We sincerely appreciate the time and effort the reviewers dedicated to evaluating our work. Their constructive comments have been invaluable in strengthening the manuscript. In response, we have reorganized the introduction to clearly present the scientific questions, added a table summarizing the sensitivity experiments, and included a table with an improved description of the aircraft campaigns. Furthermore, we have emphasized the importance of budget terms for understanding model differences. We have generated a web site for additional details on the output protocol.

Reviewer 1:

The manuscript egusphere-2025-3057 introduces a new model intercomparison project, named VOCMIP. After providing a comprehensive introduction and clear motivation, the authors describe the experimental design and the observational data used in the study. While I believe this is an excellent initiative and fully support the planned work, I must admit that I have mixed feelings about the manuscript itself.

I began reading the manuscript with great enthusiasm and found the introduction to be a well-crafted and engaging overview. Unfortunately, the experimental design does not live up to the same standard, and, more importantly, the manuscript lacks a clear scientific focus for the project. While some aspects of this focus are touched upon in the introduction, they are never fully articulated or clearly defined.

Hence, although the concept is promising, the manuscript lacks certain essential details and a clear scientific plan that are necessary for a more robust understanding and evaluation of the project. I therefore recommend that the authors revisit the scientific goals of the project and provide a more detailed and precise description of how the intercomparison project will address these goals.

Response: We have included scientific questions and added several sensitivity simulations with the goal of improving the understanding of model differences. We have also highlighted that the budget terms provide crucial insights into model differences, including variations in VOC lifetimes.

General comments:

As I mentioned earlier, I am unclear about the primary focus of this model intercomparison project. At the moment, it appears to center on "checking the differences in concentrations against observations," but the specific scientific goals are not clearly articulated. It would be very helpful to have these goals explicitly outlined.

Response: We have included scientific questions to better explain the purpose of the MIP. Again, we emphasize that the budget provide important information on understanding model differences so the aim of the MIP goes beyond checking against measurements.

If the primary aim is to investigate the chemical mechanisms, why not start by using box models to isolate the chemical processes from the noise introduced by varying atmospheric forcings? For certain models listed (e.g., EC-Earth, EMAC), different chemical decomposition schemes are available. Why not compare the results from the same model but using different decomposition schemes? This could simplify the analysis and help in pinpointing the sources of variability.

Response: This suggestion is appreciated and included as one of the Tier 2 experiments. However, we foresee that a limited set of models have this capability in terms of chemical decomposition schemes.

If the goal is to study the lifetime of VOCs within each chemical mechanism, why not use the same OH field across models? Since the models are capable of forcing methane at the surface (as mentioned in line 115), they should also be able to apply a consistent oxidative power across the atmosphere.

Response: Using the same OH fields in the models is far from being the same as setting surface methane concentration. OH is explicitly calculated in the models and very difficult to use the same fields for many of the global models.

Alternatively, if the primary objective is to investigate the different parameterizations of deposition, why not focus on intercomparisons of the deposition algorithms? If photolysis rates are the main area of interest, then why not perform an intercomparison of those as well?

Response: We have added Tier 2 simulations for investigation of importance of dry deposition of VOCs. The already included budget terms also provide important information on differences in dry deposition.

While it is certainly possible to pursue all of these objectives, each would require separate simulations. A thorough and meaningful comparison would necessitate a well-thought-out experimental design to ensure the results are attributable to specific causes. Otherwise, we risk obtaining a range of results without understanding the underlying reasons for the differences.

Response: We have included a set of Tier 2 simulations with the aim of explaining some of the differences among the models. We underscore also that budget terms are crucial in understanding model differences.

I am concerned that this exercise might turn into a "race" to see which model best matches observations. While this is undoubtedly valuable, I believe there is a greater opportunity here: to understand why the best-performing models are successful. This deeper insight could yield more valuable scientific knowledge.

Response: The goal is clearly to understand why models comparing well with observations are successful and thus we find the suggestion of adding additional experiments valuable. However, the budget terms will provide important information on model differences.

Specific comments:

*) line 42 : I would appreciate a clear definition of VOCs : when is an organic compound volatile?

