Dear Mr. Custodio,

Thank you for reviewing our manuscript, "Evaluating Weather and Chemical Transport Models at High Latitudes using MAGIC2021 Airborne Measurements," for publication in Atmospheric Measurements Techniques. The insightful comments and valuable suggestions have been instrumental in improving the quality of our manuscript.

Below, we provide a point-by-point response to your comments and concerns, with our responses indicated in blue. All page numbers refer to the revised manuscript file.

1. Major Concerns

Clarity and Presentation

The manuscript is difficult to follow due to unclear wording, undued wording, and overly dense descriptions. Some key points are buried in the text, making it challenging for readers to extract the central findings and their implications. Additionally, the plots are mazy and visually overwhelming, detracting from their effectiveness in conveying the results.

> The *Results & Discussion* sections (3&4) have been rewritten entirely with emphasis on clarity and bringing out the main results. Summary tables have been added to improve the understanding of the results and avoid getting lost in too much detail in the text.

Metrics and Model Performance Assessment

While the study evaluates model performance, the choice of metrics is not optimal. The authors should consider employing more comprehensive and widely accepted set of statistical metrics for model evaluation. Correlation Coefficients and Root Mean Square Error are good; however, I would recommend bias.

> Bias was actually employed but improper wording led to confusion. The bias calculation was done according to Willmott (1982). This has been added l. 271, and the word "bias" was used instead of our previous unclear wording.

Additionally, the manuscript could discus the implications of the metrics used. For example, while some metrics may show agreement, others may reveal discrepancies, which are worth exploring.

> The new summary tables show more clearly these different performances between models and paragraphs were added to discuss those in more detail in the *Results & Discussion* section (l. 382 to 387 and l. 517 to 523)

Figures

The figures and tables are a central issue. While they contain a wealth of information, they are too crowded and difficult to interpret. Each figure should serve a clear purpose and convey specific insights. To improve:

Use clean background, simplify the layout and make sure that the data are visible.

Use color schemes that are easy to distinguish, particularly for readers with color vision deficiencies.

Add concise and informative captions that explain the key takeaways from each figure.

Our new figure template uses a white background and colour maps specifically designed for colour deficiencies (provided by the cmcrameri matplotlib package). We also acknowledge that we relied too much on colours to distinguish between the different plots and have implemented, where possible, different approaches based symbols. Figure captions for all figures except Fig. 1 and Fig. 9have also been reviewed for clarity.

2. Others Comments

The manuscript's Figure 1 is confusing and requires clarification:

Scope of Flights: Does Figure 1 intend to show the entire MAGIC2021 campaign or only flights over Kiruna? The caption does not make this clear.

> The figure intends to show the flights from the MAGIC2021 campaign that were used for this paper. All of these were over the region of Kiruna, except for one which targeted oil platforms on the Norwegian coast. Clarifications have been added in the caption of the figure.

Flight over Norway: Why the flight over Norway is not included in the figure? Is it not part of the MAGIC2021 campaign?

> The flight over Norwegian oil platforms is indeed part of MAGIC2021. We initially chose not to show it on the map in order to distinguish in detail all the other flights. However we agree that this choice makes the map unclear so all the flights used in the study are now shown on the map

Unexplained Elements: The color blocks in the middle of Figure 1 are not explained in the caption or the text. This information seems to appear "out of the blue," and the figure lacks sufficient annotation to guide the reader. Please ensure that every feature in the figure is fully explained in the caption and supported by the text.

> Colours of the figure have been modified in order to better distinguish each category. Detail on the land cover dataset can be found in the paper whose reference is now provided in the caption text.

> New caption for Figure 1:

Location (left) and date and time (right) of MAGIC2021 measurements used in this study. Map background shows land use adapted from the Corine2018 dataset (European Environment Agency, 2019). Individual land cover types from the Corine2018 dataset are grouped in broader categories to ease map interpretation. Notably, wetlands include inland marshes, peat bogs, salt marshes, salines, intertidal flats, coastal lagoons and estuaries, i.e. both freshwater and saltwater wetlands.

