We would like to thank the referee for their very helpful and constructive feedback. They have identified some areas where we are able to greatly improve the review. Their suggestions for figures and tables, and for future directions were particularly useful.

We have responded inline to the referee's comments in blue font below.

Anonymous Referee #2

In "Machine Learning for numerical weather and climate modelling: a review," de Burgh-Day and Leeuwenburg provide a review–aimed at weather/climate model developers–of machine learning itself, the history of its application in weather/climate modeling, and contemporary uses (and challenges) including: parameterization replacement, coarse-graining, superparameterization, fluid dynamics solvers, and others.

Overall the review is written clearly and written well for the intended audience, it is comprehensive with respect to the ML literature associated with weather & climate modeling, and it correctly and adequately summarizes the relevant literature. Overall I think this will be an invaluable addition to the literature, complementing some other recent reviews in the ML+weather/climate literature.

We thank the referee for their kind words. We are heartened that they think this review will be of value and have found their feedback and suggestions to be immensely useful. They identified some gaps in the review which we will work to address before we upload our revised version.

The three main places where, in my opinion, the review could use some improvement are: (1) addition of graphics/diagrams/tables/pseudocode/anything-to-please-break-up-the-text, (2) filling some gaps in the review, and (3) more synthesis. I also have some other minor improvements to suggest. Detailed comments follow.

*Note that two students also read this paper and provided input on the review.

Thank you for these suggestions. We have responded to each of them inline below.

Major feedback

Adding figures, diagrams, tables, pseudocode, etc.

The stated goal of the manuscript, "to provide a primer for researchers and model developers to rapidly familiarize and update themselves with the world of ML in the context of weather and climate models," would be better-served if visual aids of some form were added to the manuscript, particularly in Sections 2 and 8–10, and possibly also elsewhere.

I'd also add that the current version of the paper is a lot of text without any interruption; I found that this made it difficult for me to hold my attention on the paper.

We agree that this would improve the visual appeal of the review, and note that the other referee made the same suggestion. We have identified some areas where useful visualizations can be added, although we feel that there may be a limit to the number of visualizations we can add that would be worthwhile and wouldn't just lead to the review becoming even longer. Thank you for the suggestions of visualizations for different sections – this is a very helpful guide for us to use in what we could add that would be of value.

Some specific suggestions for figures/diagrams/etc. follow.

Section 2 Readers who are totally unfamiliar with machine learning may find it challenging to derive meaning from the text-only descriptions given in Section 2. I'm not necessarily advocating for yet another elementary neural network diagram, like one could find on wikipedia, but rather something that will help this specific audience–model developers–form a reasonably good mental model of neural networks, decision trees, and the various architectures of them. This could be in the form of a diagram, but this audience also might find it simpler to digest some pseudocode: e.g., pseudocode describing a neuron as a function, or pseudocode describing how a convolutional layer works. But then again, maybe a diagram would be better.

A visual aid along the lines of what you describe is a good idea – thank you. We are developing a flowchart-style diagram which will provide readers with a framework for deciding which ML architectures and algorithms would be good candidates to try for a given problem. We are hoping to include in it considerations such as which of the simpler but often quite effective algorithms they should try initially, and then which of the more complex ML algorithms and architectures they may want to move on to, based on the nature of their problem (e.g. is it temporal, spatial, or spatiotemporal in nature?).

We have also included a brief table showing the strengths and weaknesses of the ML algorithms and architectures mentioned in this review, with a reference to a paper with another similar table in it as well.

Sections 3–7 It could be useful to have a figure somewhere that gives a pie chart of the various ML/weather+climate modeling topics (e.g., how many papers, relatively, are in each of the categories outlined in the various sections.) Also, if any of the papers here are the authors's own work, then perhaps it wouldn't be to difficult to add a variant of a figure that already exists in the literature.

Also if any of the sources are published with a Creative Commons license, like many in GMD, then it should be acceptable to actually take figures from those papers as long as they are properly attributed following the guidelines. This could be a really simple way to break up the text and help the readers get a deeper glimpse into the work that has been done.

