**Capturing Uncertainty in GHG Biogeochemical and Process Models:
A Path Forward for Ag Offset Protocols**

Goal of the white paper: Initiate discussion on how to assess uncertainty in biogeochemical process models for quantifying GHG offset projects. A draft of the paper will be presented to C-AGG (Coalition on Agricultural Greenhouse Gases) for discussion at July 20-21st meeting in Chicago. This is designed to be a living document to be discussed and revised by the C-AGG community. NOTE: This is a very rough and incomplete draft to help initiate discussion at C-AGG meeting.

Draft Authors: William Salas (Applied Geosolutions, LLC), Steven DeGryze (Terra Global Capital), Mark Ducey (University of New Hampshire) and Johan Six (University of California at Davis).

Timeline: rough draft for discussion at Chicago meeting by July 18. Final version with comments/discussion summary from C-AGG meeting by end of August.

1.   **Background:**

Agriculture represents an important near-term option for GHG offsets. Given that N_2O and CH_4 emissions from agricultural practices exhibit high spatial and temporal variability, current efforts to use emission factors to estimate emissions are lacking. It is clear that if agricultural offset projects are going to include N_2O and CH_4 reductions, then process-based biogeochemical models are potentially an important tool to quantify emission reductions within offset protocols. Success of offset projects hinges on the development of transparent tools for quantifying performance based measures that account for full GHG impacts (soil carbon, N_2O and CH_4 reductions) and for which the uncertainty of the results is known. These tools must be accompanied with unambiguous procedures detailing how to parameterize and calibrate the tool. DNDC is a prime candidate tool for this work. While there have been many validation efforts in the past, they have been done in an *ad hoc* fashion and with a goal of validating the model performance for a single site and a unique approach to parameterization and calibration. In addition, most validation efforts have focused on evaluating the model's performance to predict absolute emissions accurately. The question remains how good a model's performance is with respect to emission reductions. If DNDC, or other process models, are to be integrated into protocols for agricultural GHG offsets, then a comprehensive and systematic validation exercise is needed to statistically quantify uncertainties in model based estimates of GHG emission reductions that are obtained by standardized approach to parameterization and calibration. In addition to providing critical documentation of model performance and statistically valid uncertainty metrics, this validation work can help guide development of offset protocols and the science for improving the tools by highlighting where the model does not perform well.

2. Statistical evaluation of model performance: A Philosophy

What we assume here is that there is a model, with a documented procedure for parameterization and calibration. This procedure has been followed and now we have an independent test data set (i.e. the model has *not* been calibrated to the particular outputs we are now testing). We would like to test whether the model is performing adequately. It is very easy to get hung up here on the statistical procedures and ignore the all-important steps of asking the right questions, and defining what "perform adequately" means. Let's focus on the definition part first, because that determines the statistical measures and approaches we should use (and not vice versa!).

There are three basic levels of performance we may wish to consider. Not all are equally important for every context. As a result, there is no single "right" criterion for deciding whether a model is performing well, there is no single "right" statistical test, and there is no single "right" threshold. We will focus this discussion on the right criterion and threshold within the context of GHG offset projects.

1. Closeness of individual predictions to individual measurements. On a case-by-case basis, do model predictions match up well to observations? This is the most relevant performance test within the context of GHG emission protocols. Of course, unless we have a perfect model, the match will be less than perfect for many or all observations, and there will likely be some for which large errors occur no matter how good the model is (because of input data issues, or because processes that rarely matter and are not represented well in the model do occasionally influence the outcome in the field). What constitutes "closeness of predictions"? Is it a relative matter (e.g. within 10%) or an absolute one (e.g. within 10 kg C/ha/yr)? For what fraction of observations would it be acceptable to have out-of-tolerance results? 10% of observations, or 1% of observations? (Demanding that all predicted and observed values have very close tolerances is likely to lead to frustration with real models and real data for complex phenomena.) If one observation corresponds with one carbon project, an alternative way of posing the same question is which percentage of all projects within a carbon standard's project portfolio will have overestimated its carbon offsets? Most likely, this percentage will not be tolerable, and the percentage of projects that do not overestimate credits will have to be increased by applying an uncertainty deduction. The question then becomes, if a deduction is applied, which percentage of carbon projects does the standard allow to be overestimating its credits.

