The authors would like to sincerely thank both referees for their careful review and constructive comments that have helped to improve the revised manuscript.

RC1: Anonymous Referee

1. *"Figure 2 provides no information whatsoever: ..."*

   The efficiency metric presented in Section 4 requires a graph-representation of the data consumers and producers (nodes) and data dependencies amongst them (edges). Figure 2 was our attempt to illustrate the complexity of this graph (albeit in a simplified form) and we do think it is important. But the reviewer is right in his assessment that currently the figure is not explained very well. We have decided to retain the figure, but have revised and extended the figure caption and added a better description in the text.

2. *"there are a number of issues with English, spelling etc..."*

   Thank you very much for the detailed comments in the supplement, which we address in the revised manuscript.

RC1: R. W. Ford

1. *" . It would be good (at least for me) if you could outline how these approaches relate and therefore give a feel of what global models can learn from this work. "*

   We believe that our results apply directly to global weather and climate models employing structured grids and explicit, split-explicit or HEVI time discretizations (e.g. FV3, NICAM). Global models employing implicit or spectral solvers may have a different scaling behavior. We have added a discussion in the revised manuscript.

2. *" . I think it would be worth toning down the claims for the metric in this paper and limit it to saying that it was useful for this analysis and that it will be further analysed in future work."*

We have focused the claims of the usefulness of the metric to the analyzed application. We agree that the journal is not the best venue to argue for the metric in general and thus focus the discussion solely on the use-case studied, which is also of biggest interest for the community.

3. *" . Could you please modify this section slightly, perhaps just to say that "algorithmic optimization is having to increasingly focus [on data movement]."*

   We do not perform or propose any algorithmic optimizations. We simply change the schedule (i.e., the order) of computations in the application to improve locality.

   (a) *"It is not immediately clear whether internode comms is optimised here. "*

   While the method can be used to optimize data-movement in general (also considering internode communication), this has not been done. We mention this is the text in the revised manuscript.

   (b) *"Can MUE be used on memory bound codes that do not use STELLA and therefore do not have a CDAG?"*

   The MUE metric is a general concept and not related to the programming model. The fact that the dynamical core of COSMO was written using a DSL allowed us to generate the CDAG automatically.

   (c) *"Is the metric [MUE] expeted to be useful for CPU's as well, given their complex memory hierarchies?"*

   Yes, the metric can be used in a straightforward manner. Especially due to the complex memory hierarchies it is of paramount importance to optimize data movement and the MUE metric can be used to assess how well a code is doing this. We will add a sentence in the conclusions mentioning the applicability of the MUE metric to other hardware architectures.

   (d) *" . However, a program might have different phases with different access patterns potentially leading to different achievable bandwidths e.g. ECMWF's IFS."*

   In this case each part would require its own micro-benchmarks. Due to the homogeneity of the COSMO dynamical we decided to use an average bandwidth for simplicity.

(e) *"How does MUE relate to roofline (and other metrics)?"*

MUE is really including the schedule of operations, not simply the efficient use of the memory subsystem. It is thus a stronger but also more complex metric than the roofline model. We have added this comment in the revised manuscript.

(f) *"Is MUE really two metrics (ratios) rather than one?"*

The MUE metric is the product of two metrics that can be of interest in their own right, namely the I/O efficiency and the memory bandwidth efficiency. The I/O efficiency measures how close an application is to making only the minimal amount of necessary memory accesses. The memory bandwidth efficiency is a measure of how close an application is from using maximum achievable memory bandwidth. This is explained in Section 4 and results of the individual metrics for COSMO are presented in Section 5.5.

4. *"I also have a number of minor suggestions for changes mostly to do with improving the readability of the paper, which are given below"*

Thank you very much for your comments! We will address them in the revised manuscript.

(a) *"P13 L6: is there any benefit in using OpenMP, either to reduce comms, or for hyper- threading?"*

The GPU-enabled version does no longer contain OpenMP directives in the Fortran part of the code, we can thus not answer this question directly. Previous experiments with an OpenMP threaded version did not show any improvement in strong scalability with respect to a flat MPI version. No modifications to revised manuscript.

(b) *"P15 L4-9: Please explain this argument in more detail. It is unclear to me what the logic is to get to 5 times more power efficient."*

A lower-bound estimate (assuming their reported SYPD) and taking HPL power draws would give a factor 7. While it might be that the power draw when running their application is less than 51% of the HPL value, this is unlikely. To be on the safe side, we assumed no less than 35% of the HPL value. Comment added in revised manuscript.

(c) *"P16 L8: This conclusion is less obvious to me as you have made the assumption that the unoptimised version achieves peak bandwidth and that may not be the case."*

The bandwidth measurements of the non-optimized version confirm that we are very close to peak bandwidth. Comment added in revised manuscript. Comment added in revised manuscript.