Reply to Stephan Kindermann’s comments

We thank Stephan Kindermann for the positive evaluation of our manuscript and the valuable suggestions that will help improve the manuscript, particularly for pointing us to the IS-ENES3 project. In the following, we respond to the comments on a point-to-point basis and indicate the changes that we made to address them.

Response to specific comments

- **Referee:** “Line 48: "submission to CMIP6 closed in 2025 with more then 16 TB of distinct datasets .." should be "with more then 16 PB " !!!”

  Response: Yes, indeed it should be 16 PB, we thank the reviewer for catching the typo.

  Changes made: We have changed the line to read "with more than 16 PB" as suggested.

- **Referee:** The analysis mainly relies on data usage from ESGF nodes, yet the important distinction of different types of ESGF nodes is not provided, especially the importance and relevance of "replica ESGF nodes" and associated data pools (which are not reflected in ESGF usage patern analysis). These replica centers are not just providing data from dedicated associated modeling centers but playing an active role in replicating, re-publishing core ESGF data and making data accessible in data pool associated computing environment and thus data usage is not reflected in any ESGF statistics collection. The associated topic of widespread use of national and shared resources is just shortly mentioned in line 66 to 68 as well as the relevance of replica pool associated compute resources for data analysis activities. The importance of "data proximate computing" is only explicitly mentioned in the context of cloud data provisioning (see lines 98ff) and in the context of provisioning data from Copernicus which supports data proximate e.g. subsetting services. The aspect that copernicus data is provided by specific distributed European data centers is mentioned, but not the fact that these data centers also are providing "esgf replica pools" and thus the Copernicus data provisioning is associated to this "ESGF replica data center role" they play in the ESGF federation.
  Referee: it is mentioned line 66 ff: ".. data usage through these and smaller institutional archive or servers in not recorded in any usage statistics .. " - some information and references on this is available e.g. from the European IS-ENES3 project, see e.g. the data statistics and KPI collection deliverables accessible at https://is.enes.org/deliverables/.

  - especially D7.1, D7.3 and D7.6, which also partially rely on the ESGF statistics service information provided by CMCC used in this paper, but provide additional information.
  - e.g. they also provide info on data usage patterns over time etc. - in line 86 it is mentioned that "download statistics over time" were not included because the CMCC database was unfortunately inaccessible .. - so these reports may include relevant information which can be included in this paper in a final version …

  Response: In our initial manuscript, we did not go into details into this aspect as we lacked data. Thank you for sharing the IS-ENES3 project and your insights on the effect of the data pool on Europe’s statistics. We supplement the revised manuscript with information from those reports and we emphasize the importance of tracking the ESGF nodes serving as data pool in the future.

  Changes made: We have updated the manuscript in fitting places (e.g., sections 2, 3.4, 4.2, 6) to highlight the impact of replica nodes and associated data pools on tracked downloads. The link between the replica nodes and C3s is will also be made explicit.

- **Referee:** The IS-ENES3 project statistics collections also underlines the importance and relevance of the establishment of these "ESGF replica sites" on the overall ESGF data download and usage patterns. After the establishment of the large replica data pools in Europe at DKRZ, IPSL and UKRI, download stats from European ESGF nodes went down as users directly did analysis at the data centers which coordinated ESGF data collection and provisioning for them. So e.g. the data usage patterns collected in pages 5 ff do not include a big percentage of European users .. etc. See e.g. Figure 9 (and 10) illustrating the data volume published vs. volume downloaded per continent (and country)- and the big difference between Europe and north America. The provisioning of replica pools in Europe and providing pool associated data analysis resources to a broader user community is certainly a key point explaining the EU / North america difference depicted there.

  Response: This is an important point that impacted our analysis and we thank Stephan Kindermann for pointing us to the IS-ENES3 resources as they can help estimate some of the impact of institutional archives and data pools on download statistics, especially in Europe.

  Changes made: We have addressed the impact of institutional data pools and replica sides more explicitly throughout the manuscript, especially using European download statistics and their changes as data pools came online as an example (specifically in Sections 3.4, 4.2 and 6).

- **Referee:** Section 5 (page 20 ff): The is-enes reports may also provide additional info on the usage of citation DOIs etc. which could be included here (as the citation service for CMIP6 was provided as part of IS-ENES).

  Response: We thank Stephan Kindermann for pointing us towards this resource tracking the DOI and metadata registration over time. Since our manuscript focuses on data usage, an in-depth discussion of the DOI system is out of the scope of this paper, so we instead point readers to the resources in question.

