Published January 13, 2015

Statistical Challenges in Analyses of Chamber-Based Soil CO_2 and N_2O Emissions Data

A. N. Kravchenko*
Dep. of Plant, Soil, and Microbial Sciences
Michigan State Univ.
East Lansing, MI 48824-1325

G. P. Robertson
Dep. of Plant, Soil, and Microbial Sciences
W.K. Kellogg Biological Station
Michigan State Univ.
Hickory Corners, MI 49060

Measurements of soil greenhouse gas (GHG) emissions have gained a lot of attention in an effort to potentially increase agriculture's role in mitigating climate effects. However, it seems not well recognized that the nature of chamber-based GHG data is such that analyses require advanced statistical techniques to fully explore experimental treatment effects. Moreover, for soil GHG data some experimental design approaches can enhance while others can weaken a study's ability to detect treatment differences. Here we identify and explore the implications of key choices in experimental design and statistical analyses relevant to chamber-based soil GHG studies. In particular, we discuss (i) relative contributions of different sources of random variability in GHG field studies, (ii) relative benefits of increasing the numbers of samples at different replication levels to increase statistical power, and (iii) benefits of accounting for heterogeneous variances and using repeated measures analysis in GHG studies. Emissions data for CO_2 and N_2O collected from three experimental sites in Michigan demonstrated high spatial and temporal variability for CO_2 and N_2O fluxes. For both gases the total variability is dominated by small-scale spatiotemporal variability sources, which constituted 55% of the total variability for CO_2 and 95% for N_2O fluxes. While increasing the number of replicate plots is the main route of rising statistical power, increasing the number of subsamples (chambers and gas samples) per replicate plot can also provide substantial gains. Judicious repeated measures analysis and especially accounting for heterogeneous variances are important strategies for the efficient analysis of chamber-based GHG data.

Abbreviations: AIC, Aikaike Information Criterion; GHG, greenhouse gas; KBS, Kellogg Biological Station; LTER, Long-term Ecological Research site; MCSE, Main Cropping System Experiment; PAS, photoacoustic infrared spectrometer; RCBD, randomized complete block design; SCE, Sustainable Corn Experiment.

The number of studies that have used field-based chambers to measure CO_2 and N_2O fluxes from agricultural soils has grown exponentially in past decades. Scopus (accessed 2014), for example, reports 5 such publications in 1995, 50 in 2005, and >200 in 2012. A number of technical advances have supported this growth, and a number of authors have addressed the technical challenges of measuring GHG emissions in the field (e.g., Hutchinson and Livingston, 2002; Smith and Conen, 2004; Rochette and Bertrand, 2007; Venterea and Baker, 2008; Rochette and Eriksen-Hamel, 2008), including mathematical approaches to calculating fluxes from discrete measurements (Venterea et al., 2009; Levy et al., 2011; Parkin et al., 2012). Much less attention, however, has been paid to the statistical challenges of analyzing GHG flux differences through time and across ecosystems or experimental treatments. This is unfortunate, since by its nature

Supplemental Material is available for this article.
Soil Sci. Soc. Am. J. 79:200–211
doi:10.2136/sssaj2014.08.0325
Received 14 Aug. 2014.
*Corresponding author (kravche1@msu.edu).

GHG fluxes are among the most statistically challenging processes in soil biogeochemistry. Even less attention has been paid to methods for optimizing sampling strategies to maximize the probability of detecting differences in GHG fluxes among experimental treatments, also known as statistical power, with a study by Morris et al. (2013) being to our knowledge the only such effort to date.

Here we illustrate the value of quantifying sources of in situ flux variation to maximize the statistical power of field experiments. We use existing data from cultivated systems to highlight key decisions faced by practitioners designing and analyzing field experiments for GHG flux measurements. These include how many plots should be used per treatment, whether plots should be blocked and, if so, how blocks should be delineated, how many chambers should be placed within a single plot, and how many measurements should be taken from a chamber to calculate a flux rate estimate. All of these decisions can greatly affect the statistical power of the resulting study.

Although the statistical techniques for making many of these decisions are well developed and described in detail in statistical texts (e.g., Cochran and Cox, 1957; Cox, 1958; Kuehl, 2000), a main factor underlying such decisions is the magnitude of variability that exists at different experimental levels, for example, at the plot and subplot scale. Unfortunately, published information on the magnitudes of these sources of variability for soil GHG emissions remains scarce. This limits the utility of existing statistical tools in experimental planning and design. As we will show, knowing the approximate size of variability associated with each experimental unit (gas samples, chambers, plots) can be used to design a study with substantially increased statistical power.

In addition to a priori design decisions, the choice of statistical methods used to analyze collected data is also of crucial importance. The method most commonly used in GHG studies for comparing fluxes among studied treatments and time points is ANOVA. Two ANOVA assumptions are particularly important for optimal statistical performance: that collected samples are independent of one another and that variances are homogeneous. Both are routinely unmet by chamber-based GHG measurements.

First, in most field-based GHG experiments individual chambers within replicate plots are measured multiple times over a defined observation period. The consequence is that measurements originated from the same chamber are more similar to one another than to measurements from other chambers. Moreover, there is a temporal component to the variability: measurements that are closer in time are often more similar to one another than to measurements separated by longer time intervals. These autocorrelation problems are best solved with a repeated measures analysis, wherein relationships between measurements taken repeatedly on the same experimental unit are described by a statistical model that incorporates this knowledge. An additional potential benefit of a repeated measures analysis is more efficient statistical comparisons. Despite its advantages, however, repeated measures analyses are not common in GHG studies. Of 28 papers reporting temporal variations in N_2O fluxes published from 2008 to 2013 in three prominent soils journals that usually pay commendable attention to statistical rigor (*European Journal of Soil Science*, *Canadian Journal of Soil Science*, and *Soil Science Society of America Journal*), only seven mentioned or used repeated measures analysis.

Second, field-based GHG measurements often exhibit heterogeneous variances for measurements taken at different time points. This problem only occasionally arises for most soil and agronomic properties but is extremely common in GHG data. Traditional ANOVA assumes that the variability of the data collected at different time points remains the same (the assumption of homogeneous variances). Violation of this assumption can have a major detrimental effect on treatment comparisons. Thankfully, there is a solution that can also improve the efficiency of statistical comparisons: fitting the data with a statistical model that accounts for heterogeneous variances among time points. Again, unfortunately, this solution is insufficiently applied in most GHG studies.

An additional statistical issue for chamber-based GHG studies is the manner in which fluxes are interpolated from multiple headspace samples. Using theoretically based models vs. statistically based regression models (whether linear or nonlinear) can also affect statistical efficiency and power. We will not consider this issue here insofar as it has been amply and capably examined by a number of authors recently (e.g., Livingston et al., 2006; Venterea et al., 2009; Levy et al., 2011; Parkin et al., 2012). Here we calculate fluxes by fitting the headspace sample data with simple linear regression models, which remains the most commonly used flux rate determination method and which tends to provide higher accuracy in statistical comparisons among the studied treatments (Venterea et al., 2009; Levy et al., 2011; Parkin et al., 2012).

In this study we use data from three experiments that employed chamber-based methods to measure soil CO_2 and N_2O fluxes to (i) evaluate the relative sizes of different sources of random variability in observed fluxes, (ii) assess the relative benefits of increasing numbers of samples at different replication levels, (iii) illustrate the performance of repeated measures analysis and the analysis that accounts for heterogeneous variances, and (iv) explore under which circumstances the implementation of these statistical tools is particularly important.

MATERIALS AND METHODS

Considered Experimental Design

Our experimental design and statistical analysis considerations apply to field studies conducted at one or multiple experimental sites with a goal of comparing two or more ecosystems or experimental treatments for their effects on soil GHG fluxes over a particular time period. We assume field studies in which each treatment is applied to one or more randomly selected plots within each of the multiple experimental blocks of a randomized complete block design (RCBD). In each experimental plot there is one or more chambers used for GHG flux measurements. Measurements are conducted as follows: multiple headspace samples, typically 3 to 10, are drawn from the chamber at brief time intervals, and gas concentrations are measured to produce a flux estimated by regressing gas concentration vs. time (Fig. 1a). Chamber measurements are taken at multiple times during a season. The experimenters are interested in comparing both temporal trends within experimental treatments and differences among treatments on individual dates (Fig. 1b). While the examples that we will use in this paper are RCBD experiments, we would like to note that most of the discussed issues are not limited to RCBD and are equally applicable to any other experimental field design settings.

