the alternative protocols, and don't have the resources at hand. I think one of the most important criteria to look for in the protocols is that at least 3 (preferably more) concentrations measurements are performed during chamber closure, so that the flux estimate can be fitted through at least 3 points. I assume the minimum data standard would still mostly overrule the guidelines in those protocols, or not? It might be good to explicitly recommend a protocol for data that are yet to be collected, just for comparability and consistency purposes. I think data from eddy covariance measurements or data from automated flux chambers connected to a laser instrument for laser spectroscopic measurements of N_2O will become more abundant in future years. I have seen data comparing automated sampling with flux chambers connected to GC versus connected to laser, and the laser data is much better. This is because data is continuously collected during e.g. a 10 minute chamber closure time. The trend of increase in N_2O concentration during chamber closure is thus determined by many more sampling points (1/second instead of 3 or 4 per 40 minutes or so). I don't know too much about protocols for laser operating conditions, etc., but it might be something worth looking into.

Reserve Technical Questions 5:

I worry a bit that those criteria will disqualify some very high quality datasets and don't necessarily leave enough flexibility to optimize the sampling scheme to the characteristics of different types of cropping systems. For example, in California, the fluxes you see after fertilization and irrigation are pretty short-lived, mostly lasting 1 week to 10 days. Sometimes, the peak only lasts a few days. In such scenario, you could miss some important data when you sample only 2 times per week, while you don't really need to keep on measuring 2 times per week for 30 days. I would suggest the following: the data submitter needs to demonstrate that they identified anticipated peaks and their duration based on existing literature or preliminary tests before the start of their measurement campaign. Then, the submitter needs to provide a graph with the flux time series that clearly shows that gas sampling after an event continued until the flux had dropped back to baseline. The graph would show arrows for all main events, including tillage, fertilization, irrigation, mowing or cover crop management events, harvest, rain, snow melt. Then, the quality by which the peak was captured after such events can be easily captured visually. This allows for the highest flexibility and opportunity to optimize the temporal sampling scheme to the cropping system under consideration. You could find examples of good time series on http://www.plantsciences.ucdavis.edu/Agroecology/outreach.html.

Reserve Technical Questions 6:

I would distinguish between datasets where data collection started before release of the minimum data standard and datasets collected after. For future datasets, I would

require winter emissions to be measured. For older datasets, I would allow submission of growing-season data only and encourage complementary data and information on winter emissions (especially snowmelt or rain after dryer periods). Such data could be collected at a later date than the original dataset. Complementary data could, in my opinion, certainly be used to supplement the growing season data and make educated decisions on the robustness of emission reductions estimates.

Reserve Technical Questions 7:

I think soil type, texture, soil organic matter content, and temperature and precipitation are important attributes. I think it is hard for the data submitter to determine how far the data can be extrapolated without doing some modelling, especially for data set that are limited to 1 or 2 fields. Describing soil and climate conditions of the site and how this relates to conditions within a broader area is probably the best he/she can do. The data can also be compared to other datasets from the expected applicability region, where available.

Reserve Technical Questions 8:

I think it depends on the vision of what to do with the submitted datasets. E.g. one research group can have a solid dataset from 1 field with a project and baseline treatment. This one dataset will probably not be enough to develop a quantification approach. However, if you have 5 groups in California that submit a dataset from similar cropping systems and treatments, conducted at different field sites, you might combine the datasets to make a robust quantification approach. Furthermore, there is also a difference between data from on-farm research versus data from experimental research stations. For on-farm research, you oftentimes have no side-by-side comparisons within one field. Instead, you could have data from e.g. a couple of farms that use drip irrigation versus a couple of farms that use furrow irrigation. In that case, one data set from 2 fields (one with and one without drip) is not sufficient to test for significant differences between the treatments. Yet, it can be an indicator for the potential of the treatment and has the benefit of being on farm data. Multiple farms using each of the alternative treatments would eventually be needed to make statistically sound conclusions. Hence, there is a difference between minimal spatial sampling requirements for development of a quantification approach vs. a robust dataset for the quantification of N_2O emissions from the field in question. In terms of the number of fields to be included, I think this also depends on how the quantification approach is developed. For meta-analysis e.g., it is best to have a minimum of 20 pairwise comparisons. That is not always available in practice. I think it is good to use a hybrid approach, where the variability and extent of extrapolation possibilities is tested with biogeochemical models. In any case, there should be enough data to be able to have a separate calibration and validation dataset or take a leave-one-out approach to test the robustness of your model.

Reserve Technical Questions 11:

I think one chamber per plot and functional location is sufficient, as long as there are at least 3 replicate plots per treatment. I would not provide strict guidelines on the determination of functional locations. I would just ask to provide literature evidence or evidence from preliminary tests that the functional locations make sense. I would also ask for some photos of the field site. It is pretty hard to visually assess the heterogeneity in the field, how the rows are formed, and how the chambers are placed relative to the pattern in the field. Maybe a little schematic diagram of the setup would be helpful too. I would not recommend chambers deployed at random

within one plot when there is a clear stratified pattern, e.g. as in row crops. A random pattern could be relevant in e.g. managed grassland.

