Reviewer #2 (Comments to the Author):

Title: On the dynamics of ozone depletion events at Villum Research Station in the High Arctic

This paper describes a study on the parameters impacting ODEs in the high Artic. It uses a comprehensive set of atmospheric parameters and sea ice conditions together with back trajectory analyses to investigate the sources of ODEs. Apart from a statistical analysis, a machine-learning algorithm is used to identify the most important parameters affecting ozone.

The paper is very well written and addresses important processes in a part of the Earth's atmosphere that is most vulnerable to climate change. It is therefore of high scientific relevance and fits well into the scope of ACP. The way the data analysis is performed and the results are discussed are appropriate given the complexity and multiphase nature of halogen release and ozone depletion events in Polar Regions. There are, however, a few points that should be addressed before final publication:

> We thank the reviewer for their comments and suggestions. We have addressed each comment below with review comments in black, author response in blue, and additions to the original text in red. We have indented the author's response for clarity. Line numbers given in the author's response refer to lines in the revised manuscript.

The authors do not pay appropriate credit to former studies on tropospheric ozone depletion which apply very similar methods as the present study. In particular, Frieß et al. [2023] presents statistical analysis of multi-decadal ozone (and BrO) observations based on back-trajectories and sea ice data, just as this present study, but for the Antarctic. The paper would benefit a lot from a discussion on the possible similarities and differences between the drivers of ODEs in both hemispheres via comparison of the results from both studies.

> We thank the reviewer for bringing this valuable reference to our attention. While Frieß et al. (2023) uses a similar methodology, they do not explicitly investigate the relationship between formally defined ozone depletion events and meteorological/air mass history variables but rather focus on vertically resolved BrO mixing ratios. Nevertheless, Frieß et al. (2023) is a valuable resource, one which the authors were not aware of, therefore, we have cited Frieß et al. (2023) on lines 84, 182-184, 192-194, 802-804, 837-839, 868, 894-895, and 932-933.

I am not an expert in machine learning and I have to admit that I was quite lost while reading Section 3.4. I could imagine that other experts in atmospheric physics and chemistry, but not in machine learning, would experience the same. I therefore feel that Sections 2.6 (ML methods) and 3.4 (ML results) require substantial revision as discussed in more detail in the specific comments below.

> We originally moved the two pages of text describing the ML methods to the SI to reduce the overall length of the article. We admit this was not the correct decision in hindsight. Therefore, we have moved the entire description of the ML methods to the main text.
>
> We think that the science in these sections is sound so what needs to be improved is the readability and clarity. We have thus clarified/simplified parts of the text to make it more understandable by readers inexperienced with ML
>
> We have added the following lines to make this clearer in the text:
>
> Lines 261-265: Cross validation involves splitting the training data in 10 equally sized folds (or groups), training the model using nine folds and testing the model using the remaining fold. This was repeated 10 times to use each fold as a test set once. The final evaluation metrics were

averaged using the arithmetic mean to select the optimal hyperparameters and make an overall evaluation of the model performance.

Lines 566-582: The evaluation metrics of the ML for all spring months combined and individual months are displayed in Table 1. We use three common metrics for evaluating a binary classification ML model: accuracy, recall, and AUC ROC (Area Under Curve Receiver Operating Characteristics). Briefly, accuracy is the fraction of correctly predicted observations regardless of label (ODE vs Non-ODE), recall is the fraction of ODEs correctly predicted and AUC ROC evaluates how well a model can discriminate between positive and negative labels across all decision thresholds for binary classification. In general, the ML model can accurately reproduce ODEs over all spring months combined as evidenced by how all three metrics are close to unity (their maximum value). However, when evaluating the results on an individual monthly basis, there is an increase in the recall metric and decrease in the accuracy and AUC ROC (see Sect. 2.6 for a detailed description of the evaluation metrics) from March to May (Table 1), which is likely connected to the increasing frequency of ODEs from March to May. With increased ODE occurrence, the recall metrics would increase as positive labels (ODEs) are more likely to be identified when they occur more often and the accuracy and AUC ROC metrics would decrease with the increased occurrence of positive labels due to a concurrent increase in number of incorrectly labeled ODEs. The ML model is also free from over-fitting given the close agreement between the train and test sets. Overall, this ML model is sufficiently accurate, robust, and suitable for the investigation of ODEs.

Caption of Table 1: The accuracy gives an overview of the model performance for both labels (ODEs vs Non-ODEs), recall gives the model performance only for positive labels (ODEs), and AUC ROC evaluates the model performance over different decision thresholds, together, these three metrics give a comprehensive view of the model's performance. The three metrics range from 0 (worst) to 1 (best).