Response: The following sentence is included at the beginning of the introduction: *'Volatile Organic Compounds (VOCs) are organic molecules (i.e., carbon-containing compounds) that readily evaporate and enter the gas phase under ambient environmental conditions.'*

*) lines 45-47 : I would move the descriptions of different compounds forming VOCs just after the definition (at the beginning of the text).

Response: Modified as suggested.

*) line 77 : Would be great to add a references on this statement (i.e. nucleation of VOCs' oxidation products)
Response: The sentence flows better without references to the statement and topic not investigated in VOCMIP.

*) line 82 : "Despite" sounds odd to me in this sentence

Response: The first part of the sentence is deleted including "*Despite*"

*) line 94 : If I understood correctly, emissions are possible by conserving mass, carbon, moles or also reactivity, as described in detail few lines before.
Response: Correct.

*) line 98 : I would remove the reference to the VOCMIP, as this has not yet introduced (coming later in line 110)

Response: The introduction is reorganized and "VOCMIP" is defined much earlier in the introduction.

*) line 114 : I found quite funny that you describe methane as "the most aboundant VOC in the atmosphere and the dominant source of key VOCs" , but you will neglect it in this work. I also do not really understand the connection with methane being prescribed at the surface. Actually, this would be even an advantage, as most of the dynamical-driven processes are forced and you could simply focus on the chemistry of methane. In addition, while not a major focus, it seems (table 1 and 2) that you will need exactly the same data of all the other components. So there is no real differences from the technical point of view between methane and other VOCs.

Response: Methane will certainly not be neglected, and it was included in Table 1 (now Table 2) and thus variables in Table 2 (now Table 3) will also be included. The paragraph is deleted since we have added a Tier 2 experiment explicitly devoted to methane.

*) Line 132 : What is the motivation to select these emissions? What about different speiciation methods in the emissions? How do you assure that the same amount of VOCs (either in reactivity or mass) are actually emitted ?

Response: These are standard emissions and used in other MIPs as in CMIP7. Emissions will be provided as described in the output protocol (now Table 3). We have made a website where the output protocol is included.

*) Line 133 : "supplemented with natural emissions" is quite vague: if you have completely different emissions you cannot expect any intercomparison at all (or only partially).

Response: We have designed one experiment (see experiment 2d in Table 1) where models will use the same natural emissions.

*) Table 2: What is the definition of "Chemical destruction"? The total chemical sink (including photolysis)? Would be a speciation (e.g. OH oxidation or NO3 oxidation) be more informative?

Response: Additional details are added to the table caption and the website includes the output protocol with additional specifications.

*) Line 192 : Some of these acronyms were not introduced before (and some only in the Fig.2 label). Would be very interesting if you have a stragety to compare models with "BIGALK" with models that have a more (or less) chemical detail (i.e. VOCs speciation).
Response: The figure is meant for illustration purposes and in the first phase of VOCMIP we will contrate on some of the main VOCs. BIGALK is likely not among the main components for the first phase.

*) Line 195 : "within carious project" sounds a bit vague.

Response: This will be specified clearly when actual data are used in VOCMIP.

We sincerely appreciate the time and effort the reviewers dedicated to evaluating our work. Their constructive comments have been invaluable in strengthening the manuscript. In response, we have reorganized the introduction to clearly present the scientific questions, added a table summarizing the sensitivity experiments, and included a table with an improved description of the aircraft campaigns. Furthermore, we have emphasized the importance of budget terms for understanding model differences. We have generated a web site for additional details on the output protocol.

Reviewer 2:

The manuscript introduces the Volatile Organic Compound Model Intercomparison Project (VOCMIP), proposing simulations for 2015 and 2019 using eleven global atmospheric chemistry models with common anthropogenic emissions (CEDS, GFED) and meteorological forcing from reanalysis data. The authors plan to compare model outputs with satellite retrievals (IASI, CrIS, TROPOMI), surface station measurements (EBAS database), and aircraft campaign data (KORUS-AQ, ATom) for key VOCs including formaldehyde, methanol, acetone, isoprene, and various alkanes. Participating models include CESM2 CAM-Chem, EC-Earth, EMAC, FRSGC/UCI CTM, GISS, GFDL, LMDZ-INCA, NorESM2-LM, OsloCTM3, UCICTM, and UKCA. The paper presents requested model output (3-hourly 3D fields in Table 1, monthly budget terms in Table 2), describes available observational datasets (Table 4, Figure 3), and shows preliminary burden comparisons between two models (Figure 2).

