Response to Reviewer 2 Comments

https://doi.org/10.5194/egusphere-2025-2460-RC2

for

Understanding European Heatwaves with Variational Autoencoders

https://doi.org/10.5194/egusphere-2025-2460

Submitted to

Earth System Dynamics

by

Aytaç Paçal[1,2], Birgit Hassler[1], Katja Weigel[2,1], Miguel-Ángel Fernández-Torres[3], Gustau Camps-Valls[4], Veronika Eyring[1,2]

[1] Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

[2] University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

[3] Department of Signal Theory and Communications, Universidad Carlos III de Madrid (UC3M), Leganés, Madrid, Spain

[4] Image Processing Laboratory (IPL), Universitat de València (UV), Paterna, València, Spain

Authors' Response to Reviewer 2

Major Comment 1

This research analyses heatwaves in western Europe, during the entire year, from a spatio-temporal multi-variate perspective. To this end the authors use ML and DL techniques. They find four clusters of heatwave patterns throughout the entire year, with dynamics consistent with previous literature. Notably, they use ERA5 and extended the variables to characterize heatwaves.

While this avenue of work (Deep Learning for heatwave understanding) is very interesting, both from the methodological and climate-scientific perspective, I have my concerns regarding the novelty of the presented research. From the current work it seems that most of the methods are one-to-one copied from Happé et al. (2024), including the heatwave selection method, VAE, and the GMM clustering, including their respective (hyper)parameters. It needs to be clear throughout the entire manuscript what the novelty is of the current work and what has been reproduced or based on previous studies. Currently, the authors cite Happé et al. (2024) in some places but they do not contextualize their work as an application of the framework developed by Happé et al. (2024). If the authors see their work not as an application but rather an extension of the framework, additional developments need to be made to the current AI framework. Generally, the Abstract, Introduction, and Discussion & Conclusion need to properly reflect which part of this research is novel and which follow the framework from Happé et al. (2024). Please find below more detailed comments.

Response: We thank the reviewer for bringing this important point to our attention. We indeed followed the general heatwave detection and analysis framework introduced by Happé et al. (2024), but our study extends it in several key aspects. While Happé et al. (2024) focused on summer heatwaves in the KNMI-LENTIS model ensemble using two atmospheric circulation variables and shorter temporal windows, we apply the method to ERA5 reanalysis data, consider nine atmospheric variables, and analyze year-round heatwaves. This allows us to investigate both summer and winter heatwaves, including pre-onset conditions, with an 11-day window. This enables us to study the build-up and temporal evolution of heatwaves. Furthermore, we compare historical (1941–1990) and recent (2001–2022) periods to investigate the shift in the latent space of heatwaves under climate change. This reveals that recent heatwaves cluster differently from historical

ones, offering a perspective not addressed in Happé et al. (2024). We will revise the Introduction, Methodology and Discussion to clarify these points, emphasizing that our work applies and extends the framework of Happé et al. (2024) to an observational, multivariate, and year-round setting, while also examining its evolution under climate change.

Major Comment 2

Introduction, L60-70 Here it reads as if this is the first study that uses the framework of VAE+Clustering to characterize climate extremes (especially line 68-70). Since this is not the case, it needs to be framed clearly what the novelty is of this work with respect to previous works, and how this study is either an application or extension of previous works. Please also have a look at:

- Spuler FR, Kretschmer M, Kovalchuk Y, Balmaseda MA, Shepherd TG. Identifying probabilistic weather regimes targeted to a local-scale impact variable. Environmental Data Science. 2024;3:e25. doi:10.1017/eds.2024.29

Response: We thank the reviewer for this comment and the suggested literature. We agree that our current phrasing in the Introduction may give the impression that this is the first study to use a VAE and latent space clustering for characterizing climate extremes. We will revise the text to clarify that our research builds on previous applications of VAEs and clustering in climate science, such as Happé et al. (2024) and Spuler et al. (2024), and that the novelty of our work lies in extending this framework to an observational, multivariate, and year-round analysis of European heatwaves.

VAEs are primarily developed as generative models, able to learn complex data distributions and generate realistic synthetic samples (Kingma and Welling 2019). Beyond generative tasks, VAEs have also been widely used in anomaly detection in domains such as network security, risk management, health monitoring, and computer vision (Pang et al. 2021; Nassif et al. 2021; Albuquerque Filho et al. 2022). Among other applications, autoencoders and VAEs have been applied to learn spatiotemporal regularities from video data (Hasan et al. 2016; Fan et al. 2020). VAEs and clustering methods such as GMMs have been used to identify regimes in the latent space and analyze their dynamical behavior in climate science (Lindhe,

Ringqvist, and Hult 2021; Happé et al. 2024; Paçal et al. 2023). Spuler et al. (2024) introduced the RMM-VAE, a probabilistic machine learning method that combines variational autoencoders with clustering to identify circulation regimes targeted to local impact variables. Happé et al. (2024) applied a 3D variational autoencoder to heatwave events from the KNMI-LENTIS dataset over western Europe, showing that the latent space captures physically interpretable circulation regimes and that heatwaves are best represented in a probabilistic framework.

Building on these previous applications, our study extends their methods to provide an observational, multivariate, and year-round analysis of European heatwaves. This study aims to understand how heatwaves develop and what processes and local phenomena contribute to their evolution.

Albuquerque Filho, José Edson De et al. (2022). "A Review of Neural Networks for Anomaly Detection". In: *IEEE Access* 10, pp. 112342–112367. DOI: 10.1109/ACCESS.2022.3216007.