Lines 594-605: The SHAP approach is designed to estimate the importance of each input variable to the model output based on coalitional game theory (Molnar, 2022) (see Sect. 2.6 for a more detailed description). SHAP values represent the marginal contribution of each input variable to the model output, or in other words: how important each variable is to the model for making a prediction. SHAP values can be positive or negative, with positive values indicating a variable is more likely to contribute to an observation being predicted as an ODE while negative values mean a variable is more likely to contribute to an observation being labeled as a Non-ODE. SHAP can produce both local and global explanations. The global importance gives an overview of the most important variables to the model output. The local importance of each observation can give information about the relationship between the SHAP and input values (positive or negative relationship, linear or non-linear), or in other words how does the model output vary over the range of input values.

The abstract is quite short. It would be important to provide some more specific information on this study (e.g., measurement site, observation period, etc.).

The Atmospheric Chemistry and Physics guidelines on manuscript formatting (https://www.atmospheric-chemistry-and-physics.net/submission.html#getready) specify a maximum abstract length of 250 words, which the abstract is currently at. The measurement site is indicated on lines 12 and 13 and in the title and we have added the years on line 13.

Specific Comments

L17: It is not clear what you mean with "increasing". Is this a seasonal tendency or an increase over the years?

> Our intent was to describe the results of our trend analysis for these two parameters. We have added the words. We have amended this sentence.
>
> Line 17-18: Positive trends in ODE frequency and duration are observed during May (low confidence) and April (high confidence), respectively.
>
> We have also changed the language throughout the manuscript to indicate the direction of trends as positive or negative where appropriate instead of increasing or decreasing.

L64: Please explain what you mean with "relative rate principle"

> Relative rate principle is a standard method for investigating atmospheric chemistry and is used widely in laboratory studies of unknown reaction rate constants by reacting with a compound with a known rate constant with another compound with unknown rate constant (Finlayson and Pitts, 1986). An expression is obtained independent of the mutual reactant concentration and a fitted line is obtained consisting of ratio between the known rate constant and unknown rate constant and thus the unknown rate constant can be calculated from the slope and known constant; therefore, the name "relative rate principle". We have earlier used this approach to demonstrate that Br is reacting with O_3 and gaseous elemental mercury (GEM) rather than Cl (Skov et al., 2004, 2020).

Section 2.4: It should be pointed out that the back-trajectory analysis applied here is very similar to the methods by Frieß et al. [2023].

> We have added a sentence which shows that this methodology has been applied to previous studies including Frieß et al. (2023).
>
> Lines 192-194: This methodology has been utilized by previous studies to systematically analyze the geographic origns of air masses (Dall'Osto et al., 2017, 2018; Frieß et al., 2023; Heslin-Rees et al., 2020; Pernov et al., 2022).

Sections 2.6: The description of the machine learning algorithm is far too short. I think the reader should be able to get at least a basic understanding of the model without going through the detailed description in the supplemental material. See also my comments to Section 3.4 below.

> We admit the Methods section describing the ML algorithm is short. This is due to the majority of the text being moved to the supplement in an effort to shorten the length of the paper. We have moved the entire section describing the ML methodology to the main text as outlined above and below.

L267: It is not clear to me what you mean with "monthly hours within the same bin".

> The word "monthly" should not have been included in this sentence and we have removed it. We thank the reviewer for this good catch.

Section 3.4: This section is hard to understand for readers inexperienced in machine learning. I do not have any clue what to learn from Table 1, except that high numbers are good. What is a "cross validation score"? What is "Aera Under Curve Receiver Operating Characteristics"? What does "Recall" mean? The following discussion is mainly based on SHAP values. The explanation of this parameter should therefore be moved from the Supplemental to Section 2.6.

We have now moved the entire ML methodology section to the main text to aid inexperienced readers.

The next sections of Sect 3.4 describe the efficacy of our ML model using common evaluation metrics for classification tasks, the most important variables to the model output, and how they affect the ML model's prediction of ODEs using the SHAP methodology. Detailed descriptions of the evaluation metrics and other ML terms are now included in the methods section. We show that our model is capable of reproducing ODEs and it gives physically meaningful results. The input variable importance is not revealed from the statistical analysis demonstrating the added value of ML. The relationships between ambient and SHAP values demonstrates how individual observations contributes to the model's prediction of ODEs, which largely agrees with the statistical analysis but reveals intricacies that are not borne out of the statistical analysis (e.g., threshold values for positive prediction of ODEs) as the SHAP methodology accounts for dependencies between variables in the model something the statistical analysis does not.

