We thank the second reviewer for his thorough reading of our manuscript and insightful comments that helped clarify our manuscript and strengthen our validation. For simplicity, we rewrote the reviewer's comment below (in black) and respond to them point-by-point (in blue).

———————————————————————————————————

This paper describes an huge amount of cutting-edge work. It is well written and I have only minor comments. Basically, I think it is acceptable as-is. That said, I was a bit disappointed with the paper.

Thanks. We understand the reviewer's comments and feeling, but we want to reiterate here that this is a model development paper to present our new ICON-Sapphire version. The goals were to document the model code, to show that km-scale global coupled simulations are technically feasible and to show to which extent basic features of the climate system can be reproduced by our current set-up (see lines 102-106 in the introduction). The reviewer raises many interesting scientific questions in his major comments, like the effect of small scales on larger scales, or the impact of the coupling, questions that our new ICON-Sapphire version would indeed allow to answer in the future. But before such questions can be answered, we need to have a reference paper that describes the model and its current capability, upon which future (and more interesting studies) can build upon.

Major Comments:
1. Coupled km-scale simulations unlock a lot of interesting questions, but the results shown here seemed more of the "we made pretty pictures" variety. Having written overview papers for new model releases myself, I can commiserate – these papers are a lot of work to write and hard to make interesting to read. That said, I had a couple of questions going into the paper which might be useful to reflect on:
As replied below in detail, we reflected upon the questions raised by the reviewer. Although we showed a couple of "pretty pictures" (current Figs. 12, 13, 16 out of the eighteen figures), we believe we also presented a comprehensive validation of basic features of the climate system, even considering internal variability in observations to better assess the skill of the presented one-year simulation, where the short integration period complicates the validation.

a. What features in the coupled system are improved by storm-resolving scales? Some features can't help but become realistic as they become explicitly resolved. Orographic precipitation is an obvious example. Getting these things right should fix classic problems in coupled models like dynamic vegetation die-off in the Amazon due to precipitation biases or ocean circulation biases due to incorrect bathymetry. Identifying classic biases which you expect your GSRM to get right, then checking whether this happens would be interesting.
As this is the first paper presenting our ICON-Sapphire version, and since there was already a lot of material to cover (description of the code, presentation of

different set-ups), we decided to concentrate on an evaluation of the representation of basic features of the climate system against observations. In that context, we presented a comprehensive validation against observations, checking basic climate variables and properties of the climate system, including internal variability in observations (in current Figs. 4 to 9, 11, 15, 18). Adding a comparison to low-resolution simulations to assess which features of the coupled system are improved by storm-resolving simulations would be one obvious next step, we agree, but this exceeds the frame of the current paper and is left for future work.

b. What features in the coupled system are NOT improved by storm-resolving scales? Do you have a sense for what the canonical problems of GSRMs will be? What did you struggle to get right in your simulations?
One delicate issue is to get the TOA energy budget right. First, some of the known tuning knobs from low-resolution climate models don't exist anymore and from those which still exist, the simulations often didn't react as expected. Second, trade wind cumuli play an important role in determining the TOA energy budget and getting them right can be challenging, especially in our case as we are trying to avoid having to use a shallow convective parameterization. The second delicate issue may be the coupling between the atmosphere and the ocean. The atmosphere turned out to be (at least to us) surprisingly sensitive to small difference in initial SSTs. Throughout the development phase, we had instances where the winds, e.g. in the tropical Atlantic, switched from easterlies to westerlies. This may also explain the difficulties of our simulation to capture precipitation at the equator. The last delicate issue seems ocean mixing, e.g. we have too shallow ocean mixed layers in the tropical Atlantic (see Fig. 11) and simply tuning the mixing coefficient in the TKE scheme didn't fully alleviate that problem. In this revised version, we first expanded the discussion about the TOA imbalance (see lines 450-461). Second, we also added some details concerning the issue with ocean mixing on lines 563-565. Third, we more explicitly mentioned in the goals (see lines 105-106) that our validation can shed light onto potential remaining shortcomings of GSRMs and accordingly partly rewrote the second paragraph of the conclusions (see lines 715-723) to more explicitly mention what we are not getting right and what the canonical problems of GSRMs might be given our experience with the development of ICON-Sapphire. Finally, we also added both in the abstract (see lines 9-10) as well as at the beginning of section 4 (see lines 393-394) that, although several features of the climate system are well captured, not every feature is well captured. We agree that our previous formulation was too positive as, as already described in the previous version, not every aspect of the climate system is well captured.

c. What do we get out of global coupled k-scale models that was missing from prescribed-SST runs or regional simulations? In this context, it would be nice to see prescribed-SST companion simulations and/or regional simulations.
This is a very important question, we agree with the reviewer, but it would justify a paper on its own.