Fan, Yaxiang et al. (2020). "Video anomaly detection and localization via Gaussian Mixture Fully Convolutional Variational Autoencoder". en. In: *Computer Vision and Image Understanding* 195, p. 102920. DOI: 10.1016/j.cviu.2020.102920.

Happé, T. et al. (2024). "Detecting Spatiotemporal Dynamics of Western European Heatwaves Using Deep Learning". In: *Artificial Intelligence for the Earth Systems* 3.4, e230107. DOI: 10.1175/AIES-D-23-0107.1.

Hasan, Mahmudul et al. (2016). *Learning Temporal Regularity in Video Sequences*. arXiv: 1604.04574 [cs.CV].

Kingma, Diederik P. and Max Welling (2019). "An Introduction to Variational Autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4, pp. 307–392. DOI: 10.1561/2200000056.

Lindhe, Adam, Carl Ringqvist, and Henrik Hult (2021). *Variational Auto Encoder Gradient Clustering*. DOI: 10.48550/arXiv.2105.06246.

Nassif, Ali Bou et al. (2021). "Machine Learning for Anomaly Detection: A Systematic Review". In: *IEEE Access* 9, pp. 78658–78700. DOI: 10.1109/ACCESS.2021.3083060.

Paçal, Aytaç et al. (2023). "Detecting Extreme Temperature Events Using Gaussian Mixture Models". In: *Journal of Geophysical Research: Atmospheres* 128.18, e2023JD038906. DOI: https://doi.org/10.1029/2023JD038906. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023JD038906.

Pang, Guansong et al. (2021). "Deep Learning for Anomaly Detection: A Review". In: *ACM Computing Surveys* 54.2, pp. 1–38. DOI: 10.1145/3439950.

Spuler, Fiona R. et al. (Jan. 2024). "Identifying probabilistic weather regimes targeted to a local-scale impact variable". en. In: *Environmental Data Science* 3, e25. DOI: 10.1017/eds.2024.29.

Major Comment 3

Methods 2.1; Why do the authors take this exact grid area? Or the 15d moving window? Crucially, why do the authors take a grid of 0.7 degrees spatial resolution if ERA5 has 0.25? If these parameters are chosen because those were used in Happé et al. (2024), that needs to be stated as such. Happé et al worked with 0.7 degrees because it is the native resolution of EC-Earth, and hence appropriate for that study. It is unclear why one would work with that resolution for ERA5, instate of the native 0.25 degrees.

Response: We thank the reviewer for this question. We regridded ERA5 to ensure comparability with Happé et al. (2024) and to reduce the very high dimensionality of the multivariate spatiotemporal inputs. Although ERA5's native resolution is finer, the use of MSE as the reconstruction loss emphasizes large-scale circulation patterns while smoothing out small-scale variability. Retaining the native 0.25°grid would therefore not provide additional benefit for our approach, while substantially increasing computational cost. The 15-day climatological window is a common choice for smoothing day-to-day noise while preserving seasonal variations. This approach is consistent with other extreme-event detection studies (Sulikowska and Wypych 2020; Happé et al. 2024). In the revised manuscript, we will make these choices clearer.

Sulikowska, Agnieszka and Agnieszka Wypych (2020). "Summer Temperature Extremes in Europe: How Does the Definition Affect the Results?" In: *Theoretical and Applied Climatology* 141.1, pp. 19–30. DOI: 10.1007/s00704-020-03166-8.

Major Comment 4

Heatwave identification – the authors take the "1941-1980 daily" percentile, which will inherently cause more heatwaves in the last 4 decades, as thermodynamics lead to an increase in temperature everywhere. This is important to consider when studying dynamics of heat extremes – how meaningful are the dynamical types that are then found? Furthermore, the test-set also consists of heatwaves from the last two decades – how do the authors deal with this non-stationarity?

Response: We selected the historical period for this study based on previous research (IPCC 2021; Elguindi, Rauscher, and Giorgi 2013; Reid et al. 2016). As a result, we naturally observe an increase in events in recent decades due to global warming, which is a significant aspect of our analysis. Additionally, our focus is on the patterns of heatwaves; while there may be more occurrences of heatwaves, we are clustering them based on their structures rather than their frequency. To analyze the effect of composite maps on atmospheric patterns, we tested composite maps based on different sample sizes (e.g., $N = \{1, 5, 10, 25, 50, 100\}$), which allows us to analyze the stability of the detected patterns and assess potential cancellation effects. While composite maps with more than 5 samples exhibit similar patterns, the 1-sample composite map differs (See Figure 2). This is expected, since the 1-sample map corresponds to cluster centers and the stochastic nature of the method means that the positions of heatwave samples, or cluster centers, in the latent space may not remain stable. Composite maps are therefore more reliable for identifying robust and consistent spatial features, since similar samples are already grouped closely together in the latent space by the VAE.

We also tested our model with random train/val/test splits across the entire 1941–2022 period to observe the effects of data splitting. Figure 1 shows the latent space from the VAE model trained using a random split dataset. While this approach provides a smoother latent space, summer heatwaves still accumulate more closely together. This confirms that the clustering is not simply an artifact in the time series but reflects consistent structural patterns across periods. We will clarify this point in the revised manuscript and emphasize that our conclusions concern the relative structures of heatwaves rather than their frequency of occurrence.

