**Referee #2**

R2: The manuscript presented here deals with the model prediction/description of measured data that reflect the Zn, Cu, Cd and Pb dynamics in a long-term field trial amended with different organic amendments. Then, after evaluation, the used model is extrapolated to the future, to evaluate the possible risks of TEs by long-term application of organic amendments on agricultural fields. The advantage of this model is, according to the authors, that it has a restricted amount of input parameters. However, there are several issues with the manuscript in its current state.

We thank the referee for acknowledging the interest in the ms.

R2: I have however a range of comments that to my point of view should be addressed before addressing this MS for publication in SOIL.

Major comments My first major comment is related to the general approach followed and to what is to my point of view the main conclusion of the paper. The author concluded (lines 509-511) that "the IDDM-ag model provided an adequate description of the measured EDTA-extractable concentration trends: : :". Looking at the figure 5, this is not so obvious. There are indeed several situations for which there is a clear discrepancy between experimental data and modelling (e.g. Zn-SS, Pb-SS, Cd-COM) and also other situations where the Swiss (e.g. Zn-FYM, Pb-FYM) or the ZOFE (e.g. Cu-SS) was alternatively the hypothesis which enables the model to have the best fit. In addition, these fits are based on the modelling approach considering lateral mixing. If the principle of lateral mixing is explained and if the uncertainty related to such a computation is discussed in the MS, there is not any validation of such a computation based on experimental data. In addition to that, there are a lot of uncertainty on some major flux of trace elements for a significant part of the field experiment history, particularly on trace elements added by organic residues. To consider whether the fits obtained were adequate or not, uncertainty in model parameters and input data should be considered and compared to the uncertainty in experimental data.

We acknowledge the fact that there is great uncertainty in the input data, as well as on the mechanisms governing TEs accumulation in the soil; decoupling and/or quantifying these uncertainties is out of the scope of the ms. Yet, the approach followed to quantify the inputs was able to reproduce the observed trends with lateral mixing and under the "Idealized Trend" assumption. Said this, we recognize that stating that simulations were "adequate" would require additional considerations, therefore we won't use this term in the ms.

For what concerns lateral mixing, we report the answer to Referee#1.

Long-term experiments comprising different treatments are valuable sources of information, but here we showed a potentially unrecognized drawback: plots can be affected by soil mixing with ploughing (plots have a separation space, but eventually it is too narrow for mechanical ploughing). This effect is hardly detectable unless elements present in trace concentrations are taken into account. Therefore, lateral mixing was introduced to check whether we could fit the data with realistic mixing coefficients. The 4 treatments considered were perfectly adjacent in 3 out of the 5 repetitions (in the other 2 repetitions the compost treatment only was separated from the "block"), so we can argue that the lateral mixing is well represented by the considered treatments and collected data. Clearly, the specific conditions in the long-term experiments cannot be extrapolated to real field conditions, and this is why we did not include the lateral mixing when projecting into the future the TE accumulation. Running another sampling campaign across the whole transect (12 treatments) is not feasible at the moment and, moreover, since some data are missing, it is not clear the real benefits we will get out of such an activity.

Moreover, considering the question about the adequacy of the fits of EDTA-extractable concentration, it should be to my point of view necessary to show as the first step the fits obtained for total trace element concentration trends in soil. The idea is that if the total trace element trends are not adequately simulated, how the EDTA-extractable trend could be? The simulation of total trace element concentration trends seem to me even more necessary as the accumulation trend expected is not visible for most trace element and organic fertilisation modalities. In particular, total Cu concentration shows a strong decreasing trend that was attributed to the past application of Cu fungicide, then followed by a sharp removal of Cu from the top-soil layer. No convincing explanation is given for this as the authors said that they do not have information about Cu-fungicide applications and assumed that soil ploughing and bioturbation explained the dilution of Cu concentration in soil without any simulation to support these strong assumptions. Without any other explanation, it seems that the Cu dataset is strongly biased and should, to my point of view, be removed from the MS.

The IDMM is mainly designed to simulate labile metal concentrations, but we agree that it would be informative to show the total metal concentration simulations as well. Therefore, we will include them, but we propose not to disregard the Cu data. Indeed, we do recognize that the hypothesis of fungicide application and bioturbation/soil removal is not conclusive, but we should also consider that: i) the Cu labile concentration trends are plausible and a valuable source of information; ii) we should give evidence in the literature of those data which could not be explained satisfactorily, for future research; iii) there is no obvious reason why Cu total concentrations should be biased, while the total concentrations of the other metals should not be so. Therefore, we propose to show the total concentration simulations for Zn, Cu, Cd, Pb, leaving open the question of the large decrease of Cu, which the model will not simulate.

