Reviewer 3

Many thanks for your constructive comments on our manuscript. Please find our responses below, with your original comments in regular text and our responses underneath in green:

'The N14CP model seems to be a simple but heavily calibrated model, but it is not adequately described in the paper for the readers to fully understand the long discussion about the pattern of different model outputs'

Thank you for highlighting this issue, this is also something picked up by reviewer one. To reiterate some of our response there, we attempted here to highlight the overall workings of the model and most relevant processes here rather than repeat the full details of the model which have been published elsewhere, as is common practice with modelling research. We direct the reader to previous papers where the model is described fully, but we appreciate your comments and will seek to expand some of the salient description in this manuscript. We will add more detail to the model methodology section, in particular regarding P uptake and loss and by reducing the discussion of model outputs where these do not explicitly relate to model processes.

More detail regarding P processes will be added in '2.2.1. N14CP model summary' and explicit reference to the relevance of phosphatase enzymes to the $P_{CleaveMax}$ parameter in '2.3.3. Model parameters for the acidic and calcareous grasslands'. Superfluous detail will be cut from '4.2. Simulating grassland C, N and P pools by varying plant access to P sources'.

We would agree that the model is fairly simple by design, though not that it is heavily calibrated. In the application of the model in this study, we calibrate only the $P_{CleaveMax}$ parameter and the initial pool of weatherable P (P_{Weath0}). Aside from these two values, all other model parameters are not calibrated to the experimental site.

'However, the main deficiency that I find is the model performance against measurements in figure 2. First of all, I don't think the 1-to-1 point plot is the best way to display the results, since each point is representing a different experiment and to me it is more interesting to see the different model performance of varying scenarios rather than looking at a overall r2 of eight very different scenarios.'

Thank you for your comment. We understand your point of view here, and agree that a one-to-one plot is not the only way the model performance against measurements can be displayed and communicated. We are happy to incorporate the observations into the timeseries plots to help the reader compare the data and model by scenario.

However, we believe the 1-to-1point plot is the most concise way to present data visualising comparisons of both sizes of simulated versus observed pools and to a lesser extent, how they change with experimental nutrient manipulation. The colour coding and markers are intended to help the reader see the different model performance of varying scenarios. The r^2 then gives an indication of performance across grasslands and treatments and is an interesting measure as it considers how much the model captures variability across sites/treatments. We agree though that achieving a high r^2 is not the ultimate purpose of the study, and that the performance needs to also be interpreted on a 'scenario' basis. We will add a point of clarification on interpretation of the r^2 in the text, and ensure we have sufficiently discussed the performance on a site-by-site basis.

‘One example as the authors have already noticed, is that AGB carbon and soil C are noticeably overestimated in acidic grassland but soil N is not, and more surprisingly, total soil P is underestimated. This pattern really indicates that the model is not capturing the SOM stoichiometry, and it actually worries me about the main focus of the paper is on effects of N and P on soil C storage’

Thank you for raising this issue, these are important points to discuss.

Firstly, we acknowledge that the overestimation in biomass C and soil C alongside an underestimation in total P may imply that the model has failed to accurately capture all elements of the empirical acidic grassland. There were combinations of $P_{CleaveMax}$ and P_{Weath0} that produced simulated C, N and P pools closer to the empirical pool sizes than the pair of values presented in the manuscript section ‘3.1. Varying phosphorus source parameters’.

However, we chose to not use this parameter combination as the resulting simulated grassland was behaving in accordance with a solely P-limited grassland rather than the N-P co-limited grassland we know it to be from the empirical data. This was problematic as the empirical data show strong and clear patterns of increasing biomass and SOC with addition of N, which would not have been captured if we used the parameter combination that produced the least discrepancy between the sizes of the observed and simulated pools.

Instead, we used the parameter combination that reduced the discrepancy between the observed and simulated data the most whilst still maintaining behaviour consistent to stronger N than P limitation.