The introduction of AirCore measurements is presented in a very shallow manner. While AirCore is a critical part of the study, its role and methodology are not sufficiently explained. Readers who are unfamiliar with AirCore technology will grasp it.

> Description of the instrument has been expanded. The new expanded section can be found from l. 98 to l. 109.

The division of the atmosphere into three layers based on pressure ranges—P > 800 hPa, 300 < P < 800 hPa, and P < 300 hPa—is arbitrary and does not align with commonly accepted atmospheric definitions. The chosen pressure thresholds do not accurately correspond to the planetary boundary layer (PBL), free troposphere (FT), or lower stratosphere (LS). A more scientifically sound approach would involve:

Using PBL height (PBLH) to define the boundary layer.

Defining the tropopause to separate the troposphere from the stratosphere.

This approach would ensure that the results are more meaningful and interpretable, especially for discussions of CH_4 transport dynamics across these atmospheric layer

> We agree that this arbitrary definition was a blind spot of the study. The three analysis levels were re-defined according to this comment. The first level (BL) was redefined according to boundary layer height data from ERA5. More precisely, BL height was interpolated from ERA5 on the aircraft and balloon trajectories, then compared to flight height and flagged as 'in BL' if the flight altitude was below the BL height interpolated from ERA5. Tropopause height was determined from AirCore measurements using the

Cold Point Tropopause method (Section 2.3.2 starting l. 267 was updated). The same method was used on ERA5 data to assess its accuracy against balloon observations. As temperatures are in good agreement between ERA5 and MAGIC2021 at the tropopause, the tropopause height derived from both datasets are very close.

The table captions should be placed at the top of the tables, following standard formatting conventions. Additionally, the table labels should succinctly describe the contents of the table. For instance, it does not make sense to include information about what is not in the table. Ensure that the captions are clear and concise, helping the reader to quickly understand the data presented.

> Tables' caption placement was changed according to this comment. Superfluous information has been removed from table captions where found.

The manuscript refers to "four statistic," which is an unclear and incorrect phrasing. Likely, the authors mean "four metrics used to evaluate model performance." The use of appropriate and precise terminology is critical for clarity. This error is indicative of broader language issues in the subsection "Statistics," which should be rewritten to ensure proper English usage and a professional tone.

> The section has been rewritten according to the reviewer's comment. Additionally, the word "statistic(s)" has been replaced by "metric(s)" in other sections too.

The caption for Figure 2 is insufficient to help readers understand the plot. Captions should summarize the key information conveyed in the figure and provide any necessary context for interpretation. In its current state, the caption leaves too much ambiguity and fails to assist the reader in navigating the content.

> New Figure 2 caption:

Wind rose plots for MAGIC2021 observations as well as ERA5 and WRF simulations. The radial axis gives the proportion (in %) of wind coming from a given direction given by the angular axis. Coloured bins represent the share of speed ranges shown in the legend associated with each direction. Rows correspond to data products MAGIC2021 observations, ERA5, WRF d01 and WRF d02. Columns correspond to atmospheric layers: Boundary Layer (BL), Free Troposphere (FT) and Lower Stratosphere (LS), described in Section 2.3.2.

The manuscript's discussion of wind fields is constrained solely to advection (horizontal transport), which provides an incomplete picture. The vertical component of wind, which is critical for transport processes and atmospheric mixing, is entirely missing. Vertical transport are among the most significant challenges in atmospheric modeling. Without addressing these, the discussion remains superficial. The authors could evaluate turbulence representation and vertical wind components in the models, as these are critical to understanding transport processes.

> This paper aims at comparing campaign data to simulation data. Unfortunately vertical wind data quality was inconsistent in the MAGIC2021 dataset, particularly with the weather balloon instruments. Therefore the decision was taken to focus on horizontal winds for which data could be compared to models.

