Dear Reviewers and Editors,

We would like to express our sincerest thanks to the two anonymous reviewers for their highly constructive comments and suggestions, which can be fully accounted in revising the manuscript. However, one reviewer asks for re-calibration in a cross-validation approach, and this will take substantial computing time (more than 2 times the number of simulations done for the first version of the article). As we appreciate this comment we therefore ask for respite before submitting a revised manuscript.

Following is a list of our responses/actions for each of the reviewers' comments/suggestions. Any further comments and suggestions will be highly appreciated.

Sincerely yours,

Yanyan Yu

on behalf of the co-authors

RESPONSES TO THE COMMENTS OF REVIEWER #1:

Reviewers' Comment (RC): 1. as mentioned in the title, this paper presents sensitivity analysis and calibration of a model but not the model validation. Nevertheless I think a model has a value only if it is able to reproduce the reality. I would have appreciated to find a "model validation" in this article. Authors work on 3 pedons per site. If they do not have access to other data (which is unlikely), several solutions are available to them to present if not a true validation, at least a stopgap: i- working on 2 pedons to calibrate the model and validating it on the third one. Repeating this process on another couple of pedons and so on, to discuss the difference between simulated and observed data in each of the 3 cases. ii- at least simulating each pedons and not only the weighted average of the 3 pedons

Authors' Reply (AR): Validation on an independent test set would be ideal, but as the SoilGen model simulates many aspects of soil formation this requires substantial input, not only of soil data but also of long-term climate data. It is the reconstruction of these soil boundary inputs that makes it not easy to do additional runs at other locations. Furthermore, it is better to do validation on a series of plots for a model like SoilGen2 contrary to on an individual plot only, because it is the more general applicability of the model that we wish to improve by calibration, and the problem is that local factors may contribute to local errors (e.g. the C stored in ectorganic layer may result from wind shelter effects, which are outside the modeled system). Therefore, the first stopgap (a cross-validation approach), proposed by the reviewer, would meet the problem described above better than the second stopgap, which we interpret as "calibrate for each plot separately and discuss to what degree the site results (parameter values and RMSE) are comparable". The second stopgap will produce site-specific calibrations only. We therefore propose to redo calibrations according to a cross-validation approach. However, as the simulation time is long (one run per plot and parameter combination takes 12 hours) we will need time for this and ask for respite.

RC: 2. Belgian pedons are for the 3 of them cutanic fragic Albeluvisol (3 orders of precision) whereas Chinese pedons are for 2 of them Kastanozem (1 order of precision) and for the last one a Luvisol. On which type of soil each calibrated set of parameters can be used for? What is the range of application? by only looking at soil description, it seems that the Belgian set of calibrated parameters are restricted to very specific application whereas the Chinese one could be used for quite different types of soil (Kastanozem and Luvisol have not similar properties). Independent presentation of each pedon would perhaps contribute to address this issue.

AR: The detail of classification is greater in the Belgian sites than in the Chinese sites, which is caused by necessary conversion from the Chinese classification to the WRB. More detailed soil information for Chinese soil pedons ("*Calcic Luvisol* for LJB, *Calcic Kastanozem* for ZW2 and ZW3") will be added in the revised manuscript according to the research by Gong et al. (2003), who compared the WRB system and Chinese soil taxonomy (CST).
As for the applicability of the calibration, we do not think it is restricted to these soils, because (1) the soil names are based on characteristics developed at pedogenetic time scales (millennia) and the OC-distributions develop over centennial time scales; (2) several re-occurring qualifiers are not relevant for OC, e.g. if the albeluvisol is cutanic, while varying qualifiers (dystric and fragic) are supposedly of influence to vegetation growth and thus of OC. So the soil classification is merely added for illustrative purposes.

RC: 3. by looking at the 2.5 "calibration approach" part, reader could expect numerous run to get the right and precisely defined values. He could be disappointed by finding only 14 and 8 runs in the "3.2 - calibration" part. Could authors explain why they can minimize each function (representing important parameter) in so few runs?

