Response to reviewers:

Reviewer 1:

The manuscript fits the scope of the journal and represents clear added value to biogeochemical community. Decade-long incubation experiments with permafrost soils are indeed extremely rare and this study contributes to better understanding of main governing factors and magnitude of occurring processes during plant litter interaction with permafrost soils.

Authors:

We would like to thank the reviewer for the time invested in our manuscript and the valuable comments, which will help to improve the manuscript.

Reviewer1:

My main criticism is lack of information on soil minerals, the key components controlling C storage. The identity of these minerals should e established via combination of XRD and total chemical analysis, and their nature should be mentioned already in the Abstract (which clays, which oxides – Fe, Al, Mn?, amorphous, allophanes?)….

Furthermore, SEM observations of post-reacted minerals can be useful to identify possible changes on the surfaces of these minerals after incubation

Authors:

We agree that information on detailed clay minerology and the iron oxide composition would be very helpful for better understanding the mechanisms of organic matter stabilization and may also provide valuable information on the reasons for the observed heterogeneity of litter stabilization. However, we only had a very limited amount of sample available for our incubations, as they were retrieved from cores drilled into the permafrost in a very remote, hard-to-access region. Therefore, we were strongly restricted in the analysis that were possible after sample fractionation, and had to prioritize certain analyses. We focussed on C/N and carbon stable isotope analysis, for which most of the available mineral fraction sample was required. To determine the role of iron oxides on the stabilization of fresh litter we used the remaining material of the mineral fraction and extracted short-range-order minerals, known to play an important role in organic matter stabilization, with hydroxylamine (doi.org/10.2136/sssaj1983.03615995004700020010x). Subsequently we quantified the concentrations of iron, aluminium and DOC in the extracts. While enough material (> 1 g) was available from the oxic incubations we had to compile samples from replicates of the anoxic incubations for the extractions. The extracted DOC comprised about 4 % of the total C in the MAOM fraction without a significant difference between oxic and anoxic incubations. Aluminium concentrations were below the detection limit in most of the extracts but extracted Fe concentrations significantly correlated with the liberated DOC concentration, which indicates a role of amorphous iron in organic carbon stabilization. However, due to low extract volume and low DOC concentrations in the extracts we were not able to determine the carbon stable isotope signature of extracted DOC. Hence the contribution of *litter-C* and *permafrost-C* in the DOC could not be quantified. Therefore, the main question, i.e., the contribution of *litter-C* and *permafrost-C* to the DOC pool, could not be answered with this method and we decided not to present these data. Since we do not have any soil material left

from the MAOM fractions of most of the samples, we unfortunately are not able to conduct further analyses, which we acknowledge in the revised manuscript (lines 345-347). However, we deepened our discussion on the role of clay minerals and iron oxides in the revised version of the manuscript in an attempt to properly address this comment (lines 343-345, lines 375-378).

Reviewer1:

Another major comment is experimental setup (section 2.2): For the incubation conditions, 4 °C over 9 years is not what one expects in soils of the Samoylov Island or other places in Yakutia. Discuss the role of annual freezing cycles on laboratory modelling

Authors:

We are fully aware that such long laboratory experiments are highly artificial, and that it is impossible to mimic the in situ conditions present in the Siberian tundra with our experimental setup. Based on this, we discuss the shortcomings of using a constant incubation temperature of 4 °C. Furthermore, we discuss the potential impact of freeze-thaw cycles on the decomposition of organic matter under in situ conditions, and how this may affect the interpretation of our data (lines 427–433).

Reviewer1:

L375-380 Again, the identity of clay (or oxide?) minerals, their specific surface area and their capacity to adsorb soil DOM become the crucial issues for understanding related mechanism. See for instance some published experiments in this direction (https://doi.org/10.1016/j.geoderma.2021.115601)

Authors:

We agree that clay mineralogy and iron oxide composition are important parameters for stabilizing organic matter, but as outlined in detail in our reply above, we are unable to conduct further analyses due to insufficient soil material for further analysis. We acknowledge this shortcoming, however, and have expanded the discussion on this important topic to address this concern (lines 343–347 and 375–378).

Reviewer1:

L 409-410 What are the possible mechanisms of such binding mode?

Authors:

Indications for a clustered distribution of organic matter on mineral surfaces and a layered structure of these clusters have been described repeatedly (e.g., https://doi.org/10.1046/j.1365-2389.2003.00544.x; https://doi.org/10.2134/jeq2005.0342). The nature of these interactions is not well understood but multivalent cations and hydrogen bonding (for a review see e.g., https://doi.org/10.1007/s10533-007-9103-5), hydrophobic interactions and polysaccharide-rich microbial necromass are involved. We expanded the discussion on this topic (lines 410-422; see also response to Reviewer 3).

Reviewer1:

The role of Fe oxides, especially at the variable redox conditions, on OM binding to soil minerals, is not discussed

Authors:

As suggested, we include discussion on the important function of iron oxides in stabilising organic matter in (lines 333-335, lines 343-345, lines 436-440).

Reviewer2:

General Comments:

This study addresses an important gap in our understanding of permafrost carbon dynamics. The experimental design is strong: isotope labelling to track fresh litter versus old permafrost carbon, systematic comparison of oxic and anoxic conditions, physical fractionation to see where carbon ends up, and most impressively, the running incubations for 9 years. That timeframe distinguishes this from most decomposition studies, which only look at the labile pool and extrapolate from there.

