We thank the referees for the insightful comments, which have greatly helped us improve the manuscript. Below, we have reiterated the reviewers' remarks in red italics, followed by our corresponding responses in standard black text. The references to specific lines and figures pertain to the provided diff file.

Reviewer 1:

This paper examines how the European radiative forced responses can impact Arctic climate by using a combination method of two machine learning techniques, the k-mean clustering and convolution neural network. Specifically, the authors classified six patterns and discussed how these six patterns responds to the European radiative forcing and to impact the Arctic. This paper is interesting, and the topic is crucial for the community, especially the method constraining the k-mean clustering for auto-encoder with PMSET. Therefore, I do suggest this paper to be published in Weather and Climate Dynamics.

We greatly appreciate your positive feedback and recognition of the significance of our topic to the community. We are pleased that you found our methodology, which leverages the flexibility of Deep Learning (DL) to provide a framework for analyzing climate patterns, to be of interest.

Other point-by-point comments are following (Lines are from highlighted change version):

1. Lines 10-11, "the negative forcing over Europe"? Negative radiative forcing? Less radiation?

Thank you for your remark. We revised it to "negative radiative forcing" in line 10 of the diff file.

2. Lines 18-19, "the shifts in the mean characteristics of the atmospheric circulation patterns"? Not clear. What does it mean the shifts inf the mean characteristics? Spatial? Temporal? Or others?

Thank you for highlighting the unclear statement. We mainly refer to shifts in spatial characteristics. We have revised the sentence in line 13 of the diff file to specify that it refers to spatial characteristics.

3. Lines 159-160, how other data fields or using only one of the two fields may change the results? Please not simply use "might" but test with simple cases.

You raise a crucial point regarding the impact of including additional meteorological fields in our analysis. Indeed, understanding how much information can be gained by adding another field is important. This can be effectively analyzed using self-supervised methods in the context of data imputation. However, the primary focus of this study was not to quantify the amount of informational content one can gain for our study's purpose by adding various data fields. Instead, motivated by the work of Weyn et al. (2019) and Weyn et al. (2020), we specifically used the fields of MSLP and $\tau_{300-700}$ for our study. While the impact of adding or removing selected data fields has been analyzed in their studies for forecast contexts, our choice to use these two fields was driven by the aim to capture the most information regarding large circulation regimes with the minimum number of data fields. This has been discussed in the first paragraph of section 2.3.1.

4. Why separate Figure 6 and 7? If so, why not showing the 2m temperature and SIC mean state in Figure 6 as well?

Thank you for your comment. Initially, we aimed to keep the climatology and anomaly in separate figures. We did not include the 2m temperature and SIC mean state in Figure 6 because we anticipated that the 2m temperature climatology would show reduced temperatures toward the pole, and the SIC climatology is somewhat represented in the anomaly with the green outline. For the PMSET, we included the climatology to observe how the control run produces PMSET climatology. However, we agree that adding the 2m temperature and SIC fields and combining the climatology and anomaly figures would enhance clarity and simplify the analysis. Therefore, we have merged the two figures and added the climatology for 2m temperature and SIC. Please see Figure 6 and lines 327-333 and 340 in the diff file for these updates. Thank you for your valuable suggestion.

5. Lines 338-340, "The anomalies in Figure 7 were calculated by subtracting the respective mean fields of the Control from those of the Experiment run". Does this refer to the mean state differences between the Experiment and Control run in annual mean?

Thank you for highlighting this point. Yes, the anomalies in Figure 6 of the diff file represent the annual mean differences between the Experiment and Control runs. We have revised the text for clarity in lines 330-331 of the diff file.

6. Line 342, "an inverse correlation"? Does it mean pattern correlation statistically? Or not mathematically, just saying a similar but opposite values? Same in Line 351.

Thank you for your question. In this context, explaining the patterns in Figure 6 of the diff file, we meant that the patterns in the figures exhibit similar but opposite values. Thus, we did not perform a statistical test for this part, and the term "inverse correlation" was used descriptively rather than mathematically. We added "qualitatively" in lines 335 and 345 of the diff file to make this more clear.

7. Lines 392-396, are these a repetition of the method section?

Thank you for your comment. We acknowledge that this section contains a component of repetition. However, we believe it is necessary to maintain the storyline clearly and emphasize the connection between the methodology and the results. Including this repetition in the results section helps the reader keep track of the paper's narrative and better understand the methodological decisions. Therefore, we propose to keep this section as it is.

