**Responses to Reviewers' Comments**

for

# A Multivariate Analysis of Atmospheric Drivers for Western European Heatwaves

(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 the Editor

Dear Dr. Kai Kornhuber,

We would like to thank you and the reviewers for the valuable comments which helped improve the quality of our manuscript. Following your suggestion, we changed the title from "Understanding European Heatwaves with Variational Autoencoders" (egusphere-2025-2460) to "A Multivariate Analysis of Atmospheric Drivers for Western European Heatwaves".

We revised the manuscript according to the reviewers' comments. All changes are highlighted in the annotated version. Below is a summary of the main updates.

We revised the Abstract, Introduction, Methods, Discussion, and Conclusions to clarify where we follow the methodology proposed in Happé et al. (2024) and in which aspects we extend their work. These additions explain that our study uses the same general VAE training and the latent space clustering framework but applies it to ERA5 reanalysis, includes nine variables, uses longer temporal windows, analyzes heatwaves year-round, and compares historical and recent periods. We revised the manuscript to make a clearer distinction between methods, findings, and limitations.

We added experiments with random train/validation/test splits to address concerns about our temporal splitting. The resulting latent space still shows consistent seasonal organization and clustering of summer events. We also improved the explanation of reconstruction residuals and added a comparison of the effect of different hyperparameter choices on the reconstruction loss. We tested composite maps with different sample sizes to clarify our choice. Furthermore, we expanded the limitations section to better explain the assumptions of the method and possible sources of uncertainty.

We hope that these revisions address concerns, and we kindly ask you to consider the revised manuscript for further review in Earth System Dynamics.

Sincerely,

Aytaç Paçal, Birgit Hassler, Katja Weigel, Miguel-Ángel Fernández-Torres, Gustau Camps-Valls, Veronika Eyring

**Response to Reviewer 1 Comments**

https://doi.org/10.5194/egusphere-2025-2460-RC1

for

# A Multivariate Analysis of Atmospheric Drivers for Western European Heatwaves

(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 1

**General Comments.** This work uses a non-linear dimensionality reduction method to study heat wave characteristics in western Europe. More specifically, they train a 3D variational autoencoder to reconstruct 11-day windows of multiple atmospheric variables around historical heat wave onset dates. Afterwards, the trained VAE is used to embed heat waves from a test period temporally after the training period. Then, the embeddings are clustered, and a shift in frequency in these clusters between training and testing is observed.
Strengths:

- Non-linear method to analyze heat wave characteristics and their spatio-temporal trends.
- Relevant study area: western europe has recently experienced a series of devastating heat waves, which are assumed to often have been preceded by atmospheric blocking events.
- Detailed description of the methodology, which enables reproducibility
- If valid, the results are interesting: an observed shift of summer time heat wave characteristics in the past 20 years.

We appreciate the reviewer's thoughtful summary and positive feedback on our work. In the revised manuscript, we addressed the reviewer's specific comments to clarify the methodology further, strengthen the interpretation of the results, and enhance the overall robustness and readability of the work.

**Major Comment 1**

1. The VAE may potentially be applied in an OOD scenario: Data from 2001-2022 may be outside of the training distribution (heatwaves during 1941-1990). There is no guarantee that for out of domain samples, neural networks produce meaningful latent representations. This may lead to flawed interpretations on any such latent representations.

2. The previous comment is particularly relevant in the context of one of the main claims of this work: that there is a change in heat wave patterns after 2000. While this may very well be true, it can not be ruled out that this is just an artifact of the way this study was set up.

3. As a way forward I'd suggest to actually not do a temporal splitting of the train/val/test sets, but instead a purely random one. Ultimately you want to extract good representations for the entire time series 1941-2022

**Response:** We appreciate the reviewer's concern about potential out-of-distribution effects. Our primary objective was to assess whether recent heatwaves differ from a historical baseline (1941–1990). Not incidentally, we conducted 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 was trained on the entire period, as shown in Figure 1. Summer heatwaves still accumulate more closely together. This confirms that the clustering is not simply an artifact of non-stationarity in the time series but reflects consistent structural patterns across periods. We revised the introduction and methods to more clearly articulate why a temporal split was chosen in relation to our scientific question and added the new figure to the Appendix.

VAEs are primarily developed as generative models, capable of learning complex data distributions and generating realistic synthetic samples (Kingma and Welling 2019). Beyond generative tasks, VAEs have also been widely applied in anomaly detection across various domains, including 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 Happé et al. (2024) to a multivariate and year-round analysis of European heatwaves using reanalysis data. This study aims to understand how heatwaves develop, what contributes to their evolution, and how recent heatwaves are represented in relation to the heatwaves from the historical baseline. In contrast to Happé et al. (2024), who focused on heatwave samples extracted from KNMI-LENTIS dataset over Western Europe, we (i) include nine atmospheric variables from the ERA5 reanalysis dataset to assess their contribution to heatwave evolution, (ii) analyze spatiotemporal samples over a larger window (±5 days around the onset, i.e., 11 days in total) to characterize build-up and decay processes better, (iii) train on 11-day multivariate ERA5 heatwave samples that cover the full year with standardized anomalies, and (iv) assess generalization with an explicit historical-to-recent evaluation. To this end, we first train a spatiotemporal VAE on multivariate reanalysis heatwave samples from the historical period of 1941-1980. Then, we cluster and analyze the latent space representation of recent heatwaves (2000-2022) in Western Europe. Section 2 describes the data and methodology followed in this research. Section 3 provides the results for the heatwave cluster analysis. Finally, the main results are summarized and discussed in Section 4.

Each heatwave sample consisted of 9 climate variables for 11 days on a $64 \times 192$ spatial grid (North Atlantic, see hatched gray region in Figure 2), resulting in input tensors with the shape $(9, 11, 64, 192)$. The model was trained on heatwave samples from 1941 to 1990, validated on samples from 1991 to 2000, and tested on samples from 2001 to 2022, resulting in 1,408, 320, and 928 samples, respectively. This split enables an out-of-distribution setting where the test years reflect recent climate conditions that differ from the historical baseline because recent decades exhibit non-stationary warming and circulation changes (Chitsaz et al. 2023; Rouges et al. 2023). Because the 1941–1980 baseline precedes the strong post-1980 warming trend (Elguindi, Rauscher, and Giorgi 2013; Reid et al. 2016; IPCC 2021), this choice naturally leads to more detected heatwave events in recent decades. However, our goal is not to quantify frequency changes but to characterize the structural and dynamical patterns of heatwaves. The model focuses on circulation-related variability rather than long-term thermodynamic shifts. To assess robustness against non-stationarity, we also trained a version of the VAE using random train (0.8)/validation(0.05)/test(0.15) splits across the whole 1941–2022 period. While this approach provides a smoother latent space as expected, the resulting latent space (Figure A3) shows consistent seasonal clustering, confirming that the identified heatwave types are not artifacts of temporal trends. By training the model on historical heatwave samples, the VAE learns an internal representation of typical patterns over time, capturing the complex spatial and temporal relationships between the different climate-relevant variables. This allows the model to learn a low-dimensional representation of extreme multivariate events in the latent space.