You have direct evidence that litter carbon enters the mineral-associated fraction rapidly but doesn't necessarily stay there long, while old permafrost carbon in that same fraction persists for centuries. That observation alone challenges the standard fractionation framework where mineral association implies stability. The comparison between oxic and anoxic conditions is also well-designed. You not only measured rates but you tracked which carbon pools formed under each condition and how stable they were.

That said, the paper needs substantial revisions before it can be published. The analytical approach has problems that weaken your quantitative conclusions, and the interpretation claims more certainty than the evidence provides. These are fixable issues. With better statistical methods, more careful uncertainty characterization, and interpretation that stays within what your correlational evidence can support, this could be a strong contribution in soil biogeochemistry.

Section addressing individual scientific questions/issues ("specific comments"):

1. Concerns about the MRT analysis:

The 28% exclusion rate for anoxic samples is a problem: You excluded 10/36 anoxic replicates because the model didn't find "reasonable fitting parameters," but only 3/36 oxic replicates. The paper doesn't say what "reasonable" means here. This asymmetry suggests the two-pool model is working fine for oxic decomposition but struggling with anoxic dynamics, yet you're treating both sets of MRT estimates as equally reliable. What if the excluded samples share some characteristic (maybe they're the ones with the slowest decomposition rates, or unusual substrate chemistry)? Then your reported median anoxic MRT of 1572 years for stable permafrost-C could be biased.

Even for samples where the model converged, the results raise questions about whether it's working properly. For anoxic permafrost-C, Figure 2a shows the labile pool fraction has a median of ~1.2%. But total anoxic permafrost-C decomposition over 9 years was only 0.5-4.4%. This means the "labile pool" estimate is what was actually decomposed. The model isn't separating truly labile material from what was just simply accessible over the timeframe, it's just labelling "whatever decomposed" as labile. This suggests the two-pool partitioning may not be meaningful for anoxic permafrost-C.

Please define the criteria for "reasonable" parameters and report whether the excluded samples had anything in common. The anoxic samples that failed to fit might just not follow two-pool kinetics given that anoxic decomposition uses foundationally different metabolic pathways. There should be an acknowledgement that the anoxic MRT estimates rest on a more restricted subset of the data for full disclosure.

The uncertainty characterization needs work: You're not reporting confidence intervals for individual MRT estimates, even though nonlinear curve-fitting produces them. Given the reported median MRTs and the small fraction of permafrost-C decomposed over the incubation, those fitting uncertainties are probably huge. A reported MRT of 1572 years might have 95% confidence bounds between 500-5000+ years. The variability across samples is also extreme, for stable litter-C under oxic conditions, the IQR actually exceeds the median. And you're not reporting the humification coefficient even though it's one of the four fitted parameters and Knoblauch 2013 reports it.

Given that your Random Forest model for anoxic permafrost-C MRT explained only 6.5% of variance, you should explicitly acknowledge that the controls on this parameter remain basically unidentified. When you say permafrost-C MRT is "10-fold higher" than litter-C MRT, you're calculating from medians, but given the spread in both distributions many individual sample pairs probably show considerably smaller or larger ratios. It would be more accurate to say something like "permafrost-C MRTs are consistently higher than litter-C MRTs across samples, typically by one to two orders of magnitude, though both show considerable variation."

The interpretation needs to be clearer about which C pools these MRTs represent: The Abstract says "mean residence times of the stable litter C pool of 17.6 (22.3) yr... indicate a substantial stabilization of fresh litter C." But the stable pool is only 45% of litter-C under oxic conditions, 55% is in the labile pool with MRT below 0.5 years. Someone could reasonably think litter-C overall persists for about 17 years when actually more than half decomposes in the first year. The Conclusions state that "the mean residence time of recent plant litter C is one order of magnitude lower than that of permafrost C" without noting this comparison is between stable pool MRTs. For permafrost-C where >92% is in the stable pool, the stable pool MRT approximates whole-system turnover well but for litter-C where only 45% is stable, it doesn't.

You should specify in the Abstract and Conclusions that these MRT values refer to the stable pool fraction and what proportion of each C source that represents. Consider reporting a flux-weighted system MRT that integrates both pools: MRT_total = (fraction_labile × MRT_labile) + (fraction_stable × MRT_stable). And maybe reframe the "substantial stabilization" language to acknowledge that while 40-60% of litter-C did resist decomposition over 9 years, the majority actually decomposed quickly.

Authors:

We excluded replicates from the two-pool model analysis for which the model produced a mean residence time (MRT) of 1 million years, the upper limit of the model. This indicated that the model could not be fitted to the data because the data was already too linear after the long pre-incubation period. This information is included in the revised manuscript (lines 178-182). We also clearly state which pool the presented MRT belong to. Furthermore, we removed the random forest analyses of the MRT of the anoxic incubations from the manuscript (e.g. Fig. 4c, e). Additionally, we acknowledge in the discussion that the model results from the anoxic incubations are based on fewer replicates and are therefore less robust than the results from the oxic incubations (lines 444-445, see response to reviewer 4).

Reviewer2:

1. Comment on Statistical Methodology:

Concerns about the statistical approach used to identify controls on decomposition outcomes: questionable analytical choices, incomplete reporting that prevents evaluation of reliability, and interpretation that overstates what the correlational evidence can support.

