Referee 1 expressed a positive evaluation for ms acceptance and a few typos/suggestions that we have implemented. Regarding Referee 2 we provide below the following comments.

Total vs labile concentrations. IDMM is a model mainly devoted to labile metal simulations. In fact, total concentrations add up the contributions of labile and (prevalent) non-labile metal concentrations. From a mechanistic point of view, the dynamics of labile and non-labile metal concentrations are very different; IDMM not only treats the two pools in different ways, but relevant parameters are constrained with different dataset and experiments. The labile pool processes are mainly short-term equilibria (and so are the relevant datasets), while the non-labile processes are long-term (irreversible?) processes. We provided evidence of such processes in Figure 2. As the understanding of the two kinds of processes have different levels of confidence, models that use them will reflect the same knowledge gap. This is the reason why we presented the total metal concentrations in the Supporting Information. For example, for Pb we used fresh kinetic constants from a long-term lab experiment: NIL and FYM observations were well estimated while SS and COM were overestimated. Zn was well reproduced, except for underestimation in SS. On the other hand, Cd was overestimated in all the treatments. Surely, Cu is a different story: the total concentrations showed a strong reduction over time that we could not explain. Yet, we decided to keep Cu by offering the motivation that labile metal concentrations have reasonable (and well reproduced) trends. Therefore:

1) Models, like knowledge, need time and trials to accumulate confidence and robustness, this is the basics of research. What we presented is a quite challenging simulation of labile concentrations (which gave promising results) and less satisfactory simulations of total metal concentrations, thought total metal simulations were better in some cases and worse in others, but, except for Cu, the trends were overall matched. What we present here could be the basis for future studies and developments.
2) The argument that if total metal concentrations are not perfectly reproduced, then the labile metal simulations should be dismissed has no process-driven foundation. The mechanisms behind labile and non-labile metal dynamics are related but different.
3) We did not enter the question of whether total or labile concentrations should be used for risk assessment, nor should the Referee do so. Total and labile pools have different meanings; we concentrated on the labile pool as IDMM is fundamentally a labile-pool dynamic model.

ZOFE long-term trial. The argument provided by the Referee is that the dataset from ZOFE long-term trial should be dismissed as it conflicts with the bulk of the literature. First, we don't think so as the metal total concentrations generally increased or reached a plateau in the long-term. This was not true for copper (and we provided extensive comments that we could not justify such a loss) and in the sewage sludge treatment, for which we attributed such a behaviour to soil lateral mixing. In fact, we think that it is hard, if not impossible, to reply to the opinion that a dataset should be dismissed if it does not conform to certain expectations. Unless qualitative and quantitative justifications are provided, data should not be dismissed on the basis of unjustified feelings.

Regarding soil lateral mixing, this effect was already reported at Rothamsted long-term experiment, and there the datasets have been used for publication. While long-term experiments have unique long-term records for metal dynamics, they might come with the disadvantage of having small plots. While the soil lateral mixing effect is not significant for soil organic matter simulations, it is for trace metals. The proposal of the referee is to carry out the simulations for another long-term experiment, which should serve for validation. We argument that this is a totally different work and that there is no guarantee that the other work will have unambiguous data to validate soil lateral mixing!

FTIR and XRD. These two techniques were primarily used to exclude the contribution of OM/mineral composition variation to the decreasing trends observed in the sewage sludge amended plots. Therefore, the Referee is wrong in saying that FTIR was used to quantify SOC loss (for which data are available and used in the simulations) or make a comparison between treatments. The Referee asks whether we knew that sewage sludge, and relevant minerals, were added in substantial quantity to modify the mineral phase composition. The fact is that we didn't know, but now that we have the XRD dataset we can exclude it. On the other hand, the FTIR revealed a change in the SOM composition. How relevant it was for metal availability is difficult to quantify, but we notice that Pb availability decreased over time in contrast with the other metals. This might have an effect in TEs modelling, because the lability of the incoming metals could be tuned rather than being kept constant (this was already noticed by Bergkvist and Jarvis (2004). We suggest that, if these information are not provided, someone might ask whether the amendment itself contributed to the observed trends.

Finally, as an answer to the Editor's point, we have added a table with the limits of quantification and we have shortened and clarified the part on FTIR/XRD analysis.