We would like to thank the referee for their very helpful and constructive feedback. We feel that their advice has very much improved the quality of the review, especially relating to the Introduction and Subsection 5.3.

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

Anonymous Referee #1

General comments:

The authors review the application of Machine Learning (ML) techniques to weather and climate modelling with an emphasis on historical and current developments. A glossary of commonly used terms and basic introductions to some concepts are provided. An in-depth exchange of knowledge between the ML and geoscientific modelling communities could be immensely beneficial and reviews like this can be an important step to facilitate this exchange.

As far as I am able to judge, the authors do a great job at covering a wide range of relevant publications and explaining the questions tackled in many of these applications. In terms of presentation, the language is concise and the paper is enjoyable to read. Including tabular or schematic representations of ML concepts and/or the discussed applications could further improve the visual appeal of the paper. A stronger narrative thread linking the different subsections and applications would preempt the impression of reading through a long list of papers - although this may be unavoidable given the scope of the reviewed works.

We thank the referee for their kind words. We did indeed endeavor to have a narrative thread through the sections, however as the referee notes, this was somewhat challenging to do given the wide scope of the review. We have reworded the final paragraph of the introduction (summarizing the remaining sections) to try to clarify the logic of the order of the sections somewhat.

We also agree with the referee that some figures or tables would add to the visual appeal of the review. We will explore some relevant visualizations and add them in if we feel they increase the visual appeal and information content of the review.

My primary concern about the paper in its current form is its utility to aid researchers in the development of better geoscientific models. Due to the wide range of works that are being discussed, many concepts and models are only touched upon in brief, without further elaboration of the underlying principles and connections between different applications. References to methodological works that could support future model development are only sparsely included in the main text or the glossary. In my opinion, incorporating suitable methodological references into the glossary and introductory sections could greatly strengthen the paper!

This is a very good suggestion. We have added more explanation of a selection of good foundational papers and books in Section 2:

"Suggested starting points for interested readers, including guidance on the utility of different model architectures and algorithms, and the connections between different applications and approaches, are as follows:

- Hsieh (2023) provides a thorough textbook on environmental data science including statistics and machine learning
- Chase et al (2022a, 2022b) provide an introduction to various machine learning algorithms with worked examples in a tutorial format and an excellent on-ramp to ML for weather and climate modelling
- Russell & Norvig (2021) provide a comprehensive book regarding artificial intelligence in general
- Goodfellow et al. (2016) provide a well-regarded book on deep learning theory and modern practise
- Hastie et al. (2009) provide a book on statistics and machine learning theory"

While we think that a review which focused on the relative strengths and weaknesses of each ML algorithm and architecture would be of great value, it is not the primary focus of this review. This review seeks to provide an overview of the major developments in the research around ML for weather and climate modeling. To also include an exploration of the methodological strengths and weaknesses of different architectures and algorithms would, we believe, significantly increase the size of an already very long manuscript. We have thus left this to future work, and would view this review as being complimentary to one exploring the methodological aspects of ML for weather and climate modelling in more detail.

Specific comments:

L20 - Isn't there an ongoing research effort to extend numerical models to utilise GPU hardware?

This is true, however it is still not an easy task. We have amended this sentence to clarify that it is doable but difficult:

"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))."

The fact work is underway to make numerical models able to run on GPUs is also acknowledged in Section 7.4

L24 - What about improvements in subgrid parameterizations due to better process understanding?

We have amended this sentence to include 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."

L65/66 - Maybe include a reference? (e.g. McGovern et al 2019 [1])

Good suggestion – some references added: "(e.g., McGovern et al., 2019; Toms et al., 2020; Samek et al., 2021)"

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. Bulletin of the American Meteorological Society, 100(11), 2175-2199.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247-278.

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. Journal of Advances in Modeling Earth Systems, 12(9), e2019MS002002.

L104 - Very debatable if this is a necessary requirement for e.g. a weather prediction model?

It's not necessary if other items in the list are satisfied, but the list specifies "one or more of". We would suggest that if the model didn't provide any of the other benefits in the list, it would have to at least provide the last item in the list; insight into physical processes not provided by current numerical models or theory

L116ff - A narrative thread linking these subsections would be much appreciated!