2. I'm also disappointed by the discussion (particularly section 4.1.1) about:
a. Model tuning: 4 W/m2 is a huge imbalance. I'm confused why you didn't insist on tuning the model better before running these simulations. How can you hope to do multi-decadal simulations with such a large radiative imbalance? I don't expect you to redo the simulations, but acknowledging the problem and explaining why you didn't want to or couldn't tune the model would be interesting. Ignoring the issue leaves the reader feeling like they missed something.
We agree with the reviewer that the imbalance prevents performing long simulation as the model cools too much with time. The reasons for not insisting on tuning the model are twofolds. First we wanted to know what we get right more or less out of the box, just by trying to represent explicitly as much as we can from the climate system. Second, many scientific questions, e.g. dealing with the coupling between convection and SST, or effects of small-scale oceanic features on the large-scale circulation of the atmosphere, can be already tackled with a one-year simulation, and for the first time without being concerned that many of these features are the result of the design of a convective parameterization. We thus found important to let the community knows that such simulations already exist and can already be used for specific investigations. We added in the conclusions on lines 724-727 these considerations. Also we expanded the discussion about the TOA imbalance (see lines 450-461), now explaining our strategy to fix the energy imbalance (see reply to next comment).

b. Model drift: Fig 4 shows that TOA energy is adjusting rapidly over the course of this simulation. It would be nice to see a similar graphic for global-average surface temperature. Lack of discussion of what this drift means was conspicuously absent from the paper. Do you really think you could do a long simulation with this configuration? Do you have plans for fixing the drift?
As suggested by the reviewer, we added a panel of global-average surface temperature in Fig. 4. No we don't think that we can do a long simulation with this configuration (see our reply to previous comment), which we now acknowledge in the conclusions (lines 723-726). Yes we have a plan for fixing the drift. The TOA imbalance is mostly related to a preponderance of low clouds. These arise when using the Smagorinsky scheme, which has a mixing cutoff. Our formulation sets the eddy diffusivities to zero if the Richardson number is greater than the eddy Prandtl number. This cutoff unrealistically inhibits mixing, both because of well known limitations of the Smagorinsky scheme in simulating the transition to turbulence (Porté-Agel et al., 2000), and because of a failure to incorporate the effect of moist processes. As a result, over cold and moist surfaces, insufficient ventilation of the boundary layer occurs, causing moisture to build up and resulting in excessive low clouds. In ongoing experiments, we have explored adding a small amount of background mixing at interfaces between saturated and unsaturated layers where the equivalent potential temperature decreases upward, mimicking the effects of buoyancy reversal (Mellado, 2017). Low clouds respond sensitively to this background mixing, what provides a convenient control on their amount and on their influence on the top of the atmosphere energy budget. Ongoing work is

exploring theoretical justifications for the choice of the background mixing, but it may also be set empirically, as a way to provide a better representation of the statistics of low clouds. We added these considerations on lines 450-461.

3. I think a lot of the ocean analysis is naïve because it doesn't acknowledge that it takes the ocean a long time to drift away from its initial condition. Thus a lot of the analysis is probably more reflective of having an initial condition which looks like observations rather than that your ocean dynamics are working correctly. You can get a sense of initial condition versus equilibrated model bias by comparing the output from a coarse-resolution coupled run at initialization, after 1 year, and after 500 yrs. I bet most of your fields of interest look a lot more like the initial condition than the 500 yr value. I'm not sure this means you should throw out your ocean analysis, but I do think you need to clearly articulate the potential source of good skill.