Reserve Technical Questions 12:

Here, I think it is again important to distinguish between requirements for quantification methodologies vs. requirements for a specific data set. N_2O flux measurements are very labor-intensive, especially with manual chambers. Automated chambers on the other hand are costly (~5000$ for 1 automated chamber, ~100,000 for laser instruments that measure N_2O concentrations, etc.). Running gas samples on the GC is also time consuming. It takes on average 6 minutes for 1 sample to be run on the GC. Including standards, etc., that is about 0.5 hour of running time for 1 flux measurement in 1 chamber. Adding reps exponentially increases labor and analysis time. E.g. if you have 3 functional locations, 2 treatments, and 3 plots within a field, that is already 18 chambers per field, and 9 hours of analysis on the GC for 1 sampling date. I think most field experiments are sufficiently replicated with at least 3 plots per field and multiple functional locations per plot or one sufficiently large chamber to integrate spatial variability within a plot. I think the power analysis becomes much more important once it comes to the development of quantification approaches.

Reserve Technical Questions 15:

The description on how to deal with outliers feels pretty robust in the minimum data standard and incentivises to take out outliers as little as possible, which I think is the way to go. I am wondering if you would rather define it as a % deviation of the N_2O emission reduction. If the difference between project and baseline is 20%, a difference in 10% for one of the emissions is large. I would make it such that the N_2O emission reduction can't vary more than e.g., 10% (or another threshold that is acceptable for the Reserve).

Reserve Technical Questions 17:

I think the listed parameters are sufficient. If the Reserve intends to develop its own database that can be utilized for in-house development of quantification approaches, I would accept data from anomalous years too, but indicate that the year was anomalous. Such years do happen every e.g. 10 years. And if you have e.g. 10 field-year datasets, it is actually better to have 1 anomalous year in there.

Reserve Technical Questions 18:

I think this lists the most important attributes. The only other thing I could think of is soil management (e.g., tillage, residue management), but this is already included in the minimum standard as far as I can tell. You might want to consider rephrasing: is the soil type representative for this cropping system in this region. Some crops do better in certain soil texture types than other crops, so that a certain soil texture might select for a certain type of cropping system, regardless of how abundant that soil texture type is in the region.

Reserve Spatial Sampling Memo Question 1:

a) most experiments had 3 replicate plots per field/ or per treatment within one field.

b) 2 for tree crops (tree row and tractor row) and 3 for field crops (berm, side, furrow)
c) approx. 10 cm tall and 30 cm diameter
d) I think the stratified sampling was appropriate for the experiments in question. In all experiments, significant differences between functional locations were found, at least at some times during the measurement period.
e) Functional locations
f) I would not recommend placing the chambers at random, unless you work in a grassland system or the chambers are large enough to cover multiple functional locations in a representative fashion.
g) Tomato in California, Walnut in California, Almond in California, Vineyard in California.
h) In California, irrigation might cause much more stratification of N_2O hotspots compared to rain-fed systems that are more homogenously watered. As said before, recommendations for grasslands are likely different, especially if they are grazed. In that case, dung and urine hotspots need to be considered. This is the same for grazed cover crops, which occurs in some cropping systems in California. Given such examples, I think it is important to phrase the minimum data standard in a way that allows flexibility and optimization of the sampling regime for the system at hand. Flux time series, photos, and diagram of the sampling setup in the field might be helpful to evaluate the rigidity of the sampling scheme.

Reserve Spatial Sampling Memo Question 2:

I think stratification within plots is the way to go, unless in grassland, or when flux chambers are large enough to go across functional locations, etc. In fields that show a gradient (e.g. there is a slope or a shadow created by a tree line), it would be good to recommend randomized complete block design (RCBD) for the different treatments within a field.

Note that on page 3 in suggestion b. 85 (A) +5 (B) +15 (C) does not add up to 100. If you choose for stratified sampling, I would always analyse the data according to b. In the preliminary test, there might be no differences between the functional locations in N_2O emissions at that given time. Yet, after some particular events, differences between locations might exist, which you didn't identify in your preliminary experiment.

Reserve Spatial Sampling Memo Question 2, page 3, "A weakness of this approach…":

I don't see why this is a weakness. You simply asked two different questions and analyse your data in a different way to answer each of these questions. If anything, it is a strength, because it is more efficient use of data.

Reserve Spatial Sampling Memo Question 2, page 3, "(2) Include at least 3 chambers…":

I don't think this is feasible. This would mean that you have (3 plots) x (2 or 3 functional locations) x (3 pseudoreps) = 18 or 27 chambers per treatment in each field. Since you have at least one baseline and one project, this is 36 or 54 chambers per side-by-side comparison. I think one chamber per functional location per plot is enough. It is statistically sound to test differences in functional locations across the plots. It is called a split-plot design instead of a completely randomized design. Interpretations are valid for the field in question.