  Changes made: We point readers to the resource in Section 5.1.

- **Referee:** Section 6: lessons for the future

  - line 374: 1000TB of erronous files - not clear what the criteria for "erroeousness" is .. only the aspect "wiht metadata that did not macht existiong CVs" ? In this case many files are not really erronous from the users perspective, they still used them in their analysis activities ..

  Response: The 1000 TB that we call erroneous is the sum of downloads that have a table_id-variable_id pair that is not in the CMOR tables and table_id and frequency that are incompatible. Yes, by looking at the data, some users can feel confident enough that they know what the data is and that they can use it regardless of the CVs. However, many users can still be left unsure what the data is when they look into it or unable to use the data (ex. they downloaded data with frequency “day”, but the actual data is monthly as defined in the table id “Amon”. This is not what they were looking for). We make this clearer in the revised manuscript.

  Changes made: A clarification of “erroneous” has been added to section 3.1.

- **Referee:** the importance to include large ESGF replica data providers in future usage patterns analysis should be stressed. As here data is just downloaded once and then used by a large user community without any hint in ESGF data node statistics analysis …

  Response: We agree with the referee that large data pools linked to ESGF nodes will impact the analysis and this impact should be stressed more throughout the manuscript.

  Changes mde: As outlined above, we now address the impacts of ESGF data pools with associated computing infrastructure more throughout the manuscript (e.g., Sections 2, 3.4, 4.2, 6).

- **Referee:** A comment for future CMIP usage pattern analysis if data is more and more provided in cloud and analysis ready formats (and not in the form of individual netcdf files as in CMIP6) would be usefull. Some of the CMIP6 data on Amazon and google cloud was already provided e.g based on Zarr. The future is certainly providing data in cloud opmtimized and analysis ready formats, with access patterns not based on files but on small chunks. Is this an aspect where you can provide some early insights - or is this not possible based on the information currently available about cloud based data provisioning ? This is also a problem / an aspect which needs to be prepared for the future - how to collect data usage information not based on file download but based on chunk level access patterns …

  Response: From discussions with members of the pangeo community, our understanding is that they will not be creating zarr copies of the CMIP7 data as they did with CMIP6. The plans do not seem to be finalized yet, but the community seems to be going in the direction of using tools like virtualizarr (https://virtualizarr.readthedocs.io/en/stable/) to be able to access netcdfs “like a zarr”. In CMIP7, there are new requirements on the chunking of the netcdfs to make them more cloud-friendly (https://guidance.mipcvs.dev/CMIP7/Guidance_for_modellers/#5-model-output-requirements). These are new tools and we are not certain how logging works with them, but we have added a discussion in the revised manuscript on the difference between statistics on files vs. chunks and the expected growth of ARCO data.

  Changes made: We have added a discussion of the difference between statistics on files vs. chunks and the expected growth of ARCO data in the conclusion.

Reply to Mario Acosta’s review

We thank Mario Acosta for the thoughtful comments and valuable perspective on the manuscript. In the following, we outline how we have implemented the suggestions, which helped strengthen our manuscript.

Response to main comments

1. Interpretation of downloads versus actual usage should be handled more carefully

Referee: The manuscript is appropriately transparent in stating that downloads provide only a partial view of usage and that many factors influence the statistics, including dataset size, time resolution, spatial resolution, ensemble size, archive strategy, and publication timing. This is one of the strengths of the manuscript. However, in the interpretation of the results, the language occasionally becomes stronger than the evidence fully supports.

For example, the paper draws conclusions about which variables, experiments, or sources are most “used”, but in many cases the statistics likely reflect a mixture of true user demand and archive-side effects. The variable-level ratio RRR, defined as downloaded data over available data, is a helpful step toward normalisation, but similar corrections are not developed with the same strength for experiments, models, or institutions.

I suggest that the authors tighten the language throughout the manuscript to distinguish more explicitly between:

- observed download activity,
- actual scientific use,

- archive availability and visibility,
- and access patterns conditioned by file structure or provider capabilities.

Response: We agree that usage might be too strong a word. Usage is what we set out to study, but the data only really allows us to make conclusions on downloads. We clarified the language in the revised manuscript.

Changes made: In most of the manuscript, wherever applicable, “usage” has been replaced by “download” and “popular” by “most downloaded” .