Elguindi, N., S. A. Rauscher, and F. Giorgi (2013). "Historical and future changes in maximum and minimum temperature records over Europe". en. In: *Climatic Change* 117.1, pp. 415–431. DOI: 10.1007/s10584-012-0528-z.

IPCC (2021). *Climate Change 2021: The Physical Science Basis*. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Edited by V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou. Cambridge University Press (In Press).

Reid, Philip C. et al. (2016). "Global impacts of the 1980s regime shift". In: *Global Change Biology* 22.2, pp. 682–703. DOI: https://doi.org/10.1111/gcb.13106. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.13106.

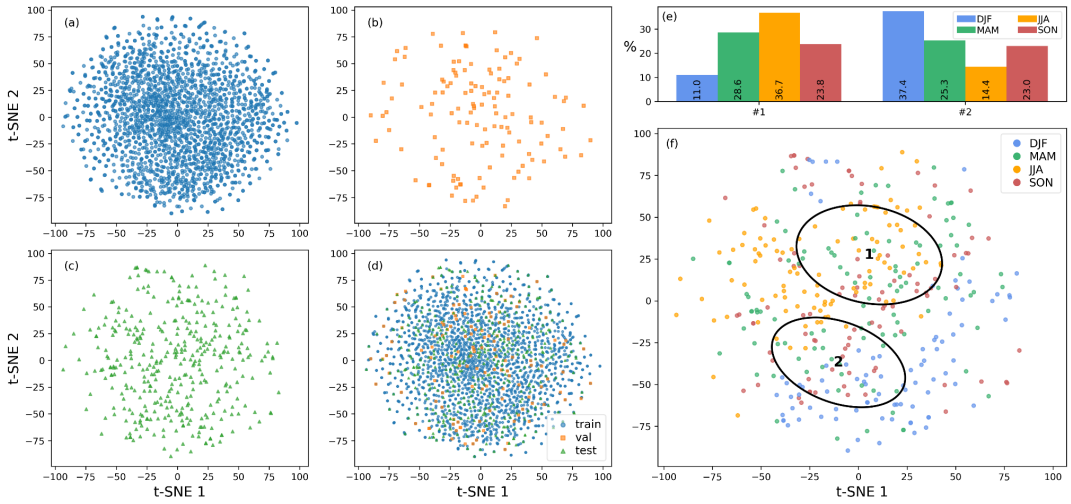


Figure 1: t-SNE representation of the latent space with random-split data (0.8/0.05/0.15). Following our workflow (testing with more than 2 components), a 2-component GMM provided the best fit, as it was closest to a single Gaussian distribution and thus better captured the smooth latent space structure.

Major Comment 5

Methods 2.2; Indeed, here the authors mention following the methods proposed by Happé et al. (2024). It would benefit the entire methods section if it would be very clear which parts of the methodology deviate from Happé et al. (2024).
Methods 2.3; As these methods as well follow Happé et al. (2024), it would be transparent to mention something like 'following Happé et al. (2024) we use a 3d VAE . . . ". Then continue explaining where your methods deviate and why the authors made those choices (e.g. improvement of training/framework/. . .). For example, the use of t-SNE is also done in Happé et al. (2024), yet this is not mentioned in your section 153-160). Additionally, the choice of 100 closest heatwaves to each centroid is also not cited as following Happé et al. (2024) – L161.

Response: Thank you very much for highlighting this point. Our study builds directly on the framework of Happé et al. (2024), but extends it in several significant ways. While Happé et al. (2024) focused on summer heatwaves in the KNMI-LENTIS model ensemble using two atmospheric circulation variables and shorter temporal windows, our work applies the approach to ERA5 reanalysis, incorporates nine variables, and examines heatwaves year-round with longer pre-onset windows. Crucially, we also analyze temporal shifts by comparing historical (1941–1990) with recent (2001–2022) periods, revealing climate-change–related changes in heatwave dynamics. These choices are also in line with the future outlook part in Happé et al. (2024) where the authors suggested use of observational datasets for further analyses of heatwaves. We will revise the Abstract, Introduction, and Discussion to make this novelty explicit and clearly separate where our study confirms previous results versus where it extends them.

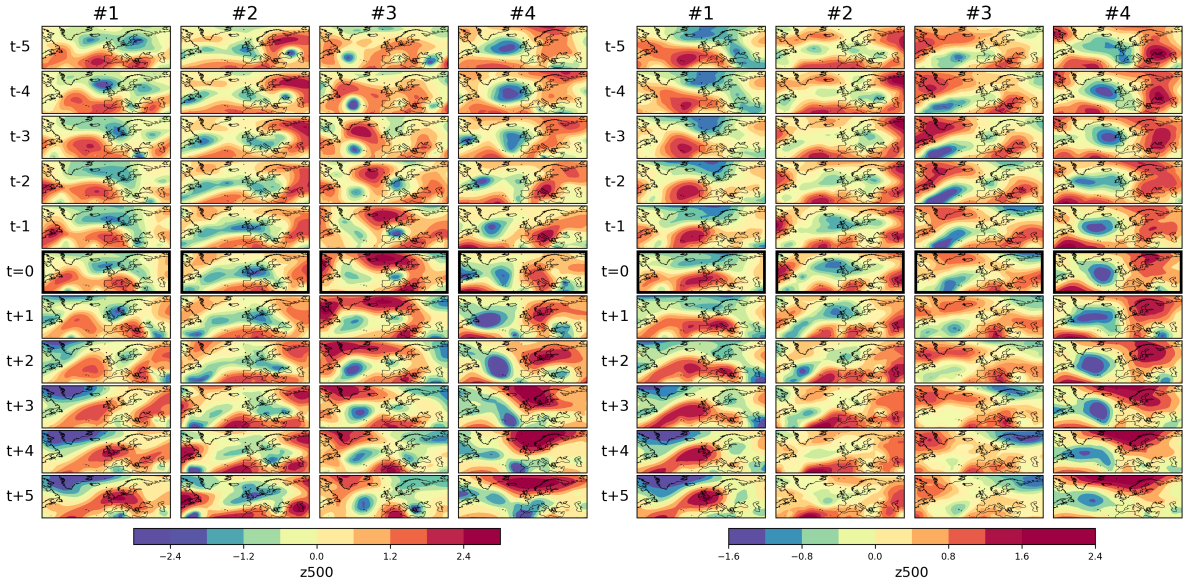
Happé, T. et al. (2024). "Detecting Spatiotemporal Dynamics of Western European Heatwaves Using Deep Learning". In: *Artificial Intelligence for the Earth Systems* 3.4, e230107. DOI: 10.1175/AIES-D-23-0107.1.

