Supporting Information for

Explaining the shortcomings of log-transforming the dependent variable in regression models and recommending a better alternative: evidence from soil CO_2 emission studies

Kao-Lee Liaw[1], Myroslava Khomik[1,2], M. Altaf Arain[1]

[1]School of Earth, Environment & Society and McMaster Centre for Climate Change, McMaster University, Hamilton, ON, L8S 4K1, Canada,[2]Hydrometeorology Research Group, Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Contents of this File

This supporting information has 15 texts, 7 tables, and 13 figures. The tables and figures are to be nested under the texts. Each text deals with a theme that is useful for gaining better understanding of the log-transformed and nonlinear approaches.

As a consequence of the review process, we added two topics that are not important for demonstrating the superiority of the nonlinear approach over the log-transformed approach but may be of interest to other empirical researchers.

One topic is the reformulation of our Gaussian-Gamma model as a Generalized Linear Model (GLM) with the assumption that the conditional CO_2 emission has a Gamma distribution. In Text S9, we do the reformulation and apply the GLM to the data of the Temerate Forest, using the GENMOD procedure of SAS software. We found that although it can also avoid the under-prediction problem of the log-transformed approach, the approach using GLM is inferior to the nonlinear approach. In Text S10, a randomized experiment is used to confirm the superiority of the nonlinear approach over the GLM approach. Nonetheless, the superiority of the nonlinear approach for the purpose of estimating the unknown coefficients does not imply that the assumption of the Gamma distribution is not an adequate one. For the purposes of simulation and forecasting, we would recommend the use of the Gamma distribution, because it is a better representation of the physical process than others, such as the Normal distribution.

Another topic is the use of the nonparametric Generalized Additive Model (GAM) to study the effects of soil temperature and soil moisture on CO_2 emission. In Text S15, we applied two versions of the GAM to the data of the Temperate Forest, using the GAMPL procedure of SAS software. The application showed that the Gaussian specification is better than the Q_{10} specification for representing the temperature function, and that the Gamma specification is better than the Gaussian specification for representing the moisture function.

Titles of Texts:

Text S1. Understanding the Useful Property of the Saturated Specification of Our Regression Model

Text S2. The Fundamental Reason for the Under-prediction Problem of the Log-transformed Approach

Text S3. The Determinants of the Severity of the Under-prediction Problem of the Log-transformed Approach

Text S4. Regressions for Assessing the Effects of CV and Skewness on the Severity of the Log-transformed Approach's Under-predictions

Text S5. Overcoming the Very Weak Explanatory Power of the Gaussian-Gamma Specification of the Regression Model for the Semiarid Mediterranneam Marsh in Southern California

Text S6. The Blindness of the Comparable R-Square to the Systematic Bias in the Predictions of the Log-transformed Approach

Text S7. Failure of Log-transformation to Normalize the Conditional Distributions of the Dependent Variable

Text S8. Statistical Inference without the Homoscedasticity Assumption for the Error Term in the Nonlinear Approach

Text S9. Application of the Gaussian-Gamma Specification of the Regression Model to the Data of the Temprerate Forest via the GENMOD Approach

Titles of Tables:

Table S5. Comparison of the descent parameter of soil moisture between (A) the nonliear approach and (B) the GENMOD approach: Based on 10 randomly selected subsamples of the field field data of the Temperate Forest in southern Ontario (n=15,523)

Table S6. Estimation results of applying two versions of the Genralized Additive Model to the data of the Temperate Forest in southern Ontario for predicting soil CO_2 emission by soil temperature and soil moisture

Titles of Figures:

Figure S1. The concave curve representing the log function of CO_2 emission (in red) and the tangent line (in blue) touching the curve at the mean CO_2 emission ($\bar{y}$ = 4.88 µmol/m^2/s for the Temperate Forest of southern Ontario).

Figure S2. CO_2 emission predicted by the nonlinear approach (Y”) versus CO_2 emission predicted by the log-transformed approach (Y’), based on the applications of the Gaussian-Gamma model to the data of four ecosystems: (A) Temperate Forest (n=15,523), (B) Temperate Grassland (n=566), (C) Desert (n=1,474), and (D) Semiarid Mediterranean Marsh (n=38,213).

Figure S3. The inability of log-transformation to normalize the distribution of CO_2 emissions observed in the Temperature Forest of southern Ontario: Evidence in two bins (Rounded Temperature=8 degrees C & $Rounded\ Moisture = 15\%$; and Rounded Temperature=12 degrees C & Rounded Moisture=15%).

Figure S4. The failure of the log-transformed approach to make the distribution of the residuals to be closer to a normal distribution: Based on the application of the Gaussian-Gamma model to the data from the Temperate Forest of southern Ontario (N=15,523).

Figure S5. The tendency for the conditional variance (or the conditional standard deviation) to increase with the square of the conditional mean (or simply the conditional mean) among the 20 non-empty bins of the 15,523 observed CO2 emissions in the field data of the Temperate Forest of southern Ontario.

Figure S6. The likelihood and log-likelihood functions of a specific observed CO_2 emission in the data set. In the model for predicting emissions, the conditional distribution of the dependent variable is assumed to be a Gamma distribution. The specific observed emission is 5 µmol/m^2/s.

Figure S7. The limited ability of the delta method in revealing the pattern of the dependence of (1) the severities of the log-transformed approach’s under-predictions of conditional mean CO_2 emissions on (2) the within-bin CVs of emissions: Based the 20 non-empty bins that were created by crossing the temperature values rounded to the nearest 4ºC and the moisture values rounded to the nearest 5% in the data of the Temperate Forest.

Figure S8. The limited ability of the delta method in revealing the pattern of the dependence of (1) the severities of the log-transformed approach's under-predictions of conditional mean CO_2 emissions on (2) the within-bin CVs of emissions: Based the 132 bins with at least 10 observations that were created by crossing the temperature values rounded to the nearest 1ºC and the moisture values rounded to the nearest 1% in the data of the Temperate Forest.

Figure S9. Predicted dependence of CO_2 emission on (A) soil temperature (holding soil moisture at 15%) and (B) soil moisture (holding soil temperature at 22 degree C), revealing the different patterns between the log-transformed and nonlinear approaches: Based on the data of the Temperate Forest in southern Ontario (n=15,523) and plotted for Chamber 1 only. By design, the shapes of these curves for other chambers are similar.

Figure S10. The spline components generated by the Generalized Additive Model with univariate splines for soil temperature and soil moisture for predicting CO_2 emission in the Temperate Forest. Link=log, Dist=gamma, n=15,523. The belt represents the 95% confidence band.

Figure S11. The bivariate spline component generated by the Generalized Additive Model with a bivariate spline for soil temperature and soil moisture for predicting CO_2 emission in the Temperate Forest.

Figure S12. The patterns of the dependence of CO_2 emission on soil temperature generated by two versions of the Generalized Additive Model: (1) with univariate splines for soil temperature and soil moisture; (2) a bivariate spline for soil temperature and soil moisture for the Temperate Forest data.

Figure S13. The patterns of the dependence of CO_2 emission on soil moisture generated by two versions of the Generalized Additive Model: (1) with univariate splines for soil temperature and soil moisture; (2) a bivariate spline for soil temperature and soil moisture for the Temperate Forest data.

Text S1. Understanding the Useful Property of the Saturated Specification of Our Regression Model

The saturated specification of our regression model is useful for revealing and understanding the shortcomings of the log-transformed approach, because it has the property of being able to predict perfectly the observed mean of the dependent variable for any combination of the values of the explanatory factors. Specifically, for any combination of the values of the explanatory factors, it can perfectly predict the mean of $\ln(y_i)$ for the log-transformed approach and the mean of y_i for the nonlinear approach. Here we want to prove this property without too many notational complications.

Before carrying out the proof, we mention a useful *property of the mean*: For a set of n values $\{Z_1, Z_2, \dots, Z_n\}$ of any variable Z, the mean ($\bar{Z} = \frac{\sum_{i=1}^{n} Z_i}{n}$) minimizes the sum of squares of the distances to all n values of Z. In other words, the mean $\bar{Z}$ solves the least squares problem: $min \sum_{i=1}^{n}(Z_i - \mu)^2$, where μ is the unknown quantity in question. This property can be easily proved by differentiating the sum of squares with respect to μ and setting the derivative to 0. In our empirical study, the variable in question is the CO_2 emission from soil. In other words, we let Z_i be either y_i or $\ln(y_i)$.

Next, in our regression model, we let each the two explanatory factors (soil temperature and soil moisture) be represented by a dummy variable assuming either the value of 1 for high level or the value of 0 for low level. In other words, all observations are classified into four distinct combinations (bins) in terms of temperature and moisture: (0, 0), (1, 0), (0, 1), and (1, 1). With this simplification, we are ready to prove in a concise way the above-mentioned property of the saturated specification of the regression model for the log-transformed and nonlinear approaches, separately.

For the log-transformed approach, the saturated specification of the model becomes

$$\ln(y_i) = \alpha + \beta D_{i1} + \gamma D_{i2} + \delta D_{i1} D_{i2} + E_i \qquad (S1.1)$$

where $\ln(y_i)$ is the natural log of the ith observation of CO_2 emission; $D_{i1}$is the ith observation of the dummy variable representing soil temperature; $D_{i2}$is the ith observation of the dummy variable representing soil moisture; $\alpha, \beta, \gamma,$ and δ are unknown coefficients to be estimated; and E_i is a random error term. Note that the number of unknown coefficients in a saturated specification of the regression model must be equal to the number of all distinct combinations of the values assumed by the explanatory factors.

The least squares method finds the estimated coefficients by minimizing the sum of squares:

$$\sum_{i=1}^{n}(\ln(y_i) - (\alpha + \beta D_{i1} + \gamma D_{i2} + \delta D_{i1} D_{i2}))^2 \qquad (S1.2)$$

A useful insight about this sum of squares is that it can be partitioned exactly into the following four partial sums:

$$\sum(\ln(y_i) - \alpha)^2 + \sum\big(\ln(y_i) - (\alpha + \beta)\big)^2 + \sum\big(\ln(y_i) - (\alpha + \gamma)\big)^2 + \sum\big(\ln(y_i) - (\alpha + \beta + \gamma + \delta)\big)^2 \qquad (S1.3)$$

The first partial sum applies to all observations with low temperature and low moisture, the second partial sum applies to all observations with high temperature and low moisture, the third partial sum applies to all observations with low temperature and high moisture, and the fourth partial sum applies to all observations with high temperature and high moisture. Applying the *property of the mean* to each of the four partial sums, we get the following results:

$$\overline{\ln(y)}_{0,0} = \ddot{\alpha} \qquad (S1.4)$$

$$\overline{\ln(y)}_{1,0} = \ddot{\alpha} + \ddot{\beta} \qquad (S1.5)$$

$$\overline{\ln(y)}_{0,1} = \ddot{\alpha} + \ddot{\gamma} \qquad (S1.6)$$

$$\overline{\ln(y)}_{1,1} = \ddot{\alpha} + \ddot{\beta} + \ddot{\gamma} + \ddot{\delta} \qquad (S1.7)$$

where $\overline{\ln(y)}_{0,0}$ is the mean of $\ln(y_i)$ conditional on $D_{i1} = 0$ and $D_{i2} = 0$; $\overline{\ln(y)}_{1,0}$ is the mean of $\ln(y_i)$ conditional on $D_{i1} = 1$ and $D_{i2} = 0$; $\overline{\ln(y)}_{0,1}$ is the mean of $ln(y_i)$ conditional on $D_{i1} = 0$ and $D_{i2} = 1$; $\overline{\ln(y)}_{1,1}$ is the mean of $\ln(y_i)$ conditional on $D_{i1} = 1$ and $D_{i2} = 1$; and $\ddot{\alpha}, \ddot{\beta}, \ddot{\gamma},$ and $\ddot{\delta}$ are the best estimates of $\alpha, \beta, \gamma,$ and δ, respectively. Thus, *the saturated specification of the regression model for the log-transformed approach has the useful property that it can predict exactly the conditional means of $ln(y_i)$ for all distinct combinations of the values of the explanatory factors.*

For the nonlinear approach, the saturated specification of the model is of the form:

$$y_i = e^{\alpha+\beta D_{i1}+\gamma D_{i2}+\delta D_{i1}D_{i2}} + \varepsilon_i \qquad (S1.8)$$

The unknown coefficients are to be estimated by minimising the following sum of squares:

$$\sum_{i=1}^{n}(y_i - e^{\alpha+\beta D_{i1}+\gamma D_{i2}+\delta D_{i1}D_{i2}})^2 \qquad (S1.9)$$

which can be partitioned exactly into the following four partial sums:

$$\sum(y_i - e^{\alpha})^2 + \sum(y_i - e^{\alpha+\beta})^2 + \sum(y_i - e^{\alpha+\gamma})^2 + \sum(y_i - e^{\alpha+\beta+\gamma+\delta})^2 \qquad (S1.10)$$

As before, each of these partial sums is for all observations with a unique combination of the values of the two explanatory factors. Again, applying the *property of the mean* to each of these partial sums, we get the following results:

$$\bar{y}_{0,0} = e^{\widehat{\alpha}} \qquad (S1.11)$$

$$\bar{y}_{1,0} = e^{\widehat{\alpha}+\widehat{\beta}} \qquad (S1.12)$$

$$\bar{y}_{0,1} = e^{\widehat{\alpha}+\widehat{\gamma}} \qquad (S1.13)$$

$$\bar{y}_{1,1} = e^{\widehat{\alpha}+\widehat{\beta}+\widehat{\gamma}+\widehat{\delta}} \qquad (S1.14)$$

where $\bar{y}_{0,0}$ is the mean of y_i conditional on $D_{i1} = 0$ and $D_{i2} = 0$; $\bar{y}_{1,0}$ is the mean of y_i conditional on $D_{i1} = 1$ and $D_{i2} = 0$; $\bar{y}_{0,1}$ is the mean of y_i conditional on $D_{i1} = 0$ and $D_{i2} = 1$; $\bar{y}_{1,1}$ is the mean of y_i conditional on $D_{i1} = 1$ and $D_{i2} = 1$; and $\hat{\alpha}, \hat{\beta}, \hat{\gamma},$ and $\hat{\delta}$ are the best estimates of $\alpha, \beta, \gamma,$ and δ, respectively. Thus, *the saturated specification of the regression model also has the useful property that for the nonlinear approach, it can predict exactly the conditional means of CO₂ emissions for all distinct combinations of the values of the explanatory factors*.

With more complicated notations, it can be proved that this useful property of the saturated specification remains true, no matter how many distinct values are assumed by the explanatory factors.

Text S2. The Fundamental Reason for the Under-prediction Problem of the Log-transformed Approach

With the saturated specification of our model, we can reveal that the fundamental reason for the under-prediction problem of the log-transformed approach is that the log function is a *concave* function. The revelation is achieved in the following two steps.

In the first step, we show that *if we first log-transform all the observed CO_2 emissions, then compute their mean, and then use the exponential function to transform the mean back to the same physical unit ($\mu mol/m^2/s$), we will find that the resulting value is less than the mean of the untransformed CO_2 emissions*. To prove this inequality, we follow the method of Blitzstein & Hwang (2015, p. 425) by drawing in Figure S1 a tangent line of the curve representing the log function at the mean of the observed values of CO_2 emission ($\bar{y}$). Let this tangent line be represented by:

$$z_i = a + by_i \qquad (S2.1)$$

where y_i represents the ith observation of CO_2 emission, and a and b represent the intercept and slope, respectively. Because the concavity of the log function implies that for any y_i, the point on the curve must be less than or equal to the corresponding point on the tangent line, it must be true that

$$\ln(y_i) \leq a + by_i \qquad (S2.2)$$

with the equality holding only for the y_i that happens to be equal to the mean. Summing across all observations and dividing by sample size, the above inequality implies the following strict inequality:

$$\frac{\sum_{i=1}^{n} \ln(y_i)}{n} < \frac{\sum_{i=1}^{n}(a+by_i)}{n} \qquad (S2.3)$$

where n is the sample size.

The right-hand-side of this inequality can be written as

$$\frac{\sum_{i=1}^{n}(a+by_i)}{n} = \frac{na+b\sum_{i=1}^{n} y_i}{n} = a + b\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = a + b\bar{y} \ = \ln(\bar{y}) \qquad (S2.4)$$

where $\bar{y}$ is the mean of y_i. The last equality in Eq. (S2.4) is due to the fact that the tangent line touches the curve at $\bar{y}$. Using Eq. (S2.4) and denoting the mean of $\ln(y_i)$ by $\overline{\ln(y)}$, we can rewrite the inequality (S2.3) as

$$\overline{\ln(y)} < \ln(\bar{y}) \qquad (S2.5)$$

Taking the exponential transformation on the two sides of this inequality, we obtain the main relation that

$$e^{\overline{\ln(y)}} < \bar{y} \qquad (S2.6)$$

Thus, the inequality in question is proved. Note that this inequality is Jensen's inequality for a sample with equal weights (Blitzstein & Hwang, 2015).

In the second step, we recognize that *with the saturated specification of our regression model, the log-transformed approach under-predicts the observed mean of CO2 emissions in any* *non-trivial bin (a bin containing at least two unequal emissions),* because Inequality (S2.6) applies to every non-trivial bin.

Remember that the nonlinear approach can predict perfectly the observed mean of CO_2 emission in every bin. Thus, for the simplified example of the saturated specification considered in Text S1, the above finding implies that $e^{\overline{\ln(y)}_{0,0}} < \bar{y}_{0,0}$, $e^{\overline{\ln(y)}_{1,0}} < \bar{y}_{1,0}$, $e^{\overline{\ln(y)}_{0,1}} < \bar{y}_{0,1}$, and $e^{\overline{\ln(y)}_{1,1}} < \bar{y}_{1,1}$. In other words, except for a trivial case, it is true that *for every distinct combination of the explanatory factors, the log-transformed approach will always under-predict the conditional mean of CO_2 emissions, while the nonlinear approach will always predict it perfectly*.

In sum, we have used the saturated specification of our regression model to reveal that the fundamental reason for the pervasive under-prediction problem is the *concavity* of the log-function used for transforming the dependent variable. Whether the dependent variable has a log-normal distribution or not is irrelevant.