We have reworded this paragraph to try to illustrate more of a narrative thread through the sections:

"The remainder of this review is structured as follows: In Section 2 a quick introduction to ML is provided, before the application of ML in weather and climate modelling is explored in the following five sections. Firstly, ML use in sub-grid parametrization and emulation, along with tools and challenges specific to this domain, are covered in Section 3. Zooming out from sub-grid scale to processes resolved on the model grid, in Section 4 the application of ML for the partial differential equations governing fluid flow is reviewed. Expanding scope yet again to consider the entire system, the use of ML for full model replacement or emulation is reviewed in Section 5. In Section 6 the growing field of physics constrained ML models is introduced, and in Section 7 a number of topics tangential to the main focus of this review are briefly mentioned. Setting the work covered in the previous sections in a broader context, a review of the history of, and progress in, ML outside of the fields of weather and climate science is presented in Section 8. In Section 9 some practical considerations for the integration of ML innovations into operational and climate models are discussed, and finally a summary is presented in Section 10. A Glossary of Terms is provided after the final Section to aid the reader in their understanding of key concepts and words."

L128 - Could it be more useful to briefly discuss the utility of the individual references rather than providing a large list?

As was already mentioned above, we have expanded briefly on the topics covered by each of the references to help guide the reader to the references most useful to them:

"Suggested starting points for interested readers, including guidance on the utility of different model architectures and algorithms, and the connections between different applications and approaches, are as follows:

- Hsieh (2023) provides a thorough textbook on environmental data science including statistics and machine learning
- Chase et al (2022a, 2022b) provide an introduction to various machine learning algorithms with worked examples in a tutorial format and an excellent on-ramp to ML for weather and climate modelling
- Russell & Norvig (2021) provide a comprehensive book regarding artificial intelligence in general
- Goodfellow et al. (2016) provide a well-regarded book on deep learning theory and modern practise
- Hastie et al. (2009) provide a book on statistics and machine learning theory"

L136 - Debatable, as recent trends in ML point strongly in the opposite direction (i.e. larger homogenised models).

We don't necessarily agree with the assertion of the referee, however we do acknowledge that it is a sufficiently nuanced and debatable premise (both in terms of our claim and the counterclaim made by the referee) that it is not suited to a single sentence at the end of the paragraph. It is not an essential point to resolve for this review, so in the interests of brevity we have simply removed the sentence.

L145 - Debatable as emphasis shifts towards self-supervised learning and better training regimes rather than architectural developments!

We do not entirely agree with the referee on this point, however we do acknowledge that alongside architectural/algorithmic improvements (e.g., Earthformer), gains have been made through improved training regimes (e.g. FengWu and assorted parametrization scheme examples). We also realize that this sentence was somewhat ambiguous in that it wasn't clear as to whether it was referring to ML in the context of weather and climate modelling, or ML in a more general context. We do acknowledge that in a boarder context there is a relatively greater focus on unsupervised learning currently. We have amended this sentence to acknowledge the other areas where research is currently focused in the weather and climate modelling context:

"A major current focus of ML research in the context of weather and climate modelling is new NN-based architectures and algorithms, and improved training regimes."

L156 - It could be important to emphasise that NN are known to interpolate within the training envelope and may not generalise well outside it (in contrast to physical laws).

We agree that this is important to point out. We have mentioned it later on in discussing applications of ML (especially for parametrization schemes) but agree that it is worth mentioning here too. We have amended this sentence to state: "NNs can therefore theoretically be candidates for accurate modelling of physical processes, although in practise they cannot always reliably predict beyond their training envelope and as such may not generalize to new regimes."

L163 - Sigmoid is a highly uncommon and suboptimal choice of activation function compared to ReLU! (ReLU is also missing from the glossary despite its ubiquity in modern models).

We have amended the text to clarify the cases in which sigmoid functions are used and have added a comment on cases where other activation functions are used:

"A commonly used activation function for a single neuron is the sigmoid function, which helpfully compresses the range between 0 and 1 while allowing a nonlinear response."

