General comments

This manuscript describes a risk assessment exercise of ozone effects on wheat production at the global scale based on modelled phytotoxic ozone dose (POD6). POD6 values are calculated following the CLRTAP methodology, using as input data the ozone concentration and other meteorological variables produced by two Earth system models, run under different shared socioeconomic pathways with contrasting radiative forcing and atmospheric pollution control policies over the XXI century. Changes in POD6 under the different scenarios are described for different regions of the world. The influence of meteorological conditions and ozone concentration on the final POD6 estimate and relative wheat yield loss are also analysed. This kind of analysis is useful to understand how climatic changes may affect the negative effects of tropospheric ozone pollution on agricultural yield in the future rather than focusing solely on changes in ozone concentration. The analysis provides also insights on the relative weight of air pollution policies and radiative forcing scenarios on the risk of ozone effects on agricultural yields by the end of the XXI century. The relative influence of these two factors changes for different regions of the world, highlighting the co-benefits of controlling both ozone precursor and greenhouse gases emissions to the atmosphere.

The manuscript is well structured and written and make a relevant contribution to the field of risk assessment effects on agricultural production under future climatic conditions. The objectives, datasets, methods and results are clearly presented and easy to follow, although I have some comments that may need minor revisions of the manuscript:

We thank the reviewer for their positive feedback and helpful comments which we address below.

Specific comments

Line 138 and Table 1. From the list of variables in Table 1, I am not able to find soil moisture. Thus, I understand that soil moisture was modelled based on GFDL and UKESM outputs. Please explain, or cite methods, on how soil moisture was modelled and how plant available water was computed from soil moisture data.

The water in the soil that is available to the plant was calculated as a water budget through a simple bucket model (following the approach by Mintz and Walker, 1993). We deemed this method appropriate, since we are simulating a wheat field, whereas the soil moisture output from ESMs takes into account many different land-use types within the same tile.

An additional explanation was added to the manuscript in Section 2.1.

Line 169. Crop geometry refers to the development of LAI and SAI over the course of the growing season, as a function of thermal time, as described in Guaita et al 2023? If that is the case, I would specify this somewhere in the manuscript, since the term "crop geometry" seems confusing to me. Was plant height changed as well over the course of the growing season?

Thanks for pointing that out. Yes, Guaita et al. (2023) is the reference also for the crop geometry. This includes LAI, SAI, root depth and plant height. The term "crop geometry" was removed from the manuscript and specification of the LAI, SAI, root depth and plant height were added in the sentence of Section 2.3.

Line 176. The Mediterranean wheat parameterization for soil moisture refers to volumetric water content while in this modelling exercise, plant available water is used. Please describe how was this modified in your calculations.

The reviewer is correct. For the mediterranean wheat, our model calculates SWC as the soil water available to the plant (i.e. the difference between the actual SWC and the SWC at wilting point) instead of PAW. This is effectively a difference compared to Guaita et al. (2023), for which only the continental parameterization of winter wheat was adopted. It is now specified at the end of the section 2.3, in the last paragraph about model updates.

For clarity, we replaced the wording "plant available water in the soil" with the more general "soil water available to the plant" across the whole manuscript.

Line 191. It seems that there is a typo in the reference Guaita et al 2023b, or the reference is missing in the bibliography.

Thanks. Exactly, that's a typo, and it has been corrected.

Line 215. I think this sentence needs some clarification. The prescribed sowing dates come from Qiao et al. (2023), while thermal times describing the phenological development of wheat were taken from González-Fernández et al., 2013 and Grünhage et al., 2012). Thus, I understand that if the thermal time at leaf senescence (according to González-Fernández et al., 2013 and Grünhage et al., 2012) was not reached before the following prescribed sowing date (according to Qiao et al 2023), then the node was excluded from POD6 calculation.

Yes, it is exactly as the reviewer said, thank you. The mentioned sentence of section 2.4 has been adjusted accordingly to make it clearer.

Line 221. Unfortunately, I am not familiar with this sort of spatial analysis to assess its use, although the concept and results obtained look reasonable to me. However, I miss some comments about the tests conducted to assess the assumptions of the ANOVA analysis.

The data in every node was tested for normality with either the Shapiro-Wilk or the Shapiro-Francia test, depending on the data being either platykurtic or leptokurtic. In case the test failed, the data in that node was log-transformed. A couple of sentences were added to the manuscript in the methodology section dealing with ANOVA (2.5) to clarify this process.

Line 239. How the accumulation period changed between the baseline and the 2100 scenarios? Was it earlier and shorter, or longer in different regions of the world?

The accumulation period shortened up to 5 days on average by the end of the century. At the same time, the accumulation period onset was noticeably earlier, up to 26 days earlier at the end of the century under SSP3-7.0 for UKESM1-0-LL.

We added some sentences on this matter at the end of Section 3.1 and, accordingly, we added more rows Table 3 displaying the duration of the accumulation period at the baseline and at the end of the century for both models. The differences in the number of days between the onset of the accumulation period at the end of the century and at the baseline were also added to the table.

Lines 244-247. This statement might be more appropriate for the discussion section.

The reviewer is correct and we appreciate the suggestion. The statement has been moved to the discussion section. Since the discussion section appeared a more appropriate part to elaborate further on the biases, we added a comment regarding the ESMs performances in simulating mean ozone concentrations.

Line 272. Desert regions or arid regions? In desert regions, it will be unlikely to find rainfed wheat crops?

Thanks, arid is more appropriate. Clearly, rainfed wheat crops cannot be find in desert regions, the reviewer is correct.