2. Correlation of modeled and measured values. Over a large number of specific cases, is it generally true that high predictions match high observations, and low predictions match low observations? If not, then the model will do a poor job of representing within-region variability and maps developed using the model will not provide good spatial representations. Ideally, we would like the modeled and measured values to fall close to a 1:1 line with an intercept near zero. Of course, unless we have the perfect model the slope of the "best fit" line between predictions and measurements will not be exactly 1.

But how much error can we tolerate? How should utilize information on model fit when the intercept is not near zero? Defining reasonable tolerance here is not a statistical question, but a management and policy question that depends on the eventual use of the model.

3. Average model performance. Over a large number of specific cases, does the average prediction given by the model match up well with the average value that is observed? If not, then in the most basic statistical sense we would call the model biased. If a model is to be used within the context of GHG protocols, a model will have to be unbiased, or, at least, not biased so that credits are systematically overestimated. Of course all models will be biased to some degree (unless we are modeling a trivially simple phenomenon associated with well-defined physical constants). How much bias can we tolerate?

4. Extending the use of models to new regions. Often we wish to implement a model in a new region. We may have a "global" validation or test data set, or at least a test set that comes from several other regions. Some reasonable questions to ask include: do we need to perform a new validation? If so, how many data points would we need, either to provide an evaluation or better yet to recalibrate our expectations about patterns of bias or variability?

Statistical Approaches

Above, we said there is no single "right" criterion for deciding whether a model is performing well, there is no single "right" statistical test, and there is no single "right" threshold. However, there are some approaches that are sensible and others that are less so.

Traditionally, statistical assessment of model performance has applied standard Fisherian hypothesis testing approaches, with varying levels of sophistication. Although new approaches may be better suited to the task, the traditional approaches are still widely used and are not completely without value, so let's review them. We'll assume we have pairs of predicted and measured values to work with, and that these pairs are in some way a representative sample of some domain of interest (such as a region or project).

Over a large number of specific cases, does the average prediction given by the model match up well with the average value that is observed? This is really a question about two means, the mean of the observed values and the mean of the predicted values. Since we have pairs of values, we can calculate differences

$$d_i = obs_i - pred_i$$

and then perform an appropriate test to see if the mean of the differences is significantly different from zero. For example, if we are willing to assume that the differences are approximately normally distributed, we could perform a simple *t* test, rejecting the null hypothesis that the average difference is zero if $p \leq 0.05$. If we do not believe the

differences are normally distributed, then a nonparametric (bootstrap) test could be used instead.

Over a large number of specific cases, is it generally true that high predictions match high observations, and low predictions match low observations? Most conventional approaches to this question involve regression. At the most elementary level, one could fit a regression and examine R^2 to see if it is close to 1. However, it is difficult to define what an acceptable value of R^2 should be (what proportion of variance explained is good enough? and how does that differ between populations that are highly variable and those where the variability is relatively small?) Moreover, one can have a high value of R^2 even if the slope is very far from 1 (or even negative, though we might hope that model predictions would at least be positively correlated with observed values!) A more sophisticated approach would be to use the estimated slope and its standard error to construct a *t* test of the null hypothesis that the slope equals 1 (which is not the same as the null in the usual test reported by statistical software, that the slope equals 0). Where the residuals of the regression are not normally distributed, a bootstrap *t* test would be a logical alternative. Where outliers are present, robust regression procedures could be used. And so on. Of course one must also answer the question of which variable – the observation or the prediction – should go on the *y* axis and which will go on the *x* axis, as this will influence the estimate of the slope and the outcome of the test. Although it is the model prediction that "has the error," and on that basis we might think predicted values should be *y* and observed *x*, the eventual use of the results suggests the opposite: down the line we will have a large number of predictions and we are trying to draw inferences about unknown true values that have not been observed. So the case is stronger for putting predicted values on the *x* axis and observed values on the *y* axis.