AR: The complexity of SoilGen2 (it is a soil genesis model simulating change in a lot of soil properties) results in a large runtime. This prohibits doing many calibration runs, but in this manuscript we only calibrate OC part of the model. Such statement was made in the manuscript to motivate the choice for the Morris' method. Additionally, the small number of runs was motivated by the "convergence" of the calibration, and the convergence was quickly reached through following two ways in our study. Firstly, the quantitative relationships between changed OC and calibrated parameters revealed by sensitivity analysis were used as the bases for variation of parameters. Secondly, a convergence criterion was set by $RMSE_{OCMS}$<10%, It means that for the tests in one plot whereby only 1 parameter varies, the calibration is stopped when the best result ($\mathrm{RMSE_{OCMS}}$) is less than 10% better than the secondly best result. Related explanations will be emphasized in the revised manuscript.

RC: 4. RMSE are very frequent in calibration methods. That's not the case for MD and DIS. Could authors explain the added-values of MD and DIS over RMSE?

AR: RMSE is more sensitive to outliers (extreme difference between measurements and simulations) than MD. Furthermore, as negative and positive differences can compensate each other in the calculation of the MD, MD gives an impression of the bias in the simulations, whereas RMSE does not. DIS gives the absolute difference as a fraction of the observed differences and is thus dimensionless (RMSE and MD are in the units of the variable), which allows to compare results over different output parameters, which is not possible with RMSE and MD. Thus, all these statistics put emphasis on different aspects of simulation quality and this is why they were applied. This reasoning will be explained in the revised manuscript.

RC: 5. I did not catch why the calibration in China accounted only for 2 parameters and not for 3 as in Belgium. KRPM remain at 0.3 for all tests.

AR: Since just a small variation of the most sensitive parameter k_{HUM} could minimize the underestimation of OC in Chinese soil pedons, no further variation of secondary sensitive parameter is needed in original calibrations. The differences between the calibrated parameters in Belgium and China have been attributed to the vegetation types and climate conditions. Related explanations will be added in the revised manuscript according to the results based on new calibrations proposed by reviewer (see RC 1).

RESPONSES TO THE COMMENTS OF REVIEWER #2:

RC: Initial conditions for simulations In the Model input data part (p. 1824), I think things should be clarified. I would intuitively understand from the objectives of the paper that OC dynamics would be modeled from incipient stages of soil formation up to today and this is not the case. Moreover, considering these input data, it seems there is a discrepancy between the initial conditions for each study area. In the case of the Belgian pedon, initial conditions correspond to the current properties, simulated and validated after after 15000 years of soil evolution since the loess deposit. For the Chinese pedon however, and if I understand well, initial conditions seem to correspond to the measurements made in the parent material. This raises two questions: why consider initial conditions for the Chinese soil (parent material properties), and current simulated conditions for the Belgian pedon ? In the first case, OM dynamics is simulated on soils with current properties, while in the second case, OM dynamics is simulated from incipient stages of soil development over 1000 years, that do not correspond to the current soil properties. In the Chinese pedon case, it seems difficult to compare current OC contents with simulated OC contents that are the result of interactions with soil properties corresponding to the parent material, that are obviously not the same as the current ones in the soil (e.g. for example the scalfac parameter that is influenced by clay content, although the significance of scalfac remains unclear, see remark 6).