R2: My second major comment is related to the second major conclusion of the paper suggesting that Cu and Zn contamination in soil can be harmful to soil organisms. To my understanding, this conclusion is based on the methodology described lines 309-313. It is however really unclear how the related computation of critical limits was effectively achieved. I looked at the cited paper of Lofts et al. 2005, from which I supposed that the free ion approach was based on EDTA-extractable concentration, pH and SOM data. If I am right, I notably wonder how the natural background concentration of trace elements in the soil was considered as regard to the fact that this specific issue is addressed by Lofts et al. (2005). Also, this methodology was tested on two dataset from UK and North America. It is thus not obvious that the methodology is relevant for the specific case and consequently the specific application of the IDDM-ag model studied here.

We propose to use a slightly different but more standardized approach to estimate trace element critical limits. This time we will apply: *Lofts et al. (2004). Deriving Soil Critical Limits for Cu, Zn, Cd, and Pb: A Method Based on Free Ion Concentrations. Environ. Sci. Technol. 2004, 38, 3623-3631.* Furthermore, in the future projections we will show the variation of the critical limits over time together with pH and SOM changes. We agree with the Referee that the background concentrations should be subtracted for the calculation of the critical limits; therefore, we propose to subtract the trace element concentrations estimated by the model before 1949 (the start of the experiment).

R2: Additional comments Lines 67-68 and 77-82: Sewage sludge is introduced differently from other organic residues (FYM and COM), notably because of the higher trace element concentration found in SS compared to FYM and COM. However, this is because the FYM and COM had relatively low trace element concentrations. For instance, I suppose that FYM is a cow manure. If a pig or poultry manure had been chosen, the concentration of several trace elements (Cu and Zn more particularly) would have been much higher and likely comparable to the concentrations observed in SS.

The organic amendments are applied at different frequencies and quantities in order to introduce the same (estimated) amount of organic matter, so that the comparison is done on the same basis. Since the TEs concentrations had not been assessed before this work, they did not contribute to the decision of applying different rates of organic amendment. Said this, the Referee is right that the FYM is from cow and the comment is valid that FYM from different sources could have much higher TEs concentrations. We will rephrase the sentence.

Line 199: To what refer the metal input? To the pool of total or available trace element?

In the model they are added to the labile pool.

Lines 199-211: It is not clear what is considered behind “geogenic input”. It is also very surprising to use data on the weathering rate of deppe layers of peat bogs to estimate the weathering rate in the present field experiment where the soil is clearly not a peat bog.

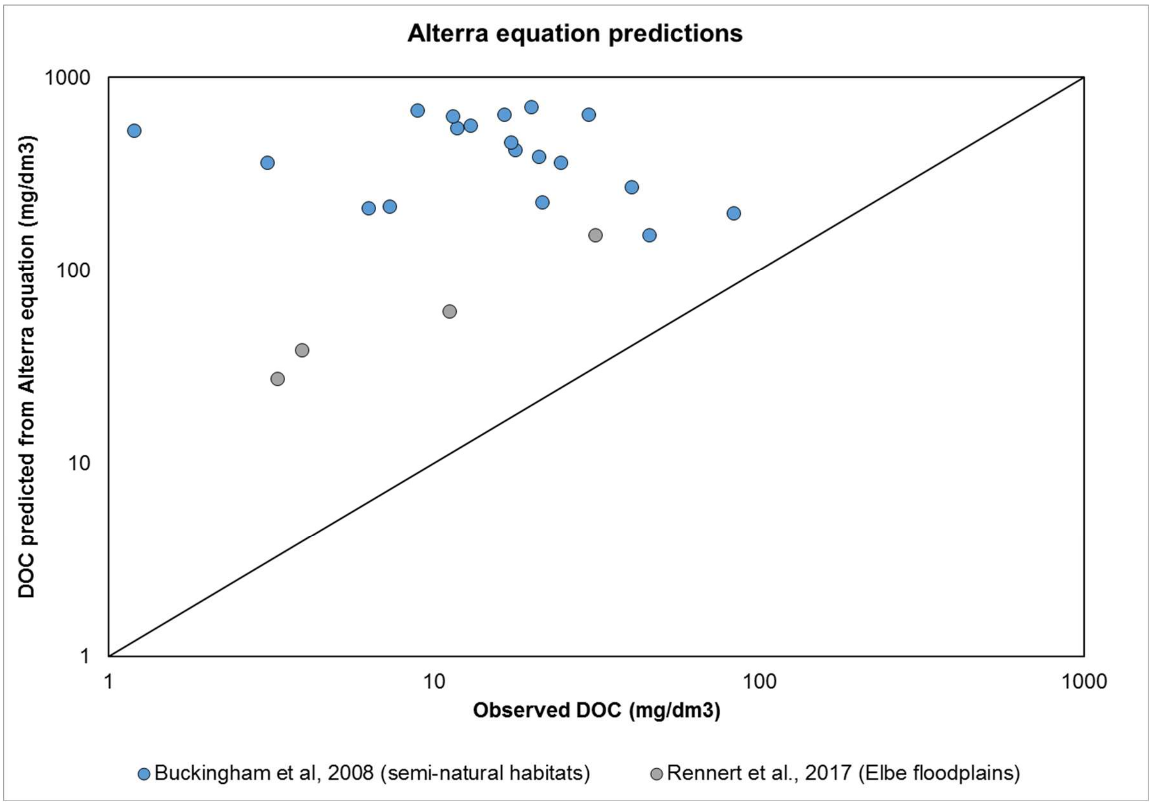
By geogenic deposition we mean the natural concentration of TEs in the atmosphere, i.e. due to eruptions from volcanoes, which give rise to TEs deposition well before anthropogenic activities became prevalent. In fact, this deposition would be detectable even at “pristine conditions”. This is why we used data from deep peat bog layers to estimate this natural deposition of TEs (please note that we did not use peat bog data to estimate mineral weathering!).