Text S3. The Determinants of the Severity of the Under-prediction Problem of the Log-transformed Approach

To look for the determinants of the *severity* of the log-transformed approach's under-prediction for each combination of the values of the explanatory factors, we first note from Figure S1 that as y_i is moved away from the mean $\bar{y}$ in either direction, the magnitude of the gap between $\ln(y_i)$ and $a + by_i$ becomes larger. Thus, an important determinant of the severity of the under-prediction should be the *dispersion* of the observed values of CO_2 emission for the combination of temperature and moisture values in question: the dispersion tends to aggravate the severity of under-prediction. To quantify dispersion, we choose the coefficient of variation (CV).

We also note from the same figure that holding $|y_i - \bar{y}|$ constant, the gap between $\ln(y_i)$ and $a + by_i$ is smaller for $y_i > \bar{y}$ than for $y_i < \bar{y}$, because the segment of the curve to the right of $\bar{y}$ is less steep than the segment of the curve to the left of $\bar{y}$. Thus, in the context of dispersion, the *skewness* of the distribution of the observed values of CO_2 emission should have a negative effect on the severity of the under-prediction. Note that positive skewness means a long tail on the right-hand-side, whereas negative skewness means a long tail on the left-hand-side.

Therefore, it is likely that *CV and skewness are two determinants of the severity of the log-transformed approach's under-prediction problem.*

Text S4. Regressions for Assessing the Effects of CV and Skewness on the Severity of the Log-transformed Approach's Under-predictions

To use regressions for assessing the effects of CV and skewness on the severity of the log-transformed approach's under-predictions, we formulate the following regression model:

$$U_j = e^{\alpha_0 + \alpha_1 V_j + \alpha_2 {V_j}^2 + \alpha_3 K_j} + \epsilon_j \qquad for\ j = 1,2, \dots, 20 \qquad (S4.1)$$

where U_j the severity of under-prediction in the jth "bin" (i.e., the jth combination of a pair of distinct values of temperature and moisture); V_j is the coefficient of variation of CO_2 emissions in the jth bin; and K_j is the corresponding skewness; $\alpha_0, \alpha_1, \alpha_2, and\ \alpha_3$ are unknown coefficients; and ϵ_j is an error term. The estimation results are shown in Table S1. When skewness is suppressed from Eq. (S1.1), the model yields an Adjusted R-square of 0.96. Without this suppression, the model yields an Adjusted R-square of 0.98. Thus, the under-prediction severity of a bin depends (1) very strongly on the dispersion of CO_2 emissions within the bin, and (2) modestly on the corresponding skewness within the bin.

Text S5. Overcoming the Very Weak Explanatory Power of the Gaussian-Gamma Specification of the Regression Model for the Semiarid Mediterranean Marsh in Southern California

The application of the GG-model to the sub-hourly data of Semiarid Mediterranean Marsh yielded a surprisingly low explanatory power. The Comparable R-square being 0.151 for the log-transformed approach and 0.184 for the nonlinear approach (Panel A of Table S2).

Based on a close examination of the time-by-day patterns of CO_2 emission, soil temperature, and soil moisture in each of the 13 calendar months, we suspected that the very low explanatory power was mainly due to a missing explanatory factor that could account for the prevalence of very high emissions (>= mean+2std) from early September to late November. Since it is likely that most of these very high emissions were generated by non-plant organisms whose life-cycle spanned this time interval, we added into the model a dummy variable, *High Season*, which assumes the value of 1 if the observation in question occurred in September-November. Thus, the structure part of the GG-model was expanded into the following form:

$$e^{f(\boldsymbol{X}_i)} = e^{\beta_0+\beta_{11}T_i+\beta_{12}{T_i}^2+\beta_{21}M_i+\beta_{22}\ln(M_i)+\gamma_0 H_i} \qquad (S5.1)$$

where $H_i$is a dummy variable if the ith observation is in September to November, and γ_0 is an unknown coefficient. The role of the dummy variable can be easier seen by breaking up Eq. (S5.1) into the following two equations:

$$e^{f(\boldsymbol{X}_i)} = e^{\beta_0+\beta_{11}T_i+\beta_{12}{T_i}^2+\beta_{21}M_i+\beta_{22}\ln(M_i)} \qquad (S5.1a)$$

for obervations not in the High Season ($H_i = 0$), and

$$e^{f(\boldsymbol{X}_i)} = e^{(\beta_0+\gamma_0)+\beta_{11}T_i+\beta_{12}{T_i}^2+\beta_{21}M_i+\beta_{22}\ln(M_i)} \qquad (S5.1b)$$

for observations in the High Season ($H_i = 1$). Comparing these two equetions, it is clear that the dummy variable allows the intercept to differ between the two seasons but does not allow the coefficients of the temperature and moisture variables to be different between the two seasons.

This expansion of the GG-model raised the explanatory power markedly. The Comparable R-square was increased to 0.425 for the log-transformed approach and 0.549 for the nonlinear approach (Panel B of Table S2).

Since the effects of temperature and moisture on CO_2 emission were likely to differ between the High Season and the rest of the study period, we then added to the model the full set of interaction terms between this dummy variable with the explanatory variables representing temperature and moisture. The structural part of the model was then expanded into the following form:

$$e^{f(\boldsymbol{X}_i)} = e^{\beta_0+\beta_{11}T_i+\beta_{12}{T_i}^2+\beta_{21}M_i+\beta_{22}\ln(M_i)+\gamma_0 H_i+\gamma_{11}T_i H_i+\gamma_{12}{T_i}^2 H_i+\gamma_{21}M_i H_i+\gamma_{22}\ln(M_i)H_i} \qquad (S5.2)$$

The role of the dummy variable can be easier seen by breaking up Eq. (S5.2) into the following two equations:

$$e^{f(\boldsymbol{X}_i)} = e^{\beta_0+\beta_{11}T_i+\beta_{12}{T_i}^2+\beta_{21}M_i+\beta_{22}\ln(M_i)} \qquad (S5.2a)$$

for obervations not in the High Season $(H_i = 0)$, and

$$e^{f(\boldsymbol{X}_i)} = e^{(\beta_0+\gamma_0)+(\beta_{11}+\gamma_{11}T_i)+(\beta_{12}+\gamma_{12}){T_i}^2+(\beta_{21}+\gamma_{21})M_i+(\beta_{22}+\gamma_{22})\ln(M_i)} \qquad (S5.2b)$$

for observations in the High Season $(H_i = 1)$. Comparing these two equetions, it is clear that the dummy variable allows not only the intercept but also the coefficients of the temperature and moisture variables to be different between the two seasons.

The estimation result based on Eq. (S5.2) is shown in Table 5 in the paper. Note that in using the REG procedure of SAS software or our SAS module to estimate the coefficients, each interaction must be represented by a variable. For example, to represent $T_i H_i$, the user must create a new variable in advance by multiplying T_i to H_i.

Text S6. The Blindness of the Comparable R-Square to the Systematic Bias in the Predictions of the Log-transformed Approach

Despite the finding that the severity of the log-transformed approach's under-prediction of the observed grand mean was greater than 10% in the applications of the GG-model to the data of all the four ecosystems, the Comparable R-square achieved by the nonlinear approach mostly turned out to be only modestly higher than the Comparable R-square achieved by the log-transformed approach: 0.551 versus 0.544 for the Temperate Forest, 0.412 versus 0.405 for the Temperate Grassland, 0.402 versus 0.371 for the Desert, and 0.486 versus 0.470 for the Semiarid Mediterranean Marsh. The four gaps are 0.007, 0.007, 0.031, and 0.016.

To understand this puzzling finding, we started by looking at the correlation coefficient between the observed emissions and the predicted emissions generated by each of the two approaches:

$$r_{o,n} = \frac{\sum(y_i - \bar{y})(y_i'' - \bar{y}'')}{\sqrt{\sum(y_i - \bar{y})^2}\sqrt{\sum(y_i'' - \bar{y}'')^2}} \qquad (S6.1)$$

$$r_{o,l} = \frac{\sum(y_i - \bar{y})(y_i' - \bar{y}')}{\sqrt{\sum(y_i - \bar{y})^2}\sqrt{\sum(y_i' - \bar{y}')^2}} \qquad (S6.2)$$

where y_i is the ith observed emission; $\bar{y}$ is the mean of the observed emissions; y_i'' is the ith predicted emission generated by the nonlinear approach; $\bar{y}''$ is the mean of y_i''; y_i' is the ith predicted emission generated by the log-transformed approach; $\bar{y}'$ is the mean of y_i'; $r_{o,n}$ is the correlation coefficient between the observed emissions and the predicted emissions generated by the nonlinear approach; and $r_{o,l}$ is the correlation coefficient between the observed emissions and the predicted emissions generated by the log-transformed approach.

One condition that can make the two correlation coefficients to assume very similar values is that the follow linear equation is nearly true:

$$y_i' = a + by_i'' \qquad (S6.3)$$

for some intercept a and positive slope b, because replacing y_i' in Eq. (S6.2) by Eq. (S6.3), we get

$$r_{o,l} = \frac{\Sigma(y_i - \bar{y})(a + by_i'' - (a + b\bar{y}_i''))}{\sqrt{\Sigma(y_i - \bar{y})^2}\sqrt{\Sigma(a + by_i'' - (a + b\bar{y}_i''))^2}} = \frac{\Sigma(y_i - \bar{y})(by_i'' - b\bar{y}_i'')}{\sqrt{\Sigma(y_i - \bar{y})^2}\sqrt{\Sigma(by_i'' - b\bar{y}_i'')^2}} = \frac{b\,\Sigma(y_i - \bar{y})(y_i'' - \bar{y}_i'')}{b\sqrt{\Sigma(y_i - \bar{y})^2}\sqrt{\Sigma(y_i'' - \bar{y}_i'')^2}} =$$

$$\frac{\Sigma(y_i - \bar{y})(y_i'' - \bar{y}_i'')}{\sqrt{\Sigma(y_i - \bar{y})^2}\sqrt{\Sigma(y_i'' - \bar{y}_i'')^2}} = r_{o,n} \qquad (S6.4)$$

Since the Comparable R-squares in question are simply the squares of these correlation coefficients, if Eq. (S6.3) is nearly true, then the Comparable R-square achieved by the log-transformed approach will be nearly equal to the Comparable R-square achieved by the nonlinear approach. The important point here is that *irrespective of the specific values of a and b*, $r_{o,l}$ will be nearly equal to $r_{o,n}$, as long as Eq. (S6.3) is nearly true. In other words, *no matter how much the log-transformed approach has under-predicted the grand mean and linearly changed the pattern of the predicted emissions, the comparable R-square of the log-transformed approach will be nearly equal to the comparable R-square of the nonlinear approach, as long as the linear equation in Eq. (S6.3) is nearly true.*

In the four panels of Figure S2, we plotted y_i' against y_i'' for the four ecosystems separately. None of the four scatter diagrams displayed a perfect linear aliagnment. The diagrams of the Temperate Forest and Temperate Grassland fitted a linear equation very well, with the R-square being 0.987 and 0.984, respectively. In other words, a linear equation could relate the predicted emissions of the log-transformed and nonlinear approaches very well for these two ecosystems. As a consequence, the Comparable R-square achieved by the log-transformed approach was only 0.007 less than the Comparable R-square achieved by the nonlinear method for both ecosystems. With greater dispersions (relative to the grand mean), the scatter diagrams of the Desert and the Semiarid Mediterranean Marsh fitted a linear equation less well, with the

R-square being 0.942 and 0.962, respectively. As a consequence, the gap in the Comparable R-square achieved by the two approaches turned out to be greater: 0.031 for the Desert and 0.016 for the Semiarid Mediterranean Marsh.

The apparently small gaps in the Comparable R-square between the two approaches misleadingly suggested that the two approaches were similarly useful. In every ecosystem, there was a clear pattern of *systematic bias*: for the combinations of temperatures and moistures that were associated with relatively high emissions, the log-transformed approach tended to under-predict rather seriously. This bias was reflected by the fact that the slopes of all four linear equations were clearly less than 1: being 0.886 for the Temperate Forest, 0.864 for the Temperate Grassland, 0.665 for the Desert, and 0.791 for the Semiarid Mediterranean Marsh. These findings confirmed that the Comparable R-square is largely *blind to the systematic bias in the predicted emissions of the log-transformed approach*.

Note that for each ecosystem, the systematic bias in question forced the slope of the linear equation to be clearly less than 1, so that the estimated value of the intercept (a) in Eq. (S6.3) was no longer a useful indicator to reflect the fact that the predicted grand mean was substantially lower for the log-transformed approach than for the nonlinear approach. In each ecosystem, the log-transformed approach's under-prediction of the grand mean by more than 10% resulted mainly from its under-predictions for the combinations of temperatures and moistures that were associated with high emissions. In other words, the serious under-prediction of the grand mean resulted mainly from the *systematic bias* in the under-prediction problem.

In sum, being largely blind to the systematic bias in the predictions by the log-transformed, *the Comparable R-square is of little usefulness for comparing the relative merits of the two approaches.* However, *for the nonlinear approach, it is definitely a useful indicator of a* *model's predictive power, because the nonlinear approach does not have the under-prediction problem of the log-transformed approach.*

Text S7. Failure of Log-transformation to Normalize the Conditioal Distributions of the Dependent Variable

An argument for log-transforming the dependent variable of a regression model is that it can normalize the distribution of the values of the dependent variable, conditional on any combination of the values of the explanatory factors. In other words, for any combination of the values of the explanatory factors, the distribution of the values of the dependent variable can be made to look more like a symmetric bell. Here we want to show that this argument is not valid for the data set of the Temperate Forest, which has a very large sample size (N=15,523) so that rather smooth and reliable conditional distributions can be generated for some bins.

We choose to look at the conditional distributions in the (8, 15)-bin and the (12, 15)-bin, which have relative large sample sizes. From the 4 panels of Figure S3, we find that in both bins, log-transformation pushes the mode too far to the right side of the range. The skewness is changed from 0.36 to -3.12 in the (8, 15)-bin, and from 0.75 to -1.07 in the (12, 15)-bin. Thus, log-transformation failed to achieve the goal of normalizing the conditional distributions of the dependent variable.

Since normalizing the conditional distributions of the dependent variable implies the normalization of the distribution of the error term, it can slao be expected that the log-transformed approach will perform less well than the nonlinear approach in making the distribution of the residuals close to being bell-shaped. The two panels of Figure S4 show the the distributions of the residuals generated by these two approaches, after applying the GG-model to the data of the Temperate Forest. With a small positive skewness (0.73), the distribution of the residuals generated by the nonlinear approach is rather close to being normal. With a very negative skewness (-2.53), the distribution of the residuals generated by the nonlinear approach is far from being normal.

In sum, for our empirical data, log-tansforming the dependent variable failed to normalize both the conditional distributions of the dependent variable and the residuals.

Text S8. Statistical Inference without the Homoscedasticity Assumption for the Error Term in the Nonlinear Approach

For notational simplicity, we let $h(\boldsymbol{x}_i, \boldsymbol{\beta}) = e^{f(\boldsymbol{x}_i, \boldsymbol{\beta})} = e^{\boldsymbol{\beta}'\boldsymbol{x}_i}$ so that the model for the nonlinear approach can be written as

$$y_i = h(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i = e^{\boldsymbol{\beta}'\boldsymbol{x}_i} + \varepsilon_i \qquad for\ i = 1, 2, \ldots, n \qquad (S8.1)$$

where the error term ε_i is assumed to have the mean being 0 but is not subject to the restriction of the homoscedasticity assumption, because of the fact that the dispersion of the observed CO_2 emissions typically tends to be greater at higher soil temperatures. Note that the first element of the column vector $\boldsymbol{x}_i$ is 1, because the first element of the row vector $\boldsymbol{\beta}'$ is the intercept. Both of these two vectors have (K+1) elements, where K is the number of explanatory variables.

The unknown coefficients are estimated by minimizing the following function with respect to $\boldsymbol{\beta}$:

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta}))^2 \qquad \text{(S8.2)}$$

Although this minimization problem does not have a closed form solution in most applications, the best solution can be found iteratively from the following formula:

$$\boldsymbol{B} = \widetilde{\boldsymbol{\beta}} + \lambda(\boldsymbol{J}'\boldsymbol{J})^{-1}[\boldsymbol{J}'(\boldsymbol{y} - \widetilde{\boldsymbol{h}})] \qquad \text{(S8.3)}$$

where $\boldsymbol{y}$ is an n-by-1 column vector whose ith element is y_i; $\widetilde{\boldsymbol{h}}$ is an n-by-1 column vector whose ith element is $h(\boldsymbol{x}_i, \widetilde{\boldsymbol{\beta}})$; $\boldsymbol{J}$ is the n-by-(K+1) Jacobian matrix (i.e. the matrix of the partial derivatives of $h(\boldsymbol{x}_i, \widetilde{\boldsymbol{\beta}})$ with respect to the unknown coefficients); $\boldsymbol{J}'$ is the transpose of $\boldsymbol{J}$; $\widetilde{\boldsymbol{\beta}}$ is a guessed value of $\boldsymbol{\beta}$; $\boldsymbol{B}$ is an improved value of $\boldsymbol{\beta}$; and λ is a positive scalar (called "step size") that can be set to less than 1 (say 0.5) by the researcher to ensure convergence with an arbitrary starting value for $\widetilde{\boldsymbol{\beta}}$. Our experience over many years is that starting with $\widetilde{\boldsymbol{\beta}} = \boldsymbol{0}$, there is always a λ small enough to achieve convergence.

Let the estimated ith error term be:

$$\tilde{\varepsilon}_i = y_i - h(\boldsymbol{x}_i, \boldsymbol{B}) = y_i - e^{\boldsymbol{B}'x_i} \qquad \text{(S8.4)}$$

where $\boldsymbol{B}$ is the best estimate of $\boldsymbol{\beta}$. Based on Angrist & Pischke (2009, p. 45), it can be shown that replacing the homoscedasticity assumption by the heteroscedasticity assumption for the error term results in the asymptotic covariance matrix of $\boldsymbol{B}$ being:

$$\widetilde{\boldsymbol{V}} = (\boldsymbol{J}'\boldsymbol{J})^{-1}(\boldsymbol{J}'\widetilde{\boldsymbol{E}}\boldsymbol{J})(\boldsymbol{J}'\boldsymbol{J})^{-1} \qquad \text{(S8.5)}$$

where $\widetilde{\boldsymbol{E}}$ is an n-by-n diagonal matrix whose ith diagonal element is $\tilde{\varepsilon}_i{}^2$.