AR: In tentative runs we found that the effect of initial OC-content and previous vegetations does no longer influence OC-contents after about 300 years, but that soil genesis can change soil texture, calcite content and hydraulic properties still. A full simulation of 15000 years would be best, but such simulation takes a long run time which would render sensitivity analysis and calibration unfeasible. To be on the safe side to, we therefore did runs of 1000 years.
As we wished to compare calibration results with today's measurements, we needed to reconstruct soil conditions at 1000 B.P. As the Belgian climate is a strongly leaching climate

with practically no dust additions, such reconstruction is necessary because there is an effect of leaching on soil properties (e.g. soil texture, calcite and hydraulic properties) over 1000 years. The Chinese climate has a large precipitation deficit and redistributions of calcite and clay towards greater depths are not as significant as in Belgium. Additionally, dust deposition is not negligible in study region of China, and fertilizes the profile with additional calcite and changes the particle size at the top. Soil properties with uncertain dust addition cannot be simulated accurately to reconstruct the 1000 B.P. situation, so we decided to assume that the current profile represents this situation fairly well. This is why the initial situation of the Chinese and Belgian simulations differed.
Regarding the parameter “scalfac”, please see the response for RC 7 below. The effect of clay content on OC cycle is included in the model and “scalfac” merely scales this effect. Based on our sensitivity analysis, it is not very sensitive.
Above explanations will be added in the revised manuscript.

RC: Bioturbation p. 1821 line 3: it is mentioned that the interaction between the OC cycle with other soil formation processes in SoilGen2 is occurring through the flow of CO_2. Is that the only interaction modeled? Indeed, p. 1821, lines 17 to 22, bioturbation (which contributes significantly to soil formation) is mentioned as a process that incorporates and mixes OC in soil. I would therefore mention this specific interaction. More importantly, Finke & Hutson (2008) mention that bioturbation can significantly influence the OC cycle in soils. Therefore, I do not understand why this process is not included in the sensitivity & parameterization analysis in this paper.

AR: Indeed the statement at p.1821 line 3 does not represent all interactions between OC cycle and soil formation processes. This we will expand in the revision to include bioturbation. In fact, processes modeled in SoilGen2 such as clay migration, (de-)calcification and bioturbation change porosity and texture and thus indirectly the hydraulic conductivity and pF-curve and water flow. Additionally, heat flow is also affected by changes in soil physical properties. As a result, depth profiles of water content, clay content and temperature at points in time are affected by these pedogenetic processes, which directly affect the modeled OC-cycle via rate-modifiers.
Finke and Hutson (2008) do not explicitly mention that bioturbation can significantly influence the OC-cycle, so this argument to include bioturbation in the sensitivity analysis does not hold. Bioturbation of course does redistribute organic pools in the topsoil and may thus influence the distribution of organic material over depth and between ectorganic and endorganic profile parts, but at natural bioturbation levels of 10s of tons ha^{-1} y^{-1} the effect will be limited as only a few promilles of the mass per year is bioturbated, which is less than the amount being transported between OC-pools. This will be different in case of tillage, but this is outside the scope of this study.
The major reason for not calibrating bioturbation however is, that it is a time series of input values (like the climatic parameters), not a process with parameters that can be calibrated. We will add this latter motivation to the revision.

Other comments of REVIEWER #2:

RC: 1. While no being a native English speaker, I have found several spelling/expression mistakes. I think the paper should be thoroughly checked in that respect before publication.

AR: We have checked the whole manuscript and corrected related errors.

RC: 2. In the abstract and introduction, the authors refer to the estimation of past terrestrial carbon pools. While this is of course of major importance, recent developments in the quantitative modeling of soil genesis can provide as well a tool to estimate future carbon pools in the soils, and therefore future atmospheric CO_2 concentration changes. I think this should be mentioned.

AR: Such description will be added in the revised manuscript. In abstract: "SoilGen2 is a useful tool to obtain aspects of soil properties (including carbon content) by simulating soil formation processes; thus it offers an opportunity for both past soil carbon pool reconstruction *and future carbon pool prediction.*"; In introduction: "Currently, with the development of soil carbon models, quantitative simulations of soil carbon storage have been widely done, but most of them focus on modern processes *and aim to predict future atmospheric* CO_2 *concentration change* (Coleman et al., 1997; Jensen et al., 1997; Kelly et al., 1997; Li et al., 1997)."