On a case-by-case basis, do model predictions match up well to observations? This question is both the simplest and the hardest to deal with. If the data set being used for validation is representative of the population we are eventually interested in, then for a great many criteria the estimation process is straightforward. For example, if we wish 90% of the predictions to be within 10% of the actual observations, and we have a representative sample, we can simply calculate the proportion of the data that meet the criterion, and even use familiar confidence limit calculations to see if our particular results might have met (or failed to meet) the specification by virtue of good (or bad) luck. Likewise, quantities such as variances can be estimated unbiasedly from the sample (and, for a sample of reasonable size, the estimates of quantities like relative root mean squared error will be nearly unbiased). The difficulty, of course, is in getting a "representative sample," especially when quality studies that can be used for validation are sparse or come from a broader, narrower, or just plain different situation than the one we wish to know about. The results of such tests and calculations should typically be taken with some humility.

Equivalence Testing. There is, unfortunately, a fundamental challenge to the traditional approaches, especially to the first and second questions. This challenge can be illustrated by a very simple example. Suppose we have only a handful of data for testing a model in a particular context, and the data are somewhat noisy. We perform a *t*-test of the

difference between the predicted and observed values, or fit a regression of observed vs. predicted values, and test whether the slope of the regression equals 1. Of course with very little data, our confidence limits on either the mean or the slope will be very wide. So even if our estimates (or the true values) are far from 0 for the mean, or far from 1 for the slope, we will be unable to reject the null hypothesis. We have too little power. Therefore, we are likely to "accept" a model not because its performance is good but because our evidence is weak. Conversely, if we have a great deal of data we may be able to reject the null that the difference is 0, even if it is so small as to be of no practical consequence, and we may do the same for the null that the slope is 1 even if it is very close. In either case, from a statistical perspective we can never prove that the model is right. The take-home message is simple: collect as little data as possible so that your model will not be rejected!

The reason for this perverse situation is that in conventional statistical practice, the roles of the null and alternative hypotheses are essentially backwards. What we want is to demand that the evidence substantiate that the model is good enough. (In conventional practice, we are asking the evidence to show that the model isn't perfect – something we already knew, but we treat perfection as the null hypothesis that must be disproven.) A lucid but somewhat theoretical review can be found in the article by Robinson et al. (2005). One solution is the set of statistical techniques known as equivalence testing (Berger and Hsu 1996, Wellek 2003). Although these approaches have been used for some time in biomedical work, they are only beginning to be explored in agricultural and forestry contexts.

Extending the use of models to new regions: These kinds of problems and questions are not unique to agricultural and environmental modeling. In the United States, for example, they arise commonly in the analysis of data from the U.S. Census Bureau. Recall that every 10 years, the U.S. Census is supposed to identify and enumerate all residents of the U.S., and this is done with a "short form." A small sample of people receives a "long form" that includes a great deal of additional identifying information. But at many levels of aggregation, for example for an individual town, there may not be enough "long form" data to estimate certain quantities or relationships accurately. The challenge is to make optimal use of the data at a regional level to identify general patterns and trends, then use the local data to calibrate or modify the regional expectation. In the statistical literature these techniques fall under the general name of "small area estimation" techniques (Rao 2003).

There are many approaches for small-area estimation problems, but a simple one can be illustrated through a mixed-effect model (Pinheiro and Bates 2000). Let's write out the usual regression equation for observed vs. predicted values:

$$Obs_i = a + b\, Pred_i + e_i$$

Here a is the intercept (which we hope is very close to 0), b is the slope (which we hope is very close to 1), and e_i is the error term for an individual observation (which we most

often model using the normal distribution, and which we hope has very small variance). But suppose the relationship between observed and predicted might change from region to region. To account for this we should add one or two terms to our equation:

$$Obs_i = (a + c_{region}) + (b + d_{region})Pred_i + e_i$$

Here c_{region} and d_{region} represent the departure of the region where observation i falls, from the "average" values of the slope and intercept. Under some mild assumptions, we can incorporate c_{region} and d_{region} into the model – treating them as random variables akin to e_i – and develop not only predictions for the regions we have in our data, but also some understanding of how c_{region} and d_{region} vary among possible new regions (for example, what their standard deviation might be, indicating how the slope and intercept vary between regions). Armed with that knowledge, we could in principle predict how much data we would need to recalibrate our expectations for a model in a new region.

