The paper is clear, well-organized, and discusses a very interesting approach, XAI methods, that could help scientists overcome epistemic opacity of ML-based models. I believe this paper can contribute in the reflection within the climate science community on the use of machine learning for modelling on the one side, and, on the other side, in the philosophy of science debates on understanding through climate models and epistemic opacity of machine learning techniques. Indeed, the two original contributions of the paper are, first, to use philosophical concepts in order to analyse the possible difficulty in the use of ML-based models in climate science, and, second, to discuss promising novel methods, i.e. XAI methods. But a number of revisions are needed. In what follows, I give some suggestions.

For the philosophy of science part:

As an intersdiciplinary researcher working with climate scientists, I find important to not introduce new terms that actually refer to already existing concepts in philosophy, but also to make connection with the relevant philosophy of science literature. Yet, there is now a rich discussion in philosophy of science on understanding through climate models and epistemic opacity of machine learning techniques which would be worth being cited and used for this paper. More precisely, I think that the following aspects of the paper should be revised:

1_ The authors argue that what they call “component-level understanding” should and can be reached with climate models but also with ML-based models with the help of XAI methods. They also argue that CMIP is a place where component-level understanding has successfully increased.

1.1_ However, this understanding — that seems similar to what Frisch (2015) calls “analytical understanding” — comes with the assumption that climate models are modular and that the interactions between the different modules (or model components) can be grasped and anticipated. But this modularity has been qualified as “fuzzy” by Lenhard and Winsberg (2010) and therefore scientists are facing what Lenhard and Winsberg call “entrenchment”. Clearly this is in conflict with what the authors are claiming in this paper. That is why I believe the authors should engage with this debate (and revise the paper accordingly, all along the paper, not only at the beginning of the paper). I don’t think that it would undermine their argument at all but will make it more nuanced and stronger; what they call “component-level understanding” might still be an ideal to pursue in the scientific practice. I also recommend the authors to read and cite the paper on modularity by Lenhard (2018).

1.2_The paper of Lenhard and Winsberg (2010) also demonstrates the failure of CMIP in making intercomparisons and thereby reaching what the authors call “component-level understanding”. In this draft, the scientific references used to support the claim that AMIP / CMIP allowed for more component-level understanding are not recent (e.g. first paragraph p. 5 for instance or Glecker et al. 1995 cited p. 8), thus it would be nice that the authors explore whether it is the aim of CMIP6/7 using recent examples/illustrations. There might be another interesting paper on this topic, the paper of Touzé-Peiffer, Barberousse and Le Treut (2020).

1.3_ Another well-discussed issue in philosophy of science that makes “component-level understanding” difficulty to reach in the case of climate models is the epistemic opacity of climate models simulations, that models be dynamical or ML-based. Here are examples of such papers: Knüsel and Baumberger 2020; Kawamleh 2021; Jebeile, Lam and Räz 2021.

2_ The authors put forward three kinds of understanding, instrumental understanding, statistical understanding and component-level understanding.

2.1_ One would expect this taxonomy to be connected to what philosophers have already said about understanding with models, or to be motivated by what scientists tell about their own practices. Thus, in (Knüsel and Baumberger 2020) and (Jebeile, Lam and Räz 2021), the authors put forward different dimensions of understanding with models that therefore comes in degree. In particular, notably following the work of de Regt and Dieks cited in the paper, Jebeile, Lam and Räz (2021) use these evaluative criteria of understanding with models: intelligibility, representational accuracy, empirical accuracy, physical consistency, delimiting the domain of validity. Is this explicitation of “understanding with models” useful for this paper? For example, the

difference between “statistical understanding” and “component-level understanding” is that only the latter meets intelligibility, no? (I am just curious here, this might not be crucial for the paper though).

2.2_ What about the concept of “process understanding” used by climate scientists? It is usually referring to the aim of fundamental research. Is it not covered by the concept of “component-level understanding”?

2.3_ In the paper, what is the role of this taxonomy after all? Couldn’t the authors simply introduce the definition of “component-level understanding” (or process understanding) and argue that it can be reached in ML-based modeling with the help of XAI methods (where we could imagine that only statistical understanding is reached)?

