Reviewer #2

This paper presents an opinion on how climate models can be significantly improved by merging traditional model development (improved parametrizations and higher resolution) with AI techniques. I think it is broadly in line with what many operational centres are already trying to do, i.e. augment models with AI instead of replacing them entirely. It is nonetheless valuable to document this in the literature for those who aren't already involved, and therefore I think the paper is an important contribution. It's also well written and I enjoyed reading it.

Thank you for the positive assessment and helpful comments, which led to a number of clarifications in the revision. (However, we are not aware of any operational center that is actually pursuing the approach we advocate, to learn about parameterizations, including functions, from data.)

I have several comments below, which I think the author should consider in producing a revised manuscript.

L56 - typo - "climate Turing test"

Fixed.

L85 - suggest re-wording this slightly - even at kilometer scale, most parametrizations (radiation, cloud macrophysics, cloud microphysics, turbulence, shallow convection, orographic drag) are still required - in fact it's only really deep convection which can plausibly be removed at this scale!

We have amended the sentence in question to read, “More recently, there have been calls to prioritize resolution increase, aiming to achieve kilometer-scale resolutions in the horizontal, with the expectation that this would alleviate the need for subgrid-scale process parameterizations, such as those for deep convection, and substantially increase the reliability of climate predictions.”

L105-108 - similar to TOA, the surface precipitation is also required to be accurate to close the water budget, which is critically important for the long-term behaviour of climate models, especially in fully-coupled Earth-systems. Might be worth mentioning this.

We now mention that “precipitation rates are of significant importance as they are part of what closes the water balance” (l. 115). (However, while a closed water balance is helpful, it is less important than a closed energy balance, given that oceans can serve as a large water

reservoir. Coupled climate models that do not conserve water have been successfully used for decades.)

L195 - this is quite a controversial statement that I'm not sure many people would agree with (I certainly don't) - progress undoubtedly is being made, cloud and convection schemes now are measurably better than 10 or 20 years ago when the cited papers were published. I would suggest rephrasing this - what I think is true to say is that progress is not as quick as we would like. The following paragraphs discuss the method by which most groups are making progress in this area (and have been for many years), so it feels slightly disingenuous to present these as new solutions to the problem, when I think they have been known as the solutions by parametrization developers for many years. The issue is that doing as suggested is hard, hence takes a long time.

Our statement is just that “the process-based approach to modeling convection and clouds is *widely perceived* as being deadlocked” (emphasis added, and we have added a more recent reference in the revision). This still seems true to us, even though, of course, we are subsequently making clear that we should reboot, not give up on, the process-based approach.

L252-256 - whilst I agree with the point being made here about developing scale-aware parametrizations, I'm not really sure the model results presented in Figures 1 & 2 are really a compelling argument for it. All major NWP centres run kilometer scale models without convective parametrizations and see huge improvements in NWP skill from doing so. Super-parametrized climate models have similarly shown measurable improvements in model skill relative to traditional parametrizations. Therefore to state on the basis of one bad model that this "approach has not achieved the anticipated success" seems like cherry picking to support the argument, when actually the weight of evidence shows that turning off the deep convective parametrization is better than including it, although could undoubtedly be improved further by scale-aware approaches. I suggest rephrasing this.

We agree that higher resolution has led to improvements in NWP. However, this does not necessarily translate into improvements in climate simulations, which are our focus. Figure 1 summarizes the current state of the art with respect to resolution benefits in climate simulations by showing the only two km-scale simulations run for at least 4 years (that we know of) to publicly release their output (in addition to two single-model resolution hierarchies from HighResMIP). This is an update compared to the previous version, that showed the previous cycle of nextGEMS, with 1 and 2 years of data instead of 4 and 5. The

new version of the figure paints these models in a slightly more positive light, but still does not show large improvements compared to the lower-resolution models.

In section 2, in the context of the discussion of Figure 1 to which the statement in question here refers, we have added: “In numerical weather prediction, enhanced horizontal resolution has led to improvements, for example, in rainfall predictions on timescales from hours to days (Clark et al., 2016). However, whereas assimilation of data at the initialization of a forecast continuously pulls numerical weather predictions close to the climate attractor, long-term climate simulations require a realistically closed energy balance to remain on the climate attractor. This balance also depends on dynamics at scales well below 1 km (e.g., in tropical low clouds, which are crucial for climate but less important for weather prediction), making it less clear that increased resolution by itself results in better climate simulations.”

Sect 4 - whilst I agree with what is being said here, it also neglects that there are important aspects of the climate which are not driven by small scale turbulence, e.g. land-sea contrasts, orography, land-surface type, SST pattern. These aspects become better resolved at higher resolution, and in turn leading to improvements in climate model skill which are not simply governed by the energy spectra. This would be worth mentioning.

In the final paragraph of section 4 summarizing the resolution discussion, we added: “While increasing resolution helps by gradually improving the resolution of turbulent dynamics and better resolving surface topography, gravity waves, and land-sea contrasts, the 1000-fold increase in the computational cost in going from O(10 km) to O(1 km) is unlikely to justify the benefits (Wedi et al., 2020).”

L337-341 - this statement is the one that worries me most in the paper. The whole point of emergent constraints is that they are emergent, i.e. they appear in climate models not because they have been programmed to be there, but because they arise as a function of the underlying model physics leading to their emergence in the same way as it does in reality. As soon as we start to pre-program emergent constraints into the model, they lose all meaning and usefulness. It may give the model a better skill when compared to past observations, but this is no guarantee of future success, since we cannot know how the emergent constraint will evolve in a changing climate.

If higher-order statistics such as covariances between SST and cloud cover are informative about a model's response to climate change, our view is that they should be used in model calibration, to obtain the best possible model and to quantify its uncertainties, in much the same way that seasonal temperature changes (and other first-order statistics) are generally used. We added footnote 3, stating, "If emergent constraint statistics are used during loss minimization, they can no longer serve as retrospective constraints on the response of the model to perturbations…"