3. **Process Model Calibration and Validation: What is sufficient?**
    a. Short summary of parameterization, calibration and validation process
    b. How do we demonstrate that models have been sufficiently calibrated?
    c. When is re-calibration needed? Differences in crop varieties, differences in cultural practices, soil biophysical and climate drivers…What statistical approaches should be used to assess the need for re-calibration? (See mixed effects discussion)

4. **Quantifying model structural uncertainty:**
    a. Approach: use independent validation data, perform statistical analysis of modeled versus measured residuals, assess heteroskedasticity, assume distribution of the residuals, etc. Note: there are non-parametric (no distribution assumed) and parametric approaches (distribution must be assumed)
    b. Should the structural uncertainty be estimated for both absolute emissions and emission reductions?
    c. What is the appropriate uncertainty metric? RMSE or % deviation. Dividing the error in different components: lack of correlation, non-unity slope, and bias
    d. How to define confidence intervals? As a factor of the average or absolute amount (related to assumption on distribution)
    e. Minimum number of validation points for deriving structural uncertainty metrics?
    f. Using local/regional/global datasets: (relates to the question regarding recalibration)

5. **Assessing uncertainty in Input Data**

a. Discuss sources of uncertainty in input data for process models: soils (using soil surveys versus measured data), crop parameters (relates to the previous section discussion regarding crop varieties)

b. Applying Monte Carlo approach to changes in emissions (given the context of offset protocols)

      i. How to estimate parameter pdfs?

      ii. Accounting for covariates in input parameters (e.g. soil SOM and clay fraction)?

c. Discuss tradeoffs between measurements and other input sources (NRCS soil surveys: cost of measurement versus discounts from input uncertainty). Examine approaches for integrating a sensitivity analysis of parameters with costs to measure parameters and thus create a graph of uncertainty vs. cost

6. **Accounting for uncertainties in Agricultural GHG Offset projects**

a. How to discount modeled reductions based on uncertainty stats

Carbon projects can have drastically different circumstances that affect how accurate calculations of emission reductions are. To ensure consistency across projects and across project types, one must standardize calculations according to the certainty of the calculations. The most logical way of doing this is by using a confidence interval and calculating the minimum amount of credits that is generated with a certain confidence. The exact value of the confidence limit is a political decision that lays with the carbon standard.

By using a deduction based on uncertainty, one does not only achieve consistency across projects, but also incentivizes better monitoring and provides flexibility for projects with a small field monitoring budget. Agricultural GHG projects calculate GHGs as the difference between project emissions and baseline emissions. The uncertainty around the difference of two emissions will be different than the uncertainty around absolute emissions. More specifically, since the baseline scenario is so similar to the project scenario (except for the variables that are affected by project activities), it is to be expected that there will be a strong correlation between baseline emissions and project emissions. The extent of this correlation affects how the uncertainty around absolute emissions is related to the uncertainty of emission reductions. A Monte Carlo analysis can be used to calculate the variation in emission reductions as input parameters are changed.

b. Accounting for interactions between model structural uncertainties and uncertainty from input parameters?

