Dear reviewer 2,

We would like to thank you for your comments and suggestions on our original manuscript. We have made edits in our original manuscript based on these comments and suggestions, and we hope that our corrections and replies are satisfactory.

I believe the manuscript could benefit from a different statistical analysis and not just considering the average across both sites and all harvest.

I agree that an argument could be made for the approach suggested by the reviewer, but there is an equally strong argument that in comparing the fertilisers that all events should be considered together as this is more realistic of the impacts at a wider scale. Before our experiment began, we decided that the results from the different fields would be combined, and to not do so would be to alter the experimental assumptions based on the results. It is true that there is a large difference in yields observed between the fields, but these results are documented in detail in the text and tables provided. To separate the fields in the statistical analysis would only reduce the replicate size and provide two very different and not very useful values for the fertiliser types. A difficulty in this area of research is that in order to make a very good comparison between the fertilisers, we would need several years of repeated experiments and many replicates more than we could financially support to come to solid conclusions. We present our data in a way that future studies can extract the data and collate it with other datasets to increase its long term value.

The introduction focuses largely on the general, big problems of fertilizer application and the part, which should guide the reader, actually comes too short: What is expected from the different fertilizer types? What are the underlying mechanism? I understand that the urea fertilizer coated with urease inhibitor is meant to reduce NH3 emissions but I do not understand "the pollution swapping", meaning that it should increase net N2O fluxes (L86-88 and L373-375).

Text added to introduction:

This is most commonly observed for nitrification inhibitors in which the slowing on the conversion of NH_4^+ to NO_3^- in soils results in a decrease in N_2O at the expense of an increase in NH_3 volatilisation (Lam et al., 2016; Zaman et al., 2009). In theory, the use of a urease inhibitor should reduce both the emission of NH_3 by reducing the rate at which urea is converted to NH_4^+ in soils, thus limiting available nitrogen in all forms. This may however, limit the rate at which crops also receive Nr and reduce yields.

The other crop quality measures, crude protein, ME, MAD and D value are barely mentioned. These variables also need to be explained: What do they mean in this context? Why did you choose them? I also suggest to explain the expected effect of fertilizer type on crop quality measures in the introduction.

These are the common metrics by which agronomists will measure the quality of the grass for animal feed. These were measured to identify if there were any large differences as a result of the fertiliser, which there was not. We add the following text to clarify:

These indicators of digestibility and energy content are commonly used to indicate the quality of the silage grass for animal feed and our study suggests that there was no significant differences between the feedstock grown using the different fertilisers.

NH3 emissions were only monitored over two weeks after each fertilization event, which I would assume could match the period when the urease inhibitor is effective. But what happens when the inhibitor becomes ineffective? Basically my question is: how long is such inhibitor effective in the

soil and how does this compare to the duration of NH3 emission observation? How would the NH3 emission look like if it would have been monitored over a longer period than two weeks?

This is a question asked by reviewer 1 as well (see reply to reviewer 1). NH_3 measurements are extremely difficult to capture from plot experiments due to the "sticky" properties of the gas. The application of the FIDES method is a novel way in which NH_3 emissions can be measured from plot scale experiments, of which this study is one of the first to report. Low emissions of NH_3 cannot be detected by any method in plot scale experiments; however, from previous experience (and the cited sources added to the methodology section), we expect more than 95 % of NH_3 emissions to occur within the first week, which we see in our study as cumulative flux stabilises over the 2nd week (or becomes negative due to low rates of deposition). If the inhibitor and the urea remained in the soil we would have seen an effect after subsequent treatments to the same plot, which we do not, even after 3 events. Based on our results, and previous studies, we do not expect measurements of NH_3 after the 2 week window to be large enough to be significant, therefore we did not expend resources to measure them.

Fig. 3: Different order of fertilizers compared to the other tables and figures is confusing.

The order is the same, with the addition of the control plots.

Fig. 4: Why is the median and not the mean plotted? Are there no error bars plotted or are they smaller than the symbols? One data point for ammonium in the urea treatment for the second fertilization event in 2016 looks like an outlier to me. If not then consider to use a break in the y-axis because the other data points are not readable due to the scaling.

This figure was also criticised by reviewer 1, and I offer the same reply here. Due to the log-normal distribution of the data and the limited number of measurements, it is exceptionally difficult to present the data (or meaningful error bars) in a clear manner in any format. When handling data that varies on a log-normal (exponential) scale, it is not statistically defensible to remove an "outlier" as it is a real value, nor do we want to "break the axis to present it. We have tried to present this data in a variety of ways, and believe that changing its current format would cause more problems than it solves. Although criticised, the figure is transparent and able to display much of the measured data in a way that is understandable to readers. In this form we can show trends in available nitrogen in the soil without specifying quantitative values for which there is no model to provide us with a way to handle the data or uncertainties.

I suggest to use repeated measures ANOVA to analyse the effect of fertilizer type on extractable ammonium and nitrate concentrations.

I disagree that this analysis is required in the text. Due to the log-normal distribution of data and small sample size, the assumptions of an ANOVA test are not met or easily applied in a meaningful way.

L25: Delete "and" after ammonia (NH3).

Corrected

L32: Delete "the" in "the urea".

Changed text to:

Overall, ammonium nitrate treatment was found to increase yields significantly (p-value < 0.05) when compared to the urea fertilisers used in this study

L33: Delete "The" in "The urea coated"

Corrected

L180: Comma missing before dC/dt.

Corrected

L238: Delete one "over" and consider re-writing this sentences.

Changed

L273: How can you conclude here that meteorological conditions were affecting yields differently in both years?

Reworded based on previous comment:

Although rainfall and temperature was similar during both years of measurement, crop yields for all treatments were substantially larger in the 2016 field plots (5.5 t ha^{-1}) than the 2017 field plots (1.48 t ha^{-1}) (Table 2).

L284: Data on ME, MAD and D value are not shown in Table 2 or anywhere else.

These were measured as part of a different work package within the project and although not really relevant to this study, we include reference to a lack of statistical significance in the differences. We have removed an earlier incorrect reference to table 2, but the description remains in the results section with some added description of its importance as a metric for quality of animal feed as it does not detract from the study either.

L327: How long were sheep grazing at the Upper Joiner Field in 2017? Were they also excluded from the plots one month before start of the experiment?

Text has been added to explain in better detail based on previous comment:

While sheep were vacated from the 2016 field a month prior to the experiment, the 2017 plots had not been grazed for more than six months before the experiment.

L334: On a log-normal scale? I do not understand this in this context.

Reworded based on previous comment to describe log-normal distribution of data, a common occurrence in scientific data handling.

L338 & 341: Nitrate does not decay. Describe it as a decrease in concentration over time.

Corrected

L355-357: That's speculation. Your experiment does not allow to conclude this.

Text changed to:

We speculate that the prior grazing of the sheep is also likely to have resulted in the residues of animal waste in the 2016 plots, which would explain the higher than expected yields and Nr in the soil measurements in these plots (Cowan et al., 2015).

L367-370 & L378-379: Here again: Description of statistical analysis is missing and no results of the statistical analysis are given.

This is true, but also of little relevance to the overall study. T-tests and p values will not add anything of value to this discussion regarding these variables.

L393-394: Re-write this sentence: ". . .fallen considerably in magnitude come harvest"?

Typically, Nr in the form of NH_4^+ and NO_3^- in the top 10 cm of soil has returned to concentrations on par with the control plots by harvest.