Note that we are not considering here the case of overall treatment differences for the entire period. For assessing the overall treatment differences, some cumulative measure of flux for the entire period should probably be used as a response variable. In this case, the response will not be affected by the temporal correlation problems, and heterogeneous variance among individual time points discussed here. Analysis of cumulative fluxes can be thus a somewhat simpler strategy that should be considered when the overall treatment differences and not the time trends and/or comparisons at different points in time are of main interest.

Data Sources

The data are from three experimental sites in Michigan: the Mason site in East Lansing at Michigan State University's (MSU) campus farm and two sites at MSU's Kellogg Biological Station (KBS) Long-term Ecological Research (LTER) site in Hickory Corners. One of the KBS sites along with the Mason site is part of the Sustainable Corn Experiment (SCE) and the other is the KBS LTER Main Cropping System Experiment (MCSE) (Robertson and Hamilton, 2014). Greenhouse gas measurements were taken at all three sites every 2 to 3 wk from early spring till November.

The Mason and KBS SCE sites are corn–soybean rotation experiments established in 2011. They are chisel plowed and receive conventional fertilizer and herbicide inputs. The two treatments used in this study are with and without a winter rye cover crop. In the cover crop treatment, winter rye is planted in fall after corn and soybean harvest and then terminated in spring approximately 2 wk before planting the subsequent corn–soybean crop. At these sites CO_2 and N_2O flux measurements were conducted by using a photoacoustic infrared spectrometer (PAS) (INNOVA Air Tech Instruments, Ballerup, Denmark) from chambers installed in individual plots, with one chamber per plot. The fluxes were assessed based on 5 to 7 gas measurement during 15 min of chamber closure following the procedures described by Iqbal and Parkin (2013). Both sites are RCBD experiments with six replications. Blocking at these two sites is driven strictly by topography, with two blocks at each site located in depression (toeslope) areas, two at side-slope locations with slopes of 3 to 6°, and two at flat terrain summits.

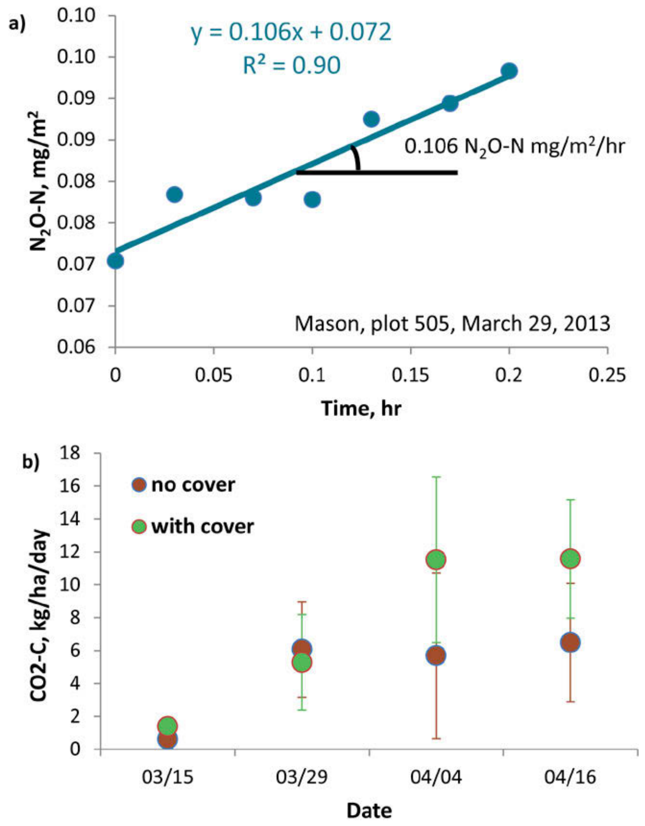

Fig. 1. Illustration of the data collected in the field experiments: (a) N_2O concentrations measured in seven gas samples collected with 2-min interval and used for estimating N_2O flux rate by fitting a linear regression equation to the data and (b) average CO_2 flux rates from four time points during the measurement season for two studied treatments (named "no cover" and "with cover"); error bars represent standard errors.

The KBS MCSE site was established in 1989 with 11 treatments, 7 of which were used in this study. The treatments that we used include four corn–soybean–winter wheat rotations and three perennial systems. The annual crop treatments include (i) a conventionally managed chisel-plowed system; (ii) a no-till system otherwise similar to the conventional system; (iii) a reduced input system with one-third of the chemical inputs, mechanical weed control, and leguminous cover crops; and (iv) a biologically based system with no chemical inputs otherwise similar to the reduced input system. The perennial system treatments include (v) continuous alfalfa, (vi) trees of a Populus clone on a 10-yr rotation cycle, and (vii) an early successional treatment, abandoned after spring plowing in 1989 and annually burned from 1996. The experimental treatments were arranged in RCBD with four replicate blocks. All of the farming operations at the study plots were performed by commercial-sized equipment similar to that used by local farmers. Detailed experimental design and agronomic details are available in Robertson and Hamilton (2014) (http://lter.kbs.msu.edu). At this site CO_2 and N_2O fluxes were measured by withdrawing four 10-mL headspace samples over a 1-h chamber closure period that were placed in 5-mL glass vials delivered to the lab where N_2O concentrations were analyzed using an electron capture detector-equipped gas chromatograph (7890A, Agilent

Technologies, Santa Clara, CA), and CO_2 concentrations were analyzed using an infrared gas absorption analyzer (LI-820 CO_2 analyzer, LI-COR, Lincoln, NE). We calculated fluxes as the change in gas concentration in the headspace of the chamber over the deployment time (ppb min^{-1}).

For the purpose of this study, we used only a portion of the GHG data collected at these sites. For assessments of variability due to blocking and the factors affecting it, and for illustrating repeated measures and heterogeneous analysis methods, we used 2011 data from the KBS MCSE site and 2012 data from the Mason and KBS SCE sites. To evaluate sources of random variability in GHG emissions, we used data from August to November 2013 at Mason when two chambers were measured per plot.

Statistical Analyses

Different sources of variability were quantified by estimating respective variances by using the restricted maximum likelihood approach (Milliken and Johnson, 2009). Effects of site, multiple time points, blocks, and plots nested within blocks were treated as random. The size of the chamber effect in 2013 Mason data was quantified by adding chambers nested within plots to the statistical model. The analyses were performed using PROC MIXED (SAS Institute, 2009).

The presence of temporal correlations was first assessed by calculating and plotting sample temporal variograms of the flux data by using PROC VARIOG (SAS Institute, 2009). Then statistical models that account for temporal correlation, including autoregressive, spherical, and exponential functions, were fitted to the data sets using REPEATED option of PROC MIXED. The Aikaike Information Criterion (AIC) was used to compare model performances. Temporal correlation was regarded to be present when AIC of at least one of the models that accounted for temporal correlation was better than AIC from the model with uncorrelated errors.

Power analysis was conducted by using procedures described by Stroup (2002) as outlined in Kravchenko and Robertson (2011). Note that while all available data from all treatments were used in estimating variances, in all power calculations it was assumed that only two treatments were present in the study. Variance estimates from Mason 2013 data with two chambers per plot were used for power calculations for scenarios with different numbers of replicated plots and numbers of chambers per plot, and for evaluating the role of multiple chambers per plot in increasing the power of detecting treatment differences.

For estimating the relative effects of different replication scenarios for increasing statistical power, we selected hypothesized differences between treatments for power calculations so as to generate relatively large power values. For most CO_2 examples, the hypothesized difference between the two treatments was set to be equal to 50% of the mean value. Because of much higher variability of N_2O data to obtain relatively large power values, we had to use much greater hypothesized differences between the treatments. Thus, for N_2O examples the hypothesized difference had to be set to one treatment's mean being as much as 5 to 6 times bigger than the other. Probability of Type I error of 0.05 was used for all power calculations.