As was discussed before, there are two components to the uncertainty of model simulations. Under the simplest (and most conservative) assumptions, these components can be combined as if there was no correlation between the two uncertainty terms. However, it is known that a model's performance is strongly dependent on the values of the input

parameters. For example, it is known that the accuracy of modeled nitrous oxide emissions decreases with increasing soil organic matter content. The stronger the correlation between these two error terms, the smaller the combined uncertainty. If enough samples in a dataset are available, one can theoretically calculate the structural uncertainty given a set of input parameters.

c. Role of aggregation in reducing model structural uncertainties. Carbon credits are always calculated and managed for an individual project that can consist of different individual fields. In fact, due to the complexity of GHG offset protocols and the transaction costs associated with offset projects, it is likely that several agricultural fields will be combined within a discrete agricultural GHG offset project. The aggregation of individual fields into one project will affect the calculation of uncertainty and the uncertainty deduction. The uncertainty will likely decrease as more fields are included within a project since the uncertainty can be averaged over individual fields. The uncertainty deduction mechanism incentivizes project aggregation since the (relative) deduction will be smaller if more fields are combined within a carbon project package.

Applying the DNDC Model for Rice Methane Offset projects: An Example

To illustrate these concepts, we compiled a dataset of 99 independent validation data for rice methane. These data were compiled from 9 peer reviewed papers and a recent validation exercise in California using data from one peer reviewed publication and a recent dataset collected at Rice Experiment Station in Biggs, CA (data courtesy of Dr. W. Horwath, UC Davis). For each validation, the DNDC model was calibrated to the local rice variety using methods similar to calibration approaches discussed in draft and existing protocols for ACR, VCS and CAR. However, it is important to point out that this dataset was compiled from several versions of the DNDC model. We are in the process of trying to compile all of the input data and field data to redo the validation of the 99 datapoints using a single version of the model. Nevertheless, we feel this dataset is useful for demonstrating the concepts outlined in this whitepaper.

On a case-by-case basis, do model predictions match up well to observations? The answer here is "it depends" – depending, of course, on what "match up well" means and how many observations are allowed to match poorly. For example, half of the individual data points have predictions within 18.0 kg C/ha/period of the observed values, and fully 75% are within 43.5 kg/ha/period. In relative terms, half of the predictions are within 19% of their corresponding observations, and 75% are within 38%. Is this good enough *at an individual site scale* – where we can probably tolerate much larger errors than we would at a regional or project-level scale?

Over a large number of specific cases, is it generally true that high predictions match high observations, and low predictions match low observations? Our answer to this question will depend on some statistical choices. If we make the observed values the *x*-axis and the predicted values the *y*-axis, then $R^2=0.88$, which we might conclude is a pretty good fit for a process model but the number itself carries little practical meaning. More importantly, the estimate of the intercept (3.41 ± 7.48) is not significantly different from zero, and with this choice of axes we would conclude that the slope (1.01 ± 0.04) is not significantly different from 1. However, as suggested above there is a very strong case that we have the axes backward here: observed values should be on the *y*-axis and predicted values on the *x*-axis. In that case the R^2 doesn't change, but the regression coefficients do. The intercept (13.39 ± 6.80) is almost significantly different from zero ($p=0.0518$), and the slope (0.87 ±0.03) shows a strong significant difference from 1 (p<0.0001). See figure below. Now we might well be concerned about the influence of possible outliers on this relationship, and could easily launch ourselves down a whole series of alternative regression analyses to deal with them, but those would change the interpretation of our results. (For example, the simplest solution – deletion of outliers – might move the slope closer to 1, but then our conclusion would be that the model corresponds well *except for some outlying cases that were deleted because they disagreed with the model* – hardly a satisfying conclusion).
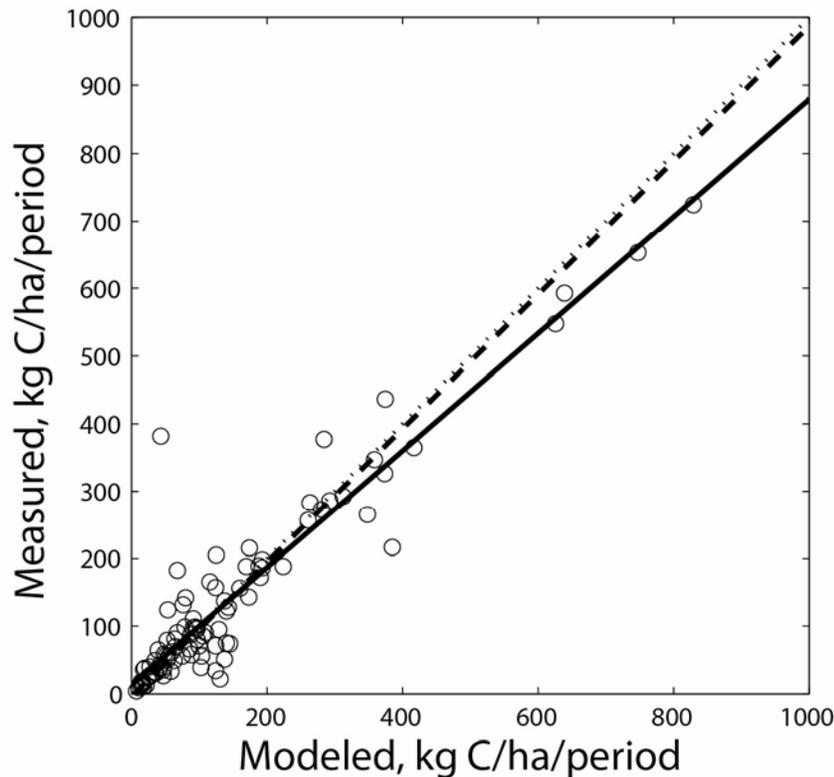
Over a large number of specific cases, does the average prediction given by the model match up well with the average value that is observed? The average value of the observed flux (130.7 kg/ha/period) is very close to the average value of the predictions (135.3 kg/ha/yr). The standard error of the paired differences is only 5.5 kg/ha/period (or about 3.5%), so the typical *t*-test approach would lead us to conclude that the difference is not statistically significant (*t*=0.85, *p*>0.05). This does **not** mean there is no difference! It only means our data are too small to resolve it, and it is probably small (within about 11 kg/ha/period, or 7% of the mean, with 95% confidence.)