Secondly, as the $P_{CleaveMax}$ parameter is poorly constrained to empirical data, due to comparatively few studies quantifying plant access to organic P, it is possible that the upper limit of $P_{CleaveMax}$ that we set in the calibration is too high. This could explain the pattern you have identified, as plants in the acidic grassland can access more organic P than they perhaps should and use it to stimulate additional growth, leading to reductions in soil P and increases in plant and soil C.

The effect of a potential overestimation of $P_{CleaveMax}$ on SOM stoichiometry may be a limitation of the modelling approach, that needs to be discussed in more detail. We don’t believe however that this suggests the model is incapable of capturing SOM stoichiometry, but rather it reflects a relatively poor quantification of organic P cycling in semi natural ecosystems.

We shall discuss these considerations in a dedicated model limitation section in the discussion and will further clarify in text what impact this may have on our understanding of carbon storage for the acidic grassland. Further detail regarding our choice of parameter combinations will be included in the methods section ‘2.3. Simulating the field manipulation experiment with the model’.

Thank you again for raising this as we feel we have perhaps not explained this sufficiently in the manuscript. Accordingly, we shall expand upon this in the aims section and the methodology section to make it apparent to the reader.

‘Secondly, it is unclear to me if all the eight experiments are calibrated or only the two unfertilized ones are calibrated.’

Data from all nutrient treatments within the experiment were included in the initial calibration. This was a concern shared by reviewer two but we feel our approach is justified:

We included SOC, SON and total P data from the 0N, LN, HN and P treatments into the cost function to determine optimal P cycling parameters. However, we excluded all biomass data across all four nutrient treatments to be used for separate blind model testing. We decided to use the variable that responded most rapidly and variably to nutrient additions to test the calibrated model, as this would have provided the most robust possible test with the available data.

Ordinarily we would agree that using only unfertilised data for the calibration would be most appropriate for a model development study. However, we should emphasise that this study was more exploratory than developmental and as such it's necessary to use data from all the various treatments to explore these uncertain variables.

'Also, the initial soil pool sizes are not clear to me either.'

Apologies but we are unsure what is meant by 'initial'. Can you please expand on this?

The only soil pool initialised in this study is the P_{Weath0} condition, which represents the initial pool of weatherable P upon soil formation. Upon mineral weathering, this enters more available soil P pools and can become available to plants and so is an important determinant not only of ecosystem nutrient limitation, but also for determining contemporary C, N and P pools.

The calibrated P_{Weath0} pools are provided for each grassland in the results '3.1. Varying phosphorus source parameters' lines 363 – 364. There is no initial pool of C and N at the beginning of soil formation.

'I find it really difficult to understand how to spin up the model for 10000 years and compare to the present day soil measurement. From figure 3 it seems that the model is still far from equilibrium in both ABG C and soil C, particularly in the acidic grassland.'

This is the benefit of spinning up over time rather than a 'spin up to equilibrium' approach. Real ecosystems are rarely in equilibrium due to constantly changing multiple conditions and so our approach avoids this assumption.

To allow this spin-up, as described in the paper, we use a variety of input data. Inputs nearer the present are more accurately defined based on site-scale measurements, and assumptions are made regarding past conditions:

- Climate: Site based temperature and precipitation data is used for the past 60 years, and prior to this, mean annual temperature was temporally varied using an anomaly based on Davis *et al.* (2003) and mean annual precipitation was maintained as constant
- N deposition: Data regarding Wardlow-specific N deposition from 2004 to 2014 was incorporated and scaled using the historical anomaly formulated by Schopp *et al.* (2003), in order to simulate site-scale background deposition.
- Land use history: A land cover history is defined that sets the plant cover type on an annual basis in the model. This was set using pollen stratigraphy data for the sites spanning the majority of the spin up phase.

We would be happy to add further clarifying details on this and the rationale for this spin-up approach (which is not a new approach) in the text.

‘It actually confuses me about the poor soil C correlation between modelled and measured soil C in the acidic soil.’