We used N_2O data from the Mason site to assess the effect of the number of gas concentration measurements on the estimation accuracy of the flux rate and the resulting power of the statistical comparisons between treatments. The original data consisted of seven concentration measurements. From each chamber's 7-point data set, we selected sets of three, four, five, and six measurements and used these for flux rate estimation and power analyses. From each chamber we averaged the results from three possible combinations of randomly selected 3- and 4-point data and two combinations of 5- and 6-point data.

To illustrate the effects that the choice of data analysis, that is, with or without accounting for heterogeneous variances and temporal correlations, can have on the study results, we reported comparisons of CO_2 and N_2O data analyses using five different methods: (i) the analysis of each time point separately and analyses of the entire data set by using (ii) homogeneous and (iii) heterogeneous variance analysis methods and (iv) methods that do not and (v) do account for the presence of temporal correlations. Note that since these comparisons were conducted for illustration purpose only, in order not to burden the results and discussion with unnecessary details from the specific data sets, here we considered only four types of variance–covariance structure models. Specifically, a compound symmetry model that does not account for temporal correlation and an autoregressive model that does account for temporal correlation; both of them were used in their homogeneous and heterogeneous variance versions. Note that for the real data analysis, testing performance of a large group of models (e.g., those listed early) is strongly advised. Repeated measures analyses were conducted following procedures outlined by Littell et al. (2006) with the PROC MIXED procedure of SAS. Denominator degrees of freedom were calculated using Kenward–Roger's approximation method (Kenward and Roger, 1997). Selection of best performing structure was conducted based on AIC values as above. As an additional criterion for assessing the efficiency of data analyses, we reported average standard errors for the differences of interest, that is, either differences between selected treatments within the same week or differences between different weeks of the same treatment (Zimmerman and Harville, 1991; Brownie et al., 1993).

The unequal variance assumption was tested using Levene's test (Milliken and Johnson, 2009). Residuals were checked for deviations from normality, and when found to be heavily skewed the data were transformed as necessary to reduce skewness (specific transformation used in each case are listed in Table 1).

RESULTS AND DISCUSSION

Relative Sizes of Different Variability Sources

The sizes of variability due to studied sources were noticeably different for CO_2 and N_2O (Fig. 2a). For CO_2 the spatial components of the variability, such as site, topography (as represented by blocks), and plot constituted almost 40% of the total variability, whereas for N_2O these components were less than 4%. Alternatively, temporal components of variability were almost twice as large for N_2O as for CO_2, that is, 21% of total variability for N_2O vs. 12% for CO_2. Please note that because of relatively low frequency of data collection used in our study sites some high flux events were likely missed, and the magnitude of the temporal variability for CO_2 and, especially, for N_2O data was likely underestimated. For both gases, however, the main component of the overall variability was the residual variance, equal to 51 and 75% for CO_2 and N_2O, respectively. The residual variance in these experimental settings reflects combined variance due to individual characteristics of chambers installed within plots and variance due to differences in performances of the individual chambers at different points in time; in other words, residual variance reflects a combination of small-scale spatiotemporal variability and measurement errors.

The main trend with respect to chamber-to-chamber variability within the same plot (Fig. 2b) is that variability due to chamber was much greater than the variability due to any other experimental source. For CO_2, the chamber effect was about 10% of the overall variability, whereas the residual, which is variability in different chambers at different time points, constituted approximately 45% of the overall variability. For N_2O, chamber related sources of variability (chamber effect and residual) constituted more than 95% of the overall variability. These results suggest that the lion's share of variability for N_2O, and to a lesser extent for CO_2, is spatiotemporal variations occurring at a scale of less than 1 m^2. Medium scale differences occurring at a scale of experimental plots (several square meters) add very little to the overall variability as compared with chambers. Large scale differences due to the topographic gradient at approximately 100 m^2 noticeably contribute to variability for CO_2, but their influence on N_2O is minor, as compared with the small-scale chamber-based variability.

The assessments of the variability sources are consistent with what is known about N_2O production in soil, which is a complex mixture of multiple processes operating in soil typically in a “hot spot” and/or ”hot moment” fashion (McClain et al., 2003; Groffman et al., 2009), where, when the conditions are right, a very large surge of N_2O flux can occur from a very small space in a very short period of time (e.g., Smith et al., 1998; Castellano et al., 2010). Emissions of CO_2, on the other hand, are more influenced by overall soil properties such as soil organic matter, soil structure, and bulk density and, additionally, are produced by a much greater diversity of microbial heterotrophs more broadly adapted to different environmental conditions (Levine et al., 2011). These soil characteristics are stable in time and vary across space substantially, but in a predictable manner.

Even though the contribution of blocks to the overall variance was relatively small, blocking by topography in both Mason and KBS SCE sites in all seasons was justified (data not shown). Moreover, analysis of data from individual time points

Table 1. Summary of information on heterogeneity of variances and temporal correlations along with average standard errors for the difference between two treatment means obtained by analyzing CO_2 and N_2O flux rate data by using five different methods.†

Experiment Size	Unequal Variances by Time–Temporal Correlation	Method	Average Standard Error for the Difference Between Two Treatments
		Large (24 plots)	
KBS MCSE-CO_2	Yes/Yes	By week	3.88
		Split-plot	4.24
		Ar(1)	4.22
		CSH	3.95
		Arh(1)	3.91
KBS MCSE-N_2O (double log-transformed)	No/Yes	By week	0.24
		Split-plot	0.25
		Ar(1)	0.25
		CSH	0.24
		Arh(1)	0.24
		Small (12 plots)	
Mason-CO_2	Yes/Yes	Split-plot	30.3
		Ar(1)	30.2
		CSH	29.2
		Arh(1)	28.3
Mason-N_2O (square root transformed)	Yes/No	Split-plot	0.89
		Ar(1)	0.89
		CSH	0.51
		Arh(1)	0.70
KBS SCE-CO_2	Yes/No	Split-plot	25.4
		Ar(1)	25.5
		CSH	22.4
		Arh(1)	24.3
KBS SCE-N_2O (square root transformed)	Yes/No	Split-plot	0.75
		Ar(1)	0.76
		CSH	0.50
		Arh(1)	0.50

† Methods include (i) the analysis of each time point separately (by week) (ii) and analyses of the entire data set using (iii) homogeneous (split-plot and Ar[1]) and heterogeneous (CSH and Arh[1]) variance analysis methods, and methods that (iv) do not (split-plot and CSH) and (v) do (Ar[1] and Arh[1]) account for the presence of temporal correlations. KBS, Kellogg Biological Station; MCSE, Main Cropping System Experiment; SCE, Sustainable Corn Experiment. Ar, autoregressive structure with homogeneous variances; CSH, compound symmetry with heterogeneous variances; Arh, autoregressive structures with heterogeneous variances.

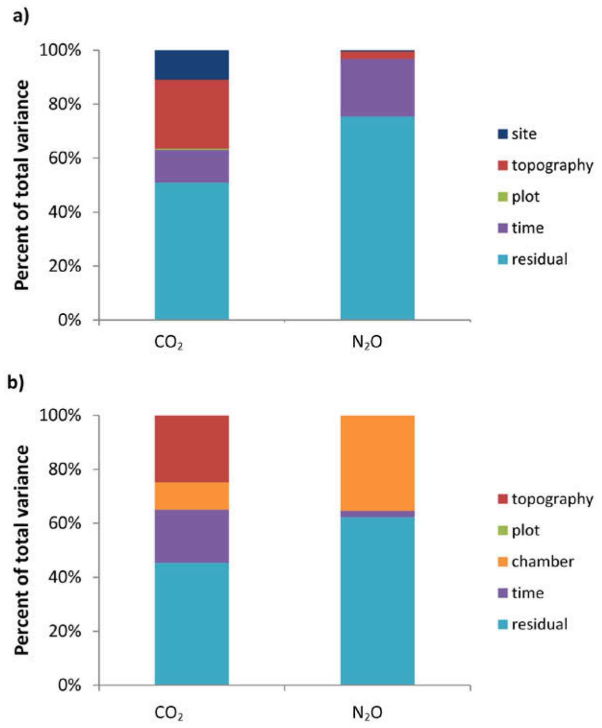

Fig. 2. Percent of total variance in CO_2 and N_2O flux rates attributed to different sources of variability in (a) combined Mason and Kellogg Biological Station (KBS) Sustainable Corn Experiment (SCE) site data of 2012 with one chamber per plot and (b) in Mason SCE site August to November 2013 data with two chambers per plot.

demonstrated that the contribution of the topographic blocks to the overall variance fluctuated substantially in time, ranging from zero to 60% for different dates for CO_2 and zero to 90% for N_2O (Fig. 3). Soil moisture appeared to be the factor that influenced the relative contribution of topography to the overall variability of N_2O and, to a somewhat lower extent, to the overall variability of CO_2. More specifically, the relative contribution of topography increased with decreasing soil moisture (Fig. 3), indicating that topographic influences on GHG emissions were more pronounced in drier soil. This influence probably results from soil moisture redistribution along the landscape, with stronger patterns along the topographical gradient under dry conditions, producing greater variability in GHG emissions. This relationship is weaker for CO_2 than for N_2O, reflecting that the influence of water for CO_2 production, while important, is not as dramatic as for N_2O. This result is consistent with the observations of Morris et al. (2013) who reported topography-driven differences in soil moisture levels to be one of the drivers of the spatial patterns in N_2O emission data.