2. The effect of archive structure is one of the most important results and deserves stronger treatment

Referee: The comparison between EC-EARTH3 and CESM2 for monthly historical tas is one of the most insightful parts of the manuscript. It clearly shows that archive choices, for example yearly files versus multi-decadal files, can dramatically alter the number of downloadable files and therefore the apparent usage statistics.

This point is central because it affects the interpretation of several later results, especially rankings by number of downloads and possibly also some by total downloaded volume. At present, this issue is discussed well, but mostly through one illustrative example. Since one of the paper’s practical recommendations is that more uniform file formatting could improve accessibility and interpretation, I think this argument would benefit from a slightly more systematic treatment.

For instance, the authors could:

- better discuss whether download volume is also substantially biased by storage conventions, not only download counts,
- clarify how representative the EC-EARTH3/CESM2 example is of broader CMIP6 publication practices,
- and emphasize more clearly that some source-level rankings cannot be interpreted independently of file chunking and storage decisions.

Response: We agree that storage choices from modelling centres, whether it is the file chunk size, the compression ratio of the output files, or the precision of the values, these factors have a significant impact on the statistics used to monitor the ESGF nodes. We made it clearer in the revised manuscript, having added suggestion of metrics that could be less dependent on these storage decisions, addressing the third point. Regarding the first point, the manuscript already mentions that when looking at the weight of a member simulation for the historical experiment, the equivalent file size for the CESM model is 3.2 times smaller than for EC-Earth3. Given the data we have, it is difficult to disentangle the impact of the storage strategy from other factors that may influence storage, such as the precision of output values or the centre choice of file compression ratio. This would require a more in-depth study which would need to account for at least these last two parameters, and take into account the impact of differences in resolution among the model components. Involving more models in this meta-analysis would address point 2 and would indeed make it possible to specifically propose advanced optimization of the storage strategies for the modelling centres. As this is beyond the scope of this paper, we clarify this point in the manuscript instead.

Changes made: We included a suggestion of monitoring metrics in the recommendations that take into account the number of years of published simulations for instance, rather than the number or size of files. This provides a better understanding of the differences between resolution-based and equivalent-variable

models, thereby enabling more effective process optimization, without limiting modelling centres' ability to use files of a manageable size based on their own criteria and to enhance the user experience. We also added a discussion of the impacts of storage decisions on download statistics to the paragraph on the CESM2/EC-EARTH3 example and clarified how representative the example is.

3. Recommendations regarding low-use variables should be framed more cautiously

Referee: One of the main conclusions is that, since for roughly a quarter of the variables more data was available than downloaded, future resource allocation could potentially be optimised by revisiting whether low-usage variables should be published at reduced frequency or with higher compression. This is a reasonable and useful discussion, especially in view of storage constraints and carbon footprint considerations.

However, I recommend that the authors frame this point more cautiously. Low download rates do not necessarily imply low scientific value. Some variables may serve specialised but essential user communities, model evaluation tasks, process studies, or reproducibility requirements. A variable can be scientifically important even if it is not widely downloaded. The paper does acknowledge some caveats, but the policy implication here remains stronger than the evidence warrants.

I therefore suggest explicitly stating that download statistics should inform prioritisation, but not be used as the sole criterion for decisions on what should or should not be archived in future CMIP phases.

Response: We agree with Mario Acosta that our phrasing was too strong and the assessment of each variable must also take into account other factors. We now frame the discussion more cautiously as we do not intend to recommend against the storage or production of certain variables. Instead, we would like to suggest the consideration of other sharing solution, since, although a variable may be of considerable scientific interest to a specific scientific community, if it is rarely downloaded via the ESGF nodes, and particularly if it takes up a lot of space, such other sharing solutions may prove less resource-intensive and potentially less burdensome for modelling centres.

Changes made: In the conclusion, we clarified the discussion on "less-downloaded variables", taking into account the various limitations of our analysis. We also suggest that downloads only be one part of the considerations for how to store variables. Work by the CMIP7 data request should also be part of the decision in order to properly understand the importance of each variable for different communities (Mackallah et al. under discussion, https://egusphere.copernicus.org/preprints/2026/egusphere-2026-1641/) .

4. The ESGF metadata inconsistency result is very important and could be highlighted even more

Referee: The finding that erroneous metadata combinations accounted for about 782,830 files and around 1000 TB of downloaded data is, in my opinion, one of the strongest and most actionable results of the manuscript. It directly illustrates the scientific and operational cost of imperfect metadata quality control, both for users and for the broader infrastructure ecosystem.