Thank you for raising this, the correlation between observed and simulated SOC certainly does appear poor at 0.01. This is likely because we have grouped the two grasslands together in the regression analysis, which may otherwise yield more reasonable r^2 values if calculated separately for each grassland (with the caveat of having half the data points). We can certainly explore calculating these regressions separately to see if it helps clarify the relationship between modelled and measured SOC.

‘Why do you choose to spin up the model for 10000 years, and how does the spin up time affect your results?’

This relates to the previous comment on spin-up above, and was also asked by reviewer one, emphasising that we should certainly add more justification to the methodology section here (section ‘2.2.2. Net primary productivity’). We refer to our response below:

‘The N14CP model is spun up from the onset of the Holocene to capture the length of time required for soil formation following deglaciation. This is not in an attempt to truly model this long time period but to form an initial condition for modern day simulations that takes in what we know about the site history and forcings.

We prefer this method over spinning up a model over an undefined time period until it matches a SOC measurement, as is common practice with other similar models, as it avoids the assumption that soils are presently in steady state (which they are not), and the biasing of results from tuning to that initial stock. If after the spin up period used here, the model can simulate the magnitude of contemporary soil C, N and P pools, it’s a good indicator that the processes used by the model and its calibration of initial conditions (P_{Weath0} for example) is suitably reflective of our empirical data.

In addition, N14CP runs on a quarterly time-step and is therefore well-suited to simulating timescales from decades to centuries, which is beneficial considering the timescales of changes in soil pool conditions and nutrient stocks, and responses to long term changes in nutrient availability.’

In reference to us choosing this time period, this is to capture the length of time required for soil formation following deglaciation in north west Europe around this time. We believe this to be the most appropriate time period to use, especially considering we simulate contemporary pools largely by varying the amount of weatherable phosphorus available at the beginning of this spin up phase.

‘A final comment, the discussion need to focus much less on the speculation of model outputs, but include some discussion about the possible caveats of model or study design and uncertainties caused by these limitations.’

Thank you for raising this and for the rest of your comments. We shall be adding a designated section into the discussion to explain some of the model limitations and caveats identified by yourself and the other two reviewers.

Specifically, we shall include detail on:

- The simplicity of the plant pool structures and N14CP's simulation of plant control over nutrient uptake, and add clarification where required, including regarding the $P_{CleaveMax}$ parameter earlier in the methodology section and its potential overestimation.
- The potential effects of CO2 enrichment on N and P availability, how these may be important and why they are currently omitted from N14CP.
- Limitations regarding the quarterly time step used by the model (that allow us to spin up from 10,000 years ago) will be discussed
- The key limitation regarding N-P co-limitation in a model using a Leibig's law of the minimum approach, which we believe may be leading to some of the previous patterns you identify.
- Additional considerations of caveats / model simplifications such as the subsurface transferal of nutrients via fungal networks and the flexibility of plant stoichiometry

Other clarifications in-text will include justification for a calibration using all experimental treatments and some clarification that the simulated grasslands are better considered as models of N and P limited semi-natural ecosystems based on empirical data, rather than perfectly modelled representations of the empirical grasslands.

Reference

Davies, J. A. C., E. Tipping, E. C. Rowe, J. F. Boyle, E. G. Pannatier, and V. Martinsen (2016), Long term P weathering and recent N deposition control contemporary plant-soil C, N, and P, Global Biogeochemical Cycles, 30(2), 231-249. https://doi.org/10.1002/2015GB005167

Davis, B. A. S., S. Brewer, A. C. Stevenson, and J. Guiot (2003), The temperature of Europe during the Holocene reconstructed from pollen data, Quat. Sci. Rev., 22(15–17), 1701–1716

Schöpp, W., M. Posch, S. Mylona, and M. Johansson (2003), Long-term development of acid deposition (1880? 2030) in sensitive freshwater regions in Europe, Hydrol. Earth Syst. Sci. Discuss., 7(4), 436–446