A practical implication of this observation is that topography can potentially provide useful auxiliary information for mapping GHG emissions at the field scale. However, because of extremely high small-scale variability in GHG emissions data, especially N_2O, it is probably unrealistic to expect high accuracy from soil GHG emission maps. Nevertheless, topographic influence on GHG emissions holds substantial continuity within the landscape, and thus can be used to better map and predict GHG fluxes, especially in drier soils.

Sample Numbers at Different Replication Levels and Their Influence on Statistical Power

Since power calculations are very sensitive to variance estimate values, to specific experimental settings, and to hypothesized differences, determination of optimal sample numbers is best conducted for every study individually. The sample numbers obtained in a course of power calculation presented here are not intended for direct use in other studies. The purpose of this section is rather to demonstrate relative potential benefits that could be achieved in a typical GHG experiment.

For a GHG study, will there be more benefit to adding more plots or more chambers per plot? Let us assume that during a single measurement session, GHG flux measurements can be collected from a total of 18 chambers in each treatment. We consider four ways in which these 18 chambers could be experimentally organized: (i) 18 replicated plots with 1 chamber per plot, (ii) 9 replicated plots with 2 chambers in each, (iii) 6 replicated plots with 3 chambers per plot, and (iv) 3 replicated plots with 6 chambers per plot. Power values for these four combinations strongly point toward the benefit of having

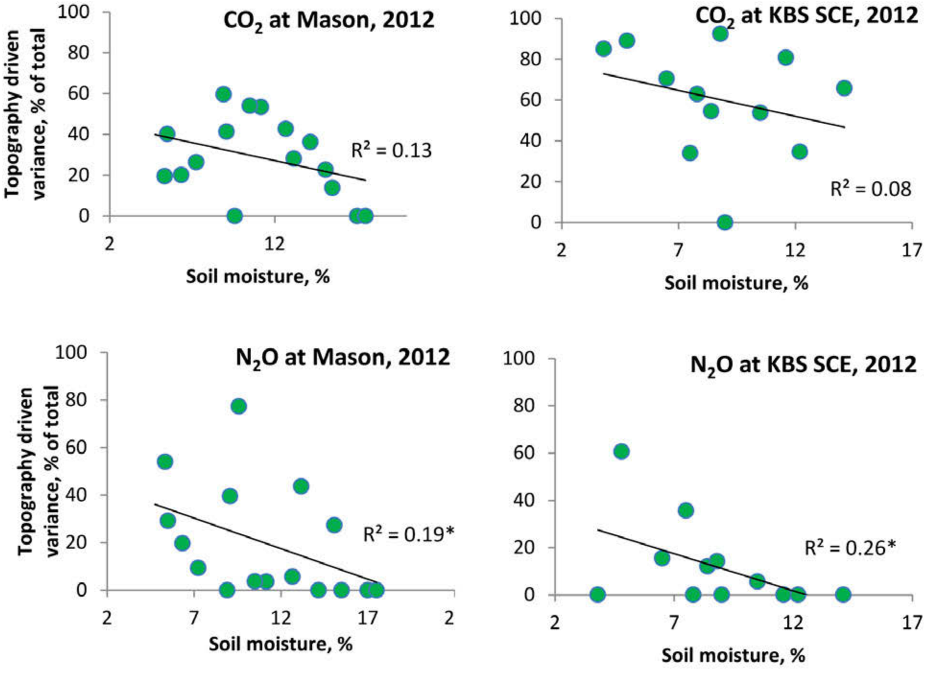

Fig. 3. Relationships between percent of total variance due to topographic blocks and soil moisture in CO_2 and N_2O flux rate data from Mason and KBS SCE sites in 2012. *Correlations statistically significant at $P < 0.1$.

more replicated plots as opposed to more chambers per plot (Table 2). With equal numbers of field measurements, that is, 18 per treatment, the design will have a statistical power of 88% if researchers choose to have 18 replicated plots with 1 chamber per each plot as opposed to a power of just 25% if there are 3 replicated plots with 6 chambers per plot.

While this result unequivocally points to a greater statistical benefit of more plots per treatment rather than more chambers per plot, this is true only when the number of replicate plots is not limiting. Once the maximum number of plots are sampled there can additionally be substantial benefit to additional chambers per plot, especially where, as is the case for GHG fluxes in our data, there is high variability associated with the individual chambers. Figure 4a illustrates the increase in power values that can be achieved by increasing the number of chambers in an experiment with six replicated plots per treatment. As mentioned earlier in Materials and Methods, the hypothesized scenarios of differences between two treatments used in power calculations here are 50% for CO_2 and 6 times for N_2O. A field study with six replications generally would be considered as a well replicated study, but given the observed variability of CO_2 and N_2O fluxes, with only one chamber per plot such a study will produce power values of only ~35% for these two hypothesized CO_2 and N_2O scenarios. With three chambers per plot the power increases to 60% and then with six chambers per plot to a more reasonable 75%. Note in Fig. 4a that the rate of increase in power with an increasing number of chambers is greater for N_2O than for CO_2. This reflects a greater chamber-to-chamber variability and greater contribution of chambers to the overall variability for N_2O as opposed to CO_2. In essence, we would expect greater benefits from increasing the numbers of chambers for N_2O flux measurements as compared with CO_2.

Yet another component that is not directly featured in the statistical analysis of GHG data, but indirectly can have a substantial effect on its power, is the number of gas samples taken to obtain the flux rate value from each chamber (Fig. 1a). The influence of the number of gas samples on the subsequent statistical comparisons among treatments stems from the accuracy in the estimation of the flux rate. The higher the number of data points the more accurately the flux rate will be estimated. When the resulting flux rates are used in further statistical analyses, the lower estimation accuracy manifests itself as greater variability of the flux rate data, thus as higher error variance values (Fig. 4b). We obtained the largest error variance when only three gas samples were used for flux estimation, producing statistical comparisons among the treatments with only 50% power. With the same numbers of plots and chambers but simply increasing the number of gas measurements to seven, the power could have been increased to as much as 75%. Levy et al. (2011) also noted the potential to considerably increase the flux estimation accuracy by increasing the number of gas concentration measurements.

Until recently, technical difficulties and measurement expenses severely limited the number of gas samples that could be taken for GHG flux determination. However, the emergence of new instrument devices such as PASs and quantum cascade lasers (I. Gelfand et al., personal communication, 2014) allow multiple gas concentration measurements per minute, thus making 5 to 10 measurements for each flux determination an easy possibility (Adviento-Borbe et al., 2007; Iqbal and Parkin, 2013). Our example demonstrates that when an opportunity for increasing the number of gas concentration measurements exists, researchers should take advantage of it.

Table 2. Illustration of the effect of the number of true replications, that is, experimental plots, and the number of chambers, that is, subsamples, on power of the statistical analysis.†

Combination		
Plots	Chambers per plot	Power %
18	1	88
9	2	75
6	3	59
3	6	25

† The illustration is for the scenario of 18 chambers per treatment. Power is calculated based on the variance estimates from the 2013 Mason site data for CO_2 assuming a 50% difference between the mean CO_2 flux rates of two studied treatments with α of 0.05.