Let $\tilde{S}_k$ be the square root of the kth diagonal element of $\widetilde{\boldsymbol{V}}$. In econometrics, $\tilde{S}_k$ is called the *robust standard error* of the estimator of the kth unknown coefficient.

Let B_k be the kth element of the best estimated coefficient vector $\boldsymbol{B}$, and let

$$\tilde{t}_k = \frac{\boldsymbol{B}_k}{\tilde{S}_k} \qquad \text{(S8.6)}$$

This $\tilde{t}_k$ is the corresponding *robust t-statistic*. When the sample size is not too small (e.g. n > 50), the coefficient associated with a robust t-statistic that is equal to or greater than 2.0 in magnitude may be considered as being significantly different from zero, because the probability in the two tails of the t-distribution beyond ± 2.0 is about 0.05.

In Table S3, the robust standard errors and the robust t-statistics are shown with the corresponding non-robust quantities that are based on the homoscedasticity assumption for the error term. Clearly, the corresponding values can differ markedly. However, for testing the null hypotheses about the coefficients, both sets of t-statistics led to the rejection of all null hypotheses, because all of them turned out to much greater than 2.0 in magnitude. Thus, for our empirical problem, the violation of the homoscedasticity assumption is not an important issue at all. The major methodological concern should be on the pervasiveness and systematic bias of the under-predictions of the log-transformed approach, rather than on whether this assumption is satisfied or not.

Text S9. Application of the Gaussian-Gamma Specification of the Regression Model to the Data of the Temprerate Forest via the GENMOD Approach

At the recommendation of a reviewer, here we show the findings of applying our GG-model to the data of the Temperate Forest of southern Ontario via the GENMOD approach. GENMOD is a procedure of SAS software that was designed to fit research data to a Generalized Linear Model, which is a highly sophisticated and analytically elegant model that subsumes many useful statistical models as special cases (a well-written document on GENMOD can be accessed at http://www.math.wpi.edu/saspdf/stat/chap29.pdf). It is highly likely that professional statisticians would recommend this model to empirical researchers. Thus, whether the approach using GENMOD is better than the nonlinear approach is an important issue to investigate.

To impliment the GENMOD approach properly, we formulate the GG-model in two step. First, the expected value of the CO_2 emission (μ_i) is linked to the explanatory factors as:

$$\mu_i = e^{\beta_0+\beta_{11}T_i+\beta_{12}{T_i}^2+\beta_{21}M_i+\beta_{22}\ln(M_i)} \qquad \text{(S9.1)}$$

where T_i is the ith observation of soil temperature; M_i is the ith observation of soil moisture; and the symbols represented by the Greek alphabets are unknown coefficients to be estimated.Next, conditional on the values of the explanatory factors, the ith observation of CO$_2$ emission (y_i) is assumed to be a random variable having the following Gamma density function:

$$g(y_i|\mu_i,\omega) = \frac{1}{\Gamma(\omega)y_i}\left(\frac{y_i\omega}{\mu_i}\right)^{\omega} e^{-\frac{y_i\omega}{\mu_i}} \qquad (S9.2)$$

where ω is yet another unknown coefficient to be estimated, and $\Gamma(\omega) = \int_0^{\infty} x^{\omega-1} e^{-x} dx$ is the well-known Gamma function (SAS Institute, undated, p. 1404). In the output of the GENMOD procedure, the estimated value of ω is called "scale". Note that in $g(y_i|\mu_i,\omega)$, the quantities to the right of the verticle line are considered as given.

For a continuous dependent variable that can assume only positive values, the GENMOD procedure also offers the alternative of letting the conditional density function be the density function of an inverse Gaussian distribution. Since we found that this distribution fitted less well to our data, we will ignore it.

In order to avoid confusion, we mention that the Gamma form of the dependency of CO$_2$ emission on soil moisture need not have anything to do with the assumption that conditional on the explanatory factors, y_i has a Gamma distribution. In the nonlinear approach, we made no distributional assumption at all. The Gamma part of the name "Gaussian-Gamma specification" came from the fact that $e^{\beta_{21}M_i+\beta_{22}\ln(M_i)}$ can be rewritten as ${M_i}^{\beta_{22}}e^{\beta_{21}M_i}$, which is the product of a power function and an exponential function. The reason for the name is that the integrant of the Gamma function $\Gamma(\omega)$ is also the product of a power function and an exponential function.

A simple propornationality property of the Gamma distribution is that

$${\sigma_i}^2 = \frac{{\mu_i}^2}{\omega} \qquad (S9.3)$$

where $\sigma_i{}^2$ is the conditional variance of CO_2 emission. In other words, the conditional variance is proportional to the square of the conditional mean, which implies that the conditional standard deviation is proportional to the conditional mean.

We now use the propornationality property to see whether the assumption of Gamma distribution is reasonable for the data of the Temperate Forest. For the 20 bins created previously from the rounded values of temperature and moisture, we computed the conditional means, the conditional variances, and the conditional standard deviations. In Figure S5, we plotted the conditional variances against the square of the conditional means in Panel A, and we also plotted the conditional standard deviations against the conditional means in Panel B. Panel B should be more useful to empirical researchers, because the axes have the original physical unit (µmol/m^2/s). Both panels show that the propernationality property is largely valid, although the linear alignment is not very tight. Thus, the GENMOD approach seems to be alright.

When a researcher has the data at hand and is ready to estimate the unknown coefficients, the density function $g(y_i|\mu_i,\omega)$ can be considered as the likelihood function in the sense that is indicates the likelihood of the observed value of y_i in the data set. To reflect this change of viewpoint, the likelihood function for the ith observation is written as

$$g(\mu_i,\omega|y_i) = \frac{1}{\Gamma(\omega)y_i}\left(\frac{y_i\omega}{\mu_i}\right)^{\omega} e^{-\frac{y_i\omega}{\mu_i}} \qquad (S9.4)$$

Here y_i is treated as given, and the challenge is to make the best guess of the values of ω and the unknown coefficients $(\beta_0, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})$ hidden in μ_i.

The GENMOD procedure estimates the unknown coefficients by using the Maximum Likelihood (ML) method that tries to maximize the natural log of the likelihood function of the whole sample:

$$L(\beta_0, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \omega \,|\, y_1, y_2, \dots y_n) = \ln(\prod_{i=1}^{n} g(\mu_i, \omega|y_i)) = \sum_{i=1}^{n} \ln(g(\mu_i, \omega|y_i)) \qquad (S9.5)$$

where n is the sample size. The ML method does not yield a closed-form solution. To estimate the unknown coefficients iteratively, this procedure uses a ridge-stabilized Newton-Raphson algorism (SAS Institute, undated, p. 1406) that can presumably achieve convergence without spending too much time.

To estimate the unknown coefficients and to generate the predicted emissions, only the following 4 statements are needed.

```
PROC GENMOD DATA = MY_DATA_SET  ;
MODEL EMISSION = TEMP SQ_TEMP MOISTURE LN_MOISTURE/DIST=GAMMA LINK=LOG;
OUTPUT OUT=OUTPUT_DATA_SET  P=P_EMISSION;
RUN;
```

In these statements, MY_DATA_SET is the name of the input data set; SQ_TEMP is the square of temperature; LN_MOISTURE is the natural log of moisture; OUTPUT_DATA_SET is the name of the data set that will contain not only the predicted emissions but also the values of all input variables; and P_EMISSION is the name of the predicted variable. The option LINK=LOG simply means that the linear-in-coefficients expression in Eq. (S9.1) is supposed to be the natural log of the expected emission. It is important to realize that this option does not ask for the log-transformation of the observed emissions.

The estimation result of the GENMOD approach is shown in Panel A of Table S4. All estimated coefficients have identical signs to the corresponding coefficients estimated by the nonlinear approach (Panel B of Table S4). The GENMOD approach performed slightly worse than the nonlinear approach in terms of the comparable R-square (0.549 versus 0.551) and the severity of the under-prediction of the grand mean emission (0.05% versus 0.01%). The intercept and slope of the linear regression of the observed emissions on the predicted emissions from the

GENMOD approach turned out to be -0.06 and 1.01 respectively, which were somewhat worse than the corresponding intercept and slope (0.00 and 1.00 respectively) of the nonlinear approach in getting close to the unbiased ideal of 0 and 1 respectively. Overall, the GENMOD approach appeared to be somewhat worse than the nonlinear approach but was also quite capable of avoiding the under-prediction problem of the log-transformed approach.

A serious shortcoming of the GENMOD approach was its failure to yield a statistically significant negative coefficient for the descent parameter of the moisture function. It turned out to be -0.0106 with a t-statistic of -1.3 (which was obtained from taking the square-root of the reported Wald chi-square) and a p-value of 0.21. In contrast, the corresponding parameter yielded by the nonlinear approach was -0.0422, with a robust t-statistic of -4.9 and a p-value of < 0.001. If we followed the statistical convention of accepting the null hypothesis whenever the p-value is less than 0.05, we would be compelled by the GENMOD approach to infer that the true descent parameter is 0, contradicting the experimental findings of Howard & Howard (1993) that the descent parameter of the moisture function was negative for all of 8 different types of soil.

To understand this shortcoming of the GENMOD approach, we tried to learn more about the nature of the likelihood and log-likelihood functions of an observed emission in the data set. Figure S6 shows the likelihood and log-likelihood functions for the observed emission of 5 µmol/m^2/s. The functions were generated from the assumption that conditional on the explanatory factors, CO_2 emission is a random variable having a Gamma distribution with a scale parameter of 4. Both functions are clearly non-symmetric around the mode of 5 µmol/m^2/s. The non-symmetric pattern has two biased properties.

Property 1: For two guessed values of emission that are of equal distance around the observed value, the Maximum Likelihood method dislikes the lower guessed value more than the higher guessed value. In other words, it dislikes the guessed value of 4 more than the guessed value of 6, because the log-likelihood is -0.3392 for the former and -0.3173 for the latter.

Property 2: The intensity of the dislike against the lower guessed value becomes greater as the guessed values get further away from the observed value. For example, its dislike against 3 in the pair of 3 and 7 is greater than its dislike against 4 in the pair of 4 and 6, because the diffence in log-likelihood between the guessed values of 4 and 6 is (-0.3392)-(-0.3173) = -0.0219, whereas the diffence in log-likelihood between the guessed values of 3 and 7 is (-0.6899)-(-0.5007) = -0.1892.

We changed the observed emission to a few other values such as 3, 7, and 9 and found that these two biased properties remained true. Thus, being the sum of the log-likelihoods of the individual observations, the log-likelihood function of the whole data set is subject to the complex effects of the biases among many individual observations.

We suspect that as a consequence of these two properties, the GENMOD approach lifted the right tail of the moisture function and made the value of the descent parameter not significantly different from 0. In contrast, none of the squared residuals in the objective function of the Least Squares method of the nonlinear approach have these two biased properties. In the objective function, any negative residual is as influential as any positive residual of the same magnitude. Thus, the nonlinear approach is less likely to yield misleading estimation results.

It is useful to note that the GENMOD approach did not yield misleading information about the effect of soil temperature. We suspect that this is related to the fact that soil temperature is much stronger in explanatory power than soil moisture. In other words, the scatter of the observed CO_2 emissions is greater around a moisture function than around a tmperature function. Therefore, the second biased property of the likelihood functions suggests that the

GENMOD approach is more likely to yield misleading information about the effect of soil moisture than about the effect of soil moisture.

In sum, we make the tentative statement that the nonlinear approach is superior over the GENMOD approach in the sense that it is less likely to yield misleading results that contradict experimental findings. To assess the robustness of this superiority, we carry out a randomized experiment in Text S10.

Text S10. Randomized Experiment to Confirm the Superiority of the Nonlinear Approach over the GENMOD Approach

To assess the robustness of the superiority of the nonlinear approach over the GENMOD approach in avoiding the risk of generating misleading information about the descent parameter of soil moisture, we carried out a randomized experiment in the following way. First, we took 10 mutually exclusive and all inclusive random subsamples from the field data of the Temperate Temperate Forest of southern Ontario (n=15,523). The sampling was made possible by the SURVEYSELECT procedure of SAS language. Second, for each subsample, we fitted the Gaussian-Gamma model via the nonlinear and GENMOD approaches, respectively. The resulting information about the descent parameter of soil moisture was reported in Table S5.

We found the following contrasts between the two approaches. First, the nonlinear approach consistently generated a negative parameter for every subsample, whereas the GENMOD approach generated a negative parameter for 8 subsamples and a positive parameter for 2 subsamples. Second, two of the parameters generated by the nonlinear approach were significantly different from 0, whereas none of the parameters generated by the GENMOD approach were significantly different from 0. Third, the mean of the parameter across the 10 subsamples differed greatly between the two approaches by a factor of about 4 (-0.0404 versus -0.0106).

Therefore, the randomzed experiment confirmed that the nonlinear approach is superior over the GENMOD approach in avoiding the generation of misleading information about the effect of soil moisture on CO_2 emission. A more general inference is that when the explanatory factor in question does not have a strong explanatory power and the observational data is quite noisy, a researcher would have a lower risk of getting physically nonsensical findings by adopting the the nonlinear approach instead of the GENMOD approach.

It is worth keeping in mind that the assumption of Gamma distribution in the GLM is highly consistent with the physical process. For forecasting and simulation, it is a better choice than other distributions (e.g. normal distribution).

Text S11. How to Use Our SAS Module to Estimate the Unknown Coefficients of the Gaussian-Gamma Model and to and Compute the Robust Statistics according to the Nonlinear Approach

Part A: SAS Program That Demonstrates the Use of the SAS Module

```
* FILE NAME: MODELLING_CO2_EMISSION_2019.SAS;
* BECAUSE LOG-TRANSFORMING THE DEPENDENT VARIABLE OF A REGRESSION MODEL
* RESULTS IN A SYSTEMATICALLY BIASED UNDER-PREDICTION PROBLEM, THIS
* PROGRAM DEMOSTRATES HOW YOU CAN ESTIMATE THE UNKNOWN COEFFICIENTS OF
* YOUR REGRESSION MODEL, WITHOUT LOG-TRANSFORMING YOUR DEPENDENT VARIABLE;
* IN THIS DEMONSTRATION, WE USE THE GAUSSIAN-GAMMA MODEL TO PREDICT SOIL
* CO2 EMISSION BY (1) SOIL TEMPERATURE AND (2) SOIL MOISTURE, USING THE
* NONLINEAR LEAST-SQUARES METHOD;
* THE ITERATIVE ALGORITHM FOR THIS METHOD IS DONE IN
"NNL_REG_ROBUST_MODULE.SAS",
* WHICH IS WRITTEN IN SAS/IML;
ods html file="D:\_SAS_DEMO_2019\MODELLING_CO2_EMISSION_2019.html";
* PLEASE SPECIFY THE PATH TO THE FOLDER OF (1)THE RAW DATA FILE AND THE AND
* (2) THE SAS MODULE FOR NONLINEAR LEAST-SQUARES ESTIMATION;
LIBNAME SD 'D:\_SAS_DEMO_2019';
PROC MEANS DATA= SD.RAW_DATA N NMISS MIN MAX MEAN STD CV RANGE;
TITLE "SUMMARY OF THE RAW DATA: HALF-HOURLY RECORDS IN THE YEAR 2014.";
TITLE2 "RS = CO2 EMISSION FROM SOIL (µMOL/M^2/S).";
TITLE3 "TS_5CM = TEMPERATURE OF SOIL AT THE DEPTH OF 5 CM (DEGREE C).";
TITLE4 "MS_30CM = MOISTURE OF SOIL AT THE DEPTH OF 30CM (VOLUME/VOLUME).";
```

```
RUN;
DATA F1;
TITLE "DATA CONTAINING ORIGINAL AND TRANSFORMED VARIABLES FOR APPLYING THE
GAUSSIAN-GAMMA MODEL:";
  SET SD.RAW_DATA_2014_S_ONTARIO;
* CREATE THE SQUARE OF TEMPERATURE;
  SQTs_5cm = Ts_5cm * Ts_5cm;
* CHANGE THE UNIT OF MOISTURE FROM PROPORTION TO PERCENTAGE;
  Ms_30cm = Ms_30cm * 100;
* LOG TRANSFORMING SOIL MOISTURE;
  LNMs_30cm = LOG(Ms_30cm);
* CREATE THE COLUMN REPRESENTING THE INTERCEPT;
INTERCEPT=1;
* YOU MUST NOT FORGET THE INTERCEPT, UNLESS YOU DO NOT WANT THE INTERCEPT TO
BE PART OF YOUR MODEL!!!;
PROC MEANS DATA=F1 N NMISS MIN MAX MEAN STD CV RANGE;
RUN;
*********************************************************************;
TITLE "GAUSSIAN_GAMMA MODEL FOR THE DATA OF THE TEMPERATE FOREST. ";
TITLE2 "DEPENDENT VAR (RS) = SOIL CO2 EMISSION ( µMOL CO2/m^2/s ).";
TITLE3 "TS_5CM = TEMPERATURE OF SOIL AT THE DEPTH OF 5 CM (DEGREE C).";
TITLE4 "MS_30CM = MOISTURE OF SOIL AT THE DEPTH OF 30CM
(VOLUME/VOLUME*100%).";
TITLE5 "RESULT OF THE NONLINEAR ESTIMATION METHOD:";
* SPECIFY THE NAME OF THE DEPENDENT VARIABLE;
%LET DEP_VAR= RS;
* SPECIFY THE NAMES OF THE EXPLANATORY VARIABLES (DON'T FORGET THE
INTERCEPT);
%LET INDEP= INTERCEPT  Ts_5cm  SQTs_5cm    Ms_30cm  LNMs_30cm  ;
DATA INPUT_FOR_MODEL;
* CREATE THE INPUT DATA MATRIX;
  SET F1  (KEEP= &DEP_VAR  &INDEP);
RUN;
PROC IML ;
use INPUT_FOR_MODEL;
%INCLUDE "D:\_SAS_DEMO_2019\NNL_REG_ROBUST_MODULE.SAS";
RUN  NNL_REG_ROBUST;
PROC MEANS DATA=Y_HAT_FILE N NMISS MIN MAX MEAN STD CV RANGE;
TITLE "THE SAS DATA SET 'Y_HAT_FILE' CONTAINS ALL VALUES OF THE PREDICTED CO2
EMISSION: DENOTED BY Y_HAT.";
TITLE2 "ITS SUMMARY STATISTICS ARE SHOWN HERE.";
RUN;
*********** END OF THE PROGRAM ********************************************;
```