This result strongly supports the manuscript's discussion of QA/QC improvements for CMIP7. I encourage the authors to highlight this even more clearly in the abstract and conclusions, because it is not merely a technical inconvenience but a real inefficiency affecting usability, reproducibility, and duplicated user effort.

Response: We agree with the reviewer and now highlight this result more strongly as suggested.

Changes made: We added a mention of QA/QC in the abstract and now highlight this result more in the conclusion.

5. Comparison across providers is useful, but should be presented as clearly asymmetric

Referee: The inclusion of AWS, Google Cloud and C3S broadens the relevance of the paper and is welcome. However, the resulting comparison is necessarily very uneven:

- Google does not provide usage statistics,
- AWS statistics are limited and difficult to interpret,
- C3S provides richer statistics but only for a constrained subset of variables, models and experiments,
- and ESGF remains by far the dominant quantitative basis for the analysis.

The manuscript already says this, but I think the framing should go slightly further. The paper is fundamentally an ESGF-based usage analysis, complemented by exploratory perspectives from other providers. Presenting it in this way more explicitly would help calibrate reader expectations and avoid overinterpreting cross-provider differences.

Response: This is accurate. At the beginning of the project, we were hoping for a more balanced comparison, but it was not possible with the data available. We now make that clearer in the manuscript.

Changes made: We now highlight that ESGF download statistics form the core of the manuscript as these are the most comprehensive throughout the manuscript from the abstract to the data descriptions and conclusions.

6. Citation analysis is interesting but should be even more clearly labelled exploratory

Referee: The attempt to compare download activity with citation counts is worthwhile, especially because it touches on the broader impact of CMIP data in science. However, the manuscript itself explains why this proxy is incomplete: dataset DOIs are inconsistently cited, some users cite model description papers instead, and the publication hub is sparse or unavailable.

Because of this, I recommend framing the citation analysis even more explicitly as exploratory and incomplete. The current discussion is mostly careful, but a reader could still come away with too strong a sense of relationship between downloads and scientific uptake. This section would benefit from a sentence stating directly that citation counts, in the present form, cannot support robust comparative conclusions across models or experiments.

Response: We agree that this cannot be more than an exploratory analysis, complementary to the other analyses in the manuscript and that this should be more evident in the manuscript.

Changes made: We have added a statement to section 5.1 that clearly labels the analysis as exploratory.

Response to minor comments

1. **Referee:** The manuscript would benefit from consistently distinguishing between “downloaded files”, “downloaded volume”, “requests”, and “users”, especially when comparing ESGF and C3S, since these metrics are not equivalent.

   Response and changes made: We use clearer language throughout the manuscript.

2. **Referee:** The wording around “most used variables” or “most popular models” could be softened in some places to “most downloaded” unless normalisation or context is explicitly applied.

   Response and changes made: We refrain from using "used" and "popular" in the manuscript and use more explicit language, usually "downloaded".

3. **Referee:** The discussion of geographical patterns is interesting and potentially important, especially regarding differences between ESGF and C3S user locations. Still, I would encourage the authors to remain cautious in attributing these patterns to resource inequality alone, since untracked institutional mirrors and local infrastructures may play an important role.

   Response and changes made: We agree with the reviewer that this requires a cautious discussion. We have added a new passage discussing the effect of untracked institutional shared resources on geographical patterns of ESGF, following a comment of the other reviewer.

4. **Referee:** It would be useful to clarify slightly more how incomplete ESGF node coverage in the CMCC monitoring system may affect the interpretation of the results.

   Response and changes made: We have added a passage on this issue. We were also contacted by Neil Swart who shared that the CCCma node was not tracked for a period of time, which will have affected the download statistics of the Canadian model, an issue that is now mentioned in the discussion.

5. **Referee:** The discussion of downscaled products is valuable and broadens the perspective of the paper beyond raw CMIP downloads. If possible, the authors could make even clearer that downstream impact through regional or statistically downscaled datasets is likely systematically underestimated by the statistics analysed here.

   Response and changes made: We agree with the reviewer that this is a salient point to highlight and have added to this section to be clearer.

6. **Referee:** The manuscript is generally well written and structured. I only suggest a final pass to ensure that the strongest conclusions always remain aligned with what the available statistics can actually demonstrate.

   Response and changes made: We agree with the reviewer that ensuring consistent and accurate language throughout the manuscript in line with the analysis results and their limitations is vital. As such, we have ensured accurate language throughout the manuscript during revision and that results and conclusions are aligned throughout.