The stepwise regression - Random Forest pipeline doesn't make sense: You say you're using stepwise regression to "obtain a sufficient number of variables for the subsequent RF analysis", but this inverts normal logic. RF handles variable selection internally: this one of its desirable features . Pre-filtering with stepwise regression just means you're propagating stepwise's well-known problems (instability, collinearity sensitivity) into your RF results. Why not let RF do the variable selection itself?

Sample sizes appear problematic for the complexity of the models: Table A2 shows models with 6-11 degrees of freedom, but your effective sample sizes look quite small. You started with 12 samples × 3 replicates per condition, then excluded replicates where the decomposition model didn't converge (3 oxic, 10 anoxic). It's not clear whether you analyzed replicates individually or used sample means. If you used means, you could have as few as 12 observations trying to fit 6-11 predictors. The anoxic permafrost-C MRT model explaining only 6.5% variance might be telling you the sample size just isn't enough. You need to state the actual sample sizes used in each analysis.

The model selection criteria need clearer justification: You selected models based on "lowest AIC, Delta value < 2, and ≥ 6 degrees of freedom." The Delta < 2 criterion is standard, but what did you do when multiple models satisfied all criteria? And the ≥ 6 df requirement seems random. Why 6 specifically?

Clarify how z-score scaling was done: Did you scale the combined dataset or separately for oxic/anoxic conditions? If you scaled separately before merging, z-scores wouldn't be comparable across conditions. If you scaled on pooled data, between-condition variance becomes part of what RF uses to make splits. You need to state which you chose and why.

Merging oxic and anoxic predictor sets requires justification: You merged the models "to include the same set of predictor variables", but your own results show these conditions produce markedly different outcomes. Forcing a common predictor structure onto systems potentially governed by different processes might hide condition-specific controls. Either explain why merging makes sense or run fully separate RF models for each condition.

Some predictor-outcome relationships are mathematically coupled: Using "final litter-C mineralization rate" to predict "fraction litter-C decomposed" is partially circular. Similarly, MRT is derived from rate constants, so finding that decomposition rates predict MRT just recovers the structure of your own calculations. You should exclude predictors that are algebraically related to response variables.

These reporting gaps prevent evaluation of reliability:

1. Report out-of-bag error estimates. The high explained variance can be optimistic with small samples. The OOB error will tell you whether the models generalize.
2. Add uncertainty to the partial dependence plots (Figure A2). Several curves show apparent nonlinearities that could reflect one or two influential observations rather than real relationships. Bootstrap confidence intervals would show which features are reliable versus noise.

3. Assess variable importance stability. Your conclusions rely heavily on importance rankings, but RF importance measures can be unstable with small samples and correlated predictors. Run the analysis multiple times and report whether rankings are consistent.

Given the sample size constraints, you might consider whether simpler analytical approaches would be more appropriate. Multiple regression with a few theoretically selected predictors, regularized regression methods like LASSO or elastic net, or even careful examination of bivariate relationships could provide more stable and interpretable results.

The interpretation crosses result limits in several places: First, the anoxic permafrost-C MRT model explaining only 6.5% variance is acknowledged as "barely converged," yet it appears in Figure 4e alongside models that actually work. Drawing conclusions from variable importance rankings in a model with such poor explanatory power is not defensible. Either exclude this model from interpretation entirely or acknowledge Figure 4e separately with explicit caveats.

Second, you slip from correlation into causal language. Saying "mineral bound permafrost-C is more important for incorporating fresh plant litter into the mineral fraction than the size of the clay fraction itself" makes a mechanistic claim your evidence can't support. Variable importance in RF indicates predictive utility, not causation. Replace phrases like "is more important for" with "is more strongly associated with" throughout the paper.

Third, the interpretation that litter-C binds to pre-existing organo-mineral clusters rather than free mineral surfaces might be plausible, but the correlation between MA-litter-C and MA-permafrost-C doesn't differentiate this mechanism from alternatives. Either acknowledge alternative explanations explicitly or provide additional evidence.

The partial dependence plots need more critical discussion: Many show sensible monotonic patterns, but several display non-monotonic responses or sharp inflections at the extremes of predictor ranges. With small samples, a single unusual observation can create apparent nonlinearities. Flag regions of low sample density and clarify which curve features are reliable enough to interpret.

Authors:

The number of replicates used to determine the mean residence times using a two-pool model is stated in the revised version of the manuscript (lines 182-184). For the oxic incubations, the number of replicates was n = 34 for the two litter-C pools and n = 31 for the permafrost-C pools. For the anoxic incubations, n was 28 for the litter-C pools and 22 for the permafrost-C pools. Our stepwise variable selection for the RF analyses approach is justified, since we wanted to identify the most relevant variables prior to RF analyses, and subsequently used RFs to rank the importance of these variables. We acknowledge limitations associated with the small sample size, but use RF and step-wise linear regression only in combination with basic statistics, e.g., correlation analyses. RF results confirm results derived by basic statistical analyses and we therefore consider conclusions robust. As mentioned above, we removed the random forest analysis of the MRT results from the anoxic incubations from the revised manuscript. Furthermore, we state that the model results from the anoxic incubations are based on fewer observations and are therefore less robust than the oxic incubation results (lines 444-445, see also response to reviewer4).