Doc concentration was fixed at 7 mg C/L. The authors further argued that Doc variation between 7 and 12 mgC/L does not impact the leaching of TE. However, considering the large variation in SOM and pH in the different fertilization modalities, I am surprised that a larger range of Doc concentration was not expected. Several authors (e.g. Araujo et al. 2019, https://doi.org/10.1016/j.envpol.2018.12.070; Cambier et al. 2014, https://doi.org/10.1016/j.scitotenv.2014.06.105; Laurent et al. (2020, https://doi.org/10.1016/j.scitotenv.2019.135927) showed drastic change in Doc concentration in soil amended with organic residues. A way to estimate the initial Doc concentration and its likely evolution over time could have been to use the empirical multi-linear regression suggested by Romkens et al. 2004 (Derivation of Partition Relationships to Calculate Cd, Cu, Ni, Pb, Zn Solubility and Activity in Soil Solutions; Alterra: Wageningen, 2004; p 75).

Please, see the answer to Referee 1 reported below.

The Referee raises an important point on the role of the porewater DOC concentration, and we acknowledge that we have not justified the assumptions made regarding its concentrations. Therefore, we propose to introduce the following considerations into the ms.

Data on porewater DOC concentrations from arable soils are scarce and contradictory. To the best of our knowledge, there are no consistent data from long-term experiments. There are a small number of meta-analyses (for example Li et al., 2019; de Troyer et al., 2014), but they are not ideal, because: 1) they do not contain data on soils amended according to all the management approaches taken in ZOFE, so some residual assumptions on DOC would be required; 2) we could not find any data on long-term time trends of field DOC concentrations under arable soils; 3) DOC concentrations obtained from laboratory soil extractions differ from data collected directly from the field using lysimeters, with the latter usually showing lower concentrations; if this is the case, most of the available DOC data are likely to be overestimates of 'true' field concentrations and thus bias the model results. We compared the predictions of the equation suggested by Referee #2 to estimate DOC from pH and SOM (Derivation of Partition Relationships to Calculate Cd, Cu, Ni, Pb, Zn Solubility and Activity in Soil Solutions; Alterra: Wageningen, 2004; p 75) with data from two studies that measured DOC sampled in field using lysimeters:




There is a consistent trend to overestimation of the observed DOC concentration (up to at least two orders of magnitude) and no relationship between observed and predicted DOC concentrations. Therefore, we strongly conclude that the Alterra equation should at best be applied with great care. Application of the equation to the ZOFE plots produced predicted DOC concentrations in the range 60-80 mg/dm3, which we contend, on the basis of the chart above, is highly likely to overestimate DOC and thus not be useful for our purposes.