Equivalence Testing. Without going into gory statistical detail, it's fair to say that the equivalence testing approach – using a nonparametric bootstrap test to avoid the impact of assuming normality – gives very similar results to the traditional approaches, in part because this is a large data set. For example, if we demand that the evidence show the modeled average is within 10% of the observed average, and use a 95% probability basis for our inference, the equivalence test rejects the null hypothesis that the model performance is worse than our specification – in other words, we can accept the average performance of the model. But if we demand that the evidence show that the slope of our observed vs. modeled relationship is between 0.9 and 1.1, we fail to reject that null – in other words, the data don't substantiate that the model meets that specification. The result is similar to that of the corresponding traditional regression test, but the

interpretation is somewhat different. Again, the equivalence testing approach is sensitive to the presence of outliers, so that is an issue we would probably explore in a detailed investigation.

Mixed Effects Example: Can our global validation data set tell us anything about likelihood of DNDC applicability for estimating methane emission from rice in new regions (like Arkansas and Louisiana, for example)? Unfortunately, with our example data set, this approach is rather anticlimactic! The test data used here suggest that c_{region} and d_{region} actually have zero variability – in other words, the slope and intercept appear to be consistent across all regions. In a sense, that is good news: it is reasonable to expect that the kinds of variability around the central trend that we see in this data, would play out similarly when implementing the model in a new region. While it would certainly enhance our confidence in the model to have some region-specific test data to confirm that, it also does not appear critical for modeling CH_4 fluxes with this particular model. On the other hand if either c_{region} and d_{region} did have an appreciable standard deviation, that would imply we really would want to acquire some data in a new region just to understand how the model performance might differ with previously studied regions. The results of estimating *a, b, c_{region}*, and d_{region} could be used in a variety of modeling approaches to simulate the recalibration process and identify an appropriate sample size.