Reviewer2:

1. Figure 2 shows important patterns that deserve more explicit discussion in the text:

You correctly note that labile litter-C MRTs are similar under oxic and anoxic conditions (around 0.5 years). However, Figure 2b also shows that some permafrost-C samples have labile pool MRTs of 30-70 years, particularly under anoxic conditions. If the "labile" pool takes decades to decompose, the two-pool partitioning may not be meaningfully differentiate fast from slow-cycling material in those samples.

The pool partitioning patterns (Figures 2a and 2c) show substantial variation that isn't discussed. For litter-C, the labile fraction ranges from 30-70% across samples, meaning individual samples show very different stabilization trajectories.

The extreme variability in stable pool MRTs (Figure 2d) is also not discussed. For stable litter-C under oxic conditions, the IQR (22.3 yr) exceeds the median (17.6 yr). For stable permafrost-C under anoxic conditions, the distribution spans from roughly 500 to 2500 years. Some of this variability likely reflects real differences in stabilization driven by soil properties. Discussing what controls this variability would add value.

Finally, you're reporting MRTs of centuries to millennia based on 9 years of observations during which less than 5% of oxic permafrost-C and roughly 1% of anoxic permafrost-C decomposed. The paper's language should suggest that these are order-of-magnitude estimates rather than precise quantitative predictions.

Authors2:

We discuss the variability of the MRT presented in Fig. 2 (lines 339-345). We assume that the variability in the size of the different pools and their MRTs is mainly caused by the wide variation in the properties of the soil samples that we incubated. This reflects the wide range of environmental conditions when these permafrost soils were formed during the Holocene. For example, the TOC content ranged from 0.6% to 12.4%, the nitrogen content from 0.04% to 0.78%, the C/N ratio from 13 to 26, the pH from 4.0 to 7.2, and the clay content from 2.9% to 21%. We intentionally selected samples with different properties to identify the effect of these parameters on carbon stabilisation.

Reviewer2:

1. Figure 3 requires more careful interpretation:

Fig. 3 shows that after 9 years, most remaining C is in the mineral-associated fraction. However, the figure shows where C ended up, not necessarily how stabilization occurred. Without baseline fractionation or temporal information, the high MAOM proportion could be either preferential incorporation into MAOM or preferential loss of POM.

You acknowledge this issue in section 4.2, noting that fractionation at the endpoint means "the size of the POM fraction would decrease more rapidly than that of the MAOM fraction." However, this caveat appears after you've already stated that the high MAOM fraction "indicates an efficient incorporation" in the results and Abstract. The endpoint fractionation bias doesn't just provide "one possible explanation." It completely affects how Figure 3 can be interpreted.

Your results actually challenge a key framework in soil carbon science. You show that more litter-C in the mineral fraction correlates with higher decomposition rates and lower MRTs. You explicitly state that your experiment "provided no evidence that the MAOM pool in permafrost is the most resistant." Fresh litter-C enters MAOM rapidly but turns over on decadal timescales, while old permafrost-C in the same MAOM fraction turns over on century to millennial timescales. This decoupling of incorporation from long-term stabilization contradicts the standard fractionation-based stability framework.

Yet Figure 3 visually treats MAOM as a single, homogeneous compartment, and the framing throughout hedges on whether this finding is a problem with your approach or a real discovery. The Abstract and Conclusions emphasize "efficient incorporation" and "stabilization mainly in the mineral associated fraction" as if MAOM dominance equates to stability, when your data show these processes are decoupled for fresh litter.

Consider positioning this finding as a contribution (I would even suggest doubling down on this a core finding and adding evidence/analyses to support it if it indeed is a real biology) rather than treating it as an unexpected complication. If the conventional MAOM = stability framework genuinely doesn't apply to fresh inputs in permafrost systems, that's worth stating clearly and prominently. The caveats you provide in sections 4.2 and 4.3 are valuable, but if this heterogeneity within MAOM is a real phenomenon, it deserves to be a headline finding integrated throughout, not just acknowledged in discussion after more conventional claims have been made.

Authors:

We are confident that the litter bound to the MAOM fraction at the end of the incubation period was incorporated into the MAOM fraction during incubation, since the litter was added as particulate organic matter. At the start of the incubation period, no litter was bound to the MAOM fraction. We emphasize this fact in the revised discussion (lines 401-403). For this reason, we do not consider a ‘baseline fractionation’ to be necessary.

Reviewer2:

1. Figure 4 presents result that require more careful contextualization:

Fig. 4e shows that the RF model for anoxic permafrost-C MRT explained only 6.5% of variance compared to 83.5% for oxic conditions. You acknowledge this model "barely converged," but 4e still displays variable importance rankings for anoxic conditions alongside the successful oxic model without visual distinction or strong caveats. Variable importance rankings are not meaningful when model explanatory power is low. Either exclude the anoxic permafrost-C results from Figure 4e entirely or present them separately with explicit warnings.

This matters for your overall narrative about controls on stabilization. You state that controls differ between oxic and anoxic conditions, but if the anoxic model essentially failed to identify any controls, you cannot make confident statements about what drives anoxic permafrost-C persistence.

Fig. 4a shows many significant relationships but does not control for covariation between predictors or address whether multiple testing correction was applied. (Bonferroni correction

"in case of multiple tests" is stated but not explicitly tied to the correlation matrix. Please clarify whether and how Bonferroni was applied to the correlation matrix specifically.)