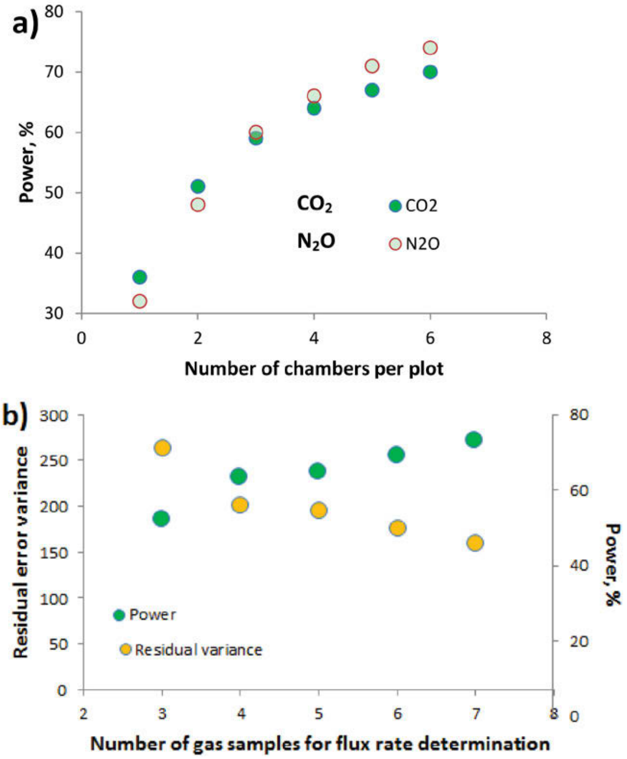

Fig. 4. (a) Illustrations of how the statistical power for comparing differences in fluxes between two hypothetical treatments is affected by number of subsample chambers per experimental plot for CO_2 and N_2O fluxes and (b) number of gas samples taken for flux determination per chamber for N_2O data. For the illustration in Fig. 4a we hypothesize a 50% difference between the two treatments for CO_2 and a sixfold difference between the two treatments for N_2O; the calculations were performed for a scenario with six replicated plots. For the illustration in Fig. 4b we hypothesized a sixfold difference between the two treatments for N_2O; the calculations were performed for a scenario with six replicated plots and one chamber per plot.

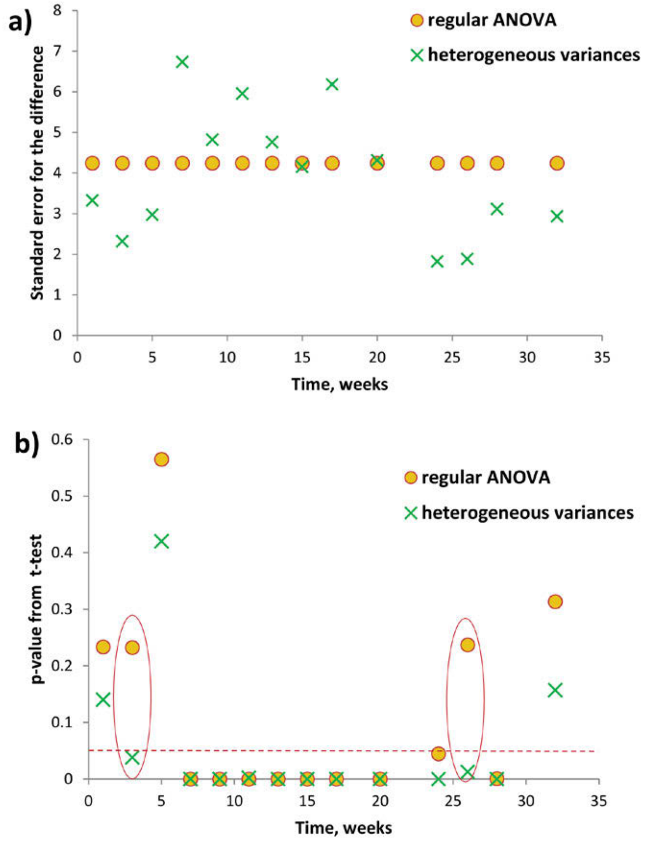

Fig. 5. (a) Standard errors and (b) *P* value from *t* tests for the difference between two treatments at different time points during the growing season of 2011 for CO_2 flux rates from the Main Cropping System Experiment (MCSE) site. Values obtained with regular ANOVA and with heterogeneous variance analyses are shown. Red dashed line on Fig. 5b marks the *P* value of 0.05; red circles mark the instances when regular ANOVA and heterogeneous variance analysis would result in different conclusions regarding treatment effects.

Performance of Unequal Variance and Repeated Measures Analyses

Heterogeneous variances

Unequal variances were present for CO_2 and N_2O data in all three of our sites (Table 1). Double log-transforming the KBS MCSE data to normalize an extremely skewed data set also solved the heterogeneous variance problem. For the Mason and KBS SCE data, however, a square root transformation provided normality but did not resolve the problems with heterogeneous variances.

Figure 5 illustrates the peril of ignoring heterogeneous variances for CO_2 flux comparisons among KBS MCSE treatments. Standard errors for differences between two treatments (Fig. 5a) and *P* values for these comparisons (Fig. 5b) were obtained from a regular ANOVA assuming homogeneous variances vs. ANOVA with heterogeneous variances (SAS codes for these analyses are provided in the Supplemental Material). The homogeneous variance analysis resulted in a standard error for the differences equal to 4.2, the same across all time points (Fig. 5a). Standard errors obtained from heterogeneous variance analysis reflected heterogeneous differences in variability that occurred at different time points; standard errors correspondingly varied from 1.9 to 6.8. A high standard error for a difference between two treatments results in a high *P* value for the corresponding *t* test. As can be seen from Fig. 5b, in all cases when standard errors from particular time points tended to be lower, the analysis with heterogeneous variances produced a lower *P* value. At least in two instances in this data set, the lowering of the *P* value due to heterogeneous variance analysis was large enough to change the conclusion regarding statistical significance of the treatment effect. In other words, a regular ANOVA would have missed detecting the differences between these two treatments.

It is important to note that typically there are physical and biological reasons for heterogeneous variances in GHG flux data at different time points during the season. For example, certain management events such as tillage and fertilizing operations and certain environmental events such as heavy rains result in substantial increases in CO_2, and especially N_2O fluxes accompanied by concomitant increases in variability. These high variability events will result in greater standard errors for different dates, and using regular ANOVA will negatively influence our comparisons at time points when data variability is low. Because of the lower statistical power of such comparisons, differences among treatments will go undetected.

On the other hand, when the standard error for the regular ANOVA is lower than the standard errors for some of the time points with high variability, the regular ANOVA can lead to overoptimistic estimates of the accuracy of the data. The comparisons between treatments at those time points might come out as statistically significant because of the artificially low standard errors, which underestimates the actual high variability present in flux data at those time points. Thus, it is important to assess the heterogeneity of variances on a data set by data set basis and, when warranted, to apply a heterogeneous variance analysis. Detailed discussion of specific statistical tests for assessing the need for heterogeneous variance analyses and performance of these analyses in SAS can be found in Milliken and Johnson (2009).

Problems with heterogeneity of variances can arise even in analyses of cumulative GHG fluxes, as some experimental treatments can have consistently trivial flux values with relatively low variability, while cumulative fluxes in other treatments can be substantial with concomitant high variability. While analysis of cumulative flux data is beyond the scope of this paper, we would like to note that accounting for heterogeneous variances among the treatments in such cases can be an advantageous strategy.