We have added the following lines to make this clearer in the text:

Lines 261-265: Cross validation involves splitting the training data in 10 equally sized folds (or groups), training the model using nine folds and testing the model using the remaining fold. This was repeated 10 times to use each fold as a test set once. The final evaluation metrics were averaged using the arithmetic mean to select the optimal hyperparameters and make an overall evaluation of the model performance.

Lines 566-582: The evaluation metrics of the ML for all spring months combined and individual months are displayed in Table 1. We use three common metrics for evaluating a binary classification ML model: accuracy, recall, and AUC ROC (Area Under Curve Receiver Operating Characteristics). Briefly, accuracy is the fraction of correctly predicted observations regardless of label (ODE vs Non-ODE), recall is the fraction of ODEs correctly predicted and AUC ROC evaluates how well a model can discriminate between positive and negative labels across all decision thresholds for binary classification. In general, the ML model can accurately reproduce ODEs over all spring months combined as evidenced by how all three metrics are close to unity (their maximum value). However, when evaluating the results on an individual monthly basis, there is an increase in the recall metric and decrease in the accuracy and AUC ROC (see Sect. 2.6 for a detailed description of the evaluation metrics) from March to May (Table 1), which is likely connected to the increasing frequency of ODEs from March to May. With increased ODE occurrence, the recall metrics would increase as positive labels (ODEs) are more likely to be identified when they occur more often and the accuracy and AUC ROC metrics would decrease with the increased occurrence of positive labels. The ML model is also free from over-fitting given the close agreement between the train and test sets. Overall, this ML model is sufficiently accurate, robust, and suitable for the investigation of ODEs.

Caption of Table 1: The accuracy gives an overview of the model performance for both labels (ODEs vs Non-ODEs), recall gives the model performance only for positive labels (ODEs), and AUC ROC evaluates the model performance over different decision thresholds, together, these three metrics give a comprehensive view of the model's performance. The three metrics range from 0 (worst) to 1 (best).

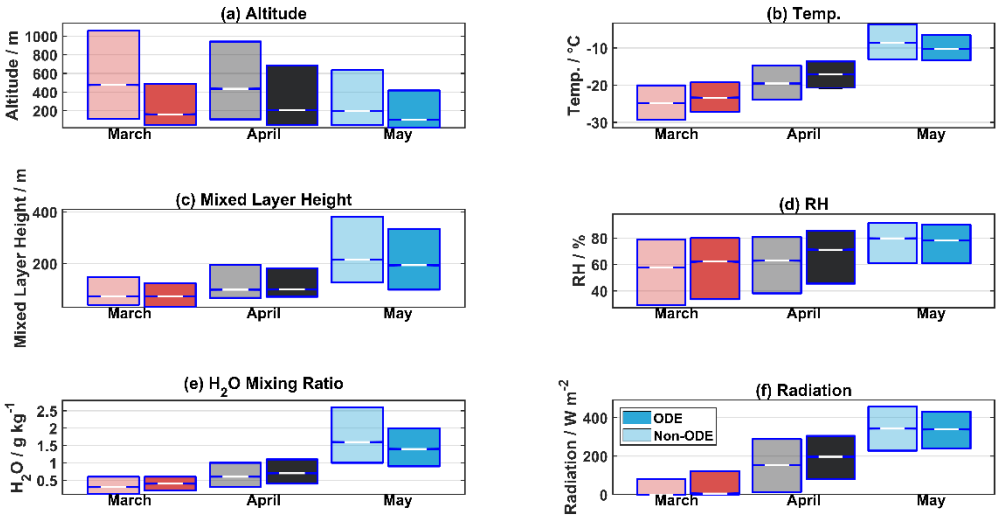
Lines 594-605: The SHAP approach is designed to estimate the importance of each input variable to the model output based on coalitional game theory (Molnar, 2022) (see Sect. 2.6 for a more detailed description). SHAP values represents the marginal contribution of each input variable to the model output, or in other words: how important each variable is to the model for making a prediction. SHAP values can be positive or negative, with positive values indicating

a variable is more likely to contribute to an observation being predicted as an ODE while negative values mean a variable is more likely to contribute to an observation being labeled as a Non-ODE. SHAP can produce both local and global explanations. The global importance gives an overview of the most important variables to the model output. The local importance of each observation can give information about the relationship between the SHAP and input values (positive or negative relationship, linear or non-linear), or in other words how does the model output vary over the range of input values.

L615ff: You state that in situ radiation measurements would not be available for the entire measurement period, and would also not be indicative for the radiation along the trajectory. Is there any reason why you do not use radiation along the trajectory, which is part of the Hysplit model output?