Your discussion sections provide important context that should be integrated into how Fig. 4 is presented. You note that MAOM size did not contribute to explaining permafrost-C stabilization, that your experiment "provided no evidence that the MAOM pool in permafrost is the most resistant," and that final rates dominated over initial conditions. These are interesting findings that challenge conventional assumptions. But Figure 4 as currently presented (especially without uncertainty estimates on importance rankings and without noting the anoxic model failure prominently) invites simpler interpretations that your discussion then has to walk back.

Consider whether Figure 4 should focus only on the successful models (oxic conditions and perhaps litter-C under both conditions) and move the failed anoxic permafrost-C model to supplementary material with appropriate caveats. Or revise 4e to visually label the failed anoxic model with clear notation that importance rankings cannot be interpreted when variance explained is negligible. The figure caption should explicitly state the variance explained for each model.

The bigger question is what Figure 4 actually says about stabilization controls. If final decomposition rates dominate predictions and soil properties have low importance, this suggests: (a) lack of appropriate baseline soil characteristics, (b) stabilization processes depend more on decomposition reactions vs. initial soil conditions, or (c) lower sample size limits detection of weaker effects. Instead of just reporting that final decomposition rates are the most important predictors, acknowledge upfront what this pattern means: "The dominance of final decomposition rates over baseline soil properties as predictors suggests that stabilization outcomes may be determined more by decomposition processes than by initial soil characteristics. This could reflect either unmeasured soil properties, dynamic interactions between substrates and soil minerals, or limitations in detecting weaker effects with our sample size."

Authors:

As mentioned above, we removed the random forest analysis of the anoxic incubation data from the revised manuscript. Furthermore, state that the RF model results from the anoxic incubations are based on fewer replicates and are therefore less robust than the results from the oxic incubations (lines 444-445). Indeed, the random forest model shows that final carbon mineralisation rates are the most important predictors of MRT for litter-C and permafrost-C, as well as the fraction of permafrost-C that was decomposed. However, decomposition rates depend on a variety of soil properties (see Fig. 4a). Therefore, the reviewer's conclusion that ‘the dominance of final decomposition rates over baseline soil properties as predictors suggests that stabilisation outcomes may be determined more by decomposition processes than by initial soil characteristics' is an oversimplification, as it neglects the dependence of carbon decomposition rates from soil properties.

Reviewer2:

1. Discussion Section has an important omission:

The limitations section acknowledges ecological and landscape-scale constraints but completely omits discussion of the analytical limitations that affect your quantitative conclusions. There's no mention of:

- The 28% exclusion rate for anoxic MRT estimates and what this implies about model applicability
- Sample size constraints that led to the anoxic permafrost-C RF model explaining only 6.5% of variance
- Endpoint fractionation bias affecting interpretation of MAOM dominance
- Uncertainties in MRT extrapolations (centuries to millennia from <5% decomposition)

Given that your conclusions rest heavily on quantitative MRT values and RF-derived variable importance rankings, the discussion should explicitly acknowledge these methodological uncertainties alongside the ecological ones.

Authors:

We have already responded to these issues, which have been raised by the reviewer in previous sections of this review.

Reviewer2:

1. Microbial community and activity needs more careful handling:

Your study attributes all decomposition and stabilization to microbial activity, but you never actually measured the microbial community. No 16S/ITS sequencing, no metagenomics, no PLFA profiles, no enzyme assays, no microbial biomass.

This creates a problem because your experimental design introduces a confound. You pre-incubated samples for 4 years, then incubated them for 9 more years at 4°C in sealed vials. The microbial community in the permafrost when you collected it is almost certainly not the community doing the work at year 9. Bottle effects, substrate depletion, redox shifts drive community succession. You're assuming microbial community composition and function aren't rate-limiting, but that's not tested.

Authors:

Indeed, we have no data on shifts in microbial community composition as this was beyond the scope of our study, and all discussion about this topic would be purely speculative.

Reviewer2:

You claim litter-C decomposition products and "microbial necromass" get incorporated into MAOM, but without biomarker data (amino sugars, enzyme assays etc.) this is speculation. We don't know if the MAOM-associated litter-C is microbial necromass, partially degraded plant polymers, or sorbed DOC intermediates.

You should acknowledge this limitation explicitly. Discuss how community succession in sealed long-term incubations probably differs from in situ thaw dynamics. Qualify your claims about necromass contributions to MAOM.

Authors:

We do not claim that microbial necromass is incorporated into the MAOM fraction, but we discuss it as a possible explanation for the low C/N ratio in the MAOM fraction (lines 352-357, lines 362-364, lines 381-383), and we explicitly acknowledged that we have no data on microbial necromass.

Reviewer2:

Anoxic decomposition is more complicated than you're treating it:

The paper treats anoxic decomposition as one simple process, but the microbiology is complex. Under anoxic conditions at 4°C, carbon mineralization involves players like fermenters, syntrophs, and methanogens operating near thermodynamic equilibrium.

Your two-pool kinetic model assumes first-order decay, but anoxic decomposition is constrained by syntrophic interactions. Hydrogen partial pressures, acetate concentrations, and thermodynamic feasibility all create non-linear feedbacks that first-order models can't account for. This probably explains why 28% of anoxic replicates failed to produce "reasonable" fits.