"Larger networks make more use of linear activations and may utilize heterogenous activation function choices at different layers."

ReLU has also been added to the glossary: "Rectified Linear Unit (ReLU). An activation function commonly used in DNNs. Defined as max(0, X). This function is used as it is computationally cheap and avoids problems of vanishing gradients."

L179 - Why are Token-sequence and Transformer models listed separately? I don't see the justification for this classification introduced as is.

There are token-sequence architectures that are not transformers. Many of the token-sequence architectures don't do dimensionality reduction, distinguishing them from transformers in that way. We have amended the list to clarify this and add some extra categories:

- "Small, fully-connected networks, which are less commonly featured in recent publications but are still effective for many tasks and are still being applied and may well be encountered in practice
- Convolutional† architectures, first applied to image content recognition, which match the connectome of the network to the fine structure of images in hierarchical fashion to learn to recognize high-level objects in images
- Recurrent token-sequence architectures, first applied to natural language processing, generation and translation; applicable to any time-series problem. Now also applied to image and video applications, and mixed-mode applications such as text-to-image or text-to-video
- Transformer architectures†, based on the attention mechanism† to provide a non-recurrent architecture which can be trained using parallelized training strategies. This allows larger models to be trained. Originally developed for sequence prediction and extended to image processed through vision transformer architectures."

L521 - ConvLSTM were introduced in 2015 also in the context of nowcasting, including a reference would be appropriate [2].

We were actually not aware that this was the origin of ConvLSTMs – a significant oversight on our part. Thank you for drawing this to our attention! We have added a sentence explaining the origins of ConvLSTMs and the reference you recommended:

"Convolutional LSTMs (ConvLSTMs), which combine convolutional layers with an LSTM mechanism, were introduced in the meteorological domain by Shi et al. (2015) for precipitation nowcasting. They have since seen wide adoption in other areas (e.g., Yuan et al., 2018; Moishin et al., 2021; Kelotra & Pandey, 2020). Their success in other domains suggests that revisiting their utility for weather and climate modelling could be worthwhile."

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, 28.

Moishin, M., Deo, R. C., Prasad, R., Raj, N., & Abdulla, S. (2021). Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm. IEEE Access, 9, 50982-50993.

Yuan, Z., Zhou, X., & Yang, T. (2018, July). Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 984-992).

Kelotra, A., & Pandey, P. (2020). Stock market prediction using optimized deep-convlstm model. Big Data, 8(1), 5-24.

L537 - Why is Sonderby et al discussed if nowcasting is supposedly omitted (L515)? Why not Espeholt et al 2021? Why is this not discussed in the context of probabilistic models?

This is a reasonable point, and we agree that Sonderby is out of scope. We have removed the sentences describing this paper.

L560 - A background reference to GNN either here or in the glossary could be beneficial! (e.g. Battaglia et al 2018 [3])

This is a good suggestion – thank you. We have added your suggested reference to the Glossary, along with Scarselli et al. (2008) and Kipf & Welling (2016).

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. IEEE transactions on neural networks, 20(1), 61-80.

L581 - I would strongly object to treating the models in this section (excluding Clare et al) as probabilistic in contrast to the ones in the previous section! These models are fundamentally deterministic, in contrast to e.g. generative models such as Ravuri et al or true probabilistic models like Sonderby et al. Discussing the different types of ensembling used in these models could be valuable on its own (also referring to Scher et al [4]).

We understand and acknowledge the basis for your objection – we were using the generation of an ensemble as a proxy for probabilistic modelling because that is the most common method for producing probabilistic outputs from physics based models in a forecasting setting, but by doing so we are definitely showing our biases, and we can see how that is an unreasonable division to make when there are many examples of intrinsically probabilistic ML models, some of which were included in the previous sections. We also agree that there is value in examining the question of ensemble ML models in more detail in any case.

We have shifted our description of Clare et al into the previous sections where it fits better in the narrative, and have significantly re-written section 5.3 as an examination of ML models used to generate ensemble predictions.