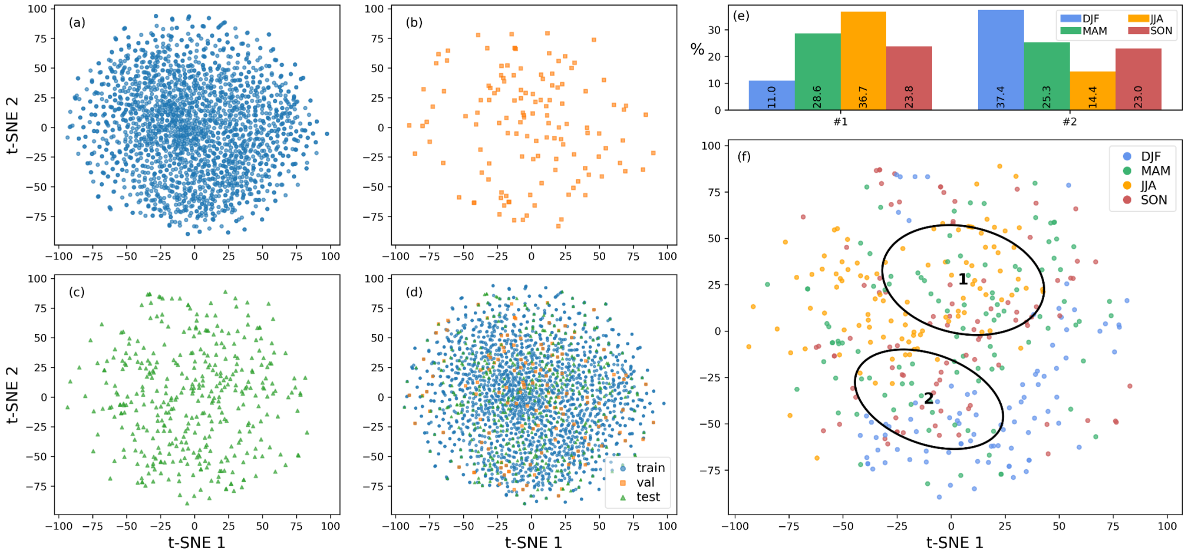


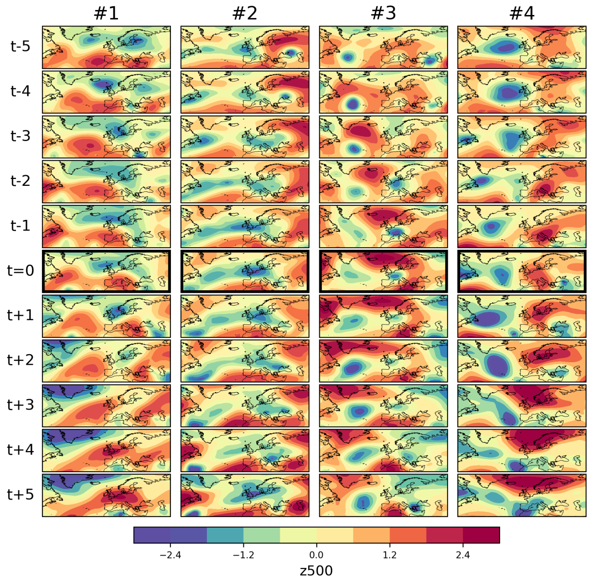

Figure 1: t-SNE representation of the latent space with random-split data (0.8/0.05/0.15). Following our workflow (testing with more than two components), a 2-component GMM provided the best fit, as it was closest to a single Gaussian distribution and thus better captured the smooth latent space structure.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

**Major Comment 2**

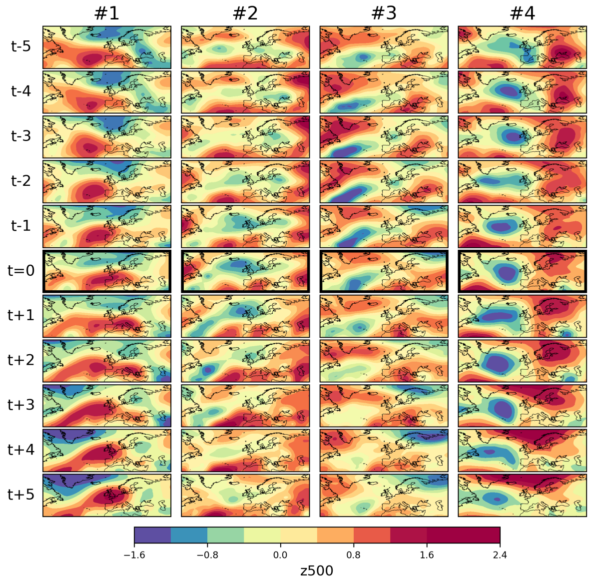
1. Why do you produce composite maps, instead of directly reconstructing the center of each cluster?
2. Could your composite maps not lead to unphysical patterns? For instance, phenomena could cancel out if they are just shifted in space slightly?

**Response:** 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 acknowledge that averaging might cancel signals if patterns are spatially shifted. To address this, 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 five 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. We added comparison figures to the Appendix and explained this in the revised manuscript.

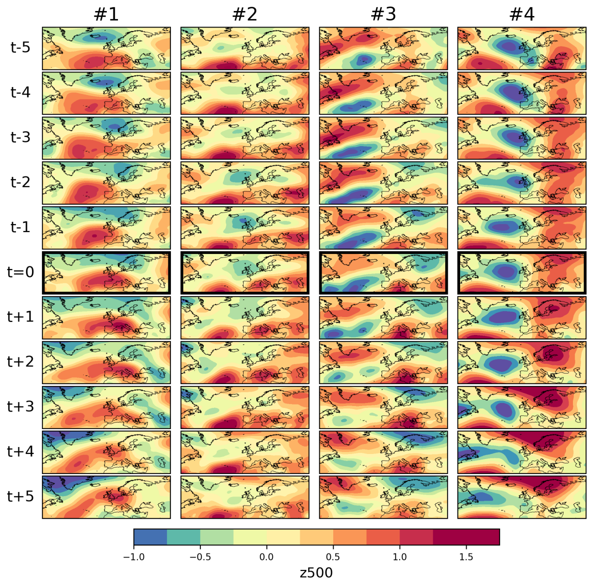
GMM is an unsupervised learning method that probabilistically clusters data points into different Gaussian components. This technique identifies the underlying groups or clusters of heatwave events in the latent space. Each cluster corresponds to a distinct type of heatwave event characterized by similar atmospheric conditions. We extracted the centroids for each GMM cluster to interpret the resulting clusters. For interpretation, we do not reconstruct the cluster centroid directly through the decoder, since the stochastic nature of the VAE means that latent positions and cluster centers can vary slightly across runs. Instead, we retrieve the 100 heatwave samples closest to each cluster centroid in the latent space and compute composite maps of the nine climate-relevant variables (Happé et al. 2024). This approach produces spatial patterns that are more stable and physically consistent, because neighboring samples in latent space already share similar structures (Kingma and Welling 2019). This process enables us to analyze the spatial patterns associated with each cluster and the conditions described by these variables, culminating in the heatwave onset (Rouges et al. 2023). A potential concern is that averaging might partially cancel features if events are spatially shifted. To assess this, we tested composite maps with different sample sizes ($N = \{1, 5, 10, 25, 50, 100\}$). Patterns remain consistent for $N > 5$, while the single-sample composite differs, as expected, since the 1-sample map corresponds to an individual event and cluster centers (see Appendix Figure A4-A9). Due to the stochastic nature of the method, the position of the heatwave sample assigned as the cluster center in the latent space may not be stable. These tests confirm that the composite maps are therefore more reliable in capturing robust spatial signals representative of each heatwave cluster.