Major Comment 6

Methods 2.3 the r^2 scores; As this section talks about reconstruction errors, I would suggest this section fits better in the result. Apart from that – are these r^2 scores based on a latent dimension size 128? Is this chosen because of Happé et al. (2024)? Why didn't the authors take a higher latent space size, since the dimensions went from 2 to 9 variables and from 5 to 11 days? The latent dimension size should be properly justified and tested. Furthermore, I have my concerns with these low r^2 scores and would be curious to see the reconstructed maps for these variables. What happens if one goes to higher latent dimension sizes? Lastly, table 2 only shows the r^2 scores for the test-subset – my suggestion would be to also include the scores of the train set; to show how well the authors' model is able to generalize. I'm especially curious to this last point, as Happé et al. (2024) showed that data augmentation was needed to avoid overfitting.

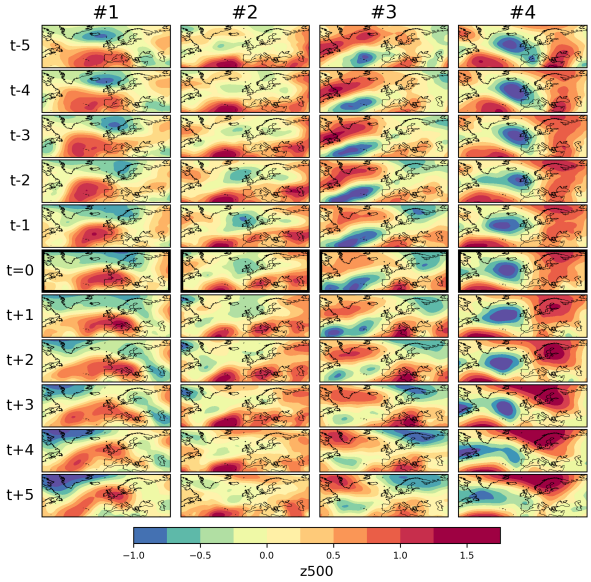
Response: We thank the reviewer for highlighting this. The latent dimension size (128) was selected based on a systematic grid search (64, 128, 256), which balanced reconstruction quality with training stability. This choice was independent of Happé et al. (2024). Our architecture also differs structurally: we employ a fully convolutional 3D encoder–decoder with pooling/unpooling layers, while Happé et al. used dense (linear) reshaping and data augmentation. Another difference between our study and Happé et al. (2024) is the dimensions of the input data. Since our focus was on assessing how well a 3D convolutional VAE model can capture large-scale atmospheric variability from raw reanalysis data without additional artificial variability introduced through augmentation, we decided not to include data augmentation.

We agree with the reviewer that reconstructed maps are helpful to illustrate model performance. We examined the model residuals to determine whether systematic large-scale biases occur in the variable fields as shown in Figure 2. The reconstructions accurately reproduce the main patterns of atmospheric variables, although they are smoother than the original inputs. Consequently, the intensity of extremes is slightly reduced. This behavior is consistent with the use of MSE as the loss function, which tends to favor average solutions over extremes. The differences are spatially varying and mainly concentrated on small scales. The model does not fail to capture the location or amplitude of large-scale anomaly patterns, although the intensity of extremes is somewhat reduced.