Figure 5 is visually confusing and "weird" in its current presentation. The layout, formatting, and choice of visualization make it difficult to follow and interpret. Clearer design and simpler representations would greatly enhance the reader's understanding of this figure. Ensure that key messages are apparent and not lost in the visual clutter.

The content of subsection 3.3 is difficult to follow due to poor organization and unclear visualizations. The comparisons presented in this section lack coherence in terms of visual representation, metrics used, and overall wording. It is essential to streamline the presentation of comparisons to make them more reader-friendly and effective.

> This section was fully rewritten, with the addition of a summary table at the end of section 3. in order to help the reader comparing the model results.

The comparison of meteorological data between models and observations is superficial, merely reporting which model or dataset is closer to observations. This approach fails to provide meaningful insights or a deeper understanding of model inter-comparisons. Readers expect a more insightful analysis of model performance, including:

Identifying potential reasons for discrepancies.

Explaining how differences in parametrizations or data assimilation processes contribute to observed biases or differences.

Suggesting ways to improve model representation of meteorological processes.

The manuscript must go beyond simply reporting agreement or disagreement to provide a more nuanced and insightful evaluation.

> Section 3.4 was largely expanded. We include a discussion on the discrepancy between WRF and ERA5 performance. We also give a possible explanation as to why the increase in domain resolution between WRF d01 and d02 simulation does not directly equate to improvements in all statistical categories, based on Mass et al., 2002; Gómez-Navarro et al., 2015. We suggest implementing data assimilation in WRF to test its impact on model performance compared to high resolution observations.

The vertical profiles presented in the manuscript are overly complicated and lack clarity. The plots are "mazy," and the text does not provide sufficient guidance to help the reader interpret them. The analysis of vertical profiles should do more than report which model performs better in specific atmospheric regions (which, as noted above, were not properly defined). A thorough discussion of the physical processes contributing to vertical variations in CH_4 and meteorological variables would enrich the article.

The conclusion that all models overestimate CH_4 at the upper troposphere-lower stratosphere (UTLS) boundary is interesting but could be influenced by the interpolation method used for data colocation. In addition:

TM3 does not have the resolution to accurately resolve the tropopause.

While IFS has more vertical level, it still struggles with tropopause representation.

The lack of proper selection for the lower-most stratosphere in this study further compounds this issue. A more refined methodology is required to draw robust conclusions about model biases in the UTLS region.

> Additionally to the redefinition of analysis layers, we added more discussion points about the stratospheric bias which we hope give insights about where it could come from and what could be done to investigate the issue further (l. 510 to 554).

The association of the overall positive CH_4 bias in the boundary layer to wetland emissions is an important finding. However, this conclusion seems premature without further testing. A sensitivity test maybe could strengthen this claim and ensure that this conclusion is robust.

> Information has been added in the Methods section about the emission products used for WRF-Chem simulations (Table 1). This information helps to link overestimates seen in the lower troposphere to the emissions products. We also added literature references and other analysis elements to toughen that claim (l. 481 to 489).

The spatial and temporal limitations of this model evaluation could be addressed by incorporating data from the CoMet 2.0 campaign over Canada in the summer of 2022. While the MAGIC2021 campaign provides valuable observations, supplementing this with additional datasets could offer a more comprehensive evaluation of model performance.

> This comment has been added in the *Conclusion* section.

References

Willmott, C. J. (1982). Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society, 63(11), 1309–1313. DOI

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A.: Does Increasing Horizontal Resolution Produce More Skillful Forecasts?, 2002

Gómez-Navarro, J. J., Raible, C. C., and Dierer, S.: Sensitivity of the WRF Model to PBL Parametrisations and Nesting Techniques: Evaluation of Wind Storms over Complex Terrain, Geoscientific Model Development, 8, 3349–3363, DOI, 2015