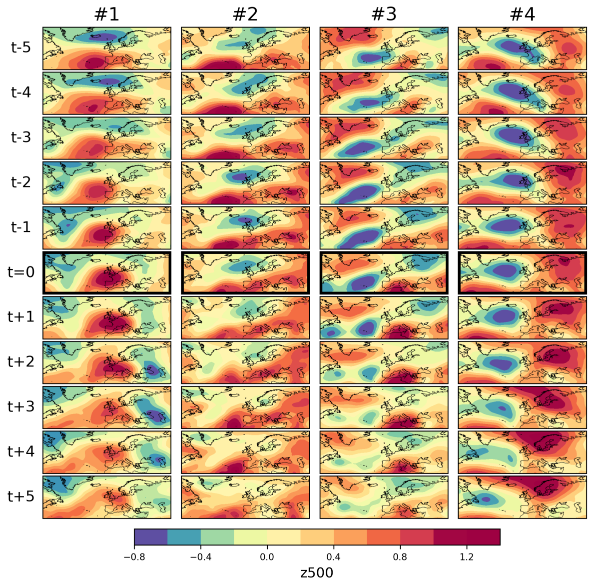

(a) 1-sample composite map



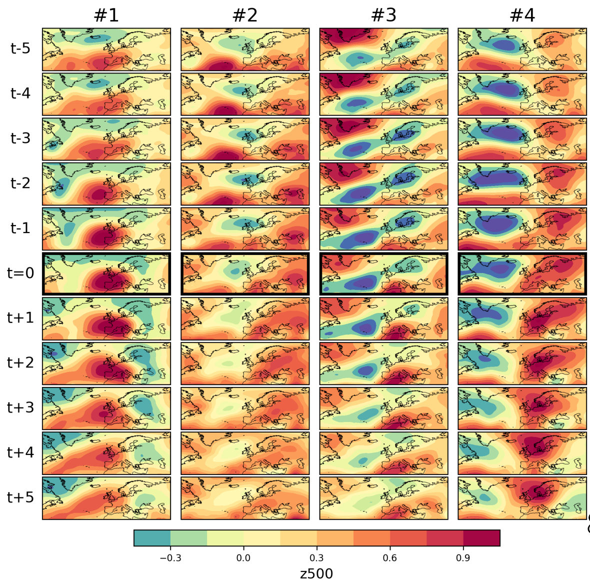

(b) 5-sample composite map



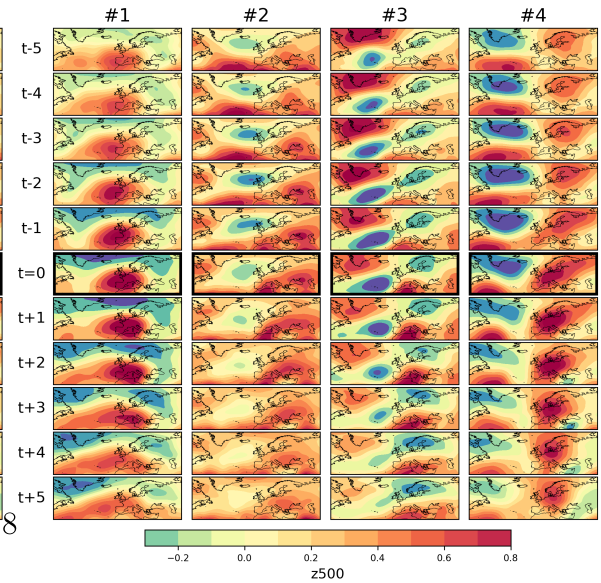

(c) 10-sample composite map




(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 3

In this study you introduce atmospheric patterns that have been observed prior to heat waves. One open question remains, if these patterns exclusively arise prior to heatwaves - or if they are not always succeeded by extreme temperature anomalies. This would be very valuable to understand the physical mechanisms but also assess potential predictive skill.

**Response:** We agree that this is an important point. In this study, we focused on identifying and characterizing the atmospheric patterns that are commonly observed before heatwaves. However, these patterns may also occur without being followed by extreme temperature anomalies, and distinguishing such cases is indeed crucial for both understanding the underlying physical mechanisms and assessing predictive skill. While a complete predictive analysis is beyond the scope of the present paper, we computed the mean bias between the summer cluster center and 11-day running window samples from the summer months of 2015–2021 for each variable, as shown in Figure 3. These biases did not reveal a strong predictive signal for heatwaves. We explained this limitation in the revised manuscript and highlighted it as an important direction for future work.

> An open question concerns whether the atmospheric patterns identified before heatwave onsets also appear during periods without subsequent temperature extremes. We computed the mean bias between the composite fields of the summer cluster center and 11-day running window samples from all summer days from 2015 to 2021 for each variable. This approach tests whether these circulation patterns exclusively arise before the heatwaves. Our preliminary results did not show a clear predictive signal, suggesting that these large-scale configurations can occur without always leading to a heatwave. Future work could investigate the potential predictive value and improve understanding of the physical mechanisms linking circulation anomalies to heatwave development.

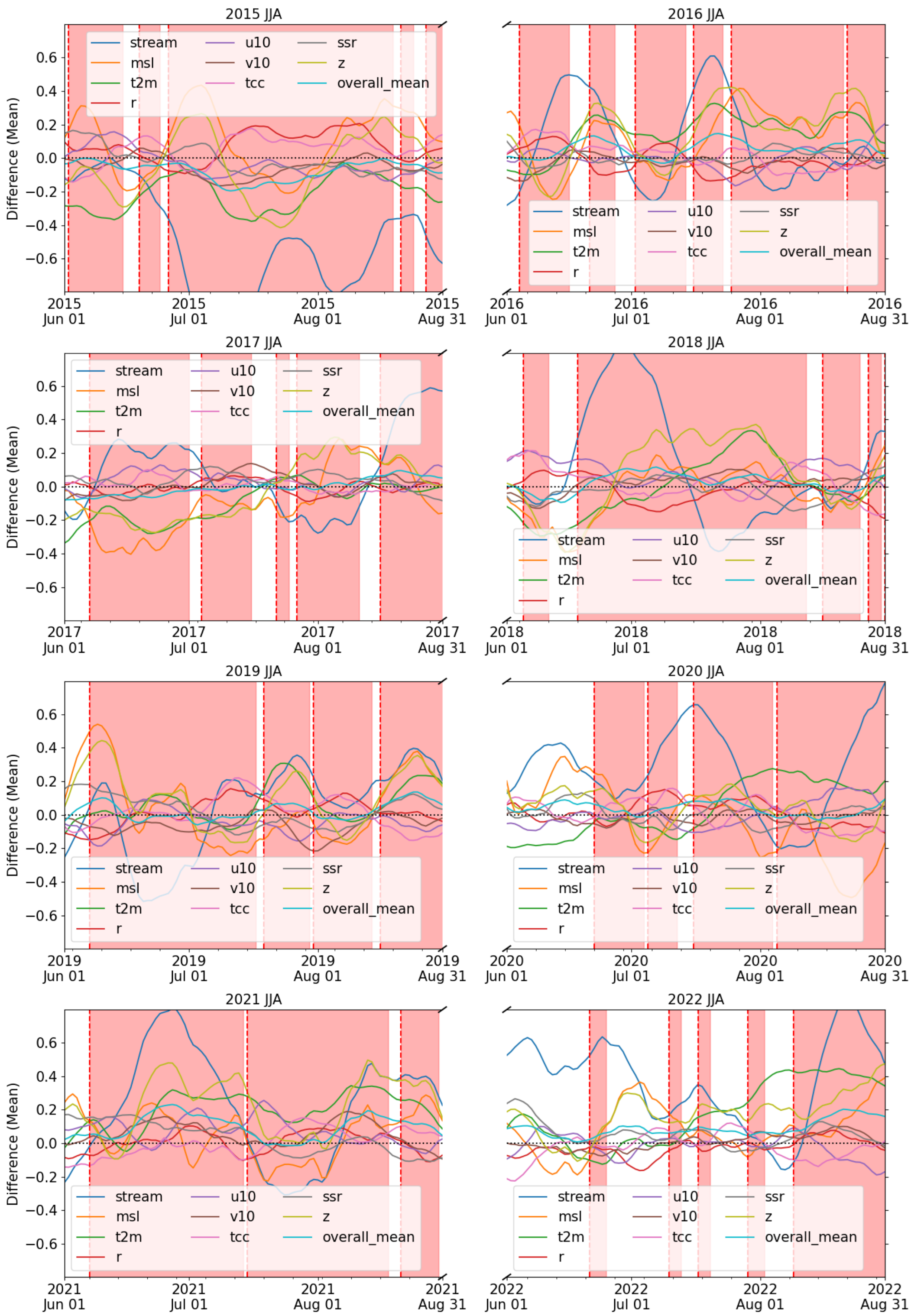



Figure 3: The mean bias between the center sample from the summer cluster and the 11-day running window samples from the summer months. The x-axis is the central date for the 11-day window. The different lines correspond to individual variables. The red shaded regions next to the dashed line indicate when heatwaves start and how long they last. To keep the plot simple, heatwaves that coincide in different locations are combined into a single representation.

## Major Comment 4

Why first t-SNE and then cluster? Why not directly cluster on the embeddings?