Figure 1. Please explain in caption the meaning of areas in white.

Thanks. The explanation has been added.

Table 4. Means presented in Table 4 are averaged between the two CMIP6 models? Please clarify this.

Yes, they are averaged between models. The specification has been added to the table.

Lines 366-369. The mean or median relative yield loss could be also a very helpful metric to describe the range of the expected risk to wheat production

It is true that it could be useful, and, when possible, the mean RYL have been indicated in the text across the whole section.

Line 379. How should the very low Pearson correlation coefficients 0.15 > r > -0.15 covering relatively large areas in the map on Figure 6 be interpreted? Is this also reflecting a lack of agreement between models? Higher uncertainty in the expected changes?

While high Pearson correlation coefficients indicate that the two models are strongly correlated (positively or negatively), very low values indicate that the data is (linearly) uncorrelated. However, this fact can be associated with one of three cases: (1) on average, there are small ΔPOD6 values (<0.65 mmol m^{-2}) in both models; (2) on average, there are large ΔPOD6 values in both models (but linearly uncorrelated among them); (3) there are large ΔPOD6 values in only one of the two models.

In the first case, correlation is small because there are small predicted POD6 changes. However, this does not indicate that there is no agreement between models, but rather that both models agree that the change is small.

In the second case, there is an interannual disagreement in the ΔPOD6 values between the two models, indicating that there might be disagreement on a year-by-year basis. In this sense, both models might agree on magnitude the expected value, although there is high uncertainty on individual years.

In the third case, there is an actual disagreement that is not highlighted by a negative correlation coefficient: in fact, one of the models predicts a large ΔPOD6 on average, while the other one expects values that are close to zero.

However, in our study, 68.6% of the nodes with a low Pearson correlation coefficient pertain to the first case, i.e. both models agree that the POD6 changes are small. 8.4% nodes fall within the second case, and 99.9% of the times the average POD6 changes have the same sign. The remaining 23.0% of the times the two models disagree, in the sense that one of the two expects small POD6 changes, while the other one expects large POD6 changes.

An additional explanation has been added at the end of section 3.3 to clarify this matter.

Lines 403-406. The description of POD6 increases under non-limiting soil moisture conditions is confusing to me. Most extensive increases in POD6 happens in Southern Europe and South Asia, but the average value of increase seems lower than for other regions like South-East Asia or North-Africa. The values reported are averaged also across models?

With the term “extensive”, we refer to the fraction of the region (with respect to the overall region surface) affected by POD6 increases under FC condition. In order to avoid confusion, we replace the word “extensive” with “widespread”, and specified the fraction of the region affected by POD6 increases. The values reported are averaged across models, and it has been specified so in the text.

Line 447. Does not

Corrected, thank you.

Line 455. The discussion should include some comments regarding the performance of the models chosen in this study to simulate ozone concentrations (as mentioned in line 245).

Thanks for the suggestion. The part relative to the model performances in simulating ozone concentrations has been expanded.

Also, uncertainties stemming from the use of particular models: in this study the general pattern matches, but one model predicts bigger changes in POD6 and higher risks of O3 effects on yield compared with the other one, and large areas show small Pearson’s correlation coefficients between modes, as shown in figure 6. I wonder if multi-model ensembles could be a useful tool for future projections.

It is a matter of fact that the UKESM1-0-LL predicts generally higher O3 concentrations with respect to GFDL-ESM4, and therefore the POD6 estimated with the first model is higher than the second one. This feature was made explicit in section 4 (discussion). Furthermore, following the second reviewer’s request, we performed a new model evaluation which can be found in the supplementary materials, and that will provide further context to interpret discrepancies across models.

The variability between the models is a source of uncertainty for the risk assessment. Hence, we agree with the reviewer that a multi-model ensemble would be welcome for this risk assessment. However, it should be noted that to date only two coupled-chemistry models provide in output hourly O3 concentrations up to 2100, and they are UKESM1-0-LL and GFDL-ESM4.

Regarding the disagreement indicated by the Pearson correlation coefficients between models, there are relatively small fractions of the globe (between 4.3% and 14.3%, depending on the scenario considered) where the two models disagree, and they can be identified either from negative Pearson correlation coefficients (Figure 6), or from small POD_6 changes predicted by one model, and large by the other. However, the position of these nodes did not reveal any specific spatial patterns, suggesting that at least a fraction of the disagreement between the models might be due more to randomness rather than by actual different ESMs features.

Therefore, despite the two models predict different absolute POD6 values, their spatial patterns well agree across the scenarios in the main risk areas.

These features were added to the discussion section.

Also, there is one variable that will change depending on SSP scenarios compared with the current situation that also affects stomatal conductance and likely PO6 but is not taken into account with this methodology, such as the CO2 concentration. This should be considered an additional uncertainty and limitation of this approach.

The reviewer is correct. The caveat has been added to the discussion section.

On the positive side, it would be interesting to comment on the advantages of assessing tropospheric ozone effects on agricultural yield under future climatic conditions using the POD approach as compared with other assessments based only on changes in ozone concentration. Finally, it could be stressed that the results presented here support the co-benefits of abating greenhouse gases and air pollution emissions jointly to help in the mitigation of air pollution effects in agriculture.

The reviewer is correct and we appreciate the suggestion. We believe that the advantage of using the POD approach resides in the ability to understand that the mitigation of ozone damage also comes as a co-benefit in controlling GHG emissions, at least in some regions of the globe. This has been added to the manuscript.

Line 479. Subscript missing in O3.

Fixed, thank you.