We have added a figure showing the breakdown of ML architectures in the papers referenced in this review (we had actually already started working on a figure like this before this feedback was posted, so it was pleasing to see this as a suggested visualization!)

Section 8 It would be great to add a graphically-rich timeline that complements the list in section 8.1. I could imagine future authors using that timeline in ML presentations, which would be a great way to get free advertising for this paper.

We think this is a really good idea, but are unsure whether we will be able to make such a timeline that will be sufficiently compact. We'll give it a go!

Section 9 Maybe I don't have an idea for a figure in this section after all.

Section 10 Consider adding a table of new advances in ML that haven't yet been employed in weather/climate modeling, but that may be useful. Also/alternatively, consider adding a figure that somehow communicates the promising new directions. It could be something as simple as a PowerPoint SmartArt that simply adds some graphical elements to highlight the text.

We have added a new section (section 11) titled “Future research directions” which summarises some potential focuses for future development and research, including identifying ML advances which haven’t yet been tested for weather and climate modelling applications. We are leaning away from including a figure, because there is too much detail to fit in a figure, but instead we plan to include a table summarizing possible future research directions.

Some gaps

We have deliberately kept the scope of the review very narrow, since without doing so we would have risked this manuscript turning into a book. We have addressed each of the suggested additions inline below.

Data Assimilation I’ve come across some literature on ML and data assimilation (some even in GMD), but the authors don’t discuss this at all here. Given that data assimilation is a critical component of weather modeling efforts, it would be a shame to overlook this. I suggest that the authors survey the literature on this.

We decided that use of ML in data assimilation is too large a topic to cover in this review – it deserves a review in its own right. We have instead focused on the step that comes afterwards – the prediction of subsequent states after the DA is complete, specifically on weather, subseasonal and seasonal, to climate prediction timescales. This is the same reason as the one that led us to not include nowcasting or to delve deeply into the use of machine learning for climate projections. Essentially, including these topics would expand the scope and length of this review beyond what is reasonable. We do, however, strongly agree with the referee that there is value in a review covering these topics and hope to see this emerge from the community in the future.

Ice sheet modeling Unless I overlooked it, this review doesn’t discuss ice sheet modeling at all, which is an increasingly important component of CMIP-class models. A quick google scholar search on ‘ice sheet model machine learning’ turns up some apparently relevant results.

This is certainly an omission on our part – we will explore the literature in this space and add it into the review.

Integrated assessment modeling / multi-sector dynamics In the lifecycle of CMIP model efforts, generation of climate scenarios (like the SSPs) is a key step. This isn’t discussed at all. There are at some efforts in this area that are worth mentioning, and I’d guess there are others, e.g: https://www.osti.gov/biblio/1769796

For similar reasons to those stated above, we have made an explicit decision to not explore too deeply the topic of climate change projections and climate scenarios, and have instead stopped at multiyear and free-running simulations. We are aware that there is a very large body of literature on the application of ML to aspects of climate scenario modelling and as stated for previous topics, we feel that this topic is better suited to its own review. We do note however that the distinction between multiyear and free-running simulations, and climate projections/scenarios, was not made explicit, and the fact that one is kept in scope while the other isn’t is not made clear.

We have added some text to the introduction to make this more clear:

"Additionally, here we consider climate modelling in the context of multiyear and free-running multidecadal simulations, but exclude the topic of ML for climate change projections, climate scenarios, and multi-sector dynamics. This is again in the interests of ensuring the scope of the review is manageable, rather than because these topics are not worthy of review. On the contrary, a review dedicated to the utility of machine learning in this area would be of enormous value to the community, but could not be done proper justice to here."

AI Ethics AI advances have ethical implications, and I think there might be some here too. It might be worth surveying some of the recent literature on ethics in AI, with a goal of summarizing the main ethical issues that come up with AI in general and what the implications of these ethical issues might be for AI in weather and climate model development.

This is a good point – the ethics of AI is indeed an important consideration. We don't feel that a full review of the ethical consideration of AI more generally is appropriate for this review, however, we do feel that it is very important for readers of this review to be made aware that there are ethical considerations associated with the use of AI. We have added a new section (section 10) discussing the ethical considerations of AI in its application to weather and climate modelling to the review, including references for readers interested in exploring the topic further.