The VOCMIP concept has scientific value, but this manuscript represents preliminary project planning rather than a completed research contribution or even a rigorous protocol document suitable for publication. Therefore, I cannot recommend this manuscript for publication in GMD in its current form. The paper has two fundamental problems: (1) it is written in a grant proposal form, and it does not contain model development and therefore does not fit GMD's scope, and (2) it lacks clearly articulated scientific objectives and a coherent experimental design to achieve them.

Specifically, journal scope mismatch: GMD publishes papers describing model development - new models (with model version numbers), significant model updates, or novel parameterizations and methodologies. This manuscript describes a protocol for running existing atmospheric chemistry models with specified inputs and comparing their outputs to observations. There is no new model code, chemical mechanisms, parameterizations, diagnostic tools, or software development. I recommend the authors consult with the editor about appropriate journal placement.

Response: From the aims & scope of GMD the manuscript is clearly within the scope: https://www.geoscientific-model-development.net/about/aims_and_scope.html

GMD has published several MIP description papers. We have included several scientific questions into the revised manuscript.

More importantly, the absence of scientific questions, which I think is the greatest problem with this paper: the introduction (lines 42-107) provides valuable background on VOC atmospheric chemistry, but this does not translate into specific, testable hypotheses. The stated aim to "identify model consistencies and discrepancies, enhance the formulation of chemical mechanisms, and advance our understanding of VOC-related processes" (lines 35-37) is too vague. What specific aspects of VOC chemistry are uncertain? What processes drive model differences in formaldehyde or methanol budgets? Can we constrain emissions using model-satellite disagreement? Without 2-4 concrete scientific questions, the project risks producing model comparisons without scientific insight.

Response: Several scientific questions are included in the manuscript. We also emphasize that the budget terms already included in the output protocol will help the understanding of model differences. This point has been included associated with the description of the table as well as in the summary section.

Also, the inadequate experimental design: the proposed experimental setup (Section 2) will produce model differences that cannot be attributed to specific causes. Lines 132-133 specify common anthropogenic emissions (CEDS, GFED) but allow each model to use different natural emissions "as defined for each participating model." This ensures models

will simultaneously vary in: natural emissions inventories, VOC speciation schemes, chemical mechanisms, deposition parameterizations, and photolysis calculations. When Figure 2 shows models differing by factors of 2-3 for acetone, ethane, and methanol, what causes these differences? Emissions? Chemistry? Deposition? Transport? Without sensitivity experiments that isolate individual processes, these questions cannot be answered.

Response: Sensitivity simulations are now included as described in Table 1 in the revised manuscript. The budget terms on emission, chemical loss and production, dry and wet deposition, and thus lifetime will provide highly useful information on explaining the model differences.

Lastly, there are also structural issues: the manuscript describes plans rather than results. It has no Results or Discussion sections. This reads as a project proposal or protocol document, not a completed research paper suitable for publication. The summary is only 3 sentences that did not actually summarize the work, and it ended on a hanging note.

Response: Many GMD papers describing a protocol, including several CMIP protocol papers. The summary section is improved.

Specific Comments:

1. Lack of process isolation in experimental design

The core flaw is that all potential sources of model uncertainty vary simultaneously. For meaningful intercomparison, you need tiered experiments. Why not test sensitivity to key processes individually? You could design experiments that:

- Prescribe identical OH, O_3, NO_3 fields to isolate VOC oxidation rates from oxidant chemistry
- Use identical deposition velocities to isolate chemical vs. depositional sinks
- Vary only emissions inventories while fixing chemistry to isolate emission uncertainty
- Compare different chemical schemes within the same model structure (EC-Earth and EMAC offer this capability per your Table 3)

Without such experiments, Figure 2's model differences become a catalog of disagreement without physical understanding.