8. Line 414-416, are they significantly different?

To evaluate the significance of these changes, we considered two factors: the consistency of clustering results independent of the k-means clustering initialization, and the intermittency of the observed signal. For the former, we performed k-means clustering 1,000 times with different initializations. The resulting changes in occurrence frequencies are reported in Table A2, which shows consistency in the change in occurrence frequencies between the Control and Experiment runs. To evaluate the intermittency of the signal, we implemented a bootstrapping strategy in the revised version. This involved creating 1,000 ensembles for each simulation by resampling with replacements from the Control and Experimental runs. The results have been incorporated into Figure 9 of the diff file and described in lines 405-407. These bootstrapped results also support the consistency of the changes in the clusters' occurrence frequencies.

Reviewer 2:

Title: Arctic Climate Response to European Radiative Forcing: A Deep Learning Approach

Authors: Sina Mehrdad, Dörthe Handorf, Ines Höschel, Khalil Karami, Johannes Quaas, Sudhakar Dipu, and Christoph Jacobi

I would like to thank the authors for their efforts in addressing my comments. The structure and storyline of the manuscript have been improved, allowing for a clearer delivery of the results. However, regarding causality, I am hesitant about the statement the authors made in their replies: 'Deriving causality in the context of regional forcing on Arctic climate is complex and multifaceted.' In a coupled climate system, it seems nearly impossible to clearly distinguish causality. Is it necessary to provide causality, or should one acknowledge that deriving it is very difficult, and/or not necessary? Overall, I am satisfied with this version but have a few additional comments for the authors to consider.

Thank you very much for your comments and perspective. Regarding the sentence you referred to, the term "derive" was perhaps not perfectly chosen; "inferring" more accurately describes our intention in the sentence making it 'Inferring causality in the context of regional forcing on Arctic climate is complex and multifaceted'. This complexity arises from the inherent uncertainties in climate modeling and the chaotic nature of the climate system itself which can be observed in the uncertainties in modeling forecast abilities beyond a few days. While it is true that climate systems exhibit long-term unpredictability, they also exhibit deterministic behavior, which allows for the study of causality within certain limits. Many studies use statistical methods to infer potential causal relationships. While these methods do not necessarily predict specific outcomes, they help identify relationships and potential causal influences between variables. These sorts of analyses like our study in this paper, are valuable for hypothesis generation and enhancing our understanding of the climate system, which may subsequently be explored further through more detailed modeling or experimental interventions. In response to your query, we believe that while proving causality in the strictest sense might be exceedingly challenging, attempting to infer and discuss potential causality is crucial for advancing our theoretical framework and guiding future research.

For the machine learning approach, I think using a convolutional auto-encoder architecture to study this topic is both exciting and promising. I also appreciate that the authors provided baseline models (e.g., SOM and EOF) for comparison. However, extending this point further, what new scientific insights can these methods bring to us? For example, the authors mentioned that 'We compare this with the traditional data space representation, which is dominated by the seasonal cycle.' These results may seem straightforward to most meteorologists or climate scientists, as some of their research is based on signals after removing the seasonal cycle.

Additionally, the stratosphere-troposphere coupling has been well-known for decades. Does using deep learning reveal something new? Related to this, the authors stated that 'a high-pressure system over Northern Eurasia and Scandinavia in autumn, observed in the Experiment run, led to reduced upward wave propagation.' Is this a new finding that previous studies did not document? The authors may want to elaborate more on this aspect and highlight the new scientific findings from this manuscript, as well as those not found in previous studies.

Thank you for expressing the value you find in using convolutional autoencoders in this analysis and in comparing it to the baseline methods. The question you raised is indeed critical, as it seeks to uncover the new scientific insights that these methods can provide. In our view, the capability of Deep Learning (DL) models to extract complex patterns from datasets holds significant promise for climate applications, though it is still at its early stage. DL models offer unprecedented flexibility, allowing us to direct the model's focus on specific aspects of the data we deem important. This project represents an effort to harness this potential in a meaningful way.