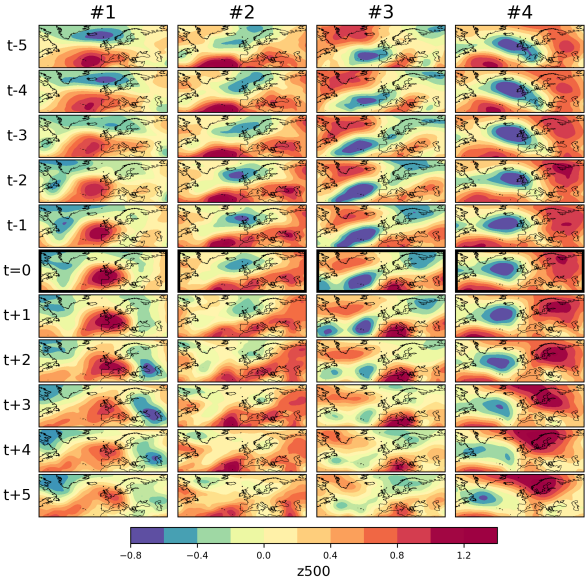


(a) 1-sample composite map

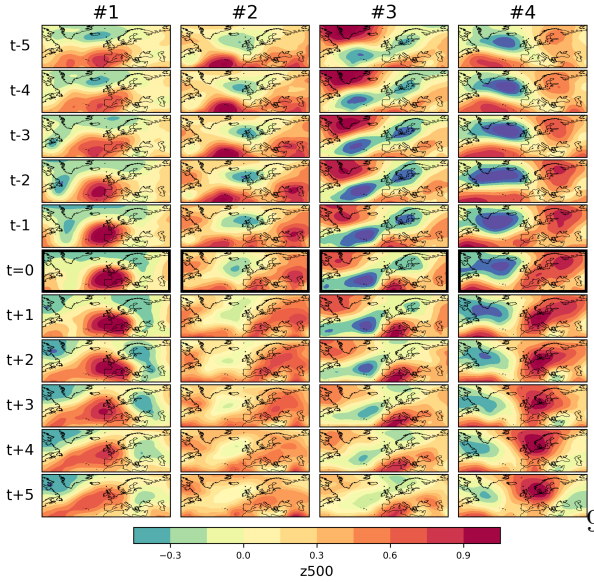
(b) 5-sample composite map



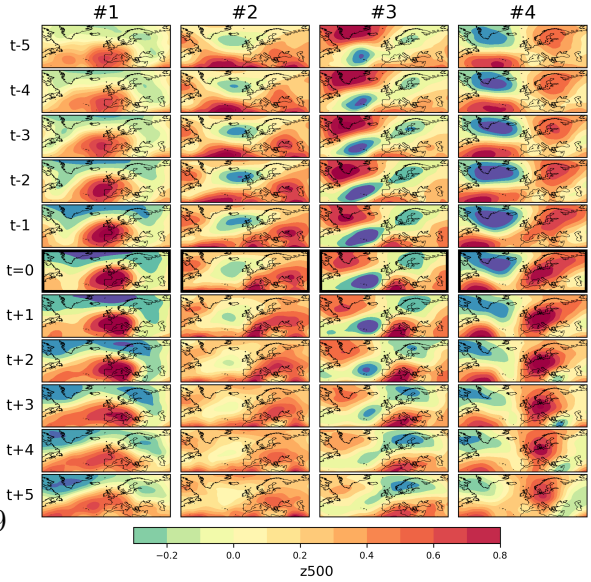
(c) 10-sample composite map



(d) 20-sample composite map



(e) 50-sample composite map



(f) 100-sample composite map

Figure 2: Composite maps for geopotential height at 500 hPa (*z500*).

Major Comment 7

Results; I'm curious as to why the authors apply PCA to go down to 50 components in the latent space – why not use PCA directly on the heatwave data? Or why not go down to 50 dimensions in the VAE latent space? What happens to the r^2 scores after doing this step?

Response: The pre-reduction step was recommended in the t-SNE paper Maaten and Hinton (2008) and the documentation for the scikit-learn function Pedregosa et al. (2011), which suggests applying a preliminary reduction step, such as PCA, before using t-SNE when working with latent spaces of higher dimension (above 50). In our case, the latent space had 128 dimensions. Applying PCA first was faster while producing similar patterns to using t-SNE alone, except for the inherent randomness in each run. We therefore adopted this approach to ensure both efficiency and robustness. We will revise this section in the manuscript to better explain the preliminary reduction step as follows:

> As an intermediate processing step, we first used Principal Component Analysis (PCA) to reduce the dimensionality of all heatwave samples from 1941-2022 to 50 components. Then, we applied the t-SNE algorithm to all heatwave samples to reduce them to 2 dimensions to visualize the latent space. This two-step approach is recommended for the analysis of a high-dimensional latent space to reduce the number of dimensions (see Maaten and Hinton (2008) and Pedregosa et al. (2011)) and helps to ensure efficiency and robustness. Because these steps involve stochasticity, we fixed the random seed to 42 (Adams 1979) for PCA, t-SNE, and GMM to ensure consistency across visualization runs.
>
> Adams, Douglas (1979). *The Hitch Hiker's guide to the Galaxy*. eng. Pan original. London: Pan Books.
>
> Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605.
>
> Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Major Comment 8

Results; I find it interesting that the authors find 4 clusters that correspond with each season. What does this mean for interpretation – did the latent dimensions clusters actually find dynamically different heatwaves or rather the dynamics of the different seasons? Would it be possible to plot composite maps within a cluster of summer-only and winter-only heatwaves? Perhaps that could show us whether these patterns are indeed found year-round or whether you find the seasonal dynamics. This would also underpin your speculative ("hints") conclusion in L383-385 better. Answering this is not trivial, as dynamics leading to heatwaves in summer (e.g. blocking) do not necessarily lead to warm anomalies in winter. Rather blocking like systems cause cold anomalies in winter. I find it therefore interesting that cluster #1 is a blocking pattern in winter, while the authors compare this cluster to UK High pattern in Happé et al. (2024) and the omega block in Rouges et al. (2023) which occur in summer [L328-241]. This as the authors show in Figure 4 that there are 0 summer heatwaves part of their cluster #1. Could it be that the fact that the authors find this pattern in winter is merely a result of the non-stationarity of the dataset? Could the authors explain this more?