Response: We have included some Tier 2 simulations to investigate several of the suggestions above. Specifying OH, O_3, NO_3 fields in complex 3d atmospheric chemistry models are complicated. Note, that our budget terms given in Table 4 (earlier Table 3) allows us understanding of differences in emissions, chemical production and loss, dry and wet deposition and consequently differences in lifetimes.

2. Methane treatment is contradictory (lines 114-116)

The text states methane is "the most abundant VOC in the atmosphere and the dominant source of key VOCs like HCHO and methanol" then immediately excludes it from detailed analysis. This creates logical problems:

- Figure 1 shows CH_4 as a formaldehyde source, yet CH_4 chemistry won't be diagnosed
- Table 1 requests CH_4 mixing ratios, contradicting "not a major focus"
- Prescribing surface CH_4 doesn't eliminate inter-model differences in vertical profiles, oxidation rates, or HCHO yields
- Methane's relatively simple chemistry (compared to isoprene or aromatics) makes it an ideal test case for understanding model differences

Either include methane with full budget diagnostics (Table 2), or explain scientifically why CH_4 chemistry differences are irrelevant to your objectives. The current justification is inadequate.

Response: CH4 was included in Table 1 (now Table 2). We have also included an additional experiment with CH4 emissions in addition to the fixed CH4 concentration simulation. We have deleted the paragraph related to methane (lines 114-116 in the submitted version) as this clearly could be misunderstood.

3. Emission specifications are insufficient (lines 132-133)

"Supplemented with natural emissions (e.g., MEGAN) as defined for each participating model" provides no standardization. This means models may use:

- Different MEGAN versions (2.0, 2.1, 3.0 have substantial differences)
- Different emission factors and environmental response algorithms
- Different land cover datasets

- Different isoprene emission estimates (you note "significant uncertainties" at line 62-64)

How will you ensure comparable VOC emission totals across models? Lines 89-95 discuss conserving mass, carbon, moles, or reactivity when lumping species - which approach should models use? This must be specified or explicitly diagnosed. For a protocol paper, emission processing must be precisely defined or you need sensitivity experiments varying emissions systematically.

Response: We have added a Tier 2 simulation with fixed MEGAN emissions (experiment 2b) to investigate the role of differences in natural biogenic emissions. Anthropogenic and fire emissions are specified and Table 3 specify that emissions are required for all species given in Table 2.

4. Diagnostic definitions need clarification (Table 2)

Table 2 requests budget terms but definitions are ambiguous: "Chemical destruction" - Does this include photolysis (listed separately) or not? Should this be total loss rate or separated by oxidant (OH vs. O_3 vs. NO_3)? For formaldehyde specifically, understanding whether OH oxidation or photolysis dominates requires separate diagnostics. "Chemical production due to photolysis" - Does this mean HCHO produced when other species photolyze? Or HCHO loss via its own photolysis? The naming is unclear.

Missing speciation: For species like formaldehyde with multiple production pathways (methane oxidation, isoprene oxidation, direct emission, etc.), requesting only "total production" limits process understanding. Consider requesting production separated by precursor VOC. Every diagnostic must have unambiguous definition for consistent model output.

Response: In the table caption of the original manuscript, it was given: '*3D total production rates and total loss rates to chemical reactions (prod_x and loss_x).*' This has been updated to: '*3D total production rates and total loss rates to chemical reactions including photolysis (prod_x and loss_x).*' This is further explained in the output protocol available at the VOCMIP web page.

In most of the models it is too complicated to achieve production separated by precursor VOC.

5.Strategy for comparing models with different chemical complexity (line 192, Figure 2)

Line 192 and Figure 2 reference "BIGALK" and other lumped species without explaining the comparison strategy. How will you compare:

- Models using BIGALK (lumped C4+ alkanes) vs. models with explicit butane, pentane, hexane, etc.?
- Different lumping approaches (by reactivity vs. by carbon number vs. by structure)?
- Oxidation products from lumped vs. explicit schemes?

This is fundamental to your stated goal of understanding "impacts of these approaches" (line 97). You need a clear methodology for cross-walking between chemical mechanisms of different complexity.