**Response:**

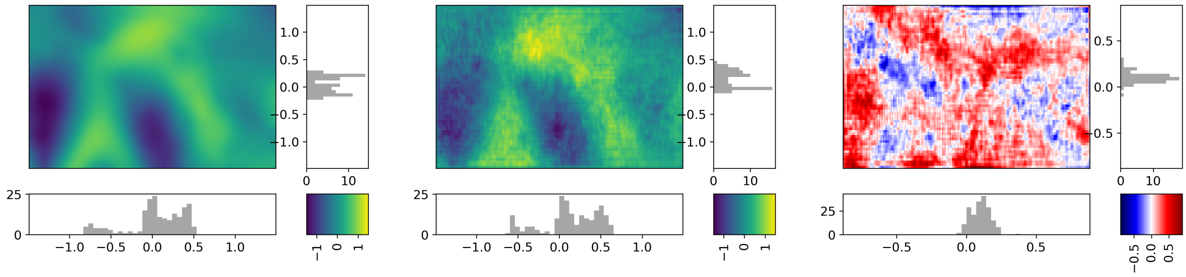
The pre-reduction step was recommended in the t-SNE paper (Maaten and Hinton 2008) and the documentation for the scikit-learn function (Pedregosa et al. 2011), which suggests applying a preliminary reduction step such as PCA before using t-SNE when working with latent spaces of higher dimension (above 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. To clarify this, we modified the manuscript 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 two dimensions for visualization of the latent space. This two-step approach is recommended by Maaten and Hinton (2008) for t-SNE visualizations of high-dimensional latent space to reduce the number of dimensions and help to ensure efficiency and robustness (Pedregosa et al. 2011). 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.

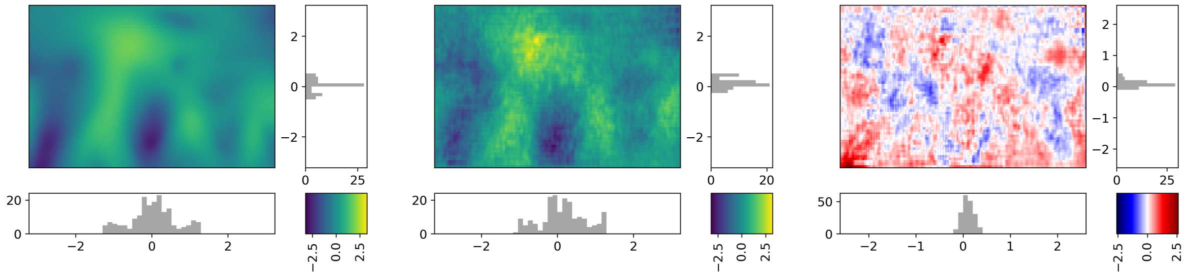
## Major Comment 5

Since your reconstruction accuracy is far from perfect (table 2), I wonder if you have looked into the model residuals. What are the errors the model makes? Does it miss any significant patterns related to heat waves? Is it unbiased? How well does it capture the extremes (i.e. the grid cells with heat waves)? This may be especially tricky since you seem to train on MSE.

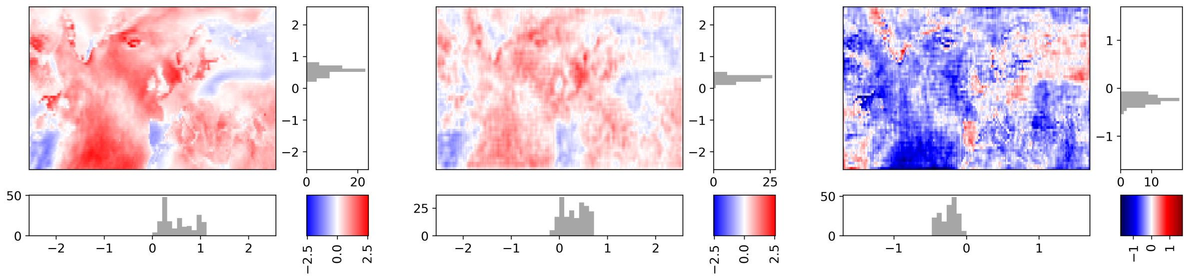
**Response:** We thank the reviewer for this suggestion. The reconstructions accurately reproduce large-scale anomaly structures but are smoother than the inputs, resulting in a slight underestimation of extreme intensities. This outcome is consistent with the use of MSE loss, which favors average solutions. Notably, the model retains the spatial location and amplitude of large-scale features, indicating that the essential heatwave-related circulation patterns are preserved despite reduced intensity of extremes (see Figure 4). We added these figures and $R^2$ scores for the training, validation, and test sets to the manuscript.



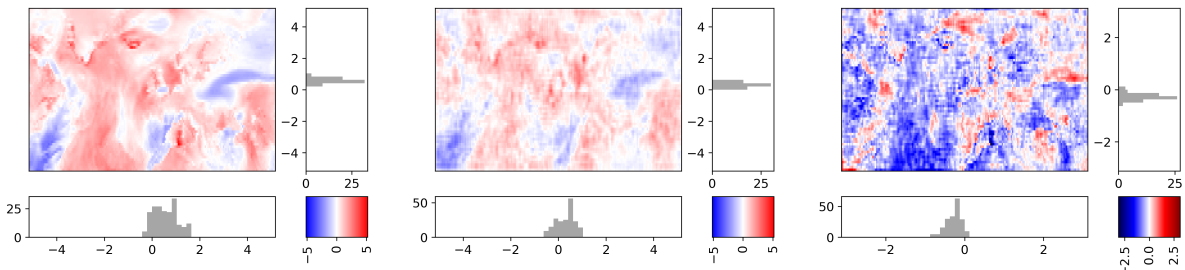
(a) Geopotential height at 500 hPa for a random sample (left), its VAE reconstruction (middle), and the resulting reconstruction bias (right), each averaged over the time dimension. The histograms below and to the right of each panel show the distributions of longitudinal and latitudinal values, respectively.



(b) Geopotential height at 500 hPa for a random sample (left), its VAE reconstruction (middle), and the resulting reconstruction bias (right), each shown at the onset date. The histograms below and to the right of each panel show the distributions of longitudinal and latitudinal values, respectively.



(c) 2-meter temperature for a random sample (left), its VAE reconstruction (middle), and the resulting reconstruction bias (right), each averaged over the time dimension. The histograms below and to the right of each panel show the distributions of longitudinal and latitudinal values, respectively.



(d) 2-meter temperature for a random sample (left), its VAE reconstruction (middle), and the resulting reconstruction bias (right), each shown at the onset date. The histograms below and to the right of each panel show the distributions of longitudinal and latitudinal values, respectively.

Figure 4: Input, reconstruction, and bias maps of geopotential height and temperature, shown as the time mean and at the heatwave onset date.

## Major Comment 6

Figure 5 would greatly benefit from actually including the maps of z500 from Lhotka & Kysely (2022) and plotting the differences.

**Response:** We thank the reviewer for this constructive suggestion. Since our focus here was on exploring the capability of the VAE, we chose to compare our results qualitatively with the published maps rather than repeating their analysis. We modified the discussion around Figure 5 (now 6) to clarify the comparison to Lhotka and Kyselý (2022).

> This procedure allows us to place these historically significant heatwaves within the structure of the learned latent space and to compare the resulting anomaly patterns directly with those documented in the literature. As shown in Figure 6, the resulting geopotential height at 500 hPa (*z500*) anomaly maps exhibit similar patterns to those reported by Lhotka and Kyselý (2022). In particular, the 2010 heatwave shows the characteristic strong positive z500 anomalies over western Russia, the 2003 event exhibits the ridge over central Europe, and the 2018 heatwave displays height anomalies over Scandinavia. These spatial features closely match the patterns reported in Figure 3 of Lhotka and Kyselý (2022), demonstrating that the VAE organizes these events in a way that is consistent with established dynamical interpretations.

## Major Comment 7

L. 342ff can you elaborate a bit more how this work improves over the previous work from Happé et al 2024? Also, as is I assume the “in their” should be replaced by “to their”?

**Response:** We appreciate this relevant 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. We revised the Abstract, Introduction, and Discussion to make this novelty explicit and clearly separate where our study confirms previous results versus where it extends them.

## Major Comment 8

L 386ff - the current limitations read somewhat superficial. This approach builds on top of many assumptions. These should be clearly stated, please extend the section on limitations.