Response: We thank the reviewer for this insightful comment. We agree that disentangling seasonal dynamics from latent-space clustering is a non-trivial task. To assess this, we examined the mean anomaly values of each cluster as shown in Figure 3. While intensity differences are visible (e.g., Cluster 2 exhibits the strongest positive t2m anomalies), the clusters also differ in other variables, indicating that the separation is not solely driven by temperature. For example, Cluster 1 corresponds to winter heatwaves characterized by weaker t2m anomalies but warm advection (positive mean v10 anomalies) under reduced solar input; Cluster 2 captures intense summer heatwaves with the highest t2m, humidity, and z500 anomalies; and Clusters 3 and 4 represent transition season events, one dominated by high solar radiation and the other by strong pressure anomalies. This suggests that the latent space clustering reflects both seasonal intensity differences and distinct dynamical configurations. Regarding possible non-stationarity, we note that our temporal split ensures that trends in the frequency of events do not drive the clustering itself, as clustering is performed on latent dimensions extracted across the entire period. Nevertheless, the non-stationarity of the climate system may influence which circulation patterns manifest as heatwaves in different seasons. We performed an additional experiment using random train/val/test splits across the entire 1941–2022

period. This allows us to evaluate whether the clustering and drift patterns persist beyond the temporal split. This approach introduces a smoother latent space representation, as the model will be trained on the entire period, as shown in Figure 1. Following our workflow (testing with more than 2 Gaussian components), a 2-component GMM provided the best fit, as it was closest to a single Gaussian distribution and thus better captured the smooth latent space structure. However, the summer heatwaves are still accumulated closer in the latent space. We will also revise the text to more clearly articulate why a temporal split was chosen in relation to our scientific question and acknowledge this explicitly in the discussion.

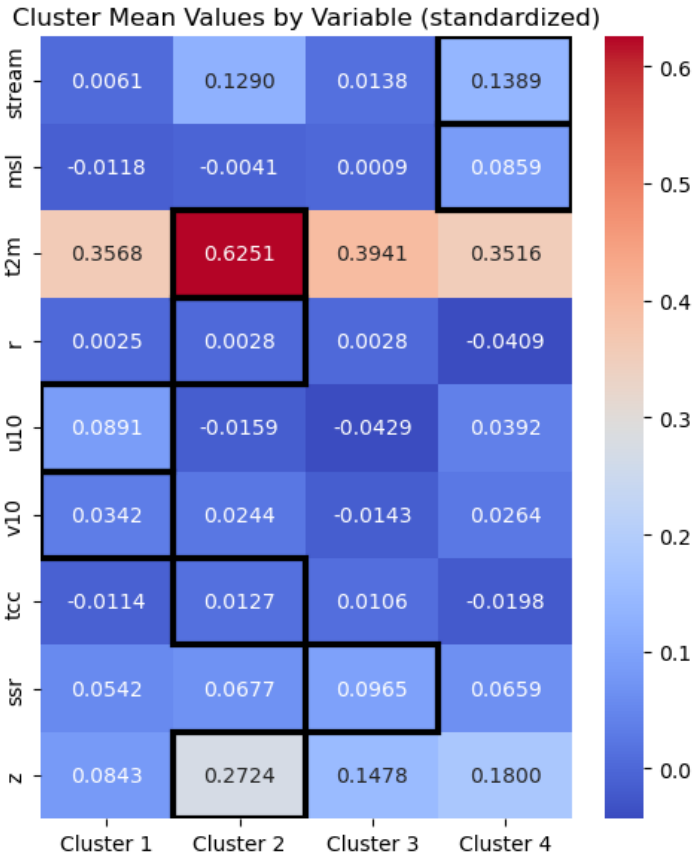


Figure 3: Mean cluster anomaly values for each variable.

Major Comment 9

In the Discussion the authors state that the VAE/GMM is sensitive to hyper-parameters; it would be good to see some of these experiments in this research. Especially the latent dimension size is essential for this research as ensuring that the latent representations are representative of your heatwave samples is not trivial – otherwise the clusters might be meaningless.

Response: We thank the reviewer for this suggestion. We agree that hyperparameters, particularly the latent dimension size, can influence both reconstruction and clustering. To address this, we performed a hyperparameter grid search (latent sizes of 64, 128, and 256; see Appendix A1), and found that 128 dimensions offered the best balance between reconstruction quality and stability. We acknowledge that presenting additional sensitivity plots would further illustrate this. While including the full set of experiments is not feasible, we will add representative comparisons of reconstruction scores and latent structures for different latent sizes to the Supplementary Material and clarify this point in the Discussion.

Major Comment 10

Discussion & Conclusion; Again, it needs to be contextualized which parts of the framework is based on previous work and which parts are novel. Using phrases such as "We confirm the results from Happé et al. (2024), by showing XYZ." Or "As opposed to Happé et al. (2024), we do/find XYZ." This helps guide the reader and highlights the novelty of the authors' work. E.g. in sentences 356-363, 392-394, and 369-400. It needs to be clear in the conclusion what the main scientific output is of your contribution.

Response: We thank the reviewer for these comments. We agree that our contribution needs to be articulated more clearly. In the revised manuscript, we will provide a clearer explanation of the difference and contribution of our study as follows:

> While we trained the VAE model with year-round heatwave samples from ERA5, in contrast to Happé et al. (2024) who used the KNMI-LENTIS ensemble dataset, the composite maps of the resulting heatwave clusters reveal atmospheric patterns

consistent with previous studies, which typically considered only summer months (Carril et al. 2008; Horton et al. 2015; Rouges et al. 2023; Krüger et al. 2023; Bischof et al. 2023; Happé et al. 2024). We observed a similar pattern in Cluster #1 to the UK High cluster described by Happé et al. (2024) and the omega block reported by Rouges et al. (2023), although our analysis considered year-round heatwave samples, whereas those studies focused only on summer events.