Response: In the first phase of VOCMIP we will concentrate on the major VOC compounds and thus BIGALK will not have a major focus. The budget terms will also help us understand model differences as written in the summary section: '*The budget terms presented in the output protocol help us understand differences in emissions, chemical production and loss, dry and wet deposition, and consequently the variations in lifetimes.'*

6. Observational comparison methodology is underdeveloped

Section 3 lists datasets but lacks comparison protocols:

e.g. satellite data (Table 4):

- Will you apply averaging kernels to account for retrieval vertical sensitivity?
- How will you handle sampling biases (clear-sky only, specific overpass times)?
- What retrieval uncertainties will be used for model evaluation?

Surface and aircraft data (lines 191-204):

- Which EBAS stations will be selected and why?
- How will you spatiotemporally match aircraft measurements (~80 m × 2 km, instantaneous) to model grid boxes (~100-200 km, 3-hourly)?
- Lines 200-204 note observational gaps for several species in Table 1 - how will models be evaluated without observations?

A protocol paper needs explicit procedures for model-observation comparison, not just a dataset catalog.

Response: We have included that averaging kernel will be applied. Further details on the model evaluations will be given in the different papers analyzing model results with observations. VOCMIP can be an activity continuing for many years and we want

to allow for flexibility to apply new methods and approaches, including machine learning techniques.

7. Missing analysis framework

The manuscript describes data collection (Tables 1-2) without explaining analysis methodology:

- How will you quantify contributions of emissions, chemistry, deposition, and transport to model spread?
- What statistical metrics will assess model skill?
- How will observational uncertainty inform model evaluation?
- What constitutes model success or failure?

Lines 137-139 request extensive budget terms but don't explain their use. Will you perform process attribution? Sensitivity analysis? Budget closure checks? The paper needs an analysis strategy.

Response: We emphasized in the summary that emissions, chemistry sources and sinks, and depositions and thus consequently lifetimes will help understanding model differences. This manuscript is intended for describing the protocol for model simulations and background for its need. We have added additional sensitivity experiments. However, analysis will be described in forthcoming VOCMIP papers.

Technical Corrections

Line 42: Define VOCs explicitly (vapor pressure threshold or other criteria) rather than assuming reader knowledge.

Response: The following sentence is included at the beginning of the introduction: *'Volatile Organic Compounds (VOCs) are organic molecules (i.e., carbon-containing compounds) that readily evaporate and enter the gas phase under ambient environmental conditions.'*

Lines 45-47: Move the compound type descriptions (hydrocarbons, oxygenated species, aromatics, etc.) immediately after defining VOCs for better organization.

Response: Modified as suggested.

Line 77: Add reference for "oxidation products…nucleate new particles"

Response: The sentence flows better without references to the statement and topic not investigated in VOCMIP.

Line 82: Replace "Despite" with "Given" or "Because of" for clarity

Response: The first part of the sentence is deleted including "*Despite*"

Line 98: Remove VOCMIP reference before the project is introduced (line 110)

Response: The introduction is reorganized and "VOCMIP" is defined much earlier in the introduction.

Line 114: "aboundant" → "abundant"

Response: "*aboundant*" not found in the submitted manuscript after a thorough search. However, a comment by Reviewer 1 has this spelling error.

Line 195: Specify which projects contribute data instead of "within various projects"

Response: This will be specified clearly when actual data are used in VOCMIP.

Table 1 footnote: Move grid specification requirements to table body with exact details (cell area, layer thickness, pressure levels)

A detailed experiment output protocol with this information is available at our website: Volatile Organic Compound Model Intercomparison Project. We refer to details available on the web page in the caption.

Figure 2 caption (lines 180-185): Move simulation details to methods text; caption should describe what the figure shows. Also the colors are not visually friendly, the fonts are squished.

Response: This figure is an example rather than methods for VOCMIP thus the details in the figure caption do not fit into the method text. Colors and fonts improved.

Table 2: Clarify which species require photolysis budget terms

Response: A detailed experiment output protocol with this information is available at our website: Volatile Organic Compound Model Intercomparison Project

Table 3: Add DOIs or URLs for model references where available

Response: Proper references are already given.