**Response:** We thank the reviewer for this point. We extended the discussion of limitations to clarify the assumptions on which the methodology is based as follows:

> Despite these promising results for identifying heatwaves and their underlying atmospheric conditions, several limitations of the current approach must be acknowledged. First, we adopt a percentile-based threshold for detecting heatwave sample onset dates using 2-meter temperature anomalies over Western Europe. However, there is no universally accepted definition of heatwaves (Barriopedro et al. 2023; Boni et al. 2023). Alternative heatwave indices could yield different heatwave samples and thus different atmospheric patterns. Second, while we used variables related to atmospheric heatwaves, other potentially relevant drivers of heatwave dynamics may not be represented. As a result, the clusters we identify should be interpreted as emerging from the chosen subset of variables rather than from the full spectrum of processes affecting heatwaves. In addition, ERA5 itself has known uncertainties (Soci et al. 2024). Future work could improve this by also analyzing other variables to capture additional feedback processes. Third, the VAE is trained on historical heatwaves (1941–1990) and tested with recent heatwaves. While this was a deliberate choice to investigate how recent heatwaves differ from historical heatwaves, it assumes that the latent representation learned from the historical period can adequately model unprecedented patterns arising from climate change. Furthermore, the VAE approach introduces some uncertainties due to the choices in model architecture. Key hyperparameters (such as the dimensionality of the latent space, learning rate, batch size, and the number of layers) can significantly influence training outcomes and the structure of the latent space. In addition to

this, dimensionality reduction using PCA and t-SNE, followed by clustering with GMM, further influences how heatwave events are grouped in the latent space, as all these steps involve stochastic elements. Finally, the VAE reconstructions tend to smooth out extremes, and the use of composite maps that average 100 nearest latent samples, while reducing randomness and providing a more stable representation of atmospheric conditions, may mask variability within clusters. These limitations highlight that the VAE clustering approach is one possible data-driven analysis of atmospheric dynamics associated with heatwaves, rather than a definitive classification.

**Major Comment 9**

Overall the storyline of the discussion needs to be streamlined. It jumps between results, limitations and outlooks. For instance, why is L 395ff after the limitations section?

**Response:** We thank the reviewer for bringing this to our attention. We restructured the discussion to remove the current back-and-forth and improve the flow. For example, we improved the text on limitations and moved some parts to methodology to have a better flow, as shown in Comment 4 and 8.

**Major Comment 10**

The choice of splitting the dataset temporally also potentially influences attribution of the breakpoint you claim to find in heat wave characteristics. How robust is this finding if you change the periods over which you aggregate the latent space samples (in figure 4)?

**Response:** We thank the reviewer for bringing this point to our attention. Our goal was to assess whether recent heatwaves (2001–2022) differ from those in a historical baseline (1941–1990), which requires a temporal split by design. Importantly, the strong warming trend in Europe in terms of heatwave occurrence and intensity after the 1980s is well-documented in the literature (IPCC 2021; Reid et al. 2016; Lo and Hsu 2010), and our findings are consistent with this broader evidence. The breakpoint signal would not be expected in a random split, where events from both periods are mixed. We modified the methodology to better explain temporal splitting and included the analyses mentioned in Comment 1.

### Minor Comment 1

L 1-2 & L. 42f not sure I can follow the logic here. To me these are two separate aspects: one is defining heatwaves, which should be done using temperatures (or if you like, including also humidity) - and the other one is understanding their dynamics, for which other variables are also important.

**Response:** Thank you very much for bringing this issue to our attention. We modified the abstract opening to state that definitions are typically based on temperature thresholds, while the study of underlying dynamics requires a broader multivariate perspective. We also adapt the introduction to separate these two aspects explicitly.

### Minor Comment 2

L. 64 reads misleading, i would say the predominant use of VAEs is generative modeling - and anomaly detection is just one application.

**Response:** Thank you very much for your comment. We modified the revised manuscript as follows:

> ...VAEs are primarily developed as generative models, capable of learning complex data distributions and generating realistic synthetic samples (Kingma and Welling 2019). Beyond generative tasks, VAEs have also been widely applied in anomaly detection across various domains, including network security, risk management, health monitoring, and computer vision (Pang et al. 2021; Nassif et al. 2021; Albuquerque Filho et al. 2022).

## Minor Comment 3

L. 87, just to be sure, you compute the climatology, after having applied the temporal operators from table 1, correct?

**Response:** We thank the reviewer for bringing this issue to our attention, and yes, that is correct. We first apply the temporal operators listed in Table 1 to the hourly ERA5 data, and then compute the climatology from these aggregated fields. We explained this in the manuscript to avoid any ambiguity as follows:

> ...Since ERA5 data is provided in hourly resolution, we first applied the temporal aggregation operators listed in Table 1 to convert the hourly ERA5 fields to daily values. From these daily fields, we then computed the 1941–1980 climatological mean and standard deviation using a 15-day moving window centered on each calendar day, which we used to standardize the variables and remove seasonal variability. 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 studies on extreme-event detection (Sulikowska and Wypych 2020). We used this period to calculate the climatology, as it precedes the stronger warming trends after 1980 in Europe (Elguindi, Rauscher, and Giorgi 2013; Reid et al. 2016). The standardized daily anomalies were then used for all subsequent analyses.

## Minor Comment 4

Have you tried a more standard 2D VAE (e.g. using an Imagenet pretrained ConvNext backbone)? In many applications the additional inductive bias on the temporal structure of the data may not translate in actual better performance, so it could be interesting to see if this is one case where it does.

**Response:** We thank the reviewer for this suggestion. We agree that comparing our approach with a more standard 2D VAE architecture would be an interesting exercise. In fact, we also experimented with a VAE with 2D convolutional encoder (decoder) layers to capture spatial features, followed (preceded) by LSTM layers to capture temporal information. However, this setup did not yield stable or satisfactory reconstructions, likely because the LSTM struggled to capture the complex spatiotemporal dynamics of heatwave events when applied after a spatially compressed representation was used. By contrast, the 3D convolutional VAE provided more robust and coherent representations. Since our focus is on demonstrating the feasibility of incorporating temporal information into the VAE for heatwaves, we therefore focused on the 3D VAE for this study.

**Minor Comment 5**

L. 191 not sure "temporal change" is the best term here, i read it first as if the temporal characteristics of heatwaves changed

**Response:** Thanks for pointing this out. Our intention was not to suggest that the temporal characteristics of individual heatwaves changed, but rather that the distribution of heatwave types shifted over time in the latent space. To avoid ambiguity, we rephrased this part.

**Minor Comment 6**

L. 200 not sure it proves that the patterns are necessarily multivariate - probably you could learn this from a single variable like air temperature also.

**Response:** We thank the reviewer for this helpful observation. While certain patterns could in principle be learned from a single variable, our goal was to demonstrate that a multivariate approach provides a more detailed interpretation of anomaly patterns associated with heatwaves. We revised the wording accordingly to avoid implying that the VAE model captures the multivariate nature, but instead emphasize that the added variables improve interpretability.

> As shown in Figure 4f, some clusters are dominated by heatwaves happening at specific seasons (e.g., Cluster *#2* with mostly June-July-August (JJA) events), indicating that the model distinguishes between heatwave patterns that align with different seasonal characteristics. While some of this separation could also emerge from individual variables such as air temperature, using multiple variables provides

a more complex representation of the atmospheric conditions associated with each cluster.

### Minor Comment 7

L. 210 - cluster 1 & 2 are clearly dominated by one season, but 3&4 not sure i agree with your assignment, e.g. cluster 4 also has 29% DJF

**Response:** We thank the reviewer for this careful observation. Our assignment of Clusters 3 and 4 to transitional seasons was based on their modal distributions, which are dominated by SON and MAM events, respectively. We acknowledge that Cluster 4 also contains a notable fraction of DJF events (29%), which indicates that the separation between seasonal and transitional clusters is not absolute. We also examined the mean anomaly values of each cluster as shown in Figure 5. For each variable, we computed the mean standardized anomaly values over all heatwave samples in a cluster. These mean values represent the typical anomaly structure captured by each cluster. For example, Cluster 1 represents winter heatwaves with weaker t2m anomalies but warm advection (positive mean v10 anomalies) and lower solar input; Cluster 2 corresponds to summer heatwaves with the highest mean t2m, humidity, tcc and z500 anomalies; and Clusters 3 and 4 capture transition season events, one with high solar radiation and the other with strong pressure anomalies, respectively. In the revised manuscript, we explained that Clusters 3 and 4 capture mixed-season dynamics, with SON/MAM as the dominant but not exclusive contributions.