You note that the oxic to anoxic mineralization ratio increased from about 3.6 initially to 8 by the end, attributed to "lower energy yield" and enzyme inhibition. But the increasing divergence more likely reflects progressive accumulation of thermodynamically unfavourable substrates. As labile compounds get consumed, the remaining organic matter requires increasingly difficult initial reactions before terminal electron acceptors can be activated. Without oxygen, there's no enzymatic shortcut for breaking complex aromatic structures.

Your discussion of oxic vs anoxic differences (Section 4.1) needs more nuanced treatment of microbial metabolic constraints under anoxia. Progressive thermodynamic limitation of syntrophic decomposition probably explains the increasing oxic to anoxic divergence better than the enzyme inhibition mechanisms you cite.

Authors:

The two-pool carbon decomposition model treats anoxic organic matter decomposition as a single process. This simplification is acknowledged in the revised manuscript (lines 178-180). However, it is unclear to us what the reviewer means with ‘thermodynamically unfavourable substrates. This could refer to substrates that yield less energy during decomposition under anoxic conditions, a potential cause of lower decomposition rates that we already discussed in the submitted manuscript (lines 327-330 in the revised manuscript).

Reviewer2:

Priming needs a more careful consideration:

The paper acknowledges priming but dismisses it in one sentence citing Knoblauch et al. (2018): "priming seems of minor importance in our permafrost samples." This deserves more scrutiny.

When you added 13C-labelled litter to pre-incubated permafrost, the litter provided labile substrates that could stimulate co-metabolic breakdown of recalcitrant permafrost-C. Your isotope approach can seperate litter-derived from permafrost-derived CO_2, so priming could be detected. But the 2018 paper assessed priming over a shorter timeframe. Over 9 years, dynamics may differ substantially. The positive correlation between MA-litter-C and MA-permafrost-C could equally mean that samples with more accessible mineral-associated permafrost-C support more microbial activity, which produces more microbial necromass (measured as MA-litter-C).

Dismissing priming based on earlier short-term observations needs revisiting. Over 9 years, low-level priming effects could effectively alter the stable permafrost-C pool.

Authors:

In our 2018 paper (http://doi.org/10.1038/s41558-018-0095-z), we estimated the importance of initial priming using a simplified approach. As we have no control incubations without added litter, it is not possible to address priming clearly. We feel it is too speculative to discuss about the importance of priming without proper data.

Reviewer2:

"Stabilization" in MAOM needs qualification:

Your central finding is that over 80% of remaining litter-C is in the MAOM fraction, which you interpret as "stabilization." But the MAOM fraction in a closed incubation system is a graveyard of microbial activity. Over 9 years, microbial cells grew on litter-C, died, and their necromass either sorbed to minerals or got consumed. The 13C label tracks carbon atoms, not molecules. So MAOM-associated litter-C could be intact microbial necromass polymers, small metabolites sorbed to mineral surfaces, partially oxidized plant polymers, or extracellular enzymes and exopolysaccharides.

Each of these has very different desorption kinetics and vulnerability to future decomposition. Calling them all "stabilized" merges fundamentally different processes. Without spectroscopic or biomarker data, it's hypothetical for this system.

The interpretation of MAOM-associated litter-C as "stabilized" needs to acknowledge the heterogeneous nature of this pool. Without molecular-level characterization, the proportion of litter-C that's truly stabilized vs. transiently associated is unknown.

Authors:

As mentioned above and in the manuscript, we have no data on microbial necromass and therefore cannot elaborate further on the composition of litter-C in the MAOM fraction. However, we demonstrated that the majority of the remaining litter-C was present in the MAOM fraction by the end of the experiment, and that the stable litter fraction, which contains the majority of the remaining litter-C, has MRTs of decades. One of the manuscript's key findings is that the MAOM pool is heterogeneous, as clearly stated in the text (lines 408-418).

Reviewer2:

The 4°C incubation has microbial implications:

Incubation at 4°C is defensible as representing permafrost thaw, but it has biological implications you don't discuss. At 4°C, microbial growth rates are extremely slow, and community turnover is minimal. The microbial community at the end is probably dominated by psychrotolerant organisms with limited metabolic diversity. *In situ*, thawing permafrost experiences temperature fluctuations, seasonal warming, and potential exposure to temperatures well above 4°C during summer.

This means the "stable" litter-C and permafrost-C pools you identified may partly reflect kinetic limitation by a thermally constrained microbial community, not the stabilization mechanisms the model attributes to the stable pool. Incubation at a higher temperature would likely yield different pool sizes and MRTs not just because of Arrhenius effects, but because different organisms and metabolic pathways become viable.

Briefly discuss how the constant 4°C incubation may constrain microbial community diversity and metabolic capacity relative to in situ conditions.

Authors:

We discuss the fact that long-term laboratory experiments may not mimic natural conditions, and acknowledge the artificial temperature conditions in the incubations in the discussion of the revised manuscript (lines 427-433, see also response to reviewer 1).

Reviewer2:

The elegant isotope-labelling design tracking carbon fate over 9 years is a substantial experimental achievement. But the paper treats microbial processes as implicit drivers without measuring them. The most impactful revisions would be (1) acknowledging the microbial community black box explicitly and discussing how it limits mechanistic interpretation, (2) providing more nuanced treatment of anoxic metabolic constraints, and (3) qualifying the "stabilization" narrative by acknowledging the heterogeneous and potentially transient nature of MAOM-associated litter-C.