Part B: The SAS Module for Estimating the Unknown Coefficient and for Generating the Robust t-statistics and the Predicted Values of the Dependent Variable

```
* MODULE FILE NAME: NNL_REG_ROBUST_MODULE.SAS;
* THIS MODULE IS FOR ESTIMATING THE COEFFICIENTS OF A MODEL IN THE
* EXPONENTIAL FAMILY BY A NONLINEAR LEAST SQUARES METHOD,
* ORIGINALLY WRITTEN BY KAO-LEE LIAW IN 2013;
* SPECIFICATION OF VARIABLES THAT CAN BE CHANGED BY THE USER:
*   N_ITER = MAXIMUM NO. OF ITERATIONS OF THE NEWTON-RAPHSON ALGORITHM.
*   DETAIL = 1 (IF YOU WANT TO SEE THE INFORMATION AT EACH ITERATION).
*   BL_SIZE = NO. OF OBSERVATIONS PER BUNCH.
*            IT MUST NOT BE BIGGER THAN THE SAMPLE SIZE. IF YOUR COMPUTER
*            HAS A SMALL MEMORY, YOU CAN REDUCE THE VALUE OF BL_SIZE.
*   STEPSIZE = A SCALAR TO ADJUST THE SIZE OF THE CHANGE IN THE PARAMETER
*              VECTOR FROM ONE ITERATION TO THE NEXT. IT MUST BE A
*              POSITIVE VALUE LESS THAN OR EQUAL TO 1. WHEN IT IS SET TO 1,
*              THE COMPUTATION TAKES THE LEAST AMOUNT OF TIME BUT THE RISK
*              OF DIVERGENCE MAY BE THE HIGHEST.
*              WHEN CONVERGENCE FAILED, TRY USING A SMALLER STEP SIZE.
*   Y = COLUMN VECTOR OF THE DEPENDENT VARIABLE.
*   INDEP = COLUMN VECTOR WITH NAMES OF EXPLANATORY VARIABLES,
*           WITH THE FIRST VARIABLE BEING THE COLUMN REPRESENTING THE
*           INTERCEPT.
*  Technical Advice:
*  IF YOU GET AN "OVERFLOW" ERROR MESSAGE, YOU SHOULD CHANGE THE SCALE OF
*  YOUR DEPENDENT VARIABLE BY DIVIDING A LARGE NUMBER (E.G. 1000) INTO IT;
START NNL_REG_ROBUST ;
N_ITER = 200;
DETAIL = 0;
STEPSIZE = 0.8;
READ all var{&DEP_VAR} into Y;
READ all var{&indep} into X;
NOBS=NROW(Y);
IF NOBS > 500 THEN BL_SIZE = 500;
ELSE BL_SIZE = NOBS;
LEFTOVER=MOD(NOBS,BL_SIZE);
NBUNCH=INT(NOBS/BL_SIZE);
Y_MEAN =SUM(Y)/NOBS;
B = REPEAT(0,NCOL(X),1); OLDB=B+1; /* STARTING VALUES */
* BEGINNING OF ITERATIONS;
DO ITER = 1 TO N_ITER;
      OLDB=B;
      XPX=REPEAT(0,NCOL(X),NCOL(X));
      XPY=REPEAT(0,NCOL(X),1);

     XPX_RB=XPX;

      SSQ=0;   SSQ0=0;
      RSRMSQ=0;
      DO IBUNCH=1 TO NBUNCH;
        N1=(IBUNCH-1)*BL_SIZE+1;
        N2=N1+BL_SIZE-1;
 Y_NULL=REPEAT(Y_MEAN, BL_SIZE);
        Y_HAT=EXP(X[N1:N2, ]*B);
       DER=X[N1:N2, ]#Y_HAT;
       * DER is part of the Jacobian;
     DIFF_I=Y[N1:N2] - Y_HAT;
       DIFF_0=Y[N1:N2] - Y_NULL;
      SSQ = SSQ + SUM(DIFF_I # DIFF_I);
   SSQ0 = SSQ0 + SUM(DIFF_0 # DIFF_0);
          DERT=(DER)`;
        DERT_RB=(DER#DIFF_I#DIFF_I)`;
            XPX_RB= XPX_RB + DERT_RB*DER;

        XPX= XPX + DERT*DER; /* BUILDING UP THE INFORMATION MATRIX */
        XPY= XPY + DERT*DIFF_I;
      END;/*IBUNCH*/
```

```
    IF LEFTOVER > 0 THEN DO;
        N1=NBUNCH*BL_SIZE+1;
        N2=N1+LEFTOVER-1;
 Y_NULL=REPEAT(Y_MEAN, LEFTOVER);
        Y_HAT=EXP(X[N1:N2, ]*B);
       DER=X[N1:N2, ]#Y_HAT;
     DIFF_I=Y[N1:N2] - Y_HAT;
       DIFF_0=Y[N1:N2] - Y_NULL;
      SSQ = SSQ + SUM(DIFF_I # DIFF_I);
SSQ0 = SSQ0 + SUM(DIFF_0 # DIFF_0);
          DERT=(DER)`;
        DERT_RB=(DER#DIFF_I#DIFF_I)`;
        XPX_RB= XPX_RB + DERT_RB*DER;

        XPX= XPX + DERT*DER; /* BUILDING UP THE INFORMATION MATRIX */
        XPY= XPY + DERT*DIFF_I;
                          END; /* At this point, the construction of the
Information Matrix is completed */
      btransp = b`;
      IF DETAIL = 1 THEN print iter SSQ btransp;
      XPX = INV(XPX);/* NOW XPX IS THE INVERSE OF INFORMATION MATRIX  */
B = B + STEPSIZE*( XPX * XPY);/* REDUCE THE STEP SIZE, IF THE MODULE DOES NOT CONVERGE */
IF MAX(ABS(B-OLDB))<1E-8 THEN DO;/* BEGINNING OF THE FINAL PART*/
DF_ESS = NOBS - NCOL(X) ;
NVAR=NCOL(X)-1;
RSRMSQ=SQRT(SSQ/DF_ESS);/* RESIDUAL ROOT_MEAN_SQUARE */
RSRMSQ0=SQRT(SSQ0/(NOBS-1));/* RESIDUAL ROOT_MEAN_SQUARE OF THE NULL MODEL */
R_SQUARE = (SSQ0 - SSQ) / SSQ0;
ADJ_R_SQ= 1 - (RSRMSQ/RSRMSQ0)**2;
PRINT  NOBS NVAR  Y_MEAN ITER ;
PRINT  SSQ SSQ0 RSRMSQ DF_ESS  RSRMSQ0 R_SQUARE ADJ_R_SQ;

   CV_RB=XPX * XPX_RB * XPX;
   STD_ERR_ROBUST = SQRT(VECDIAG(CV_RB));
   T_RATIO_ROBUST = B/STD_ERR_ROBUST;
   P_VALUE = (1 - PROBT(ABS(T_RATIO_ROBUST),DF_ESS))*2;

   STD_ERR_HOMO = SQRT(VECDIAG(XPX))*RSRMSQ;
   T_RATIO_HOMO = B/STD_ERR_HOMO;
   VAR_NAME={&INDEP}`;
   COEFFICIENT = B ;
   PRINT "ESTIMATED RESULT: STD_ERR_HOMO, T_RATIO_HOMO, and P_VALUE_HOMO are
based the HOMOSCEDASTICITY assumption.";
   PRINT "NOTE: THE INFORMATION IN THIS TABLE IS ALSO CONTAINED IN THE SAS
DATA SET 'PARM_FL'.";
   PRINT VAR_NAME COEFFICIENT[FORMAT=13.6] STD_ERR_HOMO[FORMAT=13.6]
T_RATIO_HOMO[FORMAT=10.2] P_VALUE_HOMO[FORMAT=10.6];

   PRINT "ESTIMATED RESULT: STD_ERR_ROBUST, T_RATIO_ROBUST, and P_VALUE_ROUST
are based on the HETEROSCEDASTICITY assumption.";
   PRINT "NOTE: THE INFORMATION IN THIS TABLE IS ALSO CONTAINED IN THE SAS
DATA SET 'PARM_FL'.";
   PRINT VAR_NAME COEFFICIENT[FORMAT=13.6] STD_ERR_ROBUST[FORMAT=13.6]
T_RATIO_ROBUST[FORMAT=10.2] P_VALUE_ROBUST[FORMAT=10.6];
/* CREATE THE DATA SET CONTAINING THE ESTIMATED PARAMETERS AND RELATED
STATISTICS */
```

```
   CREATE PARM_FL VAR{VAR_NAME COEFFICIENT  STD_ERR_ROBUST T_RATIO_ROBUST
P_VALUE_ROBUST STD_ERR_HOMO T_RATIO_HOMO P_VALUE_HOMO};
   APPEND;
   CLOSE PARM_FL;
/* CREATE THE DATA SET CONTAINING THE PREDICTED VALUES OF THE DEPENDENT
VARIBLE */
         CREATE Y_HAT_FILE VAR{Y_HAT};
DO IBUNCH=1 TO NBUNCH;
        N1=(IBUNCH-1)*BL_SIZE+1;
        N2=N1+BL_SIZE-1;
         Y_HAT=EXP(X[N1:N2, ]*B);
             APPEND;
 END;
 IF LEFTOVER > 0 THEN DO;
        N1=NBUNCH*BL_SIZE+1;
        N2=N1+LEFTOVER-1;
        Y_HAT=EXP(X[N1:N2, ]*B);
            APPEND;
                      END;
        CLOSE Y_HAT_FILE;
        STOP;
                                    END;/*END OF THE FINAL PART*/
   END; /* END OF ITER LOOP:  THE MAXIMUM NUMBER OF ITERATIONS IS REACHED
HERE  */
      PRINT "!!! WARNING!!!: THE ESTIMATED PARAMETERS MAY NOT BE
MEANINGFUL,";
      PRINT "BECAUSE THE MAXIMUM NO. OF ITERATIONS IS REACHED.";
                                       STOP;
FINISH ;/* END OF NNL_REG_ROBUST */
*  OUTPUT VARIABLES:
*   NOBS = NO. OF OBSERVATIONS.
*   NVAR = NO. OF SUBSTATIVE EXPLANATORY VARIABLES.
*   Y_MEAN = MEAN OF THE DEPENDENT VARIABLE.
*   ITER = THE NUNBER OF ITERATIONS AT CONVERGENCE.
*   SSQ = RESIDUAL SUM OF SQUARES.
*   SSQ0 = TOTAL SUM OF SQUARES.
*   RSRMSQ = RESIDUAL ROOT_MEAN_SQUARE.
*   RSRMSQ0 =  RESIDUAL ROOT_MEAN_SQUARE OF THE NULL MODEL.
*   R_SQUARE.
*   ADJ_R_SQ= ADJUSTED R_SQUARE.
* OUTPUT DATA SETS:
*  (1) PARM_FL:
*     VAR_NAME=A COLUMN VETCTOR CONTAINING THE VARIABLE NAMES.
*     COEFFICIENT = A COLUMN VECTOR CONTAINING THE ESTIMATED COEFFICIENTS.
*     STD_ERR_ROBUST = A COLUMN VECTOR CONTAINING THE ROBUST STANDARD ERRORS
*                OF THE COEFFICIENTS, BASED ON HETEROSCEDASTICITY ASSUMPTION.
*     T_RATIO_ROBUST = A COLUMN VECTOR OF ROBUST T-STATISTICS.
*     P_VALUE_ROBUST = A COLUMN VECTOR CONTAINING THE P_VALUES, BASED ON
*               T_RATIO_ROBUST.
*     STD_ERR_HOMO = A COLUMN VECTOR CONTAINING THE STANDARD ERRORS OF
*                    THE COEFFICIENTS, WHICH ARE BASED ON THE
*                    HOMOSCEDASTICITY ASSUMPTION.
*     T_RATIO_HOMO = A COLUMN VECTOR CONATAINING THE T_STATISTICS, WHICH ARE
*                    BASED ON THE HOMOSCEDASTICITY ASSUMPTION.
*     P_VALUE_HOMO = A COLUMN VECTOR CONTAINING THE P_VALUES, BASED ON
*                    T_RATIO_HOMO.
*  (2) Y_HAT_FILE: THIS DATA SET CONTAINS THE PREDICTED VALUES OF THE
*                    DEPENDENT VARIABLE (Y_HAT);
/*********************** END OF THE MODULE ************************/
```

Part C: Output of the Demonstration

GAUSSIAN_GAMMA MODEL FOR THE DATA OF THE TEMPERATE FOREST.
DEPENDENT VAR (RS) = SOIL CO2 EMISSION (μMOL CO2/m^2/s).
TS_5CM = TEMPERATURE OF SOIL AT THE DEPTH OF 5 CM (DEGREE C).
MS_30CM = MOISTURE OF SOIL AT THE DEPTH OF 30CM (VOLUME/VOLUME*100%).
RESULT OF THE NONLINEAR ESTIMATION METHOD:

```
                 NOBS  NVAR  Y_MEAN     ITER
                15523     4  4.8821374    24

      SSQ       SSQ0   RSRMSQ  DF_ESS  RSRMSQ0  R_SQUARE  ADJ_R_SQ
86540.224  192607.03  2.3615173   15518  3.522591  0.5506902  0.5505744
```

ESTIMATED RESULT: STD_ERR_ROBUST, T_RATIO_ROBUST, and P_VALUE_ROUST are based on the HETEROSCEDASTICITY assumption.

NOTE: THE INFORMATION IN THIS TABLE IS ALSO CONTAINED IN THE SAS DATA SET 'PARM_FL'.

```
VAR_NAME    COEFFICIENT  STD_ERR_ROBUST  T_RATIO_ROBUST  P_VALUE_ROBUST
INTERCEPT     -1.789799        0.128508          -13.93        0.000000
TS_5CM         0.130223        0.003459           37.65        0.000000
SQTS_5CM      -0.000786        0.000158           -4.99        0.000001
MS_30CM       -0.042239        0.008606           -4.91        0.000001
LNMS_30CM      1.023859        0.094738           10.81        0.000000
```

THE SAS DATA SET 'Y_HAT_FILE' CONTAINS ALL VALUES OF THE PREDICTED CO2 EMISSION: DENOTED BY Y_HAT.
ITS SUMMARY STATISTICS ARE SHOWN HERE.

```
                              The MEANS Procedure
                           Analysis Variable : Y_HAT
    N  N Miss   Minimum    Maximum       Mean    Std Dev  Coeff of Variation       Range
15523       0  1.3891025  15.3499376  4.8814432  2.6153584          53.5775658  13.9608351
```

Text S12. An R script for Estimating the Unknown Coefficients of the Gaussian-Gamma Model and for Computing the Robust Statistics according to the Nonlinear Approach

To assist users of R software to switch to the nonlinear approach, here we present an R script that can not only estimate the unknown coefficients of the Gaussian-Gamma model according the the nonlinear approach, but also compute the robust statistics (standard errors, t-statistics, and p-values) that are free from the restrictive assumption of homogenious variance for the error terms.

In our R script, we adopted the nls() function found in the stats-package of R software (R-Core Team, 2020; version 4.0.2) to estimate the unknown coefficients by the nonlinear least-squares method. We then translated the part of our SAS module that computes the robust statistics.