L661 - A large part of the affordable training is the use of much lower resolution and not due to the architecture! (1.4deg vs 0.25)

We acknowledge that this was overstating the fact, and wasn't making the main point well. We have amended the sentence to be clearer:

"Furthermore, ClimaX used novel encoding and aggregation blocks in its architecture to enable greater flexibility in the types of variables used for training, and to reduce training costs when a large number of different input variables were used."

L663 - Missing several extensions of WeatherBench (e.g. WeatherBenchProbability, RainBench) and ClimateBench.

Thank you for pointing this out – we became aware of these omissions soon after submission and have now added them in:

"Since the publication of WeatherBench, more benchmark datasets tailored to other domains have been created, including RainBench (de Witt et al., 2020), WeatherBench Probability (Garg et al., 2022), and ClimateBench (Watson-Parris et al., 2022)."

Garg, S., Rasp, S., & Thuerey, N. (2022). WeatherBench Probability: A benchmark dataset for probabilistic medium-range weather forecasting along with deep learning baseline models. arXiv preprint arXiv:2205.00865.

de Witt, C. S., Tong, C., Zantedeschi, V., De Martini, D., Kalaitzis, F., Chantry, M., ... & Bilinski, P. (2020). RainBench: towards global precipitation forecasting from satellite imagery. arXiv preprint arXiv:2012.09670.

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., ... & Roesch, C. (2022). ClimateBench v1. 0: A Benchmark for Data-Driven Climate Projections. Journal of Advances in Modeling Earth Systems, 14(10), e2021MS002954.

L981 - The activation function is applied elementwise to the result of a matrix multiplication and does not incorporate multiplication or bias addition by definition.

We have amended our definition to be more clear:

"Activation Function. The function which produces a neuron's outputs given its inputs. Commonly, this includes a learned bias term which is added to the data inputs before evaluation with a single function to produce the output value. Examples of the functions used include linear, sigmoid and tanh."

L995 - Calling it a "complex" mechanism is not necessary. Typical Attention simply computes a dot product between vectors.

This is fair enough – we have removed the word complex.

L1007 - Normalisation plays an essential role in modern NN and probably deserves its own glossary term. It does not need to be performed over the batch (i.e. LayerNorm)

This was an erroneous omission. We have added a definition of normalisation:

"Normalisation. A technique applied in many areas of mathematics, science and statistics which is also very important to machine learning and neural networks. In a general sense, this refers to expressing values within a standard range. Very often, the range of expected values is mapped onto the range 0 to 1, to allow physical variables with different measurement units to be compared on equal scale. Such normalisation may be linear or nonlinear, according to a simple or more complex function, and either drawn from known physical limits or from the variation observed in the data itself."

L1028 - "Convolutional … sliding window" seems redundant.

Given the broad scope of possible readers of this review, we think there may be readers who would be familiar with the concept of a sliding window programmatically, but would not be familiar with the term convolution. Thus we think it is good to keep both, even though they are somewhat redundant. We have amended this sentence slightly to: " Convolutional neural network. A neural network architecture commonly applied to images which utilises a convolutional (spatially connected) kernel applied in a sliding window fashion with a narrow receptive field to encourage the network to generalise from fine scale structure to higher levels of abstraction."

L1037 - Not true in general! If the network is too thin it becomes highly unstable.

We have amended this definition: "Deep NN. A neural network with many layers. Deeper, thinner networks have generally been more popular in recent times than wider, shallower ones but this is not always the case (see e.g. https://arxiv.org/abs/1605.07146)."

Technical corrections:

DNN and NN are introduced as separate abbreviations, but the distinction is not kept consistent nor does it appear beneficial.

This is a reasonable observation – we have adopted a more consistent convention of just using NN.

References:

[1] McGovern, A., R. Lagerquist, D. John Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. Bull. Amer. Meteor. Soc., 100, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

[2] Shi, Xingjian, et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." Advances in neural information processing systems 28 (2015).

[3] Battaglia, Peter W., et al. "Relational inductive biases, deep learning, and graph networks." arXiv preprint arXiv:1806.01261 (2018).

[4] Scher, Sebastian, and Gabriele Messori. "Ensemble methods for neural network-based weather forecasts." Journal of Advances in Modeling Earth Systems 13.2 (2021).