We agree that we didn't acknowledge the slow dynamics of the ocean. We also agree that we cannot say at the outset if our ocean dynamics are working correctly given the length of the simulation. Having said this, first we think it is important to show that there is no obvious bias in the ocean state as, as already said, one could already investigates interesting scientific questions just based on one year of simulation. Despite the long experience of running ICON at low resolution in a coupled mode (Jungclaus et al., 2021), that the ocean works and couples correctly to the atmosphere is not given. In fact, several bugs were actually discovered during the development phase, bugs related to the momentum coupling between ocean and atmosphere. We thus added both the note of caution related to the slow ocean dynamics and the motivation for still validating the ocean at the beginning of section 4.1 on lines 423-431. Second the ocean went through a nearly 85-year spin-up (albeit with 10-km grid spacing) and another 10 years of spin-up at 5 km. Going through this long spin-up also helped identifying issues in the ocean model. We added this on line 416. Third, the statistics considered for the validation of the ocean were salinity (Fig. 7), coupling (Fig. 8), barotropic streamfunction and water transport (Fig. 9), wind work (Fig. 10). Except for Fig. 9, all the other statistics are statistics that should respond fast when coupled to the atmosphere. Looking at Fig. 7 gives us some confidence that the biases that we see are not just a reflection of the initial state. As discussed previously in the paper, salinity biases arise at the mouth of big rivers, because we neglect river discharge, and, in the tropics, are consistent with precipitation biases. Hence, being related to the absence of river discharge and to the simulated precipitation pattern, the pattern of the salinity bias in the tropics is distinct from the one at the end of the spin-up period. This is confirmed by Fig. R1 below where we show the salinity bias in the last full month of the spin-up (December 2019) and the same month in G_AO_5km (December 2020). Except for the salinity bias in the Arctic, the pattern looks distinct. To clarify this point, we updated the discussion of Fig. 7 in the text, now mentioning which biases are inherited from the spin-up and which not, see lines 517 and 521-523. We did a similar analysis for the barotropic streamfunction and water transport (Fig. 9) by comparing values from the spin-up and from G_AO_5km. G_AO_5km

systematically shows weaker transport except in the Bering Strait (see updated Table in Fig. 9) and these smaller values are out of one standard deviation in all passages but the Indonesian Throughflow and the Mozambique Channel. Except for the Florida Bahamas Strait, the weaker transport of G_AO_5km is in better agreement with observations. The weaker transport is consistent with weaker wind stress in G_AO_5km compared to the spin-up simulation, also expressed in a weaker barotropic streamfunction. These additional considerations have been added on lines 540-544 (together with the updated Fig. 9). Hence this supplementary analysis shows that the coupling leads to systematic differences to the uncoupled spin-up.

4. I'd like to hear more about conservation properties of the model. If I understand correctly, you've designed your schemes to have decent conservation properties and don't have a mass or energy fixer. I'd really like to see a plot of the unexplained global-average water and energy leak over time. This seems to me like it could be a huge problem for your multi-decadal simulation aspirations.
We don't have a water leak, water is conserved in the model, as mentioned on line 512. We have however several areas in the model, where energy is not conserved. The dynamical core unphysically extracts energy from the flow, at an amount of about 8 W m^{-2}, and precipitation is an unphysical source of energy. The former has been documented by Gassmann (2013). The latter arises because hydrometeors are assumed to have the temperature of the cell in which they are found. Because this assumption neglects the cooling of the air that accompanies the precipitation through a stratified atmosphere, it acts as an internal energy source, roughly (and coincidentally) of about equal magnitude to the dynamic sink. This explains why these energy leaks, which are also present in ICON-ESM, were not discovered previously. Moreover, minor energy leaks related to phase changes in the constant volume grid not conserving internal energy as well as to an inconsistent formulation of the turbulent fluxes have been discovered. Fixes for all of these problems have been identified and are being implemented. These considerations have been added on lines 462-470.

Grammar, spelling, and details:

1. ~L130: it seemed odd this paragraph doesn't include citations for readers to find out more about each of these component models, but then I realized that you go into a lot more detail about each component in following sections. If convenient, citations to the overview papers on each component model would be useful here. If not, please note that details about these models are given in section 2.
As the components don't have necessarily one key paper describing them, we prefer keeping the citations in the following sections, but we added at the beginning of this paragraph that details about the components are given in section 2 (see line 131).

2. L135: You say here that the atmosphere can only be run in uniform global or regional modes, but Fig 2 and elsewhere in the text talks about nesting, which seems like it uses several different resolutions in one run. The text also seems to imply that nesting is only available in regional simulations, which seems odd. Why can't you do nested regions inside a global run? Also, I think of telescoping as identical to nesting: you divide each tile of a coarse outer grid into finer but uniform grid cells and run a regional version of the model on this patch of fine-resolution cells. This fits with the idea of a telescope extending in a few discrete segments rather than continuously deforming to extend and retract. I think the way your ocean model works is that resolution is allowed to vary smoothly throughout the domain.
We see the nesting approach in the atmosphere and in the ocean as being different. As pointed out by the reviewer, in the ocean, the grid is allowed to be non-uniform and thus to vary smoothly throughout the domain in one simulation. This is not the case in the atmosphere where one simulation can only use a uniform grid and higher resolution is achieved by combining multiple simulations with distinct grid spacings, simulations which communicate through the boundaries of each respective domain. We clarified the text accordingly on lines 138-142.

3. Section 2.1: It would be good to mention in this section that aerosols are prescribed. You say this on line 303 in the I/O section, but readers will expect to hear about aerosol treatment in the atmos description.
We added this information on line 179.