Example 1: Quantifying impact of input uncertainty: NRCS Soils

One approach for assessing the impact of model input uncertainty on modeled GHG emissions and assessment of GHG reductions from the implementation of an offset project is to use Monte Carlo methods. At apply this method for assessing the impact of uncertainty of soil conditions, the first step entails defining a possible range and probability distribution of the soil conditions. To assess a range of results for greenhouse gas emissions at a single rice field, this example demonstrates a Monte Carlo modeling approach using DNDC. Our general approach was to assume some variability in site soil attributes (clay fraction, organic matter fraction, bulk density, and pH) as modeled in the USDA NRCS SSURGO soil model. Using a Monte Carlo simulation, we modeled identical crop management practices and meteorological conditions while varying soil conditions through 1,000 iterations. Rather than using the uncertainty tools included with DNDC (which assume an even distribution of the tested parameters), our approach assumed a log-normal distribution of each of the soil attributes as well as some amount of correlation between them. Our approach involved three aspects:

1. an analysis of correlation between the four soil attributes
2. programmatic generation of DNDC inputs based on the Monte Carlo method
3. running the DNDC model in site mode and synthesizing the results

We applied this approach in two ways, the first assumes no correlation between soil parameters, which is conservative since we know that there is significant correlation. The second set of Monte Carlo runs utilized correlation statistics as part of the sampling procedure.

Soil attributes are stored within the SSURGO database according to the following relationships:

Horizon	Contains soil attribute data (low, representative, and high values) based on an assessment of soil field conditions
⇩	[one to many]
Component	The basic soil type (roughly equivalent to soil series) – soil components have many horizons and have no explicit spatial location
⇩	[one to many]
Map Unit	The smallest mapped polygon in the SSURGO model – soil map units have many components of varying fractions

To assess correlation among soils in rice growing areas of California, we selected all map units intersecting rice fields as mapped in the CA Department of Water Resources land use database. From this selection, we identified all soil components contained within the map units. Soil attribute data came from the top horizon for each component. Thus, our final database represents all soil horizons intersecting rice fields.

We calculated Pearson correlation coefficients for each set of pairs for representative values of the four soil attributes:

	Clay fraction	OM fraction	Bulk Density	pH
Clay fraction	1	-	-	-
OM fraction	0.139	1	-	-
Bulk Density	-0.526	-0.685	1	-
pH	0.263	0.098	-0.126	1

Generation of DNDC inputs

Our Monte Carlo simulation randomly generated 1,000 numbers for each of the four soil properties with the correlation matrix and with each following a log-normal distribution. This was done by using the Cholesky decomposition of the correlation matrix to transform a set of standard-normal random numbers in the logarithm space. The representative value was used as the mean, while the low and high values were transformed into log space and treated as a range of +/-3 standard deviations. This resulted in four sets of 1,000 correlated random numbers, normally distributed. The soil properties, other than pH, were finally calculated by taking the exponent of the numbers.

DNDC modeling

The model was run as a batch using the DNDC site mode. We ran two scenarios (one with a winter flood, one without a winter flood) for a single field as follows:
* Rice, planted 5/1, harvested 9/11
* Tillage on 4/23, 4/26, 4/27, 4/29, and 9/15
* Fertilizer on 4/30 (injected anhydrous ammonia), 5/1 (surface application of $(NH_4)_2HPO_4$), 5/26 (surface application of $(NH_4)_2SO_4$)
* Flooded from 5/1 to 9/1
* Winter flood from 11/15 to 1/31 (only for the winter flood scenario)
* Rice straw burned once every 8 years

We also ran a 16-year initialization period (using the baseline crop management) to stabilize soil organic matter. Only the last eight years (one year with residue removal, seven years with residue incorporation) were averaged in the final summary.

Our results indicate the modeled methane emissions and net GHG emissions are quite sensitive to soil conditions. At 90% confidence interval, the modeled CH4 and net GHG emissions were significant (over 14% in both baseline and project simulations). However, the impact of soil uncertainties on modeled changes in emissions from baseline to project conditions were quite small (<3%). The figure below shows the histogram of the Monte Carlo simulation results for the case assuming no correlation between soil input parameters. It is clear for this baseline and project scenario, that uncertainty in soil input parameters impacted both baseline and project modeled emissions in a similar degree. Accounting for correlation between soil input parameters reduced uncertainties. The table below summarizes these results.