Temporal Correlation

The presence and strength of temporal correlation varied from site to site in our study in response to soil and topographical characteristics as well as to sites' land use and management practices. For example, at the Mason site N_2O variogram values for topographic depressions substantially increased with increasing time lags, indicating the presence of temporal correlation, whereas variogram values for summits and slopes did not increase with time lags, indicating no temporal correlation, that is, variations in N_2O emissions were as large for values taken

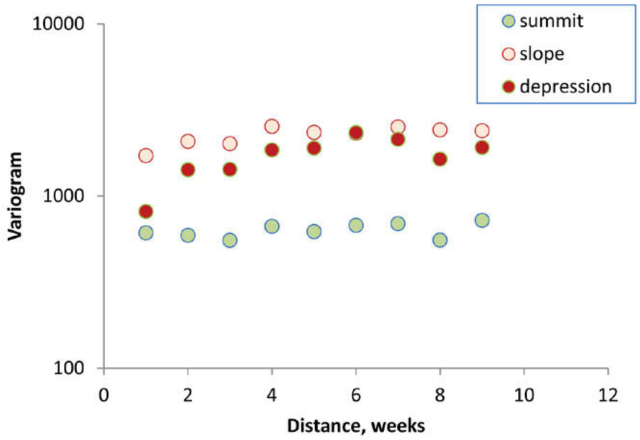

Fig. 6. Sample temporal variograms for N_2O emission data obtained from three different topographical positions of Mason site.

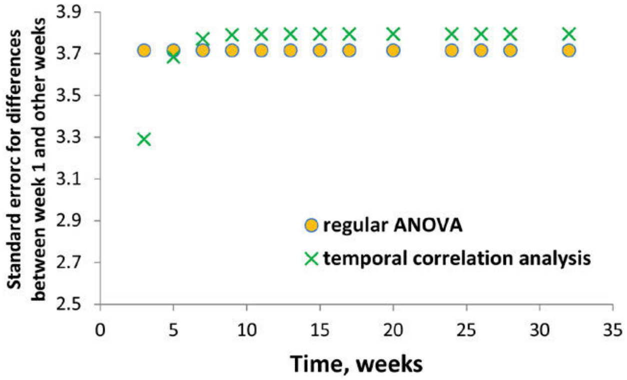

Fig. 7. Illustration of the effect of modeling temporal correlation on the standard errors for the differences between measurements at different time points. The results shown are from one of the treatments of the MCSE site for comparisons of the CO_2 emission during Week 1 with CO_2 emissions of Weeks 2 to 32.

close to one another in time as for those taken several weeks apart (Fig. 6). Temporal correlations for CO_2 fluxes were typically stronger than those for N_2O (data not shown), reflecting a well-known tendency for high temporal variability in N_2O emissions in response to rainfall events and management operations.

Taking into account temporal correlation via repeated measures analysis had a small but consistent effect of reducing the standard errors for differences in those cases when temporal correlation was present (Table 1). However, it should be kept in mind that since temporal correlation occurs in the data that originated from the same experimental plots at different time points, the strength of temporal correlation does not directly influence comparisons between treatments. Nevertheless, accounting for temporal correlation brings an indirect benefit of potentially lower estimates of plot variances, which then can at least somewhat reduce the standard errors for the comparisons among the treatments as seen in Table 1.

On the other hand, repeated measures analysis can have a considerable influence on the results of comparisons among different time points within the same treatment. Specifically, when positive temporal correlation is present, repeated measures analysis will produce lower standard errors and thus higher chances of detecting statistically significant differences among the time points that are close in time (Fig. 7). Ignoring temporal correlation would lead to comparisons in time that are based on standard errors that are too large for points close in time, while using standard errors that are too optimistically low for comparisons that are far apart. Please also note that in a study with greater sampling frequencies and shorter temporal intervals between sampling times, the strength of temporal correlations will likely be higher than the ones observed in this study. Therefore, the discussed above benefits of accounting for temporal correlations are likely to be only higher than those reported here.

Recommended Best Practices for Data Analysis

Figure 8 outlines a set of decisions that one might use to identify an optimal data analysis strategy when analyzing chamber-based GHG emissions data in field experiments. The tree is focused on comparisons among studied treatments at individual time points since these comparisons are typically the ones of most interest in field studies with multiple treatments. Note that for treatment comparisons of cumulative fluxes, neither heterogeneous variances at different time points nor temporal correlations between time points discussed here are relevant; however, judicious accounting for heterogeneous variances among different treatments or experimental sites can still lead to more efficient treatment comparisons.

Size of the experiment is the first factor to consider in deciding on the most appropriate route of data analysis. Here experiment size refers to the actual number of experimental plots or, more broadly, to the number of experimental units used in the study, not to how many times GHG measurements were collected. The size of the experiment will first of all affect the numbers of degrees of freedom for the error terms and comparisons. Thus, it will directly influence the power of the analyses and will be a deciding factor in whether or not one should analyze GHG data as a single entire-season data set or consider an analysis separately by individual time points. Moreover, the size of the experiment will affect the accuracy estimation of variance components and will define how complex a statistical model could be fit to the data, that is, how detailed can heterogeneous variances and temporal correlation components be represented in the model.

Overall, the steps of the decision tree are self-explanatory, and for each of the green boxes of the suggested analyses we have provided sample SAS codes in the Supplemental Material. The most important step in the process is to determine whether heterogeneous variances and temporal correlations are present. Once this is known, then the next step is to find an appropriate statistical model to account for the heterogeneous variances as needed and to best describe the temporal correlation as needed. If variances are found to be homogeneous and temporal correlation absent, then the analysis becomes a simple and straightforward split-plot analysis with time points as a subplot factor.

Nevertheless, additional discussion is needed for two of the suggested analyses. First, in a large experiment (one with

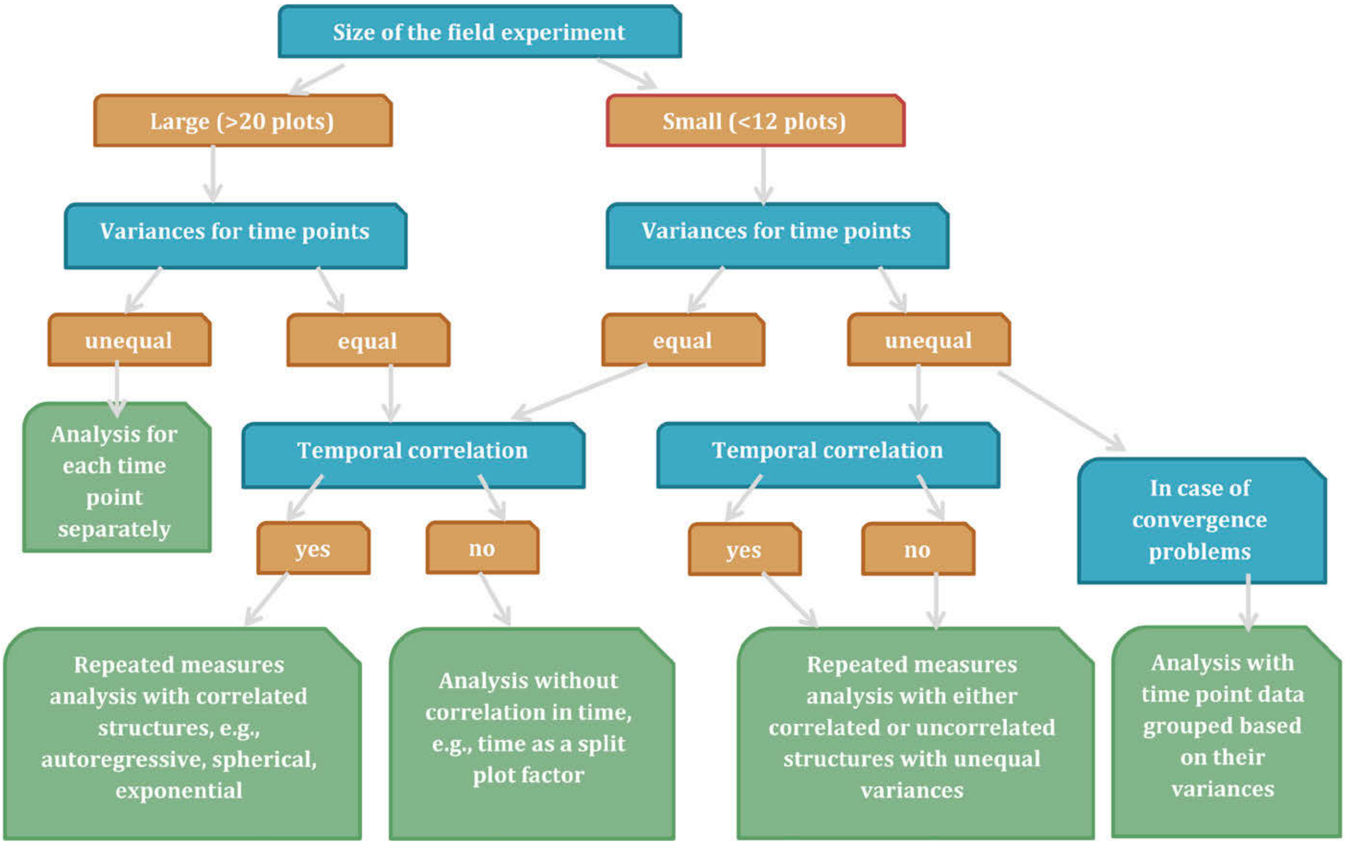
Size of the field experiment
Large (>20 plots)
Small (<12 plots)
Variances for time points
Variances for time points
unequal
equal
equal
unequal
Analysis for each time point separately
Temporal correlation
Temporal correlation
In case of convergence problems
yes
no
yes
no
Repeated measures analysis with correlated structures, e.g., autoregressive, spherical, exponential
Analysis without correlation in time, e.g., time as a split plot factor
Repeated measures analysis with either correlated or uncorrelated structures with unequal variances
Analysis with time point data grouped based on their variances