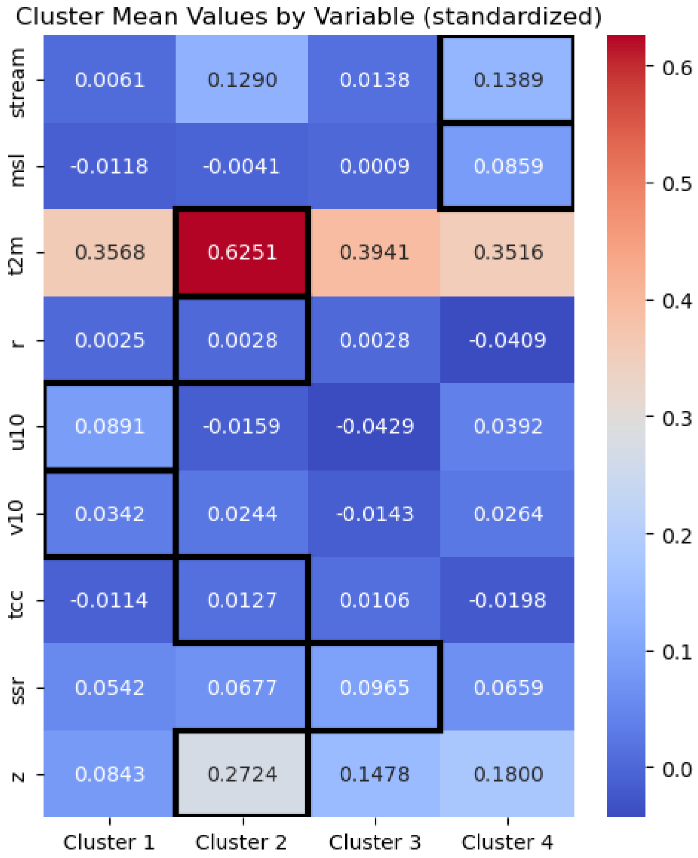



Figure 5: Mean standardized anomaly values for each variable, averaged over all heatwave samples in each latent cluster. The values summarize the typical anomaly patterns represented by the four clusters. Black boxes represent the highest value for each variable.

**Minor Comment 8**

L 356-363 this reads like methods and not discussion & conclusion

**Response:** Thank you for your comment. We moved this part to the Variational Autoencoder subsection:

> . . . A VAE, therefore, summarizes complex, high-dimensional data into a low-dimensional embedding. This low-dimensional latent space compresses the input data, allowing for the discovery of hidden patterns within the data. It can be further analyzed using visualization techniques and clustering algorithms to identify patterns and regimes in the data (Happé et al. 2024; Lindhe, Ringqvist, and Hult

2021). The reconstruction produced by the VAE can be interpreted as a learned climatological baseline or 'normal state,' against which deviations can be studied. We focused on using latent space to understand the atmospheric patterns underlying different heatwave types. Similar latent space approaches have been successfully applied in climate research (Behrens et al. 2022; Oliveira et al. 2022; Shamekh et al. 2023; Mooers et al. 2023; Camps-Valls et al. 2025).

### Minor Comment 9

Could the latent space clustering primarily reflect heatwave intensity rather than pattern type? How important is the non-linear/multimodal aspect of your method for separating these types?

**Response:** We thank the reviewer for this insightful comment. We agree that clustering in the latent space could be dominated by heatwave intensity. To address this issue, we examined the mean anomaly values of each cluster as shown in Figure 5. While the heatwave intensity does play a role in the clustering (e.g., Cluster 2 shows the strongest positive t2m anomalies), clusters differ in other variables, indicating that the separation is not only based on intensity. For example, Cluster 1 represents winter heatwaves with weaker t2m anomalies but warm advection (positive mean v10 anomalies) and lower solar input; Cluster 2 corresponds to summer heatwaves with the highest mean t2m, humidity, and z500 anomalies; and Clusters 3 and 4 capture transition season events, one with high solar radiation and the other with strong pressure anomalies, respectively. We added this analysis to the manuscript.

### Minor Comment 10

A stylistic note: The introduction might come across as overly alarmist. Consider softening the tone slightly.

**Response:** We thank the reviewer for this remark. We revised the introduction to soften the tone and maintain the focus on the scientific background. We revised the parts related to the impact of climate change on global health.

**Response to Reviewer 2 Comments**

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

for

# A Multivariate Analysis of Atmospheric Drivers for Western European Heatwaves

(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 revised 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 revised 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, capable of learning complex data distributions and generating realistic synthetic samples (Kingma and Welling 2019). Beyond generative tasks, VAEs have also been widely applied in anomaly detection across various domains, including 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 Happé et al. (2024) to a multivariate and year-round analysis of European heatwaves using reanalysis data. This study aims to understand how heatwaves develop, what contributes to their evolution, and how recent heatwaves are represented in relation to the heatwaves from the historical baseline. In contrast to Happé et al. (2024), who focused on heatwave samples extracted from KNMI-LENTIS dataset over Western Europe, we (i) include nine atmospheric variables from the ERA5 reanalysis dataset to assess their contribution to heatwave evolution, (ii) analyze spatiotemporal samples over a larger window ($\pm$5 days around the onset, i.e., 11 days in total) to characterize build-up and decay processes better, (iii) train on 11-day multivariate ERA5 heatwave samples that cover the full year with standardized anomalies, and (iv) assess generalization with an explicit historical-to-recent evaluation. To this end, we first train a spatiotemporal VAE on multivariate reanalysis heatwave samples from the historical period of 1941-1980. Then, we cluster and analyze the latent space representation of recent heatwaves (2000-2022) in Western Europe. Section 2 describes the data and methodology followed in this research. Section 3 provides the results for the heatwave cluster analysis. Finally, the main results are summarized and discussed in Section 4.

#### 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). We explained these choices and benefits in the revised methodology section.

> We used the ERA5 reanalysis dataset (Hersbach et al. 2017; Soci et al. 2024) to detect and understand heatwave dynamics. We focused only on atmospheric variables, as shown in Table 1, including those used in Happé et al. (2024) as well as additional variables relevant to heatwave dynamics identified in previous studies, such as geopotential height at 500 hPa (z500) as it captures mid-tropospheric circulation patterns and has been shown to better reproduce temperature anomalies associated with European heatwaves than sea-level pressure (Jézéquel, Yiou, and Radanovics 2018; Domeisen et al. 2022; Rousi et al. 2022; Barriopedro et al. 2023; Kim and Seo 2023; Tian et al. 2024).
>
> . . .
>
> The resulting dataset comprises daily observations from 1940 to 2022 for the North Atlantic region, spanning from 75°W to 60°E and from 30°N to 75°N on a 0.7 ° grid, as shown in Figure 2. ERA5 data was regridded from its native 0.25°resolution to 0.7°to ensure direct comparability with Happé et al. (2024), where this resolution

corresponds to the native grid of the EC-Earth3 model. Furthermore, this coarser resolution also reduces the very high dimensionality of the multivariate input while retaining the large-scale circulation patterns. Although ERA5's native resolution is finer, the reconstruction loss of the variational autoencoder emphasizes large-scale circulation patterns while smoothing out small-scale variability. Retaining the native 0.25°grid would therefore not provide additional benefit while substantially increasing computational cost. Since ERA5 data is provided in hourly resolution, we first applied the temporal aggregation operators listed in Table 1 to convert the hourly ERA5 fields to daily values. From these daily fields, we then computed the 1941–1980 climatological mean and standard deviation using a 15-day moving window centered on each calendar day, which we used to standardize the variables and remove seasonal variability. 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 studies on extreme-event detection (Sulikowska and Wypych 2020). We used this period to calculate the climatology, as it precedes the stronger warming trends after 1980 in Europe (Elguindi, Rauscher, and Giorgi 2013; Reid et al. 2016). The standardized daily anomalies were then used for all subsequent analyses.

### 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 modified the manuscript and emphasized that our conclusions pertain to the relative structures of heatwaves, rather than their frequency of occurrence.