In this study, we categorize the circulation patterns based on a target variable that we consider important for our task. Because we are interested in analyzing the effect of circulation on the Arctic climate, we selected the 3D pattern of the associated PMSET as this target variable in our approach. That is, we direct our DL model (MCAE) focus to prioritize the relationship between circulation patterns and PMSET. The MCAE is trained to identify similarities in its input domain, which includes MSLP and layer thickness from 700 to 300 hPa, that correlate with similar PMSET outcomes. The flexibility of the DL models offers various post-training analysis methods; for instance, modifications in the latent space representation can be explored through the decoder to assess changes in data space representation or the convolutional neural network (CNN) kernel weights can be examined to understand decision-making processes within the model. In our analysis, we focused on examining the model’s latent space as the post-training analysis. Initially, we identified six distinct density points within MCAE's latent space and subsequently analyzed the similarities within each of these groups to discern what the model emphasizes.

To comprehensively evaluate the performance of the MCAE, we must assess the model's reconstruction loss and its ability to represent both the circulation and PMSET patterns within its latent space. The training objective of the MCAE was to capture the structure of circulation data points and their associated PMSET in its latent space. Evaluation of the presence of the PMSET's patterns in the latent space of MCAE is straightforward and involves comparing clustering results within the MCAE latent space against those derived from a representation solely based on PMSET indices. This comparison demonstrates that the MCAE latent space retains characteristics of the PMSET indices to a certain extent. However, evaluating the representation of circulation data points in the MCAE latent space structures presents more challenges. We utilized the seasonal cycle as the metric for this purpose, given its dominance in the data space representation. Although there is a weak dependency of the seasonal cycle on PMSET indices, it is still the predominant pattern of the circulation data points, thereby serving as a crucial metric for evaluating how well the data space structure is represented in the MCAE latent space.

We retained the seasonal cycle within the dataset to assess how various data representations and methods capture this dominant pattern. Traditional clustering methods primarily categorized data based on the seasonal cycle alone, highlighting their limitations in detecting finer nuances. It was also the case when clustering on FAE latent space. However, its latent space demonstrated a more nuanced distinction by efficiently discriminating between transitional seasons, reflecting the superior quality of its feature extraction capabilities. In contrast, the MCAE also effectively managed the seasonal cycle, incorporating it into its latent space while regulating its domination based on its importance for the task. Notably, with the integration of PMSET target information, the MCAE preserved the main structural elements of the data and accounted for similarities in associated PMSET, showcasing its advanced capability in capturing both the inherent data structure and metadata nuances.

As highlighted earlier, employing DL models for climate diagnostic applications represents a promising research avenue in climate science. This work has demonstrated several advantages of applying DL models to these applications. Even though these advantages have been detailed in the different parts of the paper, we acknowledge the need to wrap all these up in a concluding manner. Thus, we have modified the conclusion section (see lines 742-754 and lines 788-791), which summarizes the key benefits of our DL model discussed here, providing a clear overview and emphasizing the potential impact on future research in the field.

Although in this work we did not focus on stratospheric coupling in deriving the latent space of MCAE and just focused on the PMSET associated with the circulation patterns, we still yielded noteworthy insights into tropospheric-stratospheric interactions. With the developed class contribution formulation, which is independent of the DL model, we attributed observed anomalies to changes in clusters behavior during experimental runs. Notably, as highlighted in your question, a cluster associated with a high-pressure system over Northern Eurasia and Scandinavia (C3) in autumn plays a role in upward wave propagation. Our quoted sentence in the context is as follows. "The decreased occurrence frequency of a cluster associated with a high-pressure system over Northern Eurasia and Scandinavia in autumn, observed in the Experiment run, led to reduced upward wave propagation. However, a slight adjustment in the mean behavior of this cluster in the Experiment run resulted in an increase in upward wave propagation". Each cluster can change in the Experiment run relative to the Control run in two primary ways: 1) changes in its seasonal occurrence frequency, and 2) minor alterations in the cluster's mean pattern. Our findings show that a reduced occurrence frequency of cluster C3 in autumn is associated with decreased upward wave propagation. This correlation is consistent with the established understanding of the influence of similar circulation patterns on upward wave dynamics. However, the extent to which this reduction in frequency shapes upward wave propagation remains an open question. Our results indicate that even minor adjustments in the mean pattern of this cluster can have a more significant effect on upward wave propagation than changes in its seasonal occurrence frequency. This analysis enables a detailed comparison of how variations in C3's occurrence frequency and adjustments in its mean pattern each contribute to the overall anomaly observed in upward wave propagation. Furthermore, our methodology offers a quantitative approach to assess the contributions of different clusters to the observed anomaly in upward wave propagation, marking a significant advancement in the analysis of complex systems such as the climate. These points have been extensively discussed in lines 671-699 and 766-780 of the diff file, which we refined further for enhanced clarity.