Monte Carlo Analysis: Impact of Uncertainties in Soil Inputs
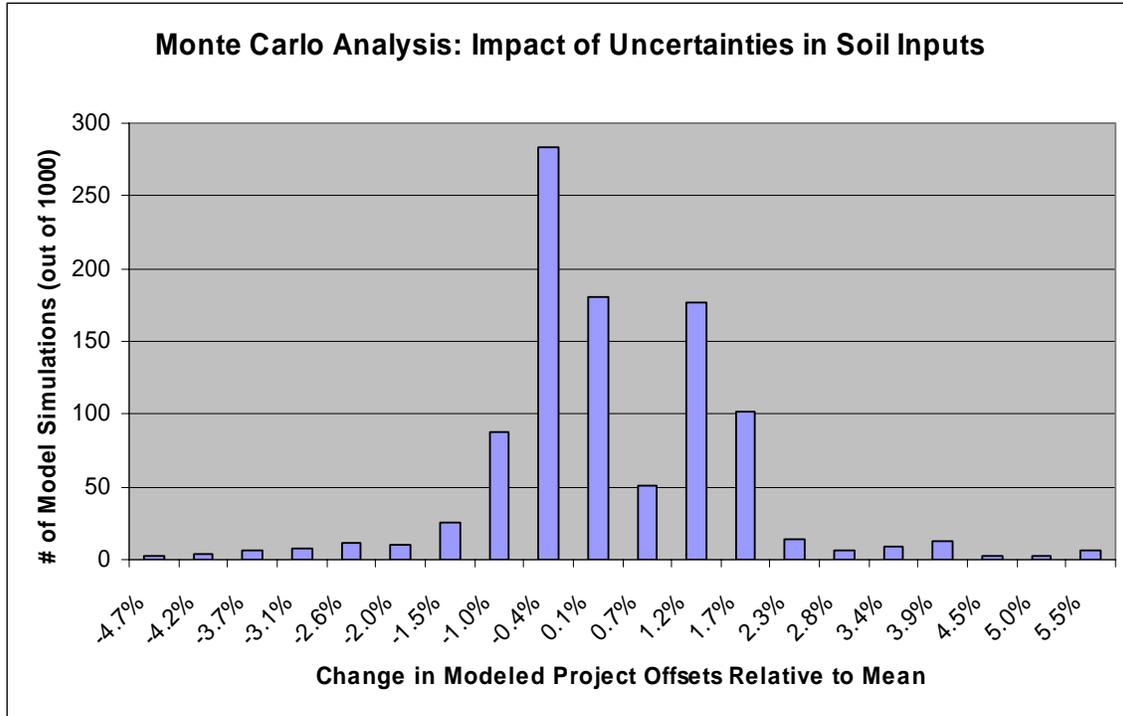
	Assuming no correlation in soil input parameters		Accounting for correlation of input soil parameters	
	CH4 GWP (90% CI / Mean)	Global Warming Potential (90% CI / Mean)	CH4 GWP (90% CI / Mean)	Global Warming Potential (90% CI / Mean)
Baseline	14.7%	14.4%	14.0%	13.7%
Project	18.5%	20.0%	17.5%	19.1%
Baseline-Project	**1.0%**	**2.2%**	**0.2%**	**1.4%**

Example 2:   Quantifying Model Structural Uncertainty

To be added…

References

Berger, R.L. and Hsu, J.C. 1996. Bioequivalence trials, intersection tests and equivalence confidence sets. Stat. Sci. 11: 283–319.

Pinheiro, J.C. and Bates, D.M. 2000. *Mixed-Effects Models in S and S-Plus*. Springer, New York.

Rao, J.N.K. 2003. *Small Area Estimation.* Wiley: New York.

Robinson, A.P., Duursma, R.A. and Marshall, J.D. 2005. A regression-based equivalence test for model validation: shifting the burden of proof. *Tree Physiology* 25: 903-913.

Wellek, S. 2003. Testing statistical hypotheses of equivalence. Chapman and Hall, London, 284 p.