Our results showed that the VAE's latent space effectively captures key atmospheric patterns that have previously been linked to extreme heat events, such as blocking highs, omega blocks, and persistent ridges similar to Happé et al. (2024) and Rouges et al. (2023).

Bischof, Sabine et al. (2023). "The Role of the North Atlantic for Heat Wave Characteristics in Europe, an ECHAM6 Study". In: *Geophysical Research Letters* 50.23, e2023GL105280. DOI: https://doi.org/10.1029/2023GL105280. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023GL105280.

Carril, Andrea F. et al. (2008). "Heatwaves in Europe: areas of homogeneous variability and links with the regional to large-scale atmospheric and SSTs anomalies". en. In: *Climate Dynamics* 30.1, pp. 77–98. DOI: 10.1007/s00382-007-0274-5.

Happé, T. et al. (2024). "Detecting Spatiotemporal Dynamics of Western European Heatwaves Using Deep Learning". In: *Artificial Intelligence for the Earth Systems* 3.4, e230107. DOI: 10.1175/AIES-D-23-0107.1.

Horton, Daniel E. et al. (2015). "Contribution of changes in atmospheric circulation patterns to extreme temperature trends". en. In: *Nature* 522.7557, pp. 465–469. DOI: 10.1038/nature14550.

Krüger, Julian et al. (2023). "Connecting North Atlantic SST Variability to European Heat Events over the Past Decades". en-US. In: *Tellus A: Dynamic Meteorology and Oceanography* 75.1. DOI: 10.16993/tellusa.3235.

Rouges, Emmanuel et al. (2023). "European heatwaves: Link to large-scale circulation patterns and intraseasonal drivers". In: *International Journal of Climatology* 43.7, pp. 3189–3209. DOI: https://doi.org/10.1002/joc.8024. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.8024.

Minor Comment 1

"This trend is projected to continue even at the lowest projected global warming scenario, and the intensity of extremes will increase proportionally with the amount of warming." L19-20 Is there a reference for this? My understanding was that this is not necessarily a proportional increase.

Response: Thank you for pointing this out. We added the references for L19-20 and also modified the text since it was not clearly distinguishing the difference between intensity and frequency.

> ... This trend is projected to continue even at the lowest projected global warming scenario (IPCC 2021). On average, the intensity of hot extremes increases approximately linearly with global warming, while the frequency of rare hot extremes increases more rapidly (IPCC 2021; Li et al. 2021; Fischer and Knutti 2014; Kharin et al. 2018; Fischer, Sippel, and Knutti 2021). ...
>
> Fischer, E. M. and R. Knutti (2014). "Detection of spatially aggregated changes in temperature and precipitation extremes". In: *Geophysical Research Letters* 41.2. DOI: 10.1002/2013GL058499.
>
> IPCC (2021). *Climate Change 2021: The Physical Science Basis*. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Edited by V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou. Cambridge University Press (In Press).
>
> Li, Chao et al. (2021). "Changes in Annual Extremes of Daily Temperature and Precipitation in CMIP6 Models". en. In: *Journal of Climate* 34.9, pp. 3441–3460. DOI: 10.1175/JCLI-D-19-1013.1.

Minor Comment 2

The motivation in the introduction seems to cover all types of extreme events and all over the globe, yet the focus of the manuscript is heatwaves over western Europe only.

Response: We appreciate this remark. Our intention with the opening paragraph was to provide a brief climate extremes context, but our scientific question and all analyses are specific to heatwaves over western Europe. In the revised version of the manuscript we will narrow the framing and state the scope more explicit in the introduction.

Minor Comment 3

In the introduction the authors motivate that heat extremes cause mortality and increased costs, in summer mostly. Then why does the study focus on year-round heatwaves? I think this is important to motivate, as heatwaves in western Europe don't cause impacts in winter.

Response:

Thank you for pointing this out. We analyze events from year-round samples to investigate the dynamical patterns that precede heatwaves not only in summer but also in winter. Since the largest impacts for heatwaves are observed in summer months, the introduction is biased towards impacts from summer heatwaves. We will clarify this distinction in the revised manuscript.

Minor Comment 4

Methods 2.4; is the model trained using r2? Or MSE? Lines 144-152; is this not better fitted in the result section? It is also mentioned here that it is difficult to capture the local surface conditions because of the course spatial resolution, but then why did the authors decide to re-grid from 0.25 to 0.7 degrees in spatial resolution?

Response: We thank the reviewer for highlighting this point. The reported r^2 values are evaluation metrics, not training objectives. We will move the r^2 values to Results and keep the Methods section focused on the training setup. Regarding resolution, we regridded ERA5 from 0.25°to 0.7° to reduce dimensionality, to emphasize large scale circulation patterns that are most relevant for our objective, to improve training stability under an MSE loss that already smooths small scales, and to facilitate comparison with

the framework of Happé et al. (2024). We will clarify these points in the revised version of the manuscript.

Minor Comment 5

Why do the authors choose MSLP, Z500, and STREAM250? Rather than different levels of Z or stream?