For the forcing and its connection to remote impacts, the authors argue that the large-scale circulation regime is the main underlying mechanism. My follow-up question is: how can one justify that the forced signal stands out from internal variability, given that large-scale circulation is intrinsically affected by internally-driven components? Many previous studies have stressed out the large internal variability could mask out the forced signal. The authors also mentioned that the forcing does not always produce uniform results. Resonating with previous comments, the authors could emphasize the role of machine learning approaches in this context and highlight what new insights these methods bring to the table and deal with the nonstationary results from the same forcing.

Thank you for highlighting this crucial aspect of the analysis. Differentiating the forced signal from internal variability, especially in large-scale circulation responses, is indeed challenging. In our study, we acknowledge the limitations posed by not having access to a very large ensemble of simulations. However, our strategy involved 30 years of climate simulation, which, while not extensive, is significant.

For the anomaly fields, we employed the t-test to define statistically significant anomalies. Additionally, to evaluate the robustness of the k-means clustering in determining seasonal occurrence frequencies, we conducted k-means clustering 1,000 times with random initializations, as detailed in Table A2 of the manuscript. However, we recognize the need to further assess the robustness of the clusters' seasonal discrepancies between the Control and Experiment runs (as illustrated in Figures 10 and E1 of the diff file), as well as the cluster monthly occurrence frequencies (Figure 9). These elements are pivotal to our class contribution analysis within this study.

To address the intermittency of observed signals, we implemented a bootstrapping strategy in the revised manuscript, creating 1,000 ensembles for each simulation (lines 405-407 of the diff file). This method involved resampling with replacement from the Control and Experimental runs, and the results have been incorporated into Figures 9 and E1 of the diff file. Importantly, these bootstrapped results support the consistency of our general conclusions. Figures 9 of the diff file now includes error bars representing the standard deviation of the monthly occurrence frequency for each cluster at the top of the bar charts. These error bars are typically smaller than the differences in monthly occurrence frequencies between the Control and Experiment runs, particularly during seasons with high differences. Figure E1 of the diff file illustrates the clusters' seasonal discrepancies between the Control and Experiment run across the 1,000 bootstrapped ensembles for both simulations (line 907 of the diff file). Compared to the last version, the areas without statistical significance are highlighted with dotted in Figure E1, and the main patterns remain consistent and show statistical significance. An exception is the summer discrepancy for cluster C4, attributed to its extremely low occurrence frequency during this season. In some ensemble members, C4 is absent, rendering the pattern statistically insignificant.

As previously discussed, our findings indicate that the impact of the forcing varies depending on the existing circulation regime, underlining the system's non-uniform response. Our approach diverges from linear analyses by capturing the climate system's nonlinear behavior, showing that different circulation clusters react distinctly to the same forcing. Such insights

underscore the advantages of our methodology, which can handle the complex responses of the climate system to external forcing. This nuanced understanding stems from our class contribution formulation, which, while independent of DL models, benefits significantly from the use of a machine learning algorithm for clustering. These points have been elaborated upon in lines 755-759 of the diff file.

As noted, the MCAE is developed to make sure that the circulation data points with similar patterns and similar associated PMSET are grouped together in the clustering. This enhances the quality of clustering ensuring it closely meets our research objectives. However, the primary purpose of employing DL in our study was to refine data classification and facilitate anomaly analysis, rather than to generate or resample model ensembles. The focus of the DL algorithm was not on resampling, down sampling, or generating model ensembles based on the simulations at hand. These methods, which focus on addressing the signal-to-noise challenges in climate forcing responses, remain a compelling motivation for further research. Nevertheless, our methodology has successfully captured the nonlinear behavior of the climate system and attributed observed anomalies to each circulation class, thereby facilitating the climate forcing analysis.

References

Weyn, J. A., Durran, D. R., and Caruana, R.: Can machines learn to predict weather? Using deep learning to predict grid1140 ded 500-hPa geopotential height from historical weather data, Journal of Advances in Modeling Earth Systems, 11, 2680–2693, https://doi.org/10.1029/2019MS001705, 2019.

Weyn, J. A., Durran, D. R., and Caruana, R.: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 109, https://doi.org/10.1029/2020MS002109, 2020.