Synthesis

In my perspective, the most impactful review articles are ones that (a) provide a comprehensive overview of the state of the literature (which this paper does quite well), and (b) synthesize what the authors have learned: even suggesting new directions that might not be immediately evident. The current version of the manuscript does great on (a) but does not do too much with respect to (b). The Conclusions section does this to some extent (e.g., "Nonetheless, there are still many challenges to overcome…This list provides a set of focus areas for future research efforts."), but the list focuses on challenges rather than promising new directions. The last paragraph starts to get at this with the sentence "Advances in the sophistication, complexity and efficiency of ML architectures are being heavily invested in…," but the manuscript then stops short of discussing these new advances or how they might point to new directions.

I recommend revamping the last section to focus on this synthesis aspect.

I'm not in the best position to give good suggestions here since I don't have as comprehensive of a knowledge of this literature as the reviewers, but after reading the paper, some untouched directions do come to mind:

- More exploration of foundation models. The authors note one recent example of a foundation model. The proliferation of foundation models in the last year (ChatGPT, for example) has this at the forefront of a lot of people's minds: what new research could contribute to the application/analysis/use of foundation models in weather and climate?
- Relatedly, it could be impactful to somehow fuse weather / climate code and data with GPT-like models. What sort of impact would it have if a model developer user could get insight from an AI model that's able to ingest and interpret high-dimensional data as well as code: e.g., "WxGPT,

why does the new change in commit 3efde6 result in a systematic cold bias in daytime maximum temperature forecasts?”

- Model emulation / tuning. There’s some literature on groups using Gaussian process models to emulate climate models, where they use the emulators for quantifying uncertainty in tuning parameters and for finding optimal tunings; other ML methods could be useful here
- Model spinup: a major barrier to use of ultra-high-resolution coupled climate models is the time required to spin-up the slow components of the system like ocean and land ice; ML methods could potentially be useful here (e.g., for learning how to translate equilibrated states from a low-res model to a high-res model)
- 3D radiative transfer for high-res models: possibly replacing a full 3D radiative transfer code with an ML approximation, or perhaps using the climate model for 1D radiative transfer and using ML to model the expected differences in fluxes due to 3D effects
- Modeling full PDFs: cutting-edge models like FourCastNET essentially emulate what dynamical models do in that they provide deterministic (albeit presumably chaotic) states. What if instead they could be trained to output a PDF of states (e.g., emulating Fokker-Planck equations) rather than deterministic states? That would be something fundamentally new relative to existing model capabilities.

My main point here is that your review already has a lot of value in establishing what has already been done, and I think this paper will be more impactful if you increase the emphasis on what could plausibly be done that has not yet been touched. I recommend thinking 5-10 years into the future rather than just incremental advances based on what’s been done. You have the unique opportunity to inspire others to try some radical new ideas.

Also, consider that this review will very likely be cited in workshop reports that inform funding agency priorities. Your last sentence states ‘academic and operational agencies will need to continue to support research in this space;’ giving specific ideas here could really have an impact.

This is a very good point and a good suggestion – thank you. We have modified the final section to provide more synthesis of the outcomes of the review. We are also adding a new section (section 11) to discuss future research directions.

Finally, I recommend also mentioning AI ethics in this last section. If you’re going to inspire researchers to think radically, it would be responsible to also admonish people to always consider the ethical issues with as much mental effort as they do the technical issues. I can’t help but quote from Jurassic Park here: “Your scientists were so preoccupied with whether or not they could, they didn’t stop to think if they should.”

Please see above for our thoughts on discussing AI ethics in this review. We agree that there is a strong ethical risk in the use of AI if it is not done with care, and we have added a new section (section 10) to emphasize this (as detailed above).