RC: 3. p. 1818, lines 4-14: I think this paragraph is very important and would request some clarification. It should be clearly mentioned that changes in soil organic carbon pools over long timescales have to account for changes in soil properties (e.g. particle size, pH…) due to soil development. Therefore, existing soil carbon models have to include the modeling of soil formation over long timescales.

AR: It should be in p 1819, lines 4-14. Related description will be changed in the revised manuscript as follows:
"*Changes of past soil carbon pools over long timescale have to account for changes in soil properties (e.g. particle size, pH) due to soil formation and development processes, which are seldomly included in existing soil carbon models (Finke, 2012; Finke and Hutson, 2008; Mermut et al., 2000), because information on past soil formation factors for different regions is unavailable (Finke, 2012; Sauer et al., 2012), and fewer models could consider all soil formation factors (Jenny, 1961: e.g. climate, organisms, relief, parent material and time, "CLORPT") in simulation at the same time (Kirkby, 1977; Minasny and McBratney, 1999, 2001; Minasny et al., 2008; Parton et al., 1987; Salvador-Blanes et al., 2007).*"

RC: 4. p.1819, line 13: References could be made to Kirkby (1977) and Salvador-Blanes et al (2007) as well.

AR: The references will be added in the revised manuscript.

RC: 5. p. 1820 lines 3-5: I would not mention this sentence at that stage of the paper.

AR: The sentence will be moved to the end of the revised manuscript.

RC: 6. p. 1820 lines 6-13: two particular soils developed on loess are selected. Does this mean that the calibration process has to be made for each soil type and each climate? This could be addressed in the discussion section. The first sentence is too long and should be split after 'SoilGen2'.

AR: We think the current study cannot answer the question if calibrations have to be repeated for all soil type/climate combinations. This we would mention in the revised manuscript:
"*In addition, the different calibrated results in Belgium and China indicate that future calibrations are needed for distinct climate conditions, possibly for different soil types as well, and uncertainty bandwidths of calibrated parameters should be given. The current study however does not allow a certain statement on this topic as only a few soil type/climate combinations have been explored.* "
The first sentence will be split after "SoilGen2" in the revised manuscript.

RC: 7. p. 1821, line 29: I did not understand what the scaling factor 'scalfac' is exactly corresponding to. I think it should be more detailed. What is the effect of clay on this parameter? Does this mean that scalfac is the only parameter varying explicitly in the model according to soil properties, and therefore with position in the soil profile?

AR: "scalfac" is the constant (1.67) used in the equation to set the CO_2/(BIO+HUM) ratio in RothC 26.3 model. The equation used is CO_2/(BIO+HUM)=1.67(1.85+1.60 exp(-0.0786%clay)), so the ratio is influenced by the clay content of the soil. In SoilGen2, the constant is transformed into a parameter that can be calibrated, and the ratio varies per soil layer with the variation of clay content in the profile. Related description of the function of scalfac will be added in the revised manuscript.

RC: 8. p. 1822, line 19: try to give a reference.

AR: The references "*(Davis et al., 2003; Verbruggen et al., 1996)*" will be added in the revised manuscript.

RC: 9. p. 1822, study regions: as soil properties can influence OC dynamics, I think it would be interesting to add basic data on soil profiles in Table 1 (particle size, pH, carbonate content,…), in particular regarding the major comment made on the initial conditions. This would be a way, for the Chinese pedon, to check if the properties of the current soil are similar or not to the underlying parent material.

AR: Information for soil properties (pH, $CaCO_3$ and clay content) will be added in Table 1 in the revised manuscript.

RC: 10. p. 1823, line 3: reference should be made to Table 1.

AR: No, the reference is correctly to Table 3 in Finke (2012), where these data were already published.

RC: 11. p. 1823: OC stocks are calculated according to bulk density estimations. As this is a key and potentially variable parameter, do we have any idea of the uncertainty of the measurement? Could it be possible to explain quickly how this parameter is modeled in SoilGen2?