An analysis of the radiation along the trajectories would indeed be an interesting analysis method. The radiation from HYSPLIT (produced from the NCEP/NCAR reanalysis data) is for the Earth's surface and is not output for an air mass' vertical position (Kalnay et al., 1996), which could affect its interpretation. The in situ meteorological variables give interpretable results which are in line with the current theory of ODEs. We produced boxplots of the meteorological variables from HYSPLIT, similar to Figure 5 in the main text, for all trajectory steps. This shows that the HYSPLIT meteorological variables display similar distributions to the in situ variables and therefore would not add new results and therefore were not included in this manuscript. We also decided to not use the meteorological data output by HYSPLIT since this would necessitate an investigation of all meteorological variables from HYSPLIT. We felt this would substantially increase the length of the manuscript. We have added text in the "Summary and Outlook" section stating that this would be an avenue for further research.

Lines 1039-1040: Analyzing meteorological conditions along the trajectory path (e.g., temperature and radiation) would help extrapolate the observations from individual stations to the larger Arctic region.



L622ff: It would be worth mentioning here that an important process that promotes bromine release at lower temperatures is carbonate precipitation from the sea ice, which reduces its buffer capacity and facilitates acidification [Sander et al., 2006]

We have added a sentence mentioning this process and reference in this paragraph.

Lines 775-777: Cold temperatures also facilitate calcium carbonate precipitation from sea ice which acidifies and lowers the buffering capacity of the salty sea ice surface thus promoting halogen release (Sander et al., 2006).

L649: It is not true that a relationship between RH and ODEs has not been reported before - see Frieß et al. [2023].

Frieß et al. (2023) reports Pearson correlations coefficients between surface measurements of RH and O_3, however, they do not explicitly investigate the relationship between RH and *ODEs* and while they observe negative correlations they do not discuss these correlations between RH and O_3 in the text.

We have added "in the Arctic" in this sentence to indicate the author's intent this was exclusively for the Arctic. We have added a sentence after this one showing that the relationship between RH and ozone has been explored in Antarctica.

Lines 802-804: However, the relationship between RH and ozone has been explored in Antarctica by Frieß et al. (2023), who showed negative correlations at Neumayer and Arrival Heights, supporting observations made in this study.

L825: This is not a new finding. Replace "Our results show..." with "Our results confirm...".

We have removed the phrase "Our results show that" from Line 845.

Technical Comments

L40: "long range" -> "long-range"

We have made this change to the manuscript.

L120: What is the meaning of "i.d."?

"i.d." stands for "inner diameter". We have added this abbreviation to Line 124.

L148: Maybe the term "accept" would be more appropriate than "require" here.

We feel the current wording is more reflective of the input data requirements for machine learning algorithms and will keep it as is.

L152: Start a new sentence after "horizontal plane".

Starting a new sentence after "horizontal plane" would result in the next sentence being too short "This includes both direct and diffuse radiation". To address the reviewer's comment while avoiding this short sentence, we have rearranged the text to:

Lines 155-157: ERA5 output of "shortwave solar radiation downwards" was used, which is the amount of shortwave downwelling solar radiation (including both direct and diffuse radiation) that reaches the Earth's surface on a horizontal plane.

L179: Either state "below mixed layer HEIGHT" or "within the mixed layer".

We have added the word “height” to line 191.

L192: "A trend analysis of trends...": please rewrite.

We have removed the words “of trends” from this sentence.

L295: Remove "For temperatures" at the beginning of the sentence.

In each paragraph of this section, we present the results for each variable separately. We have adopted a coherent, standardized structure for starting paragraphs in this section, with the variable name at the beginning to allow the reader to easily grasp the description of each variable. We discuss each variable in the same order throughout the manuscript. This will facilitate readers to easily access the pertinent information for each variable. We have adopted the same structure in the Discussion section to remain consistent. For these reasons, we have opted to keep the original text on line 403 of the revised manuscript.

Figure 10: It seems that the y-axis scale refers to the lines (SHAP values), but the histograms have different units. So probably a second y-axis on the right needs to be added for the histograms.

To satisfy the reviewers comment, we have replaced Figure 10, S9, and S10 with the ones showing the relative frequency for the histograms on the right axis. The relative frequency is proportional to the histogram count and gives a more intuitive indication of the data distribution. We originally omitted the y-axis labels for the histograms for clarity as mentioned in the caption of Fig. 10.

L572: What do you mean with "SS"? Define acronym/abbreviation.

The acronym “SS” is defined as “statistically significant” on lines 216-217 in Sect. 2.5 “Trend Analysis” of the Methods.