List of purely technical corrections:

- Stable misspelled as "Stabile" in Figure 4a
- Missing DOI for Schmidt et al.
- Line 276: Redundant "both" - "both under both oxic and anoxic conditions"

Authors

Corrected as suggested

Reviewer3:

The manuscript fits the scope of the journal well, the study provides a rare and valuable 4+9-year long-term incubation dataset that effectively bridges the gap between short-term laboratory observations and decadal ecosystem turnover times. The finding that thawing permafrost can act as a substantial sink for recent plant litter (stabilizing 39-59% of added C) is a significant contribution to the permafrost-carbon-climate feedback discourse.

Authors:

We thank the reviewer for the time spent to evaluate our manuscript and their valuable comments, which will help to improve our manuscript.

Reviewer3:

1. The second half of the abstract requires refinement to improve its readability and narrative flow. Currently, the high information density, specifically the inclusion of multiple absolute values and detailed statistical parameters such as Mean±SD and Median (IQR) detracts from the study's core message. I recommend streamlining this section by removing granular statistics and keeping only the most critical findings (e.g., representative percentages or fold-differences). Furthermore, absolute values like the mean residence times should be contextualized or simplified so that the reader can immediately grasp the significance of the results without referencing the main text.

Authors:

As suggested by the reviewer, we substantially revised the abstract, simplified the presentation of data and explained the use of the mean residence times to make the abstract clearer and more readable.

Reviewer3:

1. Lines 209-211: The sentence describing the decrease in mineralization rates is convoluted due to nested parentheses. I suggest rephrasing for clarity by separating the statistical parameters from the core narrative. For example: 'Litter-C mineralization rates declined significantly more rapidly than permafrost-C rates ($p < 0.001$). By year nine, median litter-C rates dropped to 3% (oxic) and 1.5% (anoxic) of their initial values.

Authors:

This sentence was changed as suggested (lines 218-220).

Reviewer3:

1. Lines 225-228: For a balanced comparison, quantitative data should be provided for both the labile litter-C pools and the permafrost-C pools, rather than focusing solely on the latter. Additionally, the conjunction “but also” is logically misplaced here as there is no contrasting relationship between these two observations. I suggest using “and” or “moreover” to indicate that the permafrost-C pool size aligns with the cumulative decomposition observed.

Authors:

As suggested, we completely revised the respective paragraph (lines 235-238).

Reviewer:

1. Figure 4 Caption: For better alignment with the visual layout, the label “(b)” should be moved to immediately follow the word “predictors” to clearly indicate which portion of the figure is being described.

Authors:

Done as suggested.

Reviewer3:

1. Terminology Clarity: The usage of the term “permafrost” needs to be more precisely defined and consistently applied throughout the manuscript to avoid ambiguity. Currently, “permafrost” is used interchangeably in three distinct contexts: As the general soil name used in incubation 2. As a distinct carbon source in contrast to fresh plant litter (“permafrost-C” vs. “litter-C”) 3. As a specific vertical layer, differentiating the perennially frozen ground from the seasonally thawed “active surface layer”. I suggest the authors adopt more specific descriptors, such as “perennially frozen subsoil” for the vertical layer, to ensure clarity, especially in the Discussion where these concepts overlap.

Authors:

We thoroughly revised the manuscript, taking care to use the term 'permafrost' unequivocally, in accordance with the following definition: 'Earth materials (soil, rock, moisture, gases, and organic material) that remain at or below 0°C for at least two consecutive years' (http://dx.doi.org/10.1139/e76-089). We use the term 'permafrost material' when referring to the incubated material and clearly explain the term '*permafrost-C*' as carbon originating from thawed permafrost material (lines 17-18 and lines 120-122). We write this term in italics, as we do with '*litter-C*', to indicate that it describes a distinct carbon pool rather than carbon in current permafrost.

Reviewer3:

1. The Discussion section is currently overly dense and challenging to follow. The frequent shifts between temporal scales (short-term vs. decadal) and spatial dimensions (active layer vs. deep permafrost), compounded by the ambiguous use of the term “permafrost”, significantly increase the cognitive load for the reader. I recommend:

- Streamlining Section 4.1 and 4.2 by moving some repetitive data comparisons to the Results or Supplement.
- Adding sub-headings or concluding sentences to each paragraph that explicitly state the “mechanistic takeaway”.
- Clarifying the terminology (refer to comment 5) to distinguish between the substrate, the carbon source, and the soil layer.

Authors:

We have substantially revised the Discussion section, streamlined the first two sections and added a concluding sentence to each paragraph. Furthermore, in response to suggestions from reviewers1 and 4, we have added discussion of the reasons for the high variability of MRTs and the role of clay and iron minerals. Furthermore, we clarified the use of the term 'permafrost' throughout the text (see previous comment).

Reviewer3:

1. Lines 406-420: The interpretation of the "layered organo-mineral model" is a highlight of the discussion, as it effectively reconciles the high MAOM incorporation of litter-C with its low MRT. To improve clarity for the reader, I suggest adding a visual Aid. A conceptual diagram illustrating the "inner" vs. "outer" zones would be far more effective than text alone in explaining why clay fraction size was not the primary driver.