> Each heatwave sample consisted of 9 climate variables for 11 days on a $64 \times 192$ spatial grid (North Atlantic, see hatched gray region in Figure 2), resulting in input tensors with the shape $(9, 11, 64, 192)$. The model was trained on heatwave samples from 1941 to 1990, validated on samples from 1991 to 2000, and tested on samples from 2001 to 2022, resulting in 1,408, 320, and 928 samples, respectively. This split enables an out-of-distribution setting where the test years reflect recent climate conditions that differ from the historical baseline because recent decades exhibit non-stationary warming and circulation changes (Chitsaz et al. 2023; Rouges et al. 2023). Because the 1941–1980 baseline precedes the strong post-1980 warming trend (Elguindi, Rauscher, and Giorgi 2013; Reid et al. 2016; IPCC 2021), this choice naturally leads to more detected heatwave events in recent decades. However, our goal is not to quantify frequency changes but to characterize the structural and dynamical patterns of heatwaves. The model focuses on circulation-related variability rather than long-term thermodynamic shifts. To assess robustness against non-stationarity, we also trained a version of the VAE using random train (0.8)/validation(0.05)/test(0.15) splits across the whole 1941–2022 period.

While this approach provides a smoother latent space as expected, the resulting latent space (Figure A3) shows consistent seasonal clustering, confirming that the identified heatwave types are not artifacts of temporal trends. By training the model on historical heatwave samples, the VAE learns an internal representation of typical patterns over time, capturing the complex spatial and temporal relationships between the different climate-relevant variables. This allows the model to learn a low-dimensional representation of extreme multivariate events in the latent space.

. . .

GMM is an unsupervised learning method that probabilistically clusters data points into different Gaussian components. This technique identifies the underlying groups or clusters of heatwave events in the latent space. Each cluster corresponds to a distinct type of heatwave event characterized by similar atmospheric conditions. We extracted the centroids for each GMM cluster to interpret the resulting clusters. For interpretation, we do not reconstruct the cluster centroid directly through the decoder, since the stochastic nature of the VAE means that latent positions and cluster centers can vary slightly across runs. Instead, we retrieve the 100 heatwave samples closest to each cluster centroid in the latent space and compute composite maps of the nine climate-relevant variables (Happé et al. 2024). This approach produces spatial patterns that are more stable and physically consistent, because neighboring samples in latent space already share similar structures (Kingma and Welling 2019). This process enables us to analyze the spatial patterns associated with each cluster and the conditions described by these variables, culminating in the heatwave onset (Rouges et al. 2023). A potential concern is that averaging might partially cancel features if events are spatially shifted. To assess this, we tested composite maps with different sample sizes ($N = \{1, 5, 10, 25, 50, 100\}$). Patterns remain consistent for $N > 5$, while the single-sample composite differs, as expected, since the 1-sample map corresponds to an individual event and cluster centers (see Appendix Figure A4-A9). Due to the stochastic nature of the method, the position of the heatwave sample assigned as the cluster center in the latent space may not be stable. These tests confirm that the composite maps are therefore more reliable in capturing robust spatial signals representative of each heatwave cluster.

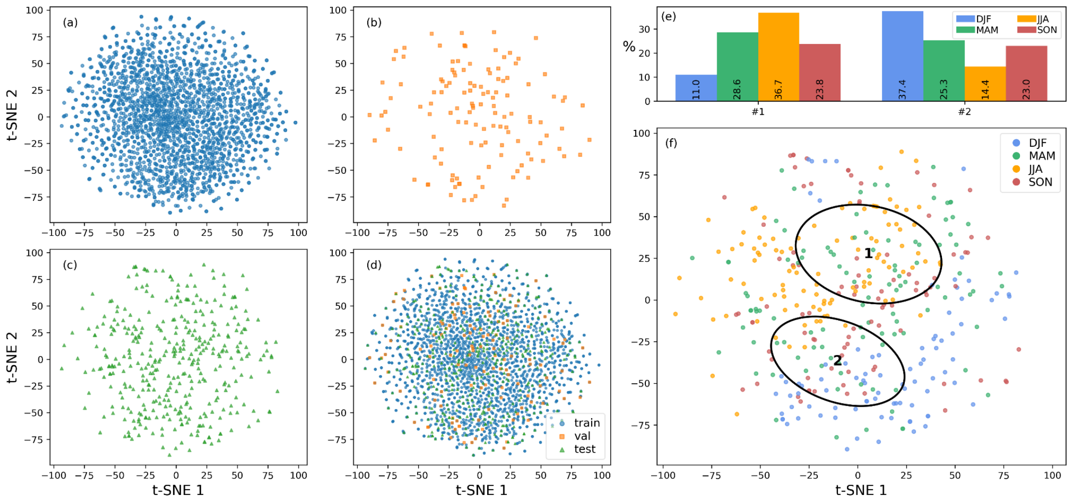
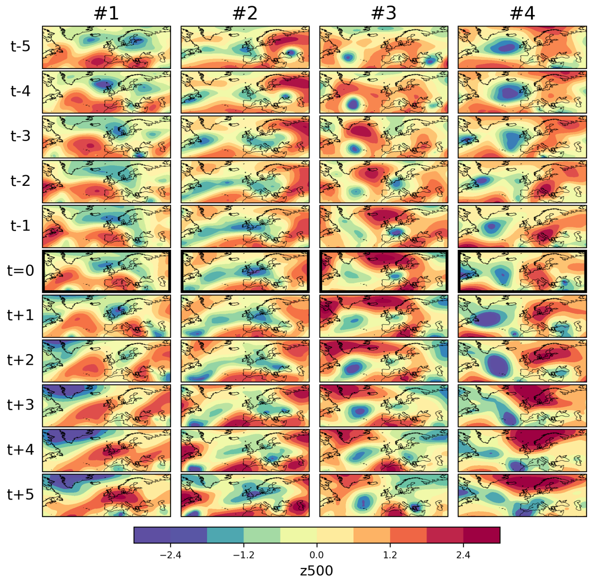
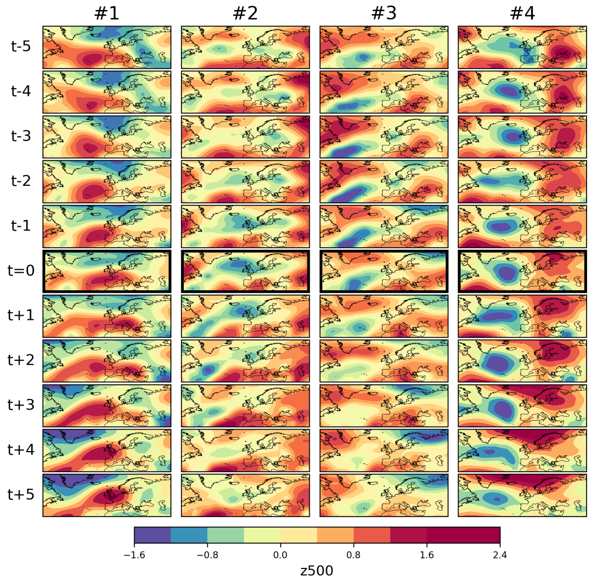


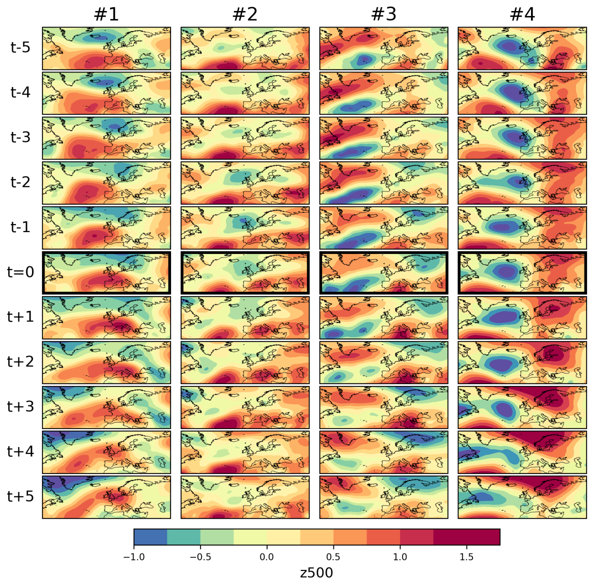

Figure 1: t-SNE representation of the latent space with random-split data (0.8/0.05/0.15). Following our workflow (testing with more than two components), a 2-component GMM provided the best fit, as it was closest to a single Gaussian distribution and thus better captured the smooth latent space structure.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