2.4_ It would also be interesting to have a characterization of this taxonomy: is instrumental understanding a weaker form of understanding than statistical understanding? Is statistical understanding in turn a weaker form than component-level understanding?
Or do they overlap?

2.5_ In the hierarchy of models envisioned by Held (2005), is he referring to model component? In the quotations given p. 11, he instead speaks about the dynamics.
Another paper on hierarchy of climate models is (Katzav and Parker 2015).

Regarding the contribution of this paper for the scientific practice:

1_ It would be worth defining what “computational efficiency” of machine learning is (introduction p. 2) as it is usually the main motivation in the use of machine learning. It would be important for this paper to clarify what it means.

2_ It seems that the authors are assuming (or have to assume) that there is some kind of isomorphism between layers in neural net and model components. Can they clarify their position on this? (cf. second paragraph p. 4).

3_ Can it not be that search for“component-level understanding” is actually search for “representational accuracy”. Trying to correct for previous idealizations and parameterizations seem to be line with the “natural” direction of scientific research, no? (This is what is assumed in Jebeile and Roussos 2023; Baldissera Pacchetti, Jebeile and Thompson 2024).

4_ As it is, section 4.3 fails to be persuasive because GAN does not to apply to climate science. The authors should explain whey they believe that, in the future, GAN will be applied to ML-driven climate science.

Minor comments:

1_ In what sense, do AI models entail “greater uncertainty”? Could you specify what you mean: is that that the outputs / predictions of models are more uncertain? (abstract, p. 2)

2_ What is actually the “functional test” that a model has to pass in order to provide instrumental understanding? (abstract, p. 2)

3_ Some technical terms should be (better) defined: “layer-wise relevance propagation”; “attribution/relevance heatmaps”; “multi-layer, convolutional recurrent neural networks”, “tree ensembles” (p. 6); distinction between “specific classificatory instances” and “global classification” (p. 7); “P-score” (p. 13)

4_ There should be no bracket after “Gettelman et al. 2019.” (p. 5).

5_ References are needed to support the claim that “In addition, there is a concurrent need to establish the trustworthiness of ML models as driven climate science potentially becomes increasingly used to inform decision makers” (p. 5).

6_ In the introduction of section 4, the authors write “we offer three examples in which ML researchers are able to acquire component-level understanding of model behaviors by intentionally designing or discovering model components that are interpretable and intelligible.” This sentence seems to suggest that “interpretable and intelligible” model components will bring component-level understanding (p. 13). Could the authors clarify this point?

7_ The authors should write what the acronym PDEs is referring to in all letters (p. 14)

8_ The authors should indicate the year of Pathak et. Al (p. 14) and add the reference in the list of references.

References indicated in this review:

Frisch, M. 2015. Predictivism and Old Evidence: a Critical Look at Climate Model Tuning. *European Journal for Philosophy of Science* 5 (2):171–190.

Lenhard, J., and Winsberg, E. 2010. “Holism, Entrenchment, and the Future of Climate Model Pluralism”. *Studies in History and Philosophy of Science* Part B, 41(3):253–262.

Lenhard, J. (2018). Holism, or the erosion of modularity: a methodological challenge for validation. *Philosophy of Science, 85*(5) 832–844.

Jebeile, J., Lam, V. & Räz, T. (2021) Understanding climate change with statistical downscaling and machine learning. *Synthese* 199, 1877–1897. https://doi.org/10.1007/s11229-020-02865-z

Kawamleh, S. (2021). Can machines learn how clouds work? The epistemic implications of machine learning methods in climate science. *Philosophy of Science*, *88*(5).

Knüsel, B., & Baumberger, C. (2020). Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A*. https://doi.org/10.1016/j.shpsa.2020.08.003.

Baldissera Pacchetti, M., J. Jebeile, and E. Thompson, 2024: For a Pluralism of Climate Modelling Strategies. *Bull. Amer. Meteor. Soc.*, https://doi.org/10.1175/BAMS-D-23-0169.1, in press.

Touzé-Peiffer L, Barberousse A, Le Treut H. The Coupled Model Intercomparison Project: History, uses, and structural effects on climate research. *WIREs Clim Change*. 2020; 11:e648. https://doi.org/10.1002/wcc.648

Katzav, J., & Parker, W. S. (2015). The future of climate modeling. Climatic Change, 132, 475–487.