Authors:

We considered adding a conceptual graph on the organo-mineral model but finally decided to rather expand the discussion and explanation of such a model (Lines 410-422, see also comment on reviewer1) and give references to the relevant original literature. Since our data rather support an already described model of organic matter stabilisation, we would prefer to reference to the existing original literature rather than presenting a conceptual graph of an already described concept in the main text.

Reviewer4:

General comments:

The manuscript aligns well with the scope of Biogeosciences and is of clear scientific interest. The long-term incubation experiments (9 years) are a major strength because they allow short-term responses, likely reflecting microbial adjustment to changing conditions and a transient increase in organic matter decomposition, to be distinguished from longer-term, more stable trends. The authors' conclusions regarding a potential "compensation" of old permafrost carbon decomposition by inputs of fresh, litter-derived organic material are also noteworthy.

Authors:

We would like to thank the reviewer for their time and appreciate their valuable comments which will help to improve our manuscript.

Reviewer4:

As a suggestion, I would encourage the authors to place greater emphasis on those results that can be interpreted most robustly. At present, the manuscript can feel somewhat overburdened by statistical material that is not always straightforward to interpret or sufficiently reliable, which may distract from the key messages and primary conclusions.

Authors:

In the revised manuscript, we removed the random forest analysis of the MRTs from the anoxic incubations due to the fact that the model could only explain a small fraction of the variability in the data. Furthermore, we substantially revised the Discussion section, focusing on the primary conclusions, e.g., the difference in decomposability between *litter-C* and *permafrost-C*, and the interaction between these two carbon pools.

Reviewer4:

Specific comments:

1. Establishing anoxic conditions using N_2. Flushing with molecular nitrogen (N_2) to create anoxic conditions is an effective and widely used method. However, it remains unclear the extent to which results obtained under complete oxygen removal can be extrapolated to the field. In many soils—especially peat, but also mineral soils—fully anoxic conditions may be difficult to maintain all year long. Even in wetlands, suppressed aerobic microbial communities may persist in soil pore spaces due to oxygenated air microsites, as well as oxygen-rich precipitation arrive throughout the year. In addition, seasonal thaw and soil moisture fluctuations (including those associated with permafrost degradation) can intermittently reintroduce oxygen into the most biologically active soil layers. Therefore, experimental designs based on complete O_2 displacement may underestimate carbon decomposition rate under anoxic conditions relative to aerobic conditions, particularly over long timescales.

Authors:

We thank the reviewer for this comment. We are aware that under natural conditions, oxygen concentrations can fluctuate strongly over the year, particularly in highly dynamic tundra soils affected by permafrost thaw. We used the two most extreme conditions that are relatively easy to keep constant in the laboratory during long-term incubation experiments: either atmospheric oxygen concentrations or the complete absence of oxygen. To discuss the impact of fluctuating oxygen concentrations in the soil, which are difficult to simulate in such long laboratory experiments, on carbon decomposition in the revised manuscript (lines 433-440).

Reviewer4:

2. Replication in the model. The number of replicates used for parameter estimation is not clearly stated, and it is also unclear how sensitive the results are to the exclusion of 10 replicates in the anoxic treatment (line 174).

Authors:

Thank you for pointing this out. We clearly state the number of replicates used for determination of the mean residence times by a two-pool model in the revised version of the manuscript (lines 182-184). For the oxic incubations it was n=34 for the two litter-C pools and n=31 for the permafrost-C pools. For the anoxic incubations it was n=28 for the litter-C pools and n=22 for the permafrost-C pools. As mentioned above, we removed the random forest analysis of the MRT data from the anoxic incubations from the revised manuscript since, at least in case of permafrost-C, it only explains a fraction of the variability in the mean residence times. Furthermore, we acknowledge that the model results from the anoxic incubations are less robust than the results from the oxic incubations (lines 444-445, see also response to reviewer2).

Reviewer4:

3. Variability in Fig. 2. Figure 2 shows substantial scatters across all four graphics. A clearer explanation of the sources of this variability is needed, or the interpretation should be framed more cautiously.

Authors:

We assume that the variability in the size of the different pools and their mean residence times is mainly caused by the wide variation in the properties of the soil samples we incubated, reflecting a wide range of environmental conditions when these permafrost soils have been formed during the Holocene. For instance, the TOC content ranged from 0.6% to 12.4%, the nitrogen content from 0.04% to 0.78%, the C/N ratio from 13 to 26, the pH from 4.0 to 7.2, and the clay content from 2.9% to 21%. We intentionally selected samples with different properties to identify the effect of these parameters on carbon stabilisation. We added discussion on the potential reasons of the substantial variation in the data presented in Fig. 2 (lines 339-345, see also response to reviewer2).

Further changes in the revised manuscript:

- We changed the designation of the two pools of the two-pool carbon decomposition model from ‘labile’ and ‘stable’ to ‘fast’ and ‘slow’ since the two pools are characterized by a higher and a lower decomposition rate constant.
- We write the two carbon pools ‘*permafrost-C*’ and ‘*litter-C*’ in italics to emphasize that we use these designations as names. In the strict sense, the carbon in the *permafrost-C* pool does not belong to permafrost anymore since the samples were incubated at 4 °C.
- The order of the co-authors was changed, with the names now listed in alphabetical order.