Following is the R script for applying the nonlinear approach to the data of the Temperate Forest of southern Ontario.

```
## START OF SCRIPT ##

## clear work space.

rm(list=ls())

## Read in the values of Rs(CO2 emission), Ts(soil temperature), and Ms(soil moisture).

## NOTE: YOU MUST MAKE SURE THAT THERE ARE NO MISSING VALUE FOR YOUR VARIABLES.

input.data <- read.csv (file="d:/rtest2020/raw_data_2014_s_ontario.csv", header=TRUE, sep = ",")

## create the additional variables sq.Ts and ln.Ma.

input.data$sq.Ts_5cm <- input.data$Ts_5cm * input.data$Ts_5cm

input.data$ln.Ms_30cm <- log(input.data$Ms_30cm)

## HERE YOU HAVE TO RE-ARRANGE THE COLUMNS OF THE INPUT DATA MATRIX TO BE

## IDENTICAL TO THE SEQUWNCE OF THEIR APPEARANCE IN THE MODEL.

input.data <- input.data[c("Rs","Ts_5cm", "sq.Ts_5cm", "Ms_30cm","ln.Ms_30cm")]

##use the nls() function to estimate the coefficients of the Gaussian-Gamma model,

##using the default algorithm = GaussNewton for the nonlinear least square method.

GG.model.fit <- nls(Rs ~ exp(b0 + b1*Ts_5cm + b2*sq.Ts_5cm + b3*Ms_30cm + b4*ln.Ms_30cm),

          data=input.data,

          start = list(b0 = 0, b1 = 0, b2 = 0, b3 = 0, b4 = 0))

## if you want to see the non-robust statistics, then activate the following statement.

#coef(summary(GG.model.fit))
```

```
## NOW YOU HAVE TO PUT THE VALUES OF THE DEPENDENT VARIABLE INTO THE COLUMN VECTOR y.

## NOTE: YOU MUST NOT CHANGE y TO ANYTHING ELSE.

y <- matrix(input.data[,1], ncol=1)

## find the number of column for the matrix x that is to be constructed next.

ncol.x <- ncol(input.data)

## NOW YOU HAVE TO PUT THE VALUES OF THE EXPLANATORY VARIABLES INTO THE MATRIX x,
STARING FROM COLUMN 2.

## NOTE: THE FIRST COLUMN OF THE MATRIX x MUST BE FILLED WITH 1.

## NOTE: YOU MUST NOT CHANGE x TO ANYTHING ELSE.

x <- as.matrix(cbind(1,input.data[,2:ncol.x]))

## NOW YOU HAVE TO PUT THE VALUES OF THE ESTIMATED COEFFICIENTS INTO THE COLUMN VECTOR
b.col.

## NOTE: YOU MUST NOT CHANGE b.col TO ANYTHING ELSE.

b.col <-  coef(GG.model.fit)

################################################################################

## Beginning of the Module for generating the robust statistics

################################################################################

nobs <-  nrow(y)

# Print the number of observations

nobs

y.mean <-  sum(y) / nobs

## y.hat is a column vector containing the predicted emissions.

y.hat <-  exp(x %*% b.col)

y.hat.mat <- matrix()

der <- matrix(0,nrow=nobs,ncol=ncol.x)

for (j in 1:ncol.x) {

der[,j] <-  x[,j]  * y.hat}

dert <- t(der)
```

```
# der is the Jacobian
diff.i <-  y - y.hat
diff.0 <-  y - y.mean
ssq <-  sum(diff.i  * diff.i)
ssq0  <-  sum(diff.0  * diff.0)
der.rb <- der
for (j in 1:ncol.x) {
der.rb[,j] <-  (der[,j] * diff.i  * diff.i)}

dert.rb <- t(der.rb)
xpx.rb <-  dert.rb %*% der
# xpx is the information matrix
xpx <-  dert %*% der
# now we change xpx to the inverse of the information matrix
xpx <-  solve(xpx)
df.ess <-  nobs - ncol(x)
nvar <-  ncol(x) -1
# rsrmsq is the Residual Root Mean Square
rsrmsq <-  sqrt(ssq / df.ess)
# rsrmsq0 is the Residual Root Mean Square of the Null Model
rsrmsq0 <-  sqrt(ssq0 / (nobs -1))
r.square <-  (ssq0 - ssq) / ssq0
# print r-square
 r.square
adj.r.square <-  1 - (rsrmsq/rsrmsq0)*(rsrmsq/rsrmsq0)
# print Adjusted R-square
adj.r.square
cv.rb <-  xpx %*% xpx.rb %*% xpx
```

```
std.err.rb <-  sqrt(diag(cv.rb))

t.ratio.rb <- b.col / std.err.rb

p.value.rb <-  2 * (1 - pt(abs(t.ratio.rb), df.ess) )

var.names <- colnames(input.data)

var.names <- replace(var.names, c(1),c("intercept"))

coefficient <- b.col

robust.stat <-  cbind(coefficient, std.err.rb, t.ratio.rb, p.value.rb)

rownames(robust.stat) <- c(var.names)

# Print out the table of robust statistics

robust.stat

##################################################################################

## End of the Module for generating the robust statistics

##################################################################################

## keep a csv file of the table of robust statistics

write.csv(robust.stat,file="d:/rtest2020/nonlinear.approach.ontario.csv")

## keep all useful output information in a text file.

##Note: The sink statements work as a pair.

sink(file="d:/rtest2020/nonlinear.approach.ontario.txt")

print("GAUSSIAN-GAMMA MODEL FOR TEMPERATE FOREST (SOUTHERN ONTARIO): NONLINEAR
APPROACH.")

print(robust.stat)

print(c("sample size=",nobs))

print(c("r.square=",r.square,"adj.r.square=", adj.r.square))

sink()

## no more statement beyond this line.

## END OF R- SCRIPT
```

The Following is the Output, saved as a text file from the R script above.

```
[1] "GAUSSIAN-GAMMA MODEL FOR TEMPERATE FOREST (SOUTHERN ONTARIO): NONLINEAR
APPROACH."

              coefficient     std.err.rb    t.ratio.rb   p.value.rb

intercept    -1.7897905650 0.1285082301 -13.927439  0.000000e+00

Ts_5cm        0.1302233187 0.0034590165  37.647499 0.000000e+00

sq.Ts_5cm   -0.0007864373 0.0001576197  -4.989461   6.120334e-07

Ms_30cm     -0.0422387602 0.0086057805  -4.908185   9.285671e-07

ln.Ms_30cm   1.0238529218 0.0947383355  10.807166  0.000000e+00

[1] "sample size=" "15523"

[1] "r.square="        "0.550690210150088" "adj.r.square="    "0.550574393733062"
```

Text S13. Assessment of the Usefulness of the Delta Method for Gaining Insights into the Under-prediction Problem of the Log-transformed Approach

When a new random variable is created by applying a nonlinear function to another random variable, the delta method is sometimes used to find *approximate* formulas for the mean and variance of the new variable. This method has three simple steps: (1) expansion of the nonlinear function into a Taylor series around the mean of the given random variable; (2) removal of higher-order terms in the series; and (3) application of the expectation opperator (or the variance opperator) to the truncated series (Cantor, 2003, p. 200). In response to a reviewer's comment about its usefulness, here we want to assess the usefulness of the delta method for gaining insights into the under-prediction problem of the log-transformed approach.

To be concrete, we consider the data of the Temperate Forest. We will group the 15,523 observations into two sets of bins that are created by crossing the rounded values of soil temperature and the rouded values of soil moisture. Let Y_j be the random variable representing the CO_2 emission that can be realized as the observed values in the jth bin of one of these two sets. Also let the expected value (i.e. the mean) of Y_j be μ_j. It is important to recognize μ_j as a *conditional mean*, because its value is conditional on the rounded values of temperature and moisture of the jth bin.

Let $Z_j = \mathrm{g}(Y_j) = \ln(Y_j)$ be a new random variable that is obtained by log-transforming Y_j. To find an approximate formula for the mean of Z_j by the delta method, we first make the following Taylor series expansion:

$$Z_j = \mathrm{g}(\mu_j) + \mathrm{g}'(\mu_j)(Y_j - \mu_j) + \mathrm{g}''(\mu_j)(Y_j - \mu_j)^2/2 + \cdots \qquad (S13.1)$$

where $\mathrm{g}'(\mu_j)$ and $\mathrm{g}''(\mu_j)$ are the first and second derivatives of $\mathrm{g}(Y_j)$ evaluated at μ_j. Keeping only the first three terms of the series in Eq. (S13.1), we then get the following approximation:

$$Z_j \cong \mathrm{g}(\mu_j) + \mathrm{g}'(\mu_j)(Y_j - \mu_j) + \mathrm{g}''(\mu_j)(Y_j - \mu_j)^2/2 \qquad (S13.2)$$

Applying the expectation opperator to both sides of Eq. (S13.2), we get

$$\zeta_j \cong \mathrm{g}(\mu_j) + \mathrm{g}''(\mu_j)\sigma_j{}^2/2 \qquad (S13.3)$$

where ζ_j is the mean of Z_j, and σ_j is the standard deviation of Y_j. Since the second derivative of $\ln(\mu_j)$ is $\frac{-1}{\mu_j{}^2}$, we get from Eq.(S13.3)

$$\zeta_j \cong \ln(\mu_j) - 0.5(\frac{\sigma_j}{\mu_j})^2 \qquad (S13.4)$$

Let $\ddot{\mu}_j$ be the mean CO_2 emission that is recovered by applying the inverse of log-transformation to ζ_j. In other words, we let

$$\ddot{\mu}_j = e^{\zeta_j} \qquad (S13.5)$$

Applying the exponential transformation to both sides of Eq.(S13.4) and substituting Eq. (S13.5) into the resulting expression, we get

$$\ddot{\mu}_j \cong \mu_j e^{-0.5(\frac{\sigma_j}{\mu_j})^2} \qquad (S13.6)$$

Except for a trivial bin in which all emissions happened to be identical, σ_j must be positive, so that the factor $e^{-0.5(\frac{\sigma_j}{\mu_j})^2}$ must be less than 1. In other words, *to the extent that the approximation yielded by the delta method is good enough, the log-transformation of Y_j will result in the under-prediction of the conditional mean of the emissions in every non-trivial bin.*

From Eq.(S13.6), we find that the severity of the underprediction of μ_j is:

$$\psi_j = \left(\frac{\mu_j - \ddot{\mu}_j}{\mu_j}\right) * 100\% \cong \left(1 - e^{-0.5\left(\frac{\sigma_j}{\mu_j}\right)^2}\right) * 100\% \qquad (S13.7)$$

Here we also see that *to the extent that the approximation yielded by the delta method is good enough, the severity of under-prediction in the jth bin (ψ_j) is a monotonically increasing function of the CV (coefficient of variation) of the emissions in the jth bin*.

Since the saturated speficiation of the regression model allows the log-transformed approach to predict perfectly the means of the log-transformed emissions (i. e. the sample version of ζ_j) in all bins, we can assess whether the analytical result of the delta method is good enough for revealing the pattern of the dependence of (1) the under-prediction severities of the log-transformed approach on (2) the within-bin CVs. For the assessment, we use two sets of bins. The first set contains 20 non-empty bins that are created by crossing the temperature values rounded to the nearest 4ºC and the moisture values rounded to the nearest 5%. The second set contains 132 bins with at least 10 observations that are created by crossing the temperature values rounded to the nearest 1º C and the moisture values rounded to the nearest 1%. Our findings are presented in Figure S7 for the first set and Figure 8 for the second set.

From Figures S7 and S8, we find that the delta method is of limited usefulness. For both sets of bins, the dependence of the severities of under-predictions on the within-bin CVs is mostly understated by the delta method. Despite its limited usefulness, the delta metnod also reveals that for our data, no single adjustment factor can adjust properly the systemtically biased under-predictions of the log-transformed approach.

Text S14. Assessment of the Log-transformed and Nonlinear Approaches while Controlling for Chamber Effects

A reviewer commented that if there are multiple chambers, a random effect for that is definitely needed. Since the CO_2 emission data of the Temperate Forest were collected using 8 chambers, here we exapanded the Gaussian-Gamma model for the data so that the chamber effects could be controlled. The expansion involved the addition of 7 dummy variables to allow the full distinctions among the 8 chambers. Using Chamber 1 as the reference chamber, the 7 dummy variables are denoted as Chamber2, Chamber3, …, Chamber8.

The estimated results from the log-transformed and nonlinear approaches are showed in Table S6. We found that after controlling for the chamber effects, the nonlinear approach remained superior to the log-transformed approach. The Comparable R-square was 0.774 for the nonlinear approach and 0.763 for the log-transformed approach. The observed grand mean of CO_2 emision was under-predicted by 0.7% by the nonlinear approach and by 6.4% by the log-transformed approach.

A noteworthy difference between the two approaches was that the log-transformed approach failed to generate a negative value for the descent parameter of the moisture function, while the nonlinear approach succeeded in generating a negative value of large magnitude for this parameter. This difference is clearly reflected by the curves representing the moisute effect in Figure S9.

The reviewer also commented that since our data are time-series data, we should check the residuals for temporal autoccorelation. We computed the first-order autocorrelation for the residuals of the observed emissions in each of the 8 chambers. From the residuals computed from the nonlinear approach, we found that the autocorrelation ranged from a minimum of 0.32 for Chamber 4 and a maximum of 0.84 for Chamber 7. The pattern of autocorrelation based on the residuals from the log-transformed approach was similar, ranging from a minimum of 0.32 for

Chamber 4 and a maximum of 0.86 for Chamber 7. In other words, we found clear evidence of temporal autocorrelation.

The autocorrelation should not have any effects on the estimated coefficients generated by either of the two approaches, because the temporal order of observations played no role in the linear and nonlinear least-squares methods. Thus, it is unlikely that the existence of the autocorrelations could negate the finding that the nonlinear approach is superior to the log-transformed approach.

In light of the very short time interval of half an hour used for measuring CO_2 emssions, it is not surprising that the residuals were found to have rather high first-order autocorrelations. The autocorrelation information may be useful for the purpose of forecasting, if longer time intervals are used to compute autocorrelations.

Text S15. Application of the Generalized Additive Model (GAM) to the Data of the Temprerate Forest: A Nonparametric Approach

For empirical researchers who are not interested in getting physically interpretable parameters, but want to obtain the predicted values that do not have the sysmatically biased under-prediction problem of the log-transformed approach, here we explain and demonstrate the use of a nonparametric approach, in which we apply the Generalized Additive Model (GAM) to the data of the Temperate Forest. Our main purpose of applying the GAM is to see whether our Gaussian-Gamma model is a reasonable model for predicting CO_2 emissions by soil temperatures and soil moistures.

For our empirical problem, we formulate two versions of the GAM. In the first version, the ith expected value of the CO_2 emission (μ_i) is linked to the explanatory factors as:

$$\mu_i = e^{\beta_0 + f_1(T_i) + f_2(M_i)} \tag{S15.1}$$

where T_i is the ith observation of soil temperature; M_i is the ith observation of soil moisture; β_0 is the unknown intercept to be estimated; and $f_1(T_i)$ and $f_2(M_i)$ are univariate spline functions of T_i and M_i, respectively. In the second version, the link is in the following form:

$$\mu_i = e^{\beta_0 + f_3(T_i, M_i)} \tag{S15.2}$$

where $f_3(T_i, M_i)$ is a bivariate spline function of T_i and M_i. For both versions, conditional on the values of the explanatory factors, the ith observation of CO_2 emission (y_i) is assumed to be a random variable having the following Gamma density function:

$$g(y_i|\mu_i, \omega) = \frac{1}{\Gamma(\omega) y_i} \left(\frac{y_i \omega}{\mu_i}\right)^{\omega} e^{-\frac{y_i \omega}{\mu_i}} \qquad (S15.3)$$

where ω is an unknown dispersion parameter to be estimated, and $\Gamma(\omega) = \int_0^\infty x^{\omega-1} e^{-x} dx$ is the Gamma function. Note that in $g(y_i|\mu_i, \omega)$, the quantities to the right of the verticle line are considered as given. The model is considered to be nonparametric, because the values of the coefficients in $f_1(T_i)$, $f_2(M_i)$ and $f_3(T_i, M_i)$ do not remain constant among the observations. The model is additive in the sense that the intercept and the spline functions are combined together by addition.

The unknown parameters and the varying coefficients in the spline functions are estimated by the penalized least-square method that minimizes the objective function of the following form:

$$\sum_{i=1}^{n} (y_i - \mu_i)^2 + \lambda L \qquad (S15.4)$$

where L is a measure of the roughness of the spline functions, and λ is the smoothing parameter. By default, the roughness is quantified as an integral of the squares of the second derivatives of the spline functions. The smoothing parameter is a positive constant imposed by the user or by default. It is used as a trade-off between the goodness-of-fit and the smoothness: a high λ results in smoother spline curves and poorer fit. To make sure that the unknown intercept is estimable, all spline functions are constrained to have the mean of 0. It is important to note that *this estimation method does not involve the log-transformtion of* y_i*, so that the under-prediction problem of the log-transformed approach does not occur in this approach.*

We used the GAMPL procedure of SAS software (SAS Institute, 2017) to apply the two version of the GAM to the data of the Temperate Forest of southern Ontario. The SAS codes for the application of Version 1 are the following.

```
ods graphics on;
PROC GAMPL DATA=MY_DATA_SET  PLOTS ITDETAILS SEED=12345;
TITLE3 " PROC GAMPL: LINK=LOG, DIST=GAMMA, WITH 2 UNIVARIATE SPLINES." ;
MODEL EMISSION = SPLINE(TEMPERATURE/DETAILS) SPLINE(MOISTURE/DETAILS) /
LINK=LOG DIST=GAMMA;
ID EMISSION TEMPERATURE MOISTURE;
OUTPUT OUT=GAMPL_OUT_1 P=P_GAMPL_1   XBETA / COMPONENT;
RUN;
```

The SAS codes for the application of Version 2 are the following.

```
PROC GAMPL data=MY_DATA_SET PLOTS ITDETAILS SEED=12345;
TITLE3 "PROC GAMPL: LINK=LOG, DIST=GAMMA, WITH A BIVARIATE SPLINE." ;
model EMISSION = SPLINE(TEMPERATURE  MOISTURE /DETAILS) / LINK=LOG
DIST=GAMMA;
ID EMISSION TEMPERATURE MOISTURE;
OUTPUT OUT=GAMPL_OUT_2 P=P_GAMPL_2  XBETA / COMPONENT;
RUN;
```

The letters in blue are key words. MY_DATA_SET is the name of the input data set that contains the variables EMISSION, TEMPERATURE, and MOISTURE. GAMPL_OUT_1 and GAMPL_OUT_2 are the names of the output data sets. P_GAMPL_1 and P_GAMPL_2 are the names of the predicted emissions. 12345 is the random seed for ensuring that the random sampling from observations to form spline knots will remain the same when the program is run again.

The two estimated spline components in Version 1 of the model are shown in Figure S10. The temperature spline has a monotonically increasing pattern, whereas the moisture spline has an overall concave pattern with several waves. The estimated bivariate spline component in Version 2 of the model is shown as a "map" in Figure S11. At each moisture level, the spline also shows a nomotonically increasing pattern as temperature increases. At each temperature level, the spline shows an overall concave pattern without waves as moisture increases.

An advantage of Version 1 is that it provides useful information about the relative predictive powers of temperature and moisture. For example, the observed CO_2 emission of 1.306 $\mu mol/m^2/s$ at 16:30 on April 1, 2014 in Chamber 1 is predicted as 1.278 $\mu mol/m^2/s$ by Version 1 of the model in the following way.

$$\hat{\mu}_i = e^{\hat{\beta}_0+\hat{f}_1(T_i)+\hat{f}_2(M_i)} = e^{1.413-1.293+0.125} = e^{1.413-1.168} = e^{0.245} = 1.278 \qquad (S15.5)$$

In this formula, the negative value of the temperature component indicates the effect of the low temperature of -0.275 C, whereas the the positive value of the moisture component indicates the effect of the high moisture of 15.7%. Since the magnitude of -1.293 is much greater than the magnitude of 0.125, the negative effect of the low temperature was much stronger than the positive effect of high moisture so that the model properly predicted a low emission.

Since the sums of the two splines across all observations are conveniently set at 0, their standard deviations can be used as a measure of their relative predictive powers. In Table S7, we find that the standard deviations are 0.742 for $\hat{f}_1(T_i)$ and 0.186 $\hat{f}_2(M_i)$, implying that temperature was much more powerful than moisture. In Table S7, this large difference in predictive powers between temperature and moisture is also reflected by the very large difference in F-statistic (10,179.4 versus 956.6).