Reviewer’s References

Bognar, K., Zhao, X., Strong, K., Chang, R. Y.-W., Frieß, U., Hayes, P. L., McClure-Begley, A., Morris, S., Tremblay, S., and Vicente-Luis, A.: Measurements of Tropospheric Bromine Monoxide Over Four Halogen Activation Seasons in the Canadian High Arctic, Journal of Geophysical Research: Atmospheres, 125, e2020JD033015, https://doi.org/https://doi.org/10.1029/2020JD033015, 2020.

Frieß, U., Kreher, K., Querel, R., Schmithüsen, H., Smale, D., Weller, R., and Platt, U.: Source mechanisms andtransport patterns of tropospheric bromine monoxide: findings from long-term multi-axis differential optical absorption spectroscopy measurements at two Antarctic stations, Atmospheric Chemistry and Physics, 23, 3207–3232, https://doi.org/10.5194/acp-23-3207-2023, 2023.

Sander, R., Burrows, J., and Kaleschke, L.: Carbonate precipitation in brine - a potential trigger for tropospheric ozone depletion events, Atmos. Chem. Phys., 6, 4653–4658, https://doi.org/10.5194/acp-6-4653-2006, 2006.

References from Author’s Reply

Dall'Osto, M., Beddows, D. C. S., Tunved, P., Krejci, R., Ström, J., Hansson, H. C., Yoon, Y. J., Park, K.-T., Becagli, S., Udisti, R., Onasch, T., O´Dowd, C. D., Simó, R., and Harrison, R. M.: Arctic sea ice melt leads to atmospheric new particle formation, Sci. Rep., 7, 3318, https://doi.org/10.1038/s41598-017-03328-1, 2017.

Dall'Osto, M., Geels, C., Beddows, D. C. S., Boertmann, D., Lange, R., Nojgaard, J. K., Harrison, R. M., Simo, R., Skov, H., and Massling, A.: Regions of open water and melting sea ice drive new particle formation in North East Greenland, Sci. Rep., 8, https://doi.org/10.1038/s41598-018-24426-8, 2018.

Finlayson, B. J. and Pitts, J. N.: Atmospheric chemistry: Fundamentals and experimental techniques, John Wiley and Sons Inc.,New York, NY, United States, 1986.

Frieß, U., Kreher, K., Querel, R., Schmithüsen, H., Smale, D., Weller, R., and Platt, U.: Source mechanisms and transport patterns of tropospheric bromine monoxide: findings from long-term multi-axis differential optical absorption spectroscopy measurements at two Antarctic stations, Atmospheric Chemistry and Physics, 23, 3207–3232, https://doi.org/10.5194/acp-23-3207-2023, 2023.

Heslin-Rees, D., Burgos, M., Hansson, H. C., Krejci, R., Ström, J., Tunved, P., and Zieger, P.: From a polar to a marine environment: has the changing Arctic led to a shift in aerosol light scattering properties?, Atmos. Chem. Phys., 20, 13671–13686, https://doi.org/10.5194/acp-20-13671-2020, 2020.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society, 77, 437–472, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996.

Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022.

Pernov, J. B., Beddows, D., Thomas, D. C., Dall´Osto, M., Harrison, R. M., Schmale, J., Skov, H., and Massling, A.: Increased aerosol concentrations in the High Arctic attributable to changing atmospheric transport patterns, npj Clim Atmos Sci, 5, 1–13, https://doi.org/10.1038/s41612-022-00286-y, 2022.

Sander, R., Burrows, J., and Kaleschke, L.: Carbonate precipitation in brine – a potential trigger for tropospheric ozone depletion events, Atmospheric Chemistry and Physics, 6, 4653–4658, https://doi.org/10.5194/acp-6-4653-2006, 2006.

Skov, H., Christensen, J. H., Goodsite, M. E., Heidam, N. Z., Jensen, B., Wahlin, P., and Geernaert, G.: Fate of elemental mercury in the arctic during atmospheric mercury depletion episodes and the load of atmospheric mercury to the arctic, Environ. Sci. Technol., 38, 2373–2382, https://doi.org/10.1021/es030080h, 2004.

Skov, H., Hjorth, J., Nordstrøm, C., Jensen, B., Christoffersen, C., Bech Poulsen, M., Baldtzer Liisberg, J., Beddows, D., Dall'Osto, M., and Christensen, J. H.: Variability in gaseous elemental mercury at Villum Research Station, Station Nord, in North Greenland from 1999 to 2017, Atmos. Chem. Phys., 20, 13253–13265, https://doi.org/10.5194/acp-20-13253-2020, 2020.