Fig. 8. Outline of the decision process in selecting an optimal statistical analysis of greenhouse gas flux data when comparisons between treatments at individual time points are of primary interest. Plots refer to the total number of plots in the experiment (across all treatments and blocks). Sample SAS codes for conducting analyses from each green box step are provided in the Supplemental Material.

more than 20 or so plots across all treatments and blocks) with heterogeneous variances, we recommend that the analysis of data from each individual time point be conducted separately, as the easiest and most powerful approach to addressing heterogeneous variances. This is probably welcome news for practitioners since conducting repeated measures analysis can be nontrivial even for an experienced analyst. Unequal variances among time points is not a problem when data from every time point are analyzed separately. In large data sets the numbers of degrees of freedom will be sufficiently high for conducting reliable comparisons among treatments in individual time point data sets. However, a potential for inflation of the probability of a type I error while conducting data analyses at individual time points should not be overlooked. In the analysis of the entire data set, the *F* tests for main effects and interactions provide some protection against the inflation. Since such protection is obviously lacking in the analyses of individual time points, some means of controlling experimentwise error rates should be employed (e.g., Westfall et al., 1999).

A question might arise as to why we do not recommend conducting data analysis by individual time points when variances are equal, or why not always do it when the size of the experiment is relatively large. The reason is that when the entire season's data set is analyzed by using an equal variance approach, it produces degrees of freedom for error and for comparisons that are considerably higher than those in

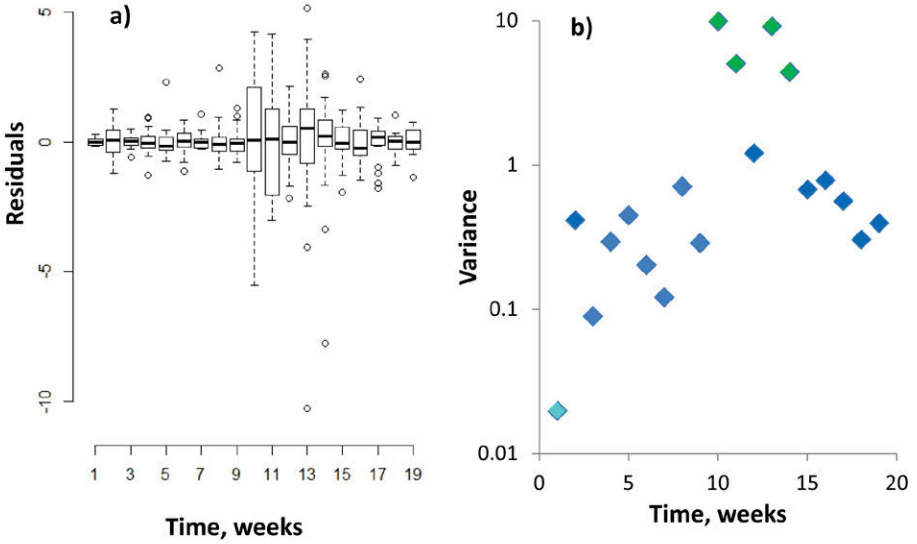

Fig. 9. Side-by-side box plots of the residuals from log-transformed 2012 N_2O data of Mason site at the studied time points (a) and variances of the residuals at the studied time points (b) with three colors indicating the three groups that were used to combine the variances in the heterogeneous variance analysis.

the individual time point analyses, thus providing higher statistical power. However, this benefit almost disappears in the analysis with heterogeneous variances where degrees of freedom are reduced by fitting a complex heterogeneous variance model to such an extent that they may not be much larger than the degrees of freedom obtained in individual time point analyses. However, analysis by individual time points is not recommended for small experiments (one with less than 12 or so plots across all treatments and blocks) because of their low degrees of freedom. In heterogeneous variance analysis of small experiments, even a slight increase in the degrees of freedom can go a long way toward increasing the power of an analysis. Note that the individual time point analysis is only appropriate where there is no interest in comparisons among different time points within the same treatments. For experiments with intermediate number of plots (12–20), either path of data analysis can be taken, but it probably better to explore the more conservative "small experiment" route first.

A potential problem with conducting analyses with unequal variances is an occasional difficulty for the complex heterogeneous variance model to converge; this is especially a problem with relatively small experiments. Besides multiple tools for improving convergence as discussed in SAS's PROC MIXED help feature, a possible solution for this problem is to group the data from time points with similar variances. This will result in a statistical model with fewer parameters to fit but with more data available for estimating each variance parameter. Such a model will have a higher chance of convergence than the model with individual variances for every time point, while still providing the benefits of the heterogeneous variance analysis.

The 2012 N_2O data from the Mason site (Fig. 9) illustrates this: variances from different time points range across three orders of magnitude, from 0.015 to ~15. Thus, analysis with heterogeneous variances is highly advisable. However, fitting a statistical model with 19 separate variances in this relatively small experiment is not possible. Thus, we put time points into three groups, as indicated in Fig. 9, and conducted an analysis with heterogeneous variances by the groups. The analysis produced an average standard error for the differences between the treatments much lower than those for the analysis with equal variances, that is, 0.77 compared with 0.89.

Conclusion

The statistical analysis of chamber-based GHG data requires care to avoid missing significant treatment differences or overstating insignificant differences. In the field experiments examined, the combination of temporal and spatial small-scale (within plot) variability dominated other sources; spatial variability was important at larger scales especially for CO_2 fluxes, and temporal variability was important especially for N_2O fluxes. Topographic sources of variability, though small, are significant for both gases, especially in drier soil; thus, blocking by topography can be important. While increasing the number of treatment replicates (plots per treatment) is a common and important strategy for increasing statistical power, in GHG analyses increasing the intensity of subsampling (chambers per plot and headspace measurements per flux determination) can also bring notable increases in power. Judicious application of repeated measures analysis and accounting for heterogeneous variances is crucial for the efficient analysis of GHG data. Use of a logical decision tree can provide valuable guidance for maximizing the power of statistical analysis of chamber-based GHG data.

ACKNOWLEDGMENTS

Support for this research was provided in parts by USDA-NIFA award No. 2011-68002-30190 Cropping Systems Coordinated Agricultural Project (CAP), by the U.S. National Science Foundation LTER Program (DEB 1027253), by the U.S. Department of Energy Great Lakes Bioenergy Research Center (DE-FCO2-07ER64494), and by MSU AgBioResearch. the authors thank Richard Price, Wakene Chewaka, and undergraduate student assistans for collecting GHG data at SCE sites and KBS LTER staff for conducting GHG measurements at MCSE site.

REFERENCES

Adviento-Borbe, M.A.A., M.L. Haddix, D.L. Binder, D.T. Walters, and A. Dobermann. 2007. Soil greenhouse gas fluxes and global warming potential in four high-yielding maize systems. Glob. Change Biol. 13:1972–1988. doi:10.1111/j.1365-2486.2007.01421.x

Brownie, C., D.T. Bowman, and J.W. Burton. 1993. Estimating spatial variation in analysis of data from yield trials, a comparison of methods. Agron. J. 85:1244–1253. doi:10.2134/agronj1993.00021962008500060028x

Castellano, M.J., J.P. Schmidt, J.P. Kaye, C. Walker, C.B. Graham, H. Lin, and C.J. Dell. 2010. Hydrological and biogeochemical controls on the timing and magnitude of nitrous oxide flux across an agricultural landscape. Glob. Change Biol. 16:2711–2720. doi:10.1111/j.1365-2486.2009.02116.x

Cochran, W.G., and G.M. Cox. 1957. Experimental designs, 2nd ed. Wiley, New York.