Minor feedback

The NVIDIA group has just put up a preprint of the newest version of FourCastNet, which allows them to perform year-length simulations: https://arxiv.org/abs/2306.03838

Thank you for drawing our attention to this. We are aware of a number of relevant publications which have come out since submitting this version of this review, for example the preprint for FengWu (https://arxiv.org/abs/2304.02948), for SwinRDM (https://arxiv.org/pdf/2306.03110.pdf), and several interesting papers making progress in the use of ML for atmospheric parameterization schemes (e.g. https://gmd.copernicus.org/articles/16/2355/2023/).

We will need to assess the latest additions to the literature since the initial submission of this review and decide whether it is worthwhile adding them in, or leaving them to a future update. Our argument for this approach is that if there are only a limited number of relevant additions, it could be reasonable to include them here, but with the field moving so fast we are worried about getting stuck in a situation where adding them in is sufficient to trigger a new round of review, which in turn sets us back long enough for more new additions to be made to the literature, and so on. Since we anticipate a continued high rate of publication in this area, additional review papers in the future to synthesize these new papers will be warranted, with potentially more focused reviews of the most promising research directions.

(lines 8, 9) Please be consistent about the spelling of parameteriz(s)ation

Fixed

In section 1, there are no references at all, even though there are numerous statements that would normally warrant references; this was quite distracting until I understood why. I now understand why this was done, since the references for those statements are given extensively in the sections that follow. My suggestion would be to make a statement early on in the introduction that states something like “In this introduction, we overview the state of machine learning in weather and climate research without providing references; we instead provide references for these statements in the detailed sections that follow.”

This is a good suggestion, and we have added words to this effect after the first paragraph of the introduction: “In the remainder of this introduction, we overview the state of machine learning in weather and climate research without always providing references; we instead provide relevant references in the detailed sections that follow.”

(line 20) “numerical weather and climate forecasts..are not amenable to transfer to specialized compute resources such as GPUs” … I’m not sure that’s strictly true. There’s been quite a lot of effort to refactor and port major codes to GPUs, demonstrating that it can be done in principle (for example consider the US Department of Energy’s E3SM / SCREAM model, which has a dynamical core that now runs on GPU; https://climatemodeling.science.energy.gov/technical-highlights/simple-cloud-resolving-e3sm-atmosphere-model-scream). It might be more accurate to say that it requires person-decades of effort to transfer these codes to GPUs.

The first referee made a similar observation, and we have amended the text to soften it and acknowledge that it is doable, albeit hard: “These numerical weather and climate forecasts are computationally costly and are not easy to implement on specialized compute resources such as GPUs (although there are efforts underway to do so, for example in LFRic (Adams et al. 2019)).”

Adams, S. V., Ford, R. W., Hambley, M., Hobson, J. M., Kavčič, I., Maynard, C. M., ... & Wong, R. (2019). LFRic: Meeting the challenges of scalability and performance portability in Weather and Climate models. *Journal of Parallel and Distributed Computing*, *132*, 383-396.

(line 24) “improve the representation of sub grid-scale processes…a computationally costly exercise” <– this isn’t necessarily true. Yes, for something like boundary layer turbulence or aerosol physics, modeling higher order moments or doing bin microphysics is more costly. But for something like convection, improvements could come simply through better physics-based theories about how convection works.

The first referee also commented on this. We acknowledge that this is a fair point, and we have modified the text to account for this: “An additional pathway to improve skill is to improve the understanding and representation of sub grid-scale processes, however this is again a potentially computationally costly exercise.”

(lines 106-107) “Furthermore, in many cases…the work is led by data scientists and ML researchers with limited expertise in weather and climate model evaluation” <– this wording risks alienating and insulting colleagues who have done work in this area who in fact have extensive expertise in weather and climate modeling. For example, consider research cited in this paper from the groups of Libby Barnes, Mike Pritchard, and Chris Bretherton – all three of them are definitively experts in weather and climate model evaluation. I suggest revising “in many cases” to “in some cases”.

This is a very good point, and we certainly do not want to undermine the expertise of those contributing to this field who do have expertise in weather and climate model evaluation. We have amended this as you suggest: “Furthermore, in some cases of ML approaches... the work is led by data scientists and ML researchers with limited expertise in weather and climate model evaluation.”