AR: First a remark: the SoilGen model models OC-stocks at a mass basis per soil compartment of fixed volume; only to calculate OC% bulk density is used. The uncertainty of measured bulk densities is not known for the sites as there were no duplicates taken. We expect that most of the uncertainty will not be due to measurement error but due to variability. Related description on how bulk density is modeled in SoilGen2 will be added in the revised manuscript as follows:
"*Bulk density, as the key parameter for calculating OC fractions, is calculated in SoilGen2 by the division between mass of solid phase and volume per compartment. Mass of solid phase per compartment, as a result of mass redistribution due to bioturbation, clay migration, OC cycle and solution/dissolution of calcite and gypsum, is recalculated at time steps varying from hours (solution/dissolution, clay migration), one day (OC cycle) to one year (bioturbation).The volume per compartment is considered as a constant.*"

RC: 12. p. 1823: OC content are determined by two different analytical methods in both regions (weight loss-on-ignition in Belgium, potassium dichromate in China). Does this influence OC contents?

AR: The different methods using for analyzing OC content have been selected according to their regional characters and no significant influence exists. Because the Belgian soil has been decalcified and contains fairly low amounts of clay in the top, the bias of the LOI-method, overestimation of OC by ignoring loss of water from various clay minerals, calcite and gypsum, is negligible. The potassium dichromate method, which is not sensitive to the high $CaCO_3$, is suitable for OC measurement in Chinese soils. Above explanations will be added in the revised manuscript.

RC: 13. p. 1824, line 10 (and so p. 1825, line 15): what is a 'typical year' for climate? Please define what is meant here.

AR: Related explanations will be added in the revised manuscript as follows:
"*Over this year, the characters of precipitation (e.g. total amount of the rainfall, the number of days with rain, frequency of rainfall intensity and monthly distribution of rainfall) are similar to multi-years average conditions.*"

RC: 14. p. 1824, lines 16-17 and p. 1825, lines 22-24: a given rate of bioturbation is assumed.

Please give some references here. These rates are very different in the two regions, and relate to very different soil thicknesses in each case. Again, I think it would have been interesting to test this parameter as well on the calibration of OC dynamics.

AR: Generally, measured data are scarce for most ecosystems. Reference "*(Gobat et al., 2004)*" will be added. The rates in the Belgium Beech forest were assumed similar to values given for German Beech forest (Gobat 2004: p.131). Because the Chinese soils are non-acidic we presumed that bioturbation would be a factor 2 higher. As for the explanation for no calibration of bioturbation, please see the second response for reviewer 2 above.

RC: 15. p. 1824, line 19: I think ref. to table 3 is incorrect.

AR: No. it is correct to Table 3 in Finke (2012).

RC: 16. p. 1825 lines 11-12: simplify mean temperature significant digits as for Belgium.

AR: Done (simplified to "*-6/22 ˚C, -5/23 ˚C and -5/23 ˚C*").

RC: 17. p. 1825, line 14: inverse distance interpolation is perhaps not the best interpolation method for climate data.

AR: If the spatial autocorrelation between measurement stations exists, a kriging approach might be better. However, the number of stations in study region in China is not sufficient to determine the autocorrelation structure, hence kriging approaches would not be suitable as well. An alternative approach might have been choosing the most nearby weather station (Nearest neighbor interpolation). Loss of variance due to different interpolation techniques is probably not a big issue. The Chinese weather stations represent a geographic climatic trend and we only wish to reproduce this trend to estimate at intermediate locations. We interpolated annual amounts of precipitation, potential evapotranspiration, and January and July temperatures. To obtain daily values (within-yearly patterns) we chose a characteristic weather station. By doing this, no variance is lost.

RC: 18. p. 1826, lines 3 to 6: this sentence is not only valid for the Chinese pedon, but for both pedons. I would put it in the introductory part of the Model input data section.

AR: Done.

RC: 19. p. 1827, line 7: I do not understand the expression between brackets after 'plausible parameter value range'. Please clarify it.