Cox, D.R. 1958. Planning of experiments. Wiley, New York.

Groffman, P.M., K. Butterbach-Bahl, and R.W. Fulweiler. 2009. Challenges to incorporating spatially and temporally explicit phenomena (hotspots and hot moments) in denitrification models. Biogeochemistry 93:49–77. doi:10.1007/s10533-008-9277-5

Hutchinson, G.L., and G.P. Livingston. 2002. Soil-atmosphere gas exchange. In: J.H. Dane and G.C. Topp, editors, Methods of soil analysis. Part 4. SSSA Book Ser. 5. SSSA, Madison, WI. p. 1159–1182.

Iqbal, J.M.J.C., and T.B. Parkin. 2013. Evaluation of photoacoustic infrared spectroscopy for simultaneous measurement of N2O and CO2 gas concentrations and fluxes at the soil surface. Glob. Change Biol. 19:327–336. doi:10.1111/gcb.12021

Kenward, M.G., and J.H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53:983–997. doi:10.2307/2533558

Kravchenko, A.N., and G.P. Robertson. 2011. Whole profile soil carbon stocks: The danger of assuming too much from knowing too little. Soil Sci. Soc. Am. J. 75:235–240. doi:10.2136/sssaj2010.0076

Kuehl, R.O. 2000. Design of experiments: Statistical principles of research design. Duxbury/Thomson Learning, Pacific Grove, CA.

Levine, U., T.K. Teal, G.P. Robertson, and T.M. Schmidt. 2011. Agriculture's impact on microbial diversity and associated fluxes of carbon dioxide and methane. ISME J. 5:1683–1691. doi:10.1038/ismej.2011.40

Levy, P.E., A. Gray, S.R. Leeson, J. Gaiawyn, M.P.C. Kelly, M.D.A. Cooper, K.J. Dinsmore, S.K. Jones, and L.J. Sheppard. 2011. Quantification of uncertainty in trace gas fluxes measured by the static chamber method. Eur. J. Soil Sci. 62:811–821. doi:10.1111/j.1365-2389.2011.01403.x

Littell, R.C., G.A. Milliken, W.W. Stroup, R.D. Wolfinger, and O. Schabenberber. 2006. SAS system for mixed models. 2nd ed. SAS Institute, Cary, NC.

Livingston, G.P., G.L. Hutchinson, and K. Spartalian. 2006. Trace gas emission in chambers: A non-steady-state diffusion model. Soil Sci. Soc. Am. J. 70:1459–1469. doi:10.2136/sssaj2005.0322

Milliken, G.A., and D.E. Johnson. 2009. Analysis of messy data, Vol. 1: Designed experiments. 2nd ed. Chapman & Hall/CRC Press, Boca Raton, FL.

McClain, M.E., E.W. Boyer, and C.L. Dent. 2003. Biogeochemical hot spots and hot moments at the interface of terrestrial and aquatic ecosystems. Ecosystems 6:301–312. doi:10.1007/s10021-003-0161-9

Morris, S.G., S.W.L. Kimber, P. Grace, and L. Van Zwieten. 2013. Improving the statistical preparation for measuring soil N_2O flux by closed chamber. Sci. Total Environ. 465:166–172. doi:10.1016/j.scitotenv.2013.02.032

Parkin, T.B., R.T. Venterea, and S.K. Hargreaves. 2012. Calculating the detection limits of chamber-based soil greenhouse gas flux measurements. J. Environ. Qual. 41:705–715. doi:10.2134/jeq2011.0394

Robertson, G. P., and S. K. Hamilton. 2014. Conceptual and experimental approaches to long-term ecological research at the Kellogg Biological Station. In: S. K. Hamilton et al., editors. The ecology of agricultural ecosystems: Long-term research on the path to sustainability. Oxford Univ. Press, New York. (in press).

Rochette, P., and N. Bertrand. 2007. Soil-surface gas emissions. In: M. Carter and E.G. Gregorich, editors, Soil sampling and methods of analysis. 2nd ed. CRC Press, Boca Raton, FL. p. 851–861.

Rochette, P., and N.S. Eriksen-Hamel. 2008. Chamber measurements of soil nitrous oxide flux: Are absolute values reliable? Soil Sci. Soc. Am. J. 72:331–342. doi:10.2136/sssaj2007.0215

SAS Institute. 2009. SAS user's guide. Version 9.2. SAS Institute, Cary, NC.

Smith, K.A., P.E. Thomson, H. Clayton, I.P. McTaggart, and F. Conen. 1998. Effects of temperature, water content and nitrogen fertilization on emissions of nitrous oxide by soils. Atmos. Environ. 32:3301–3309. doi:10.1016/S1352-2310(97)00492-5

Smith, K.A., and F. Conen. 2004. Measurement of trace gases: I. Gas analysis, chamber methods, and related procedures. In: K.A. Smith and M.C. Cresser, editors, Soil and environmental analysis: Modern instrumental techniques. 3rd ed. Marcel Dekker, New York. p. 433–437.

Stroup, W.W. 2002. Power analysis based on spatial effects mixed models: A tool for comparing design and analysis strategies in the presence of spatial variability. J. Agric. Biol. Environ. Stat. 7:491–511. doi:10.1198/108571102780

Venterea, R.T., and J.M. Baker. 2008. Effects of soil physical nonuniformity on chamber-based gas flux estimates. Soil Sci. Soc. Am. J. 72:1410–1417. doi:10.2136/sssaj2008.0019

Venterea, R.T., K.A. Spokas, and J.M. Baker. 2009. Accuracy and precision analysis of chamber-based nitrous oxide gas flux estimates. Soil Sci. Soc. Am. J. 73:1087–1093. doi:10.2136/sssaj2008.0307

Westfall, P.H., R.D. Tobias, D. Rom, R.D. Wolfinger, and Y. Hochberg. 1999. Multiple comparisons and multiple tests using the SAS System. SAS Institute, Cary, NC.

Zimmerman, D.L., and D.A. Harville. 1991. A random field approach to the analyses of field-plot experiments and other spatial experiments. Biometrics 47:223–239. doi:10.2307/2532508

SUPPLEMENTAL MATERIAL

SAS codes for the analyses from the decision tree on Fig 8:

```
****************************************;
* Analysis for individual weeks;
proc sort data = ghg;
by week;
run;
proc mixed data = ghg;
by week;
class rep treatment;
model n2o = treatment/ddfm = kr;
random rep treatment*rep;
lsmeans treatment/pdiff;
run;
****************************************;
* Analysis as a split-plot;
proc mixed data = ghg;
class rep treatment week;
model n2o = treatment week treatment*week/ddfm = kr;
random rep treatment*rep;
lsmeans treatment*week/pdiff slice = week;
run;
****************************************;
* Analysis as a repeated measures data set, autoregressive variance-covariance structure (ar(1))
is used as an example;
proc mixed data = ghg;
class rep treatment week;
model n2o = treatment week treatment*week/ddfm = kr;
random rep treatment*rep;
repeated week/subject = treatment*rep type = ar(1);
lsmeans treatment*week/pdiff slice = week;
```

```
run;
****************************************;
* Analysis as a repeated measures data set with a variance-covariance structure that accounts for
unequal variances among individual weeks (arh(1));
proc mixed data = ghg;
class rep treatment week;
model n2o = treatment week treatment*week/ddfm = kr;
random rep treatment*rep;
repeated week/subject = treatment*rep type = arh(1);
lsmeans treatment*week/pdiff slice = week;
run;
****************************************;
* Analysis with time points grouped based on variances;
proc mixed data = ghg;
class rep treatment week groupweek;
model n2o = treatment week treatment*week/ddfm = kr;
random rep treatment*rep;
repeated/group = groupweek;
lsmeans treatment*week/pdiff slice = week;
run;
```