(line 113) “A review of the application of, and progress in, ML in these areas would be of great value…” <– FYI, a review paper by Maria Molina was just accepted in the AMS journal AI4ES, titled “A Review of Recent and Emerging Machine Learning Applications for Climate Variability and Weather Phenomena.” If her paper appears online before this manuscript is finalized, I recommend citing it here. Full disclosure: I’m one of the authors of that paper.

Thank you for drawing this to our attention – it has just appeared in our alerts, and we have included it in the introduction:

“Molina et al. (2023) have provided a very useful review of ML for climate variability and extremes which is highly complementary to this review. They draw similar lines of delineation in the earth system modelling (ESM) value chain to those mentioned above; describing them as “initializing the ESM, running the ESM, and postprocessing ESM output”. They examine each of these steps in turn, with a focus on the prediction of climate variability and extremes. Here we take a different approach, focusing on one part of the value chain (running the ESM), but looking in more detail at this one part."

It looks like a really useful resource, and it complements this review well.

(line 238) Was GCM defined as an acronym before this? (It's defined later on line 253)

Oops, no it wasn't. Thank you for picking up on this. We have moved the definition to the first instance of "GCM" being used (and removed the definition from further down).

throughout the paper There is some odd formatting in the footnotes…they should probably be superscripts. Likewise, the dagger symbol, that indicates a vocab word defined in the glossary, should consistently be a superscript (sometimes it isn't.)

We have attempted to resolve the odd formatting issues. MS Word has been doing some strange things during conversion to PDF but we will take extra care to resolve these issues before our final submission. Hopefully anything we miss will be resolved by the journal's editorial team!

one of my students comments, and I agree: "Personally speaking, I found the paper's pace a bit choppy at times. For example, subsections 3.6-3.9, 5.1, 5.6, and 7.1-7.2 are only a paragraph long. Especially when these occurred back-to-back, the paper felt very "stop go stop go stop go", made even worse when the sequential subsections had little to do with each other. I'm not entirely sure of a solution here, but I wish the authors could find a way to make these subsections flow more together, or at least give us a bit more time with them. It's hard for me to digest their information when each paragraph is immediately moving on to something almost completely different. This could totally just be a me-thing though."

We received similar feedback from the first reviewer, and attempted to address the sense of the sections not flowing from each other with a bit more of an explanation of the logical flow in the introduction. That said, we aren't sure how to address it better than that either – the feeling of choppiness in the short subsections in Section 7 we would argue is due to the topics being touched on having a lot of depth that isn't covered. These subsections are intended to be tasters of these areas which an interested reader could explore further by following the references. To give a better sense of flow between subsections here would probably require expanding them significantly, which we don't feel is feasible for such an already long review. Hopefully the sense of choppiness isn't too off-putting.

Section 7.2: There's a bit more work in this area than just Mudigonda et al. (2017). Here are a few additional relevant papers (again full disclosure: I'm a co-author on two of these):

Prabhat, Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., et al. (2021). ClimateNet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. Geoscientific Model Development, 14(1), 107–124. https://doi.org/10.5194/gmd-14-107-2021

O'Brien, T. A., Risser, M. D., Loring, B., Elbashandy, A. A., Krishnan, H., Johnson, J., et al. (2020). Detection of atmospheric rivers with inline uncertainty quantification: TECA-BARD v1.0.1. Geoscientific Model Development, 13(12), 6131–6148. https://doi.org/10.5194/gmd-13-6131-2020

Rupe, A., Kashinath, K., Kumar, N., Crutchfield, J. (2023). Physics-Informed Representation Learning for Emergent Organization in Complex Dynamical Systems. arXiv. https://doi.org/10.48550/arXiv.2304.12586

https://ai4earthscience.github.io/neurips-2020-workshop/papers/ai4earth_neurips_2020_55.pdf

Thank you for pointing these out. It's clear from reading these papers that the area of extreme event identification is one where you have a great deal more expertise than we do!

We have extended the section on object detection (now section 7.4) significantly to include a brief description of each of these papers and have updated the section to take the extra information into account.

We also moved the sub-section further down in section 7, as it seemed to flow more logically that way.