Next, let’s compare the performance of the two versions of the model. Version 2 achieved a slightly higher Comparable R-square than Version 1 (0.562 versus 0.555). Both versions yielded very close predictions of the grand mean of the observed emissions (4.882), with Version 2 being slightly better than Version 1 (4.883 versus 4.886). Clearly, both versions did not have the under-prediction problem of the log-transformed approach.

Due to the strong negative correlation between temperature and moisture, it is not easy to draw the patterns of the predicted emissions, because it is impossible for the predicted emissions at any level of moisture to span the full range of temperature, and because it is also impossible for the predicted emissions at any level of temperature to span the full range of moisture. To show the patterns for the full ranges of temperature and moisture, we first computed the mean predicted emissions in all the bins created by rounding temperature to the nearest 1°C and rounding moisture to the nearest 1%. We then selected two reference moistures (16% and 10%) to draw the temperature curves and two reference temperatures (20°C and 2°C) to draw the moisture curves.

Figure S12 shows the dependence of CO_2 emission on soil temperature at 16% and 10% of soil moisture, respectively. We find that at both levels of moisture, the predicted emissions generated by the two versions of the model are highly similar for a wide range of temperature from 2°C to 17°C, with the pattern being a smooth upward trend. Beyond 17°C, the gap between the corresponding curves of the two versions widen to a maximum of about 2 μmol/m^2/s at 20°C. Beyond 20°C, the curves of both versions show a clear pattern of flattening out. Thus, we infer from this figure that the Gaussian specification with the flexibility to flatten out at higher temperature is more suitable than the Q_{10} (simple exponential) specification for representing the temperature function.

Figure S13 shows the dependence of CO_2 emission on soil moisture at 20°C and 2°C of soil temperature, respectively.We find that the curves for temperature = 20°C differ markedly between the two versions of the model. The curve generated by Version 1 shows several waves along an upward trend from 6% to 19% of moisture, whereas the curve generated by Version 2 shows a sharp rise from 6% to 9% and then essentially remains at a high plateau up to 19% of moisture. The waves do not seem to make physical sense and probably resulted from the effects of some missing expalatory factors. We also find that the curves for temperature = 2°C are very similar between the two versions within the moisture range from 15% to 24%: a plateau from 15% to 19% followed by a gentle decline towards 24%. Based on the smoother pattern generated by Version 2 of the model, we infer that the non-symmetric Gamma specification is better than the symmetric Gaussian specification for representing the moisture function.

In sum, the findings from this nonparametric approach suggest that the Gaussian-Gamma model used in our nonlinear approach is a reasonable parametric model for studying the dependence of CO_2 emission on soil temperature and soil moisture. It is worth noting that although the degrees of freedom used by our Gaussian-Gamma model (5) are much less that those used by the GAM (17 for Version 1 and 19 for Version 2), the Comparable R-square achieved by the former via the nonlinear approach (0.551) turned out to be only slightly less than those achieved by the GAM (0.555 for Version 1 and 0.562 for Version 2). It is better to use the GAM mostly for exploratory puposes, because the shape of some spline, especially its tail, could change markedly when the value of λ (the trade-off between goodness-of-fit and smoothness) is changed.

Table S1

Results of regressing the severity of the log-transformed approach's under-prediction on (1) the coefficient of variatrion and (2) the skewness of the distribution of CO2 emissions within each of the 20 non-empty bins that were created from the data of the Temperate Forest of southern Ontario: Based on the model shown in Eq. (S4.1)

Explanatory	Panel A: Without Skewness			Panel B: With Skewness	
Variable	Coefficient	Robust		Coefficient	Robust
		t-statistic			t-statistic
Intercept	-1.1444	-5.0		-0.4082	-1.2
Coefficient of Variation	0.1082	17.2		0.0841	8.3
(Coefficient of Variation)^2	-0.00059	-13.9		-0.00037	-4.6
Skewness	----	----		-0.2520	-3.1
Adjusted R-square	0.96			0.98	

Note: Each bin represents a distinct combination of a rounded value of soil temperature and a rounded value of soil moisture. The unknown coefficients are estimated by the nonlinear least squares method, without the assumption of homoscedasticity for the error term. Robust t-statistic is defined in Text S8.

Table S2

Usefulness of introducing "High Season" as an explanatory variable into the Gaussian-Gamma specification of the regression model for explaining soil CO_2 emission by soil temperature and soil moisture: Based on the sub-hourly records of the Semiarid Mediterranean Marsh in southern California (N=38,213)

Explanatory	**Log-transformed Approach**		**Nonlinear Approach**	
Variable	Coefficient	t-statistic	Coefficient	Robust
				t-statistic
	Panel A: Without High Season Effect			
Intercept	-22.39783	-52.1	-40.77299	-68.7
Temperature	0.12863	28.2	0.14735	30.0
Temperature ^2	-0.00287	-29.4	-0.00368	-35.1
Moisture	-0.37230	-46.0	-0.72649	-67.1
Ln(Moisture)	9.52386	48.1	18.17530	67.4
R-square	0.080		0.179	
Adjusted R-square	0.080		0.179	
Comparable R-square	**0.151**		**0.184**	
Severity of Overall Under-prediction (%)	**30.6**		**3.1**	
	Panel B: With High Season Effect, without			
Intercept	-23.69267	-64.4	-30.05999	-63.1
Temperature	0.09189	23.5	0.09527	21.4
Temperature ^2	-0.00106	-12.4	-0.00188	-20.1
Moisture	-0.31317	-45.1	-0.51109	-54.4
Ln(Moisture)	9.31925	55.0	13.07932	58.8
High Season	1.21877	118.4	0.96836	122.1
R-square	0.327		0.458	
Adjusted R-square	0.327		0.458	
Comparable R-square	**0.425**		**0.459**	
Severity of Overall Under-prediction (%)	**19.9**		**0.9**	

Note: The unit of temperature is degree C. The unit of moisture is volumic ratio in %.

High Season is a dummy variable assuming 1 for observations in September-November.

Severity of Overall Under-prediction is the severity of under-predicting the observed grand mean.

Table S3

Estimation results of the Gaussian-Gamma specification of the model for explaining soil CO 2 emission in the Temperate Forest of southern Ontario (n=15,523), based on (1) the heteroscedasticity assumption and (2) the homoscedasticity assumption, respectively

Explanatory Variable	Estimated Coefficient	Standard Error	t-statistic
	Panel A: Based on the Heteroscedasticity Assumption		
Intercept	-1.789799	0.128508	-13.9
Temperature	0.130223	0.003459	37.7
Temperature ^2	-0.000786	0.000158	-5.0
Moisture	-0.042239	0.008606	-4.9
Ln(Moisture)	1.023859	0.094738	10.8
	Panel B: Based on the Homoscedasticity Assumption		
Intercept	-1.789799	0.092374	-19.4
Temperature	0.130223	0.003686	35.3
Temperature ^2	-0.000786	0.000134	-5.9
Moisture	-0.042239	0.006359	-6.6
Ln(Moisture)	1.023859	0.068260	15.0

Note: In Panel A, the standard errors and t-statistics are the robust standard errors and the robust t-statistics, respectively. The unknown coefficients are estimated by the nonlinear least squares method.

Table S4

Estimation results of the Gaussian-Gamma specification of the regression model for explaining soil CO_2 emission by soil temperature and soil moisture via (A) the GENMOD approach and (B) the nonlinear approach: Based on the data of the Temperate Forest in southern Ontario (n=15,523)

Explanatory Variable	**(A) GENMOD Approach**		**(B) Nonlinear Approach**	
	Coefficient	t-statistic	Coefficient	Robust t-statistic
Intercept	-1.31740	-10.2	-1.78980	-13.9
Temperature	0.14240	54.8	0.13022	37.7
Temperature ^2	-0.00136	-10.4	-0.00079	-5.0
Moisture	-0.01060	-1.3	-0.04224	-4.9
Ln(Moisture)	0.65550	7.0	1.02386	10.8
R-square	----		0.551	
Adjusted R-square	----		0.551	
Comparable R-square	**0.549**		**0.551**	
Observed Grand Mean	4.882		4.882	
Predicted Grand Mean	4.880		4.881	
Severity of Overall Under-prediction	**0.05**		**0.01**	

Note: Comparable R-square is computed by running a linear regression of the observed emissions on the predicted emissions obtained from each approach.

Severity of Overall Under-prediction =
(Observed Grand Mean - Predicted Grand Mean)/(Observed Grand Mean)*100%.

Table S5

Comparison of the descent parameter of soil moisture between (A) the nonliear approach and (B) the GENMOD approach: Based on 10 randomly selected subsamples of the field field data of the Temperate Forest in southern Ontario (n=15,523)

	(A) Nonlinear Approach				**(B) GENMOD Approach**		
Sample Size	Estimated Descent Parameter	Robust t-statistic	p-value		Estimated Descent Parameter	t-statistic	p-value
	Based on All Field Data						
15,523	-0.0422	-4.9	0.00		-0.0106	-1.3	0.21
Based on the 10 Mutually Exclusive and All Inclusive Random Subsamples of the Field Data							
1,552	-0.0659	-2.7	0.01		-0.0375	-1.4	0.16
1,553	-0.0596	-2.3	0.02		-0.0113	-0.4	0.66
1,552	-0.0531	-1.9	0.06		-0.0335	-1.3	0.19
1,552	-0.0464	-1.8	0.08		-0.0035	-0.1	0.89
1,553	-0.0440	-1.6	0.12		-0.0152	-0.6	0.58
1,552	-0.0425	-1.7	0.09		-0.0235	-0.9	0.37
1,552	-0.0371	-1.4	0.17		0.0179	0.6	0.53
1,552	-0.0335	-1.2	0.25		-0.0071	-0.3	0.79
1,553	-0.0164	-0.7	0.50		0.0184	0.7	0.47
1,552	-0.0060	-0.2	0.83		-0.0103	-0.4	0.70
Minimum	-0.0659				-0.0375		
Maximum	-0.0060				0.0184		
Mean	-0.0404				-0.0106		

Table S6

Estimation results of the Gaussian-Gamma specification of the regression model for explaining soil CO_2 emission by soil temperature and soil moisture via (1) the log-transformed and (2) the nonlinear approaches: Based on the field data of the Temperate Forest in southern Ontario: controlled for chamber effect (n=15,523)

Explanatory Variable	**Log-transformed Approach**		**Nonlinear Approach**	
	Coefficient	t-statistic	Coefficient	Robust t-statistic
Intercept	-1.17722	-8.8	-1.65893	-19.7
Temperature	0.17012	63.9	0.14288	54.1
Temperature ^2	-0.00248	-18.1	-0.00143	-12.9
Moisture	0.00405	0.5	-0.04874	-8.5
Ln(Moisture)	0.39979	4.2	0.93286	15.1
Chamber2	0.35751	23.1	0.26120	18.4
Chamber3	-0.54185	-34.5	-0.48615	-35.9
Chamber4	-0.48527	-24.3	-0.37680	-21.8
Chamber5	0.27117	16.5	0.28624	32.4
Chamber6	-0.22607	-14.3	-0.17888	-11.4
Chamber7	0.39650	32.6	0.45404	65.8
Chamber8	0.15357	3.6	0.15605	6.5
R-square	0.664		0.774	
Adjusted R-square	0.664		0.774	
Comparable R-square	**0.763**		**0.774**	
Observed Grand Mean	4.882		4.882	
Predicted Grand Mean	4.568		4.848	
Severity of Overall Under-prediction (%)	**6.4**		**0.7**	

Note: Comparable R-square is computed by running a linear regression of the observed emissions on the predicted emissions obtained from each approach.

Severity of Overall Under-prediction =
(Observed Grand Mean - Predicted Grand Mean)/(Observed Grand Mean)*100%.

Table S7

Estimation results of applying two versions of the Genralized Additive Model to the data of the Temperate Forest in southern Ontario for predicting soil CO_2 emission by soil temperature and soil moisture. (Link=Log, Dist=Gamma, n=15,523)

Parameters and Components	**Version 1: with 2 Univariate Splines**			**Version 2: with a Bivariate Spline**		
Parameters	Estimate	Std Error	Chi-square	Estimate	Std Error	Chi-square
Intercept	1.4131	0.0041	120298.7	1.4127	0.0041	120591.4
Dispersion	3.8810	5.2697		3.8928	5.2863	
Components	D of Freedom	Std Deviation	F-value	D of Freedom	Std Deviation	F-value
Univariate Spline (Temperature)	8	0.742	10170.4	----	----	----
Univariate Spline (Moisture)	9	0.186	956.6	----	----	----
Bivariate Spline (Temperature, Moisture)	----	----	----	19	0.631	24035.9
Comparable R-square	**0.555**			**0.562**		
Observed Grand Mean	4.882			4.882		
Predicted Grand Mean	4.886			4.883		
Severity of Overall Under-prediction (%)	**-0.08**			**-0.01**		

Note: Comparable R-square is computed by running a linear regression of the observed emissions on the predicted emissions obtained from each version of the Generalized Additive Model.

Severity of Overall Under-prediction = (Observed Grand Mean - Predicted Grand Mean)/(Observed Grand Mean)*100%.

Link=Log does not imply that the observed emissions were log-transformed in the estimation method.

D of Freedom is the degrees of freedom for the F-test.

For each parameter, the smaller the "standdard error" of the estimator, the better.

For each spline conponent, the greater the "standard deviation" of the spline component across all observations, the better.

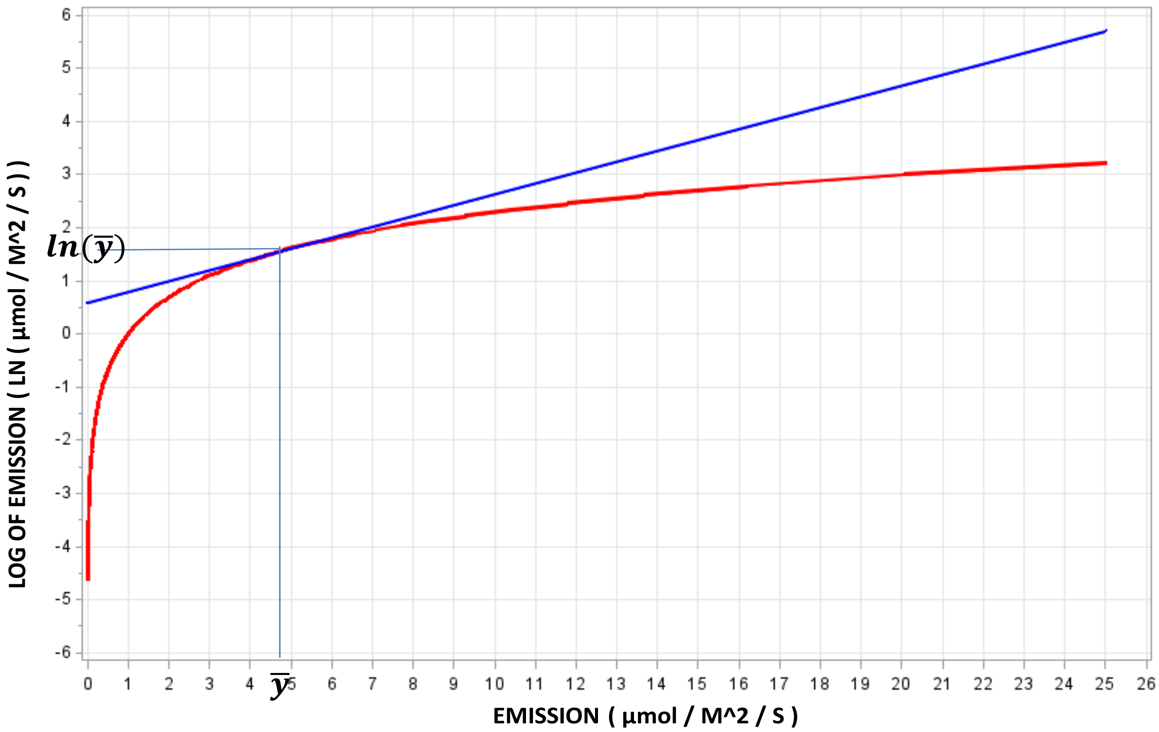

Figure S1. The concave curve representing the log function of CO2 emission (in red) and the tangent line (in blue) touching the curve at the mean CO2 emission ($\bar{y}$ = 4.88 µmol/m^2/s for the Temperate Forest of southern Ontario).

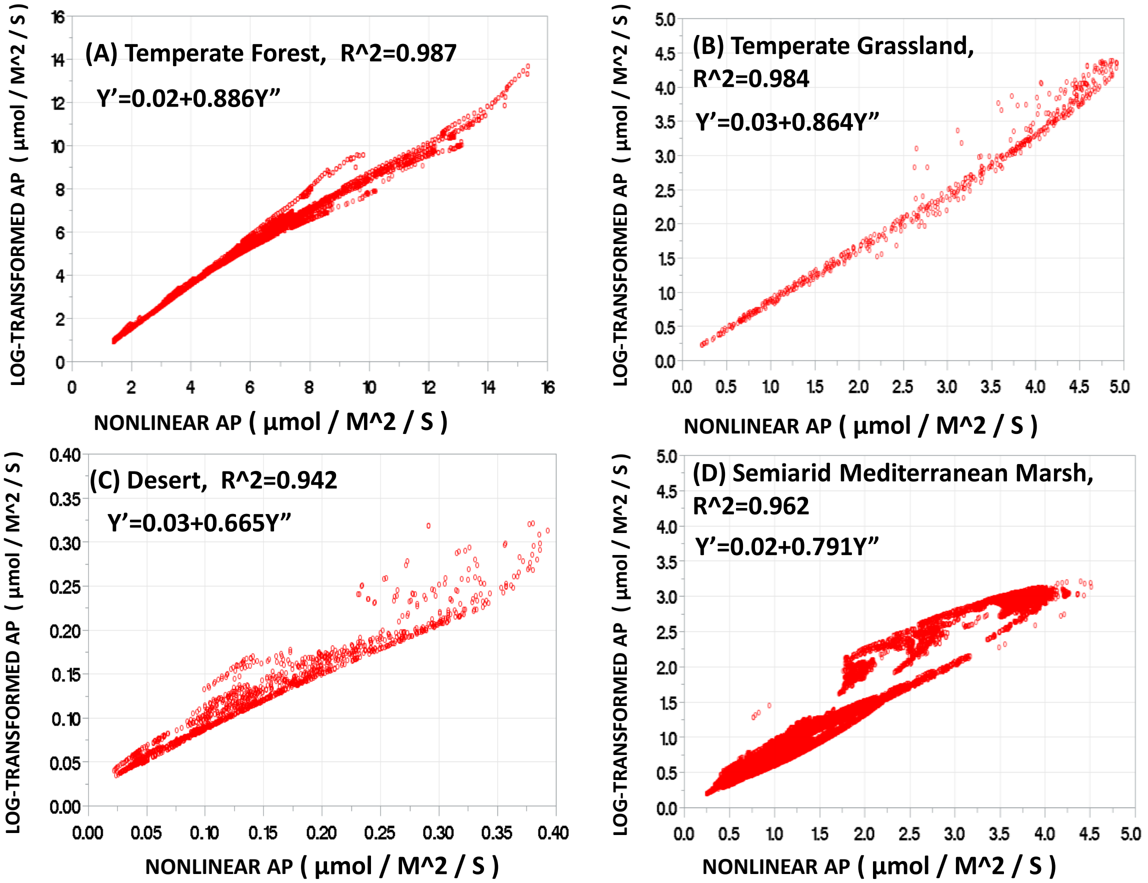

Figure S2. CO2 emission predicted by the nonlinear approach (Y")versus CO2 emission predicted by the log-transformed approach (Y'), based on the applications of the Gaussian-Gamma model to the data of four ecosystems: (A) Temperate Forest (n=15,523), (B) Temperate Grassland (n=566), (C) Desert (n=1,474), and (D) Semiarid Mediterranean Marsh (n=38,213). For Panel D, the Gaussian-Gamma model was the expanded version.