Response: These fields were selected because they are used in other studies to describe European heatwaves. We used STREAM250 and MLSP to compare our results with Happé et al. (2024). Jézéquel, Yiou, and Radanovics (2018) found that Z500 reproduces temperature anomalies more accurately than SLP. Other variables are chosen based on other heatwave studies (Domeisen et al. 2022; Rousi et al. 2022; Barriopedro et al. 2023; Kim and Seo 2023; Tian et al. 2024)

Barriopedro, D. et al. (2023). "Heat Waves: Physical Understanding and Scientific Challenges". In: *Reviews of Geophysics* 61.2, e2022RG000780. DOI: https://doi.org/10.1029/2022RG000780. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022RG000780.

Domeisen, Daniela I. V. et al. (2022). "Prediction and projection of heatwaves". en. In: *Nature Reviews Earth & Environment*. DOI: 10.1038/s43017-022-00371-z.

Happé, T. et al. (2024). "Detecting Spatiotemporal Dynamics of Western European Heatwaves Using Deep Learning". In: *Artificial Intelligence for the Earth Systems* 3.4, e230107. DOI: 10.1175/AIES-D-23-0107.1.

Jézéquel, Aglaé, Pascal Yiou, and Sabine Radanovics (2018). "Role of circulation in European heatwaves using flow analogues". en. In: *Climate Dynamics* 50.3, pp. 1145–1159. DOI: 10.1007/s00382-017-3667-0.

Kim, Jin-Yong and Kyong-Hwan Seo (2023). "Physical mechanisms for the dominant summertime high-latitude atmospheric teleconnection pattern and the related Northern Eurasian climates". en. In: *Environmental Research Letters* 18.10, p. 104022. DOI: 10.1088/1748-9326/acfa13.

Rousi, Efi et al. (2022). "Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia". en. In: *Nature Communications* 13.1, p. 3851. DOI: 10.1038/s41467-022-31432-y.

Tian, Yinglin et al. (2024). "Characterizing heatwaves based on land surface energy budget". en. In: *Communications Earth & Environment* 5.1, pp. 1–9. DOI: 10.1038/s43247-024-01784-y.

Minor Comment 6

Figure 4. If I understand correctly these samples are from all year round? Is the t-SNE trained only on train-data or on all samples?

Response:

We thank the reviewer for bringing this issue to our attention, and yes, that is correct. We used all samples for t-SNE analysis. We modified the text as follows to make this point clearer:

As an intermediate processing step, we first used Principal Component Analysis (PCA) to reduce the dimensionality of all heatwave samples from 1941-2022 to 50 components. Then, we applied the t-SNE algorithm to all heatwave samples to reduce them to 2 dimensions to visualize the latent space. This two-step approach is recommended for the analysis of a high-dimensional latent space to reduce the number of dimensions (see Maaten and Hinton (2008) and Pedregosa et al. (2011)) and helps to ensure efficiency and robustness. Because these steps involve stochasticity, we fixed the random seed to 42 (Adams 1979) for PCA, t-SNE, and GMM to ensure consistency across visualization runs.

Adams, Douglas (1979). *The Hitch Hiker's guide to the Galaxy*. eng. Pan original. London: Pan Books.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Minor Comment 7

Section 3.4 can use some more literature comparison, especially when discussing the dynamics leading to heatwaves (the causal pathways).

Response: We thank for this suggestion. We will modify this section to better explain the patterns we observed in Cluster #3 ans #4 and compare with findings from the literature.

Minor Comment 8

Some sentences need rephrasing, for example:
L195-297 "they show a negative tendency" –> positive? Tendency towards what?
L342-344 "the key difference in their study . . . " –> our study? Now it reads as if they (Happé et al.) used 11d multivariate data instead of you.

Response: We thank the reviewer for pointing this out. We revised the text so that it reads now as follows:

> . . . This negative anomaly pattern over the North Atlantic is consistent with results described by Krüger et al. (2023), Bischof et al. (2023), and Lipfert, Hand, and Brönnimann (2024), who showed that colder-than-usual North Atlantic sea surface temperatures (SSTs) with a negative tendency are associated with persistent negative anomalies. These conditions promote a deep North Atlantic trough and the subsequent formation of a European ridge, which in turn favors stronger and longer-lasting heatwaves as well as a shift toward positive summer temperature anomalies over Europe Krüger et al. (2023). . . .

> . . . Although our clustering approach is based on Happé et al. (2024), the key difference is that in our study, we train a VAE on 11-day multivariate heatwave samples from ERA5 reanalysis data to reveal interpretable heatwave regimes. . . .

Bischof, Sabine et al. (2023). "The Role of the North Atlantic for Heat Wave Characteristics in Europe, an ECHAM6 Study". In: *Geophysical Research Letters* 50.23,

e2023GL105280. DOI: https://doi.org/10.1029/2023GL105280. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023GL105280.

Happé, T. et al. (2024). "Detecting Spatiotemporal Dynamics of Western European Heatwaves Using Deep Learning". In: *Artificial Intelligence for the Earth Systems* 3.4, e230107. DOI: 10.1175/AIES-D-23-0107.1.

Krüger, Julian et al. (2023). "Connecting North Atlantic SST Variability to European Heat Events over the Past Decades". en-US. In: *Tellus A: Dynamic Meteorology and Oceanography* 75.1. DOI: 10.16993/tellusa.3235.

Lipfert, Laura, Ralf Hand, and Stefan Brönnimann (2024). "A Global Assessment of Heatwaves Since 1850 in Different Observational and Model Data Sets". In: *Geophysical Research Letters* 51.3, e2023GL106212. DOI: https://doi.org/10.1029/2023GL106212. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023GL106212.