AR: This expression is needed to certify that both the parameter value and its perturbation will be inside the plausible range between "lo" and "hi", and $[x_{i,lo};x_{i,hi}-\Delta x_i]$ between brackets will be changed to "$[x_{i,low};x_{i,high}-\Delta x_i]$" for clarification.

RC: 20. p. 1827, line 10: I did not understand the calculation 2×1/(8-1). What is the '2' corresponding to?

AR: In Morris' approach, Δ = 2×1/(p-1), where p is equal to the number of levels for which p/2 elementary effects are computed. The number"2" is an arbitrary choice in the application of Morris' method (Morris, 1991 also proposes it), it determines what fraction of the plausible parameter range is covered by each elementary effect. Above explanation will be added in the revised manuscript.

RC: 21. p. 1830 line 9: Brackets seem to miss in the equation.

AR: Corrected.

RC: 22. p. 1835, lines 1-8: I think this paragraph should be rewritten. The second sentence is very long and unclear, and I did not understand the explanation regarding leaching and bioturbation processes.

AR: The paragraph has been rewritten as follows:
"*The calibrated fr_{ecto} is also lower than default value (0.580) in SoilGen2 from measurement data from literature (Kononova, 1975). In realistic soil carbon cycle process, part of litter carbon pool in ectorganic layer leaches to endorganic layers in the form of dissolved organic carbon (DOC). However, this process is not simulated in SoilGen2, while only little carbon is being exchanged between two layers by bioturbation in the model. Therefore, fr_{ecto}, as the ratio of carbon pool in ectorganic layer to the total pool, was expected to be decreased as this decrease mimics the effect of DOC-leaching.*"

RC: 23. p. 1835, Comparison between two regions section: the litter composition as well as climate data are mentioned as factors explaining the differences in parameter values between the two sites. Could the specific properties of the two soils (e.g. particle size, pH, carbonate content…) explain these differences to some extent as well?

AR: Related discussion will be added in section 3.3.2 in the revised manuscript as follows:
"*Finally, because the effect of some soil properties (e.g. $CaCO_3$ content, pH) on OC cycle is not incorporated in SoilGen2 (except clay and water content), the difference of these properties in two regions may explain part of the variation in calibrated parameters. This relationship could be quantified via regression analysis based on many simulated and calibrated plots, however, the small number of calibrated plots in our study does not allow for such analysis.*"

RC: 24. p. 1835, line 18: Quideau et al. Is not in the references section.

AR: The reference (*Quideau, S. A., Chadwick, O. A., Benesi, A., Graham, R. C., and Anderson, M. A.: A direct link between forest vegetation type and soil organic matter composition, Geoderma, 104, 41-60, 2001.*) will be added in the revised manuscript.

RC: 25. p. 1836, Conclusions, lines 18-19: I think the sentence is not clear. A comparison is made between 'deeper soils' and 'surface soils' I think the terms are not correct. Is 'thick' and 'shallow' soils meant here? This aspect was not mentioned in the discussion section. This raises the question of the value of the parameters along the soil profile: could we imagine an implementation of the model here they could vary with depth (e.g. shielding effect of clay), or is this already accounted for in the scalfac parameter?

AR: We don't mean "thick" and "shallow" soils here, but instead refer to different soil layers. In order to make it clear, the sentence will be revised as follows:

"*The calibrated parameters follow the law that deeper soil layers are more resistant to decomposition than surface soils layers, which is induced by the age of carbon and an unfavorable environment for soil micro-organisms in the deeper layers.*"

Related discussion was mentioned in section 3.3.1 in the original manuscript, but the implementation of SoilGen2 was not described clearly, which will be added in the section 2.1 in the revised manuscript as follow:

"*Basically, the soil profile is divided into a number of compartments with equal thickness, and the OC-routines of SoilGen2 operate on every compartment separately. Therefore, changes of OC are modified by different water content, clay content and soil temperature in corresponding compartments.*"

We would like to express our sincerest thanks again to both reviewers for the highly valuable comments and suggestions that greatly improved the manuscript. Any further suggestions will be highly appreciated.