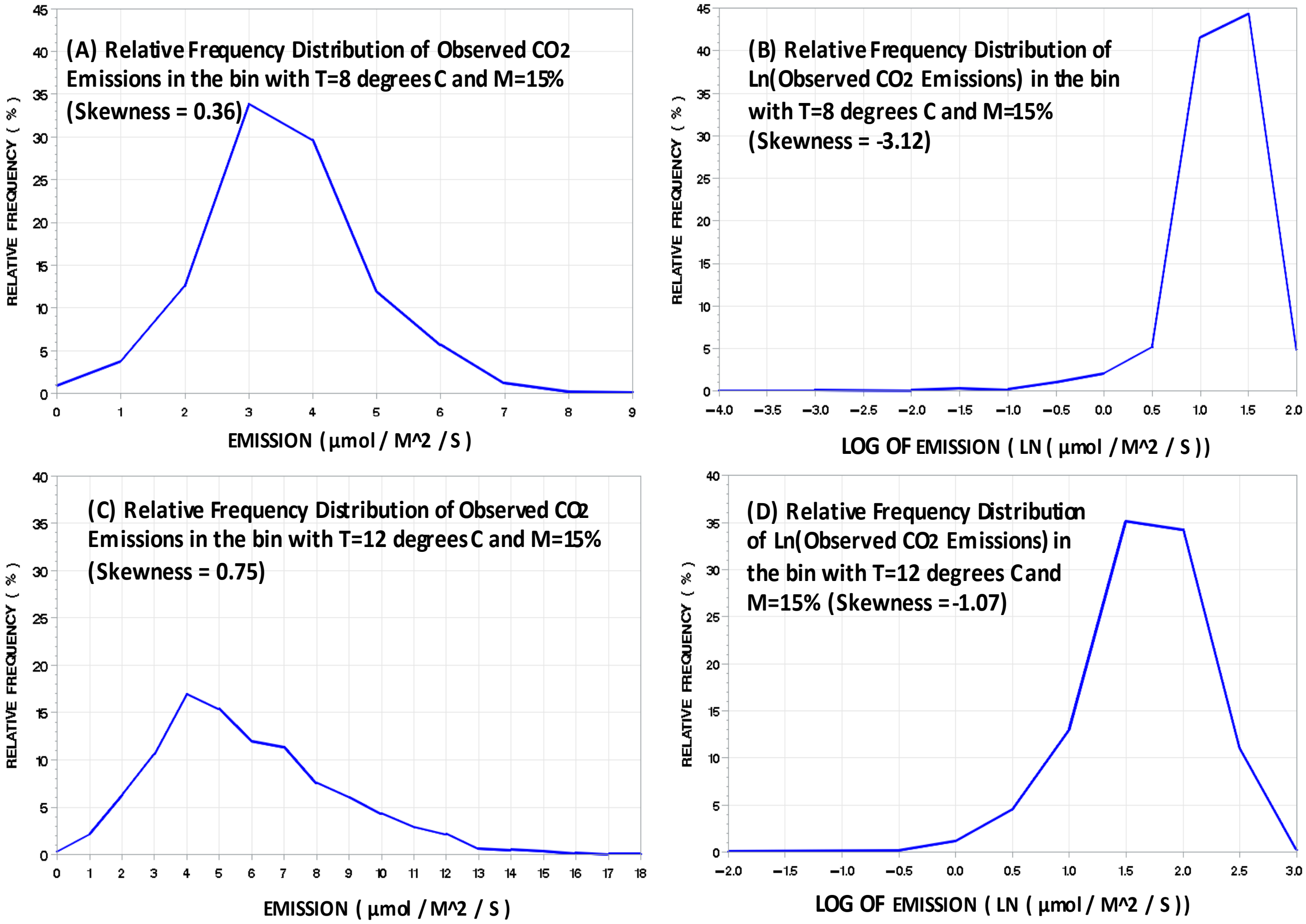

Figure S3. The inability of log-transformation to normalize the distribution of CO2 emissions observed in the Temperate Forest of southern Ontario: Evidence in two bins (Rounded Temperature = 8 degrees C & Rounded Moisture=15%; and Rounded Temperature=12 degrees C & Rounded Moisture=15%). The transformation makes a positively skewed distribution to a highly negatively skewed distribution.

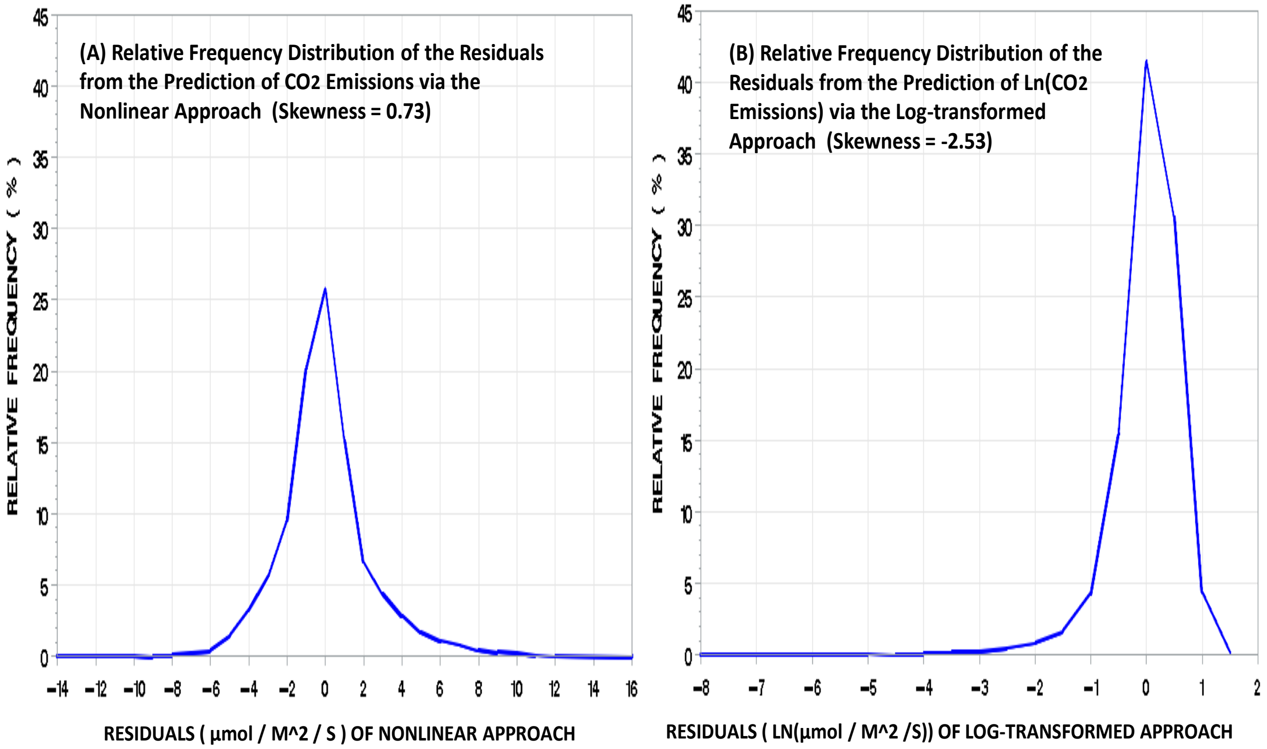

Figure S4. The failure of the log-transformed approach to make the distribution of the residuals to be closer to a normal distribution: Based on the application of the Gaussian-Gamma model to the data from the Temperate Forest of southern Ontario (N=15,523). The width of the intervals for creating the relative frequencies is 1 µmol/m^2/s for Panel A and 0.5 ln(µmol/m^2/s) for Panel B.

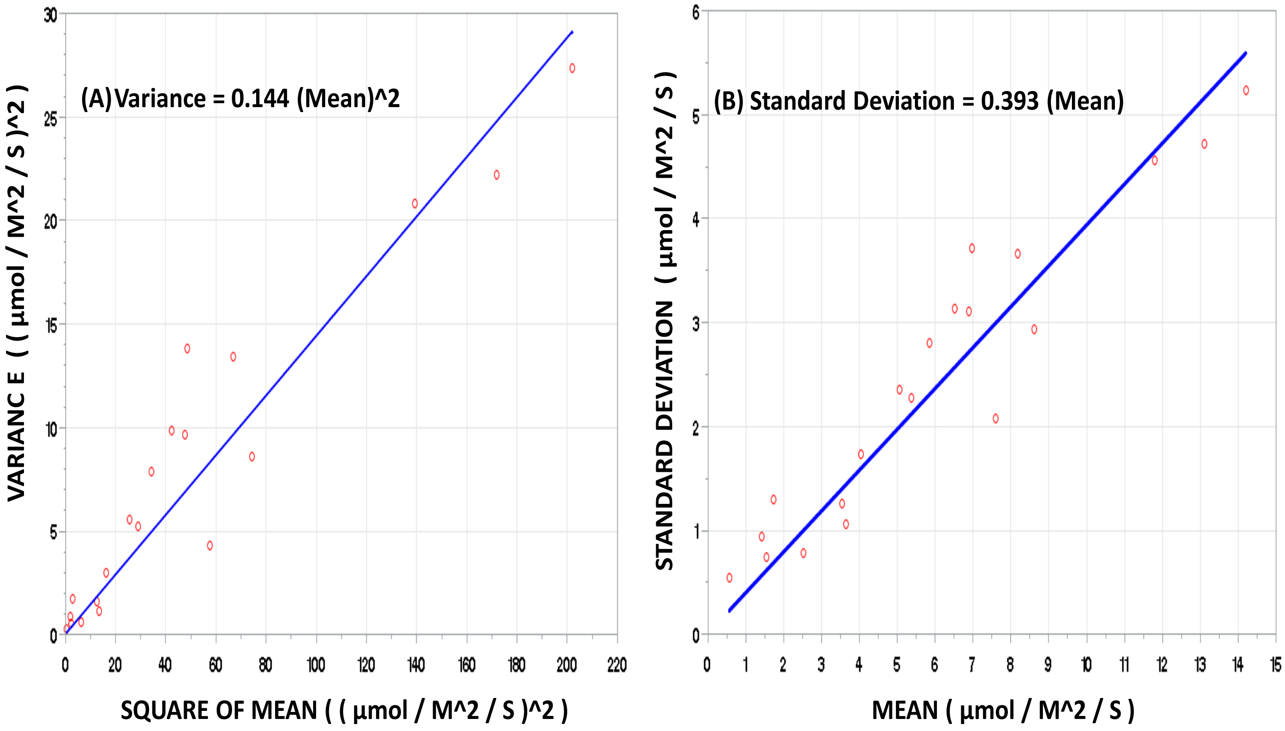

Figure S5. The tendency for the conditional variance (or the conditional standard deviation) to increase with the square of the conditional mean (or simply the conditional mean) among the 20 non-empty bins of the 15,523 observed CO_2 emissions in the field data of the Temperate Forest of southern Ontario. The 20 bins were created from the rounded values of soil temperature and soil moisture. The circles show the observed pattern. The blue line was obtained by fitting a linear regression model without an intercept. Note that the suppression of the intercept implies that both total and explained sums of squares are computed from the deviations from 0 (rather than the grand mean), resulting in the inflation of the values of the Adjusted R-square. Although the values of the Adjusted R-square (0.95 and 0.97) are quite close to 1, the dispersions around the regression lines are clearly not quite small.

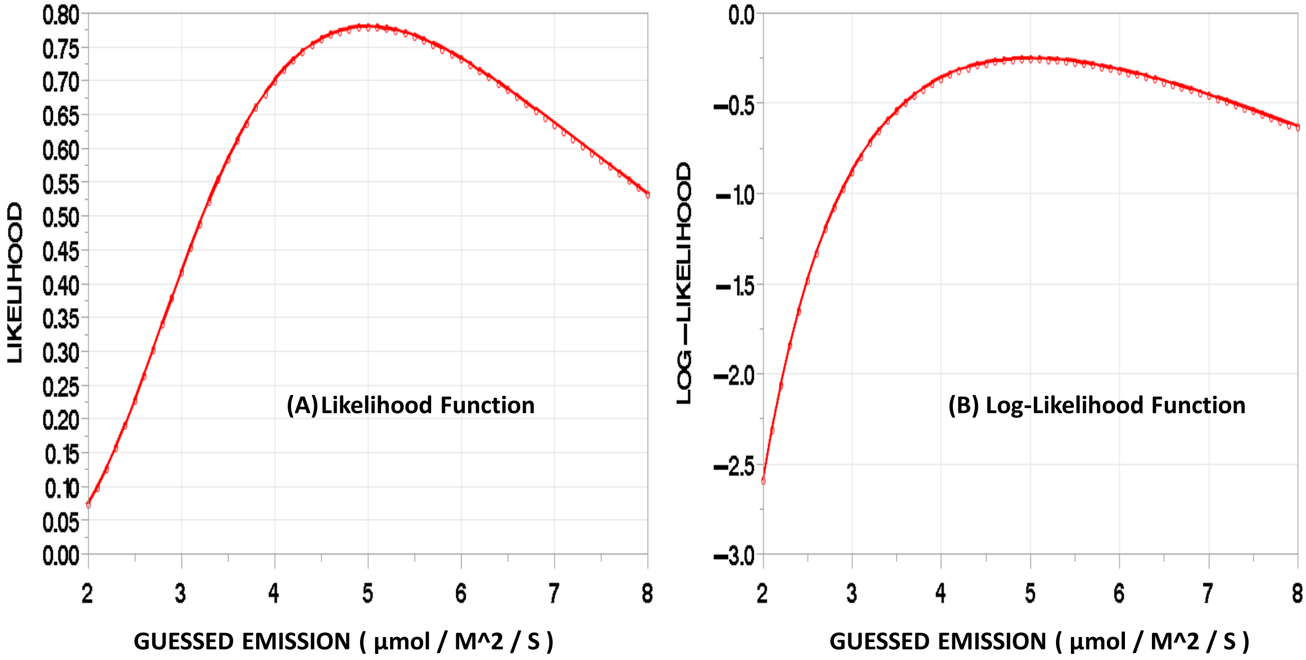

Figure S6. The likelihood and log-likelihood functions of a specific observed CO2 emission in the data set. In the model for predicting emissions, the conditional distribution of the dependent variable is assumed to be a Gamma distribution. The specific observed emission is 5 µmol/m^2/s. The scale parameter is set at 4. Both functions are clearly non-symmetric around the mode of 5 µmol/m^2/s. For gaining greater analytical insights, it is useful to note that the decline from the mode is sharper on the left than on the right, and that this difference becomes greater as the guessed emission moves further away from the mode.

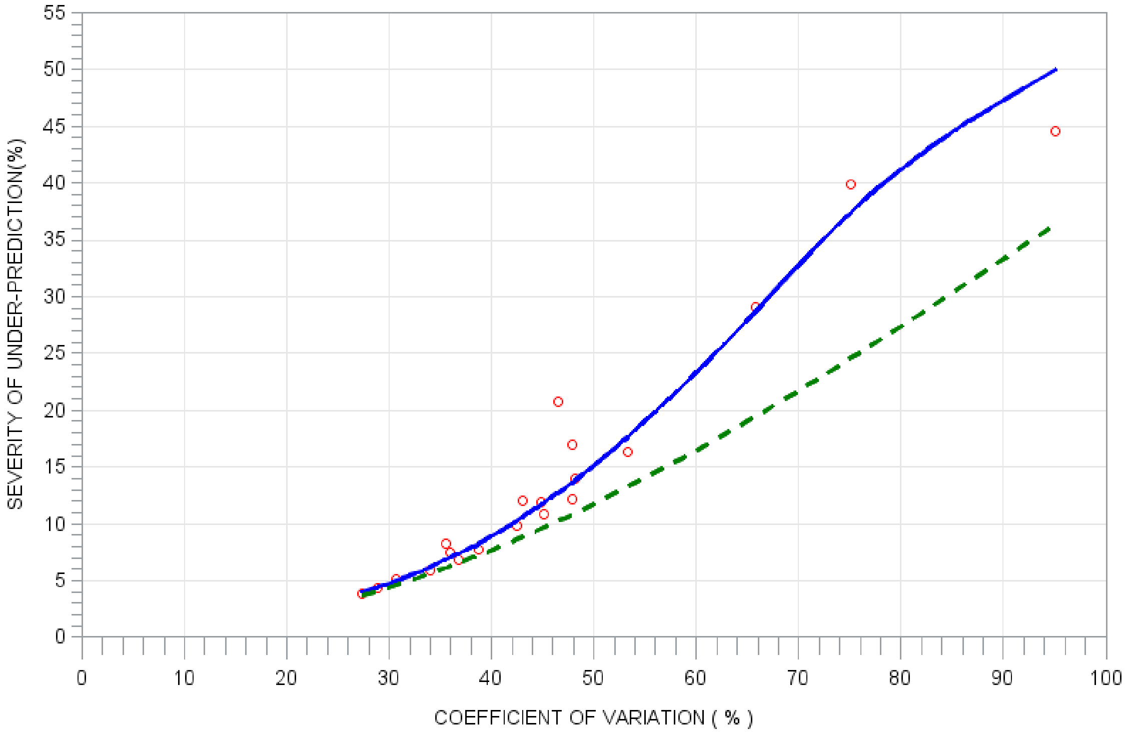

Figure S7. The limited ability of the delta method in revealing the pattern of the dependence of (1) the severities of the log-transformed approach's under-predictions of conditional mean CO2 emissions on (2) the within-bin CVs of emissions: Based the 20 non-empty bins that were created by crossing the temperature values rounded to the nearest 4ºC and the moisture values rounded to the nearest 5% in the data of the Temperate Forest. Red circles represent the actual severities versus CVs. The green line shows the pattern derived from the delta method. The blue line is the Gaussian curve best fitted to the actual pattern (Adjusted R-square = 0.96). The curve-fitting method is the weighted nonlinear least-squares method with the weights being the wthin-bin sample sizes.

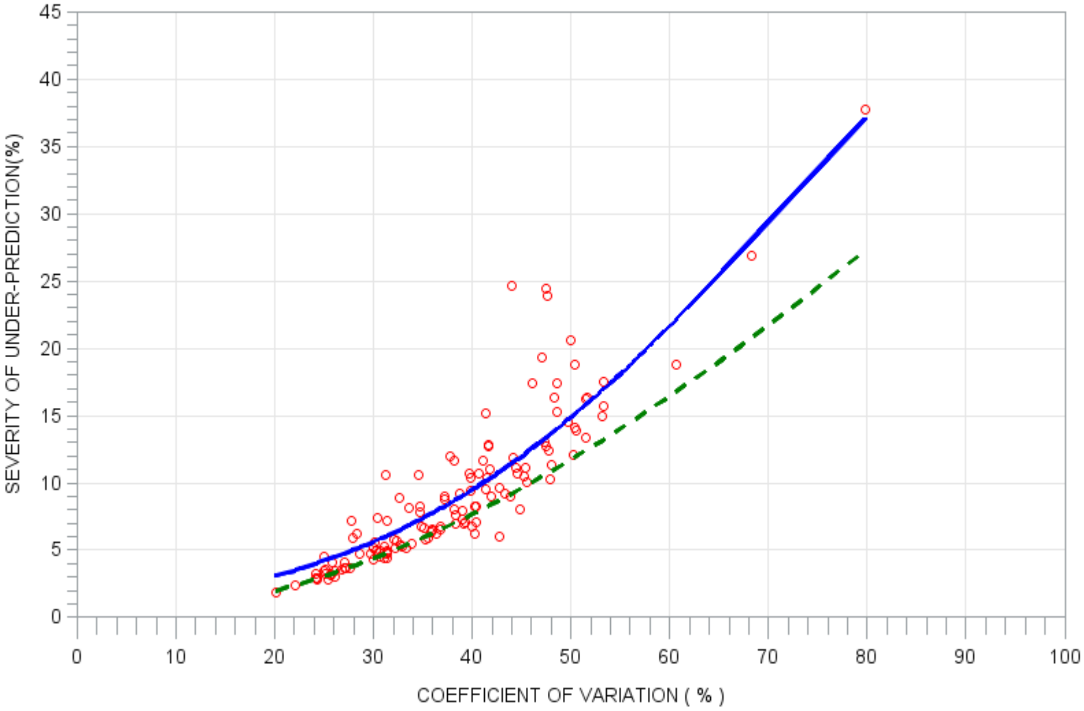

Figure S8. The limited ability of the delta method in revealing the pattern of the dependence of (1) the severities of the log-transformed approach's under-predictions of conditional mean CO2 emissions on (2) the within-bin CVs of emissions: Based the 132 bins with at least 10 observations that were created by crossing the temperature values rounded to the nearest 1ºC and the moisture values rounded to the nearest 1% in the data of the Temperate Forest. Red circles represent the actual severities versus CVs. The green line shows the pattern derived from the delta method. The blue line is the Gaussian curve best fitted to the actual pattern (Adjusted R-square = 0.94). The curve-fitting method is the weighted nonlinear least-squares method with the weights being the wthin-bin sample sizes.

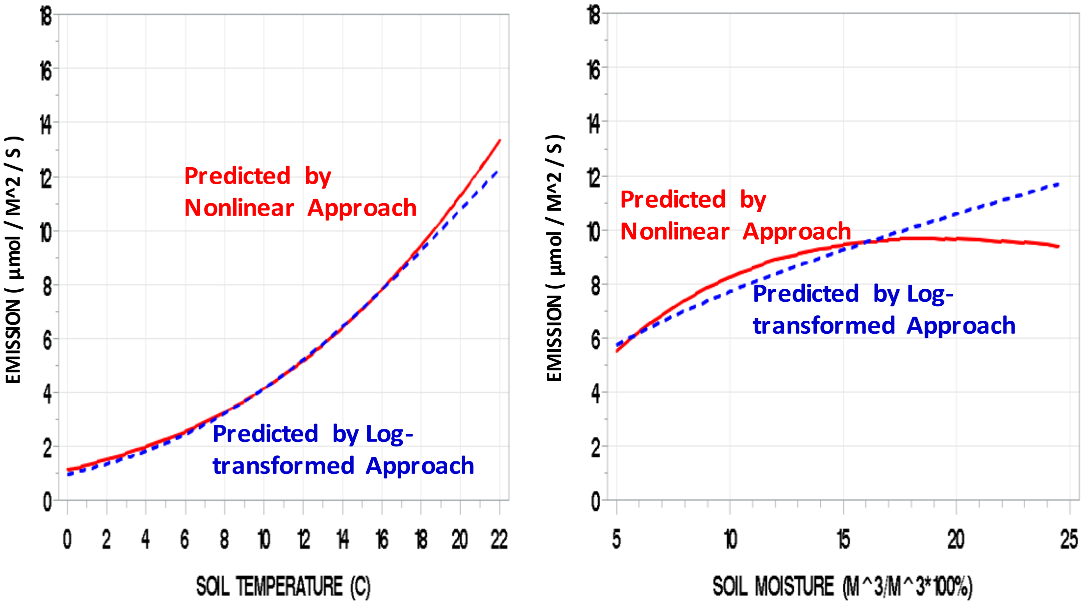

Figure S9. Predicted dependence of CO2 emission on (A) soil temperature (holding soil moisture at 15%) and (B) soil moisture (holding soil temperature at 22 degree C), revealing the different patterns between the log-transformed and nonlinear approaches: Based on the data of the Temperate Forest in southern Ontario (n=15,523) and plotted for Chamber 1 only. By design, the shapes of these curves for other chambers are similar.

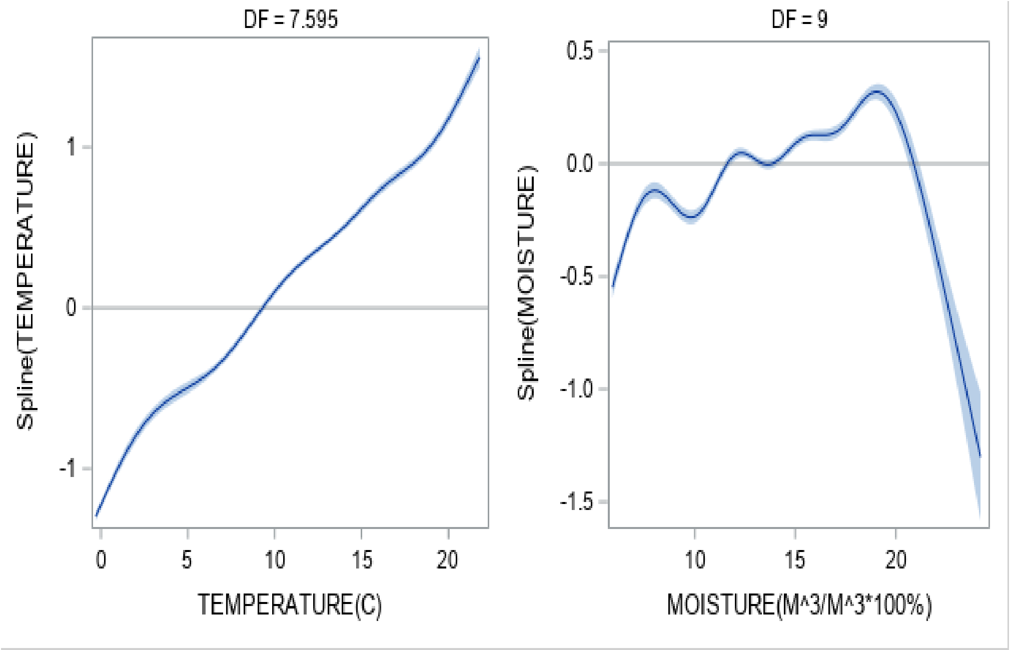

Figure S10. The spline components generated by the Generalized Additive Model with univariate splines for soil temperature and soil moisture for predicting CO2 emission in the Temperate Forest. Link=log, Dist=gamma, n=15,523. The shaded belt represents the 95% confidence band.

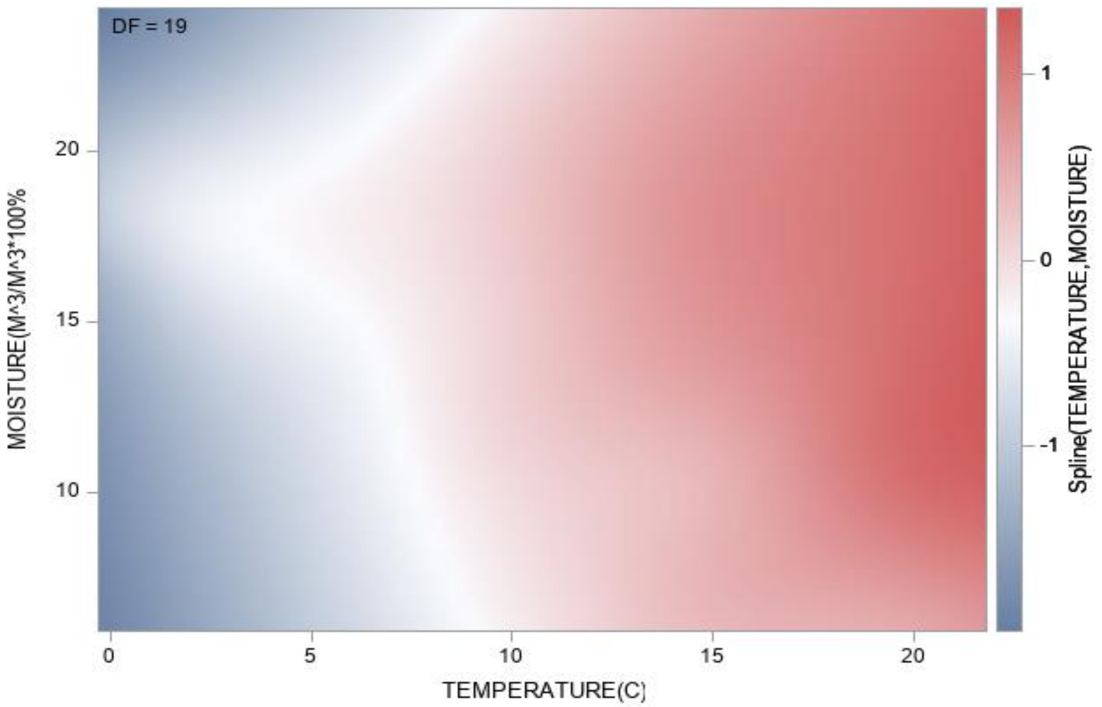

Figure S11. The bivariate spline component generated by the Generalized Additive Model with a bivariate spline for soil temperature and soil moisture for predicting CO_2 emission in the Temperate Forest. Link=log, Dist=gamma, n=15,523.

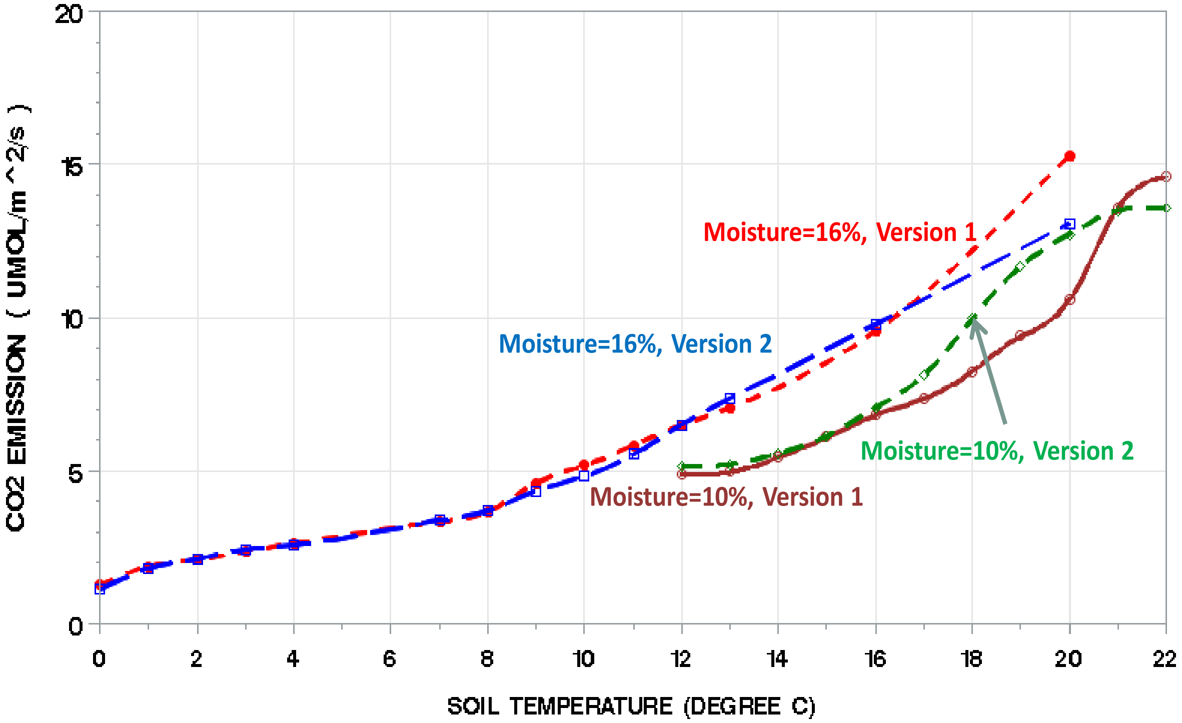

Figure S12. The patterns of the dependence of CO2 emission on soil temperature generated by two versions of the Generalized Additive Model: (1) with univariate splines for soil temperature and soil moisture; (2) a bivariate spline for soil temperature and soil moisture for the Temperate Forest data. Link=log, Dist=gamma, n=15,523. Red curve is from Version 1 at moisture=16%. Blue curve is from Version 2 at moisture=16%. Brown curve is from Version 1 at moisture=10%. Green curve is from Version 2 at moisture=10%.

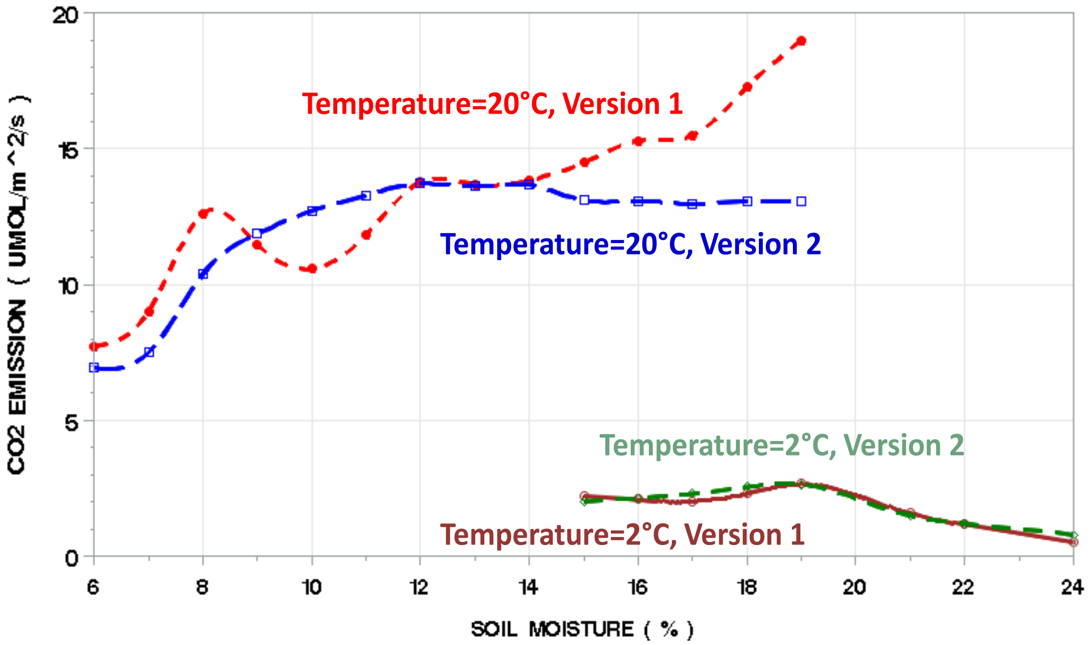

Figure S13. The patterns of the dependence of CO2 emission on soil moisture generated by two versions of the Generalized Additive Model: (1) with univariate splines for soil temperature and soil moisture; (2) a bivariate spline for soil temperature and soil moisture for the Temperate Forest data. Link=log, Dist=gamma, n=15,523. Red curve is from Version 1 at temperature=20° C. Blue curve is from Version 2 at temperature=20° C. Brown curve is from Version 1 at temperature=2° C. Green curve is from Version 2 at temperature=2° C.