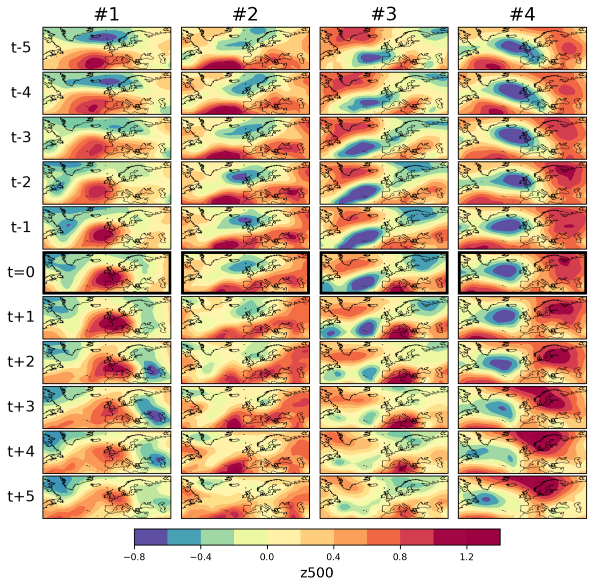

(a) 1-sample composite map



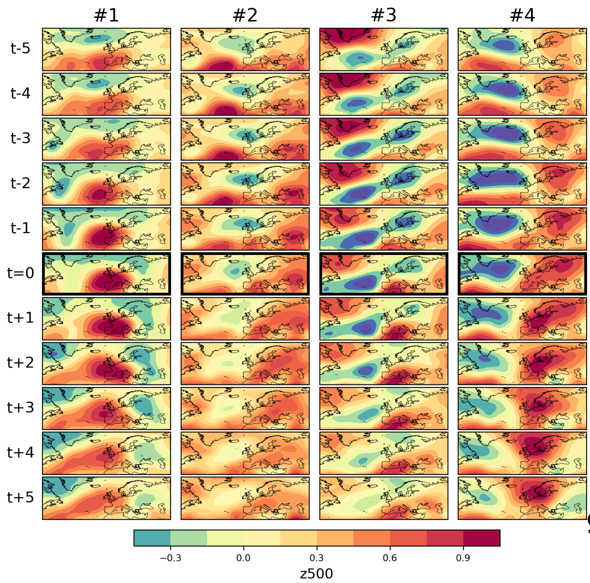

(b) 5-sample composite map



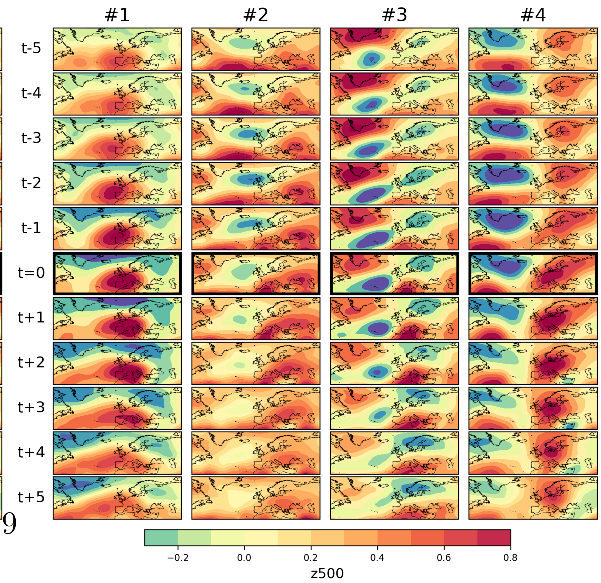

(c) 10-sample composite map




(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 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.

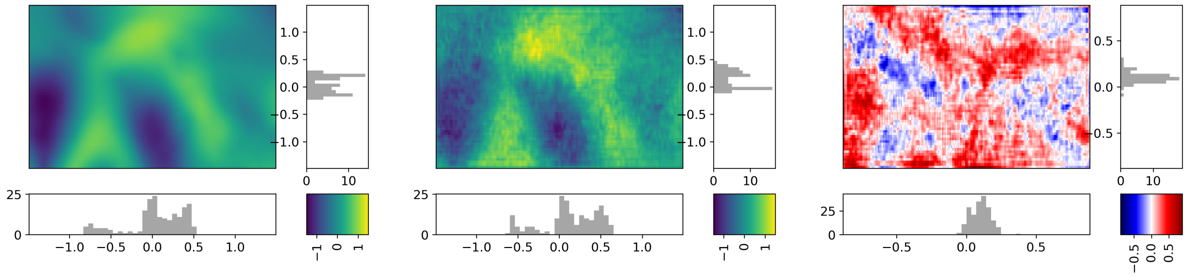
**Response:** Thank you very much for bringing this point to our attention. Our study builds directly on the framework of Happé et al. (2024), but extends it in several significant ways. While Happé et al. (2024) focused on summer heatwaves in the KNMI-LENTIS model ensemble using two atmospheric circulation 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 revised the Abstract, Introduction, and Discussion to make this novelty explicit and clearly separate where our study confirms previous results versus where it extends them.

#### Major Comment 6

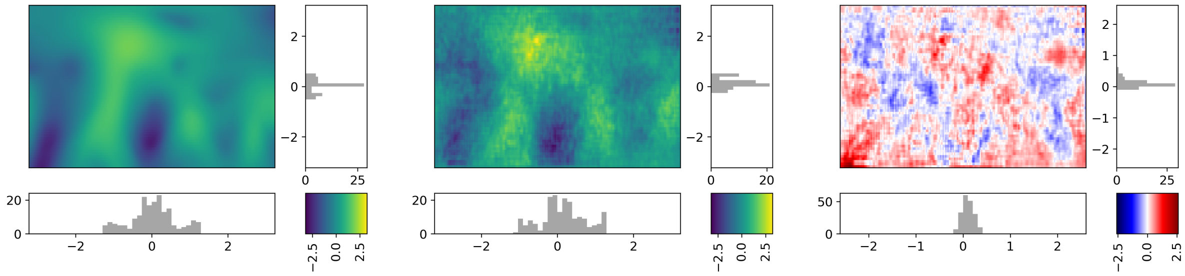
Methods 2.3 the $r^2$ scores; As this section talks about reconstruction errors, I would suggest this section fits better in the result. Apart from that – are these $r^2$ scores based on a latent dimension size 128? Is this chosen because of Happé et al. (2024)? Why didn't the authors take a higher latent 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 added a reconstruction loss comparison plot for different hiddem dimension and latent dimension sizes to the Appendix. We also added $R^2$ for test and validation scores to Table 2. The model achieves overall $R^2$ scores of 0.76, 0.73, and 0.72 for the train, validation, and test data, respectively, indicating stable performance and good generalization without signs of overfitting.

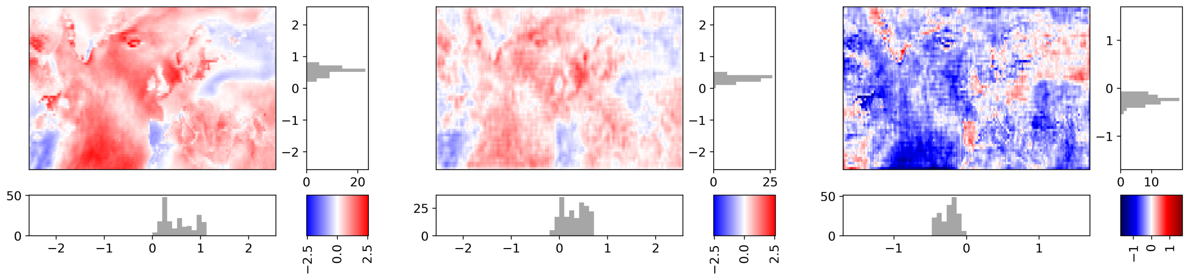
We agree with the reviewer that reconstructed maps are helpful to illustrate model performance. We examined the model residuals to determine whether systematic large-scale biases occur in the variable fields, as shown in Figure 3. The reconstructions accurately reproduce the main patterns of atmospheric 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.