In conclusion, we think that sources of reliable DOC concentrations for agricultural fields are effectively missing, and so should be identified as key research priorities. Therefore, we suggest that (i) this knowledge gap should be openly confronted and emphasised in the ms; (ii) given this lack of knowledge, the most pragmatic approach is to carry out a systematic sensitivity analysis within a plausible range of DOC concentrations, incorporating also time trends. We will do a mini-review of field measurements of annual DOC fluxes for temperate grassland sites (before 1949) and arable with/without organic amendments (after 1949) and we will perform simulations with the minimum, maximum and midpoint of our established range of DOC fluxes. To incorporate the time trends in the sensitivity analysis, we propose to apply to each plot a simplified SOC model (for example, a two-pool model like the one in Menichetti et al. (2016), which was previously applied to ZOFE), assuming that DOC correlates with the decomposition fluxes out of these pools. Clearly, this estimation of DOC is affected by the assumptions of C input decomposability when shifting the site management from grassland to arable in 1949; in addition, this approach does not include the effect of soil acidification, which is reported to control the adsorption/desorption of DOC. However, it could be considered a first estimation of DOC time trend.

We have already made some preliminary checks and the main conclusions are the following: increasing the DOC concentration in the control treatment has the effect of increasing the additional flux of metals that we hypothesized to be mineral weathering from the coarse fraction – this is to expected as the assumed DOC and weathering fluxes are not independent since the DOC controls the predicted soil labile concentration at steady state. Also, the impact of increasing five-fold the DOC concentration in the control treatment has a modest effect on the simulated labile concentrations (always <10%). On the other hand, increasing the DOC concentrations in the organic-amended treatments relative to the control treatment has the effect of lowering the metal concentration time-trends in these treatments, thus improving the overall simulation; yet, the effect is also modest.

Finally, the Referee suggests that DOC might peak after amendment application to slowly decrease until next application; this "cycle trend" is likely to happen, however since IDMM-ag has an annual time step, such short term effects are not modelled and an annual average DOC concentration, corresponding to the annual DOC flux, is what the model requires.

References:

Li et al. (2019). Effect of land management practices on the concentration of dissolved organic matter in soil: A meta-analysis". Geoderma 344 (2019) 74–81

De Troyer et al (2014). Factors Controlling the Dissolved Organic Matter Concentration in Pore Waters of Agricultural Soils. Vadose Zone J.

Menichetti et al. (2016.). Parametrization consequences of constraining soil organic matter models by total carbon and radiocarbon using long-term field data. Biogeosciences, 13, 3003-3019.

Line 303: The choice to fix pH and SOM in soil at the value found in 2014 for predictive modelling is really disputable, when considering how these two parameters are strongly impacted by the long-term applications of organic residues and particularly in the context the field experiment studied as showed in figure S4. This point should at least be discussed.

The Referee is right that even after >65 yr of soil management, none of the plots here considered have reached a new equilibrium condition: pH and SOC are still decreasing, sometimes at lower rates than initially. Therefore, keeping fixed pH and SOC can be a crude assumption. We propose to apply a SOC model to predict the future SOC (and DOC) changes over time (see answer just above). For pH, since its value will depend upon a number of factors, not least the speciation and cycling of added N, we propose to use simple extrapolation of the observed trends.

Lines 324-338 and 348-364: These two paragraphs are really too descriptive and speculative. As related to my first main comment, the simulation of total trace element concentration trends seem a prerequisite to assess the adequacy of the model used. But, as far trace element availability in soil is concerned, some (usually found in the literature, but nevertheless strong) hypotheses on the soil parameters explaining the change over time in the EDTA-extractable concentration of trace elements. These hypotheses should be checked, for instance by looking for multi-linear regression between trace element EDTA-extractable concentration or lability and some important parameters such as the input of trace metal in soils, total trace metal concentration in soil, SOM and pH in soil.

We will delete the description of the measured data and give space to the additional work that these comments have raised (i.e. DOC sensitivity analysis). Though it would be interesting to apply a multi-linear regression to explain the measured data, we feel it would be out of the scope of this ms, which is focused on model application and future predictions of TEs bioavailable concentrations.

Lines 425-445: Basically, the consideration of lateral mixing is interesting. However, the comparison of simulations with and without lateral mixing should be showed clearly (at least in supporting information) to support the conclusion that accounting for lateral mixing is important.

Yes, thanks. We will show the comparison with/without lateral mixing in the SI.

Section 3.4: It is really unclear to me what is the added value of the FTIR and XRD datasets. To my point of view, these datasets should be removed.

The spectroscopic analysis is useful for two important perspectives: i) confirming the general reduction of organic matter in soil for long term treatments in all samples; ii) confirming the importance to know the nature of organic material in terms of high affinity for TE's and its possible consequence in affecting the model. The X-Ray Diffraction analysis on soil samples give the possibility to establish that the organic treatments have not introduced exogenous mineral material, especially, in the case of the sewage sludge application. Differently, there are no other data in the manuscript that can establish this statement. These statements are functional to the discussion and give the possibility to have further insights on the TE dynamics in soils fertilized with organic amendments.

We propose to keep these sections, eventually shortening them.