4. Fig 3: Wow, this plot is cool – it has so much info. I'd prefer if dynamics and transport dots had their own line in the legend since closed and open dots obviously mean something uniquely different. Also, maybe add titles for the left and right columns of the legend since left is atm and right is ocean?
We split dynamics and transport on two lines in the legend (see new Fig. 3). We decided not to add titles for the left and right columns since in principle this information can be derived from the ICON-A, ICON-L, ICON-O labelling on the right of the plot, also given the fact that the figure is already quite busy with text.

5. Is land seen as just another atm process? It seemed odd that some components coupled via YAC but land doesn't. A sentence explaining why would be useful.
JSBACH is coupled implicitly to the atmosphere, and hence is tightly tied to it, something that doesn't work with YAC. The reason for using implicit coupling was that, at the time of development, many years ago, low-resolution simulations were in focus and there were stability concerns if using explicit coupling. For ICON-Sapphire, we are now working on rewriting the interface between the atmosphere and the land and are now using explicit coupling. In that case, one could actually use YAC, something that we might do in the future as this would allow JSBACH to be run on a different horizontal grid than the atmosphere. We don't want to go into all these details in the text, as those issues are not fully settled yet, but added that because of the implicit coupling, the land is not coupled via YAC (except for discharge), see line 135.

6. I thought Bjorn said at Pan-GASS that you have another turbulence option (Deardorff?) which was unintentionally acting as a shallow convection scheme. Is that worth mentioning here in the same spirit as you mention 2 moment microphysics and RTE-RRTMGP but don't use it?
We have indeed another turbulence option, which is the TTE scheme that we inherited from the ICON-ESM model. The TTE scheme was not active in the simulations presented in this paper. We are not mentioning the TTE scheme as we only would like to have one turbulence scheme in the future, and this will be a slightly modified version of the Smagorinsky scheme. In contrast, we will likely keep the two microphysics schemes (one moment and two moment). For the radiation, we are also only keeping one scheme, RTE-RRTMGP, but since we did the simulations with PSrad, we had to mention that scheme. We added a sentence in the text to make clear that PSrad will not be part of future releases of ICON-Sapphire, see lines 184-185.

7. Does ICON include horizontal turbulent mixing, or just vertical? It seems like horizontal mixing will be important at the hectometer scales you run at.
Yes, the reason for using Smagorinsky in place of the TTE scheme was that Smagorinsky also performs horizontal turbulent mixing. We added this information on lines 200-201.

8. L195: citation for Richtmyer and Morton numerical scheme?
The reference is Richtmyer and Morton, 1967: Difference methods for initial-value problems. We added the reference on line 212.

9. L246 – unclear what "latter" refers to.
The sentence was unclear. We rewrote it, see lines 253-254.

10. L278: What do "processes" refer to here? I think you mean the land model, atmosphere model, etc. I tend to think of these as "component models" with "processes" being particular physics schemes within a component... but that might be idiosyncratic of me. The concept of "neighboring processes" seems odd since processes have no spatial relationship to each other. I think you mean the process called before or after in sequential time splitting?
Here, we meant MPI processes and compute domains (local partitions of the horizontal grid) rather than physical processes. We rephrased this sentence to avoid misunderstandings, see lines 289-291.

11. L281: Doesn't the atmosphere just compute wind stress and provides that to whatever land model wants it? The way it's written, it sounds like the atmosphere. provides a different wind stress to sea ice versus ocean.
The atmosphere indeed provides a different wind stress over sea ice and over ocean. The atmosphere computes the wind stress for each of the surface tiles

separately. Although the velocity over ocean and sea ice is the same, the drag coefficient is not so that the wind stress is also different. We clarified the sentence, see lines 292-293.

12. L286: I think you should delete “and” between “wind” and “vectors”
We rephrased simply to “The interpolation of the wind is done”, see line 297.

13. L298: It would be handy to point out that 30” is equal to ~900m at the equator.
Added (line 309).

14. L325: “single-precision 32-bit float arrays are now kept in memory”. I think you mean that “output is now stored in single instead of double precision, reducing memory requirements by a factor of 2”. “kept in memory” sounds like you mean the data is kept in cache instead of slower-access disk and “single-precision 3d bit” is redundant.
Yes and no. The issue is that we store ICON output as 32-bit array, whereas CDO employs double precision for calculation. So before, when using CDO, we were transforming 32-bit data to double precision, which can significantly slow down the calculations for memory-intensive operations. Now we don’t do this transformation anymore. We updated the text to clarify, see lines 337-339.

15. L175 – I’m confused how you obtain good performance on GPUs if you use PSrad for all calculations in this paper and PSrad only runs on CPUs – I would have thought having some processes on CPU and others on GPU would result in excessive communication overhead and slow runs. Is radiation running in parallel with other atm processes? Or is it just that you call radiation so infrequently (every 15 min!) that it doesn’t matter? I suspect you are forced to run radiation so infrequently precisely because it is on CPU.
We apologize, this was confusing indeed. All the simulations presented in Table 1 use PSrad but the experiments performed by Giorgetta et al. (2022), where the performance on GPU was assessed, employed RTE-RRTMGP. We clarified the text, see lines 371-374.

16. L339: I think 40 TB/month is for *5 km dx*. It would be useful to point this out and to also say how much storage space you’re using for the 2.5 km grid (which I expect is 4x more = 120TB/month!).
Yes the 40 TB/month is for 5km dx and indeed, for the 2.5 km grid, we generated about 135 TB of output in a month. We added this on lines 349-351.

17. L369: “half productive” is bad grammar.
We rephrased to less productive, see line 379.

18. Adding another panel to Fig 4 showing the annual cycle of global-average surface T would be useful.
We agree and did so (see new Fig. 4).

19. L447: observations misspelled.
Corrected (line 494).

20. L452: I'm unclear how negative TOA radiative imbalance would lead to enhanced radiative cooling. First, what's your sign convention? Does negative radiative imbalance imply that the planet is losing heat? If so, I would think enhanced radiative cooling would cause the radiative imbalance. But TOA radiative imbalance could also be caused by an excessive planetary albedo.
We apologize, our formulation was confusing and incorrect. We wanted to say that the too large precipitation amounts are indicative of a too strong radiative cooling and a too strong radiative cooling would be consistent with a negative TOA radiative imbalance. We clarified the text, see lines 497-499.

21. Fig 11: caption skips panels e and g.
We reformulated the caption to correct this (see new caption Fig. 11).

22. Fig 12: panel a and e seem redundant.
We don't think so as through the shading of the rain in panel e, the small-scale structure of the salinity field is hard to recognize in panel e whereas it is nicely visible in panel a.

23. L554: "my" should be "may"
Corrected (line 606).

24. Fig 15: I thought sea breeze was a weak example of 2.5 km resolution since it is also captured pretty well at 25 km resolution. Also, this graphic would be a lot better using wind vectors rather than colors for just zonal wind.
We agree and since we already have many figures, we decided to remove this figure and just now shortly mention in the text that ICON-Sapphire can capture mesoscale circulations and their effects on convection (see lines 600-601).

25. L586: "latter" rather than "later"?
Corrected (line 639).

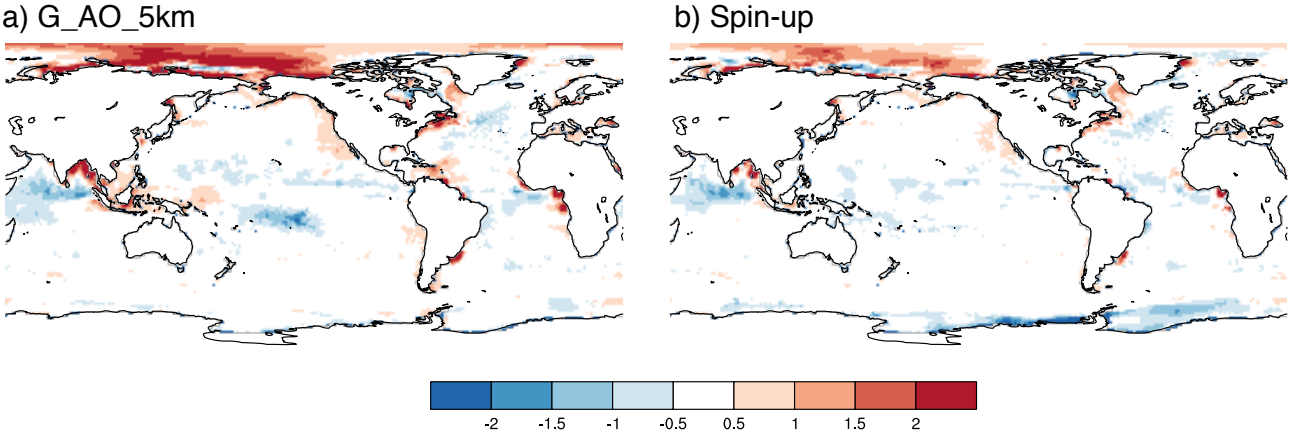

Fig. R1: Monthly mean (December) salinity bias (g kg^{-1}) in G_AO_5km and in the spin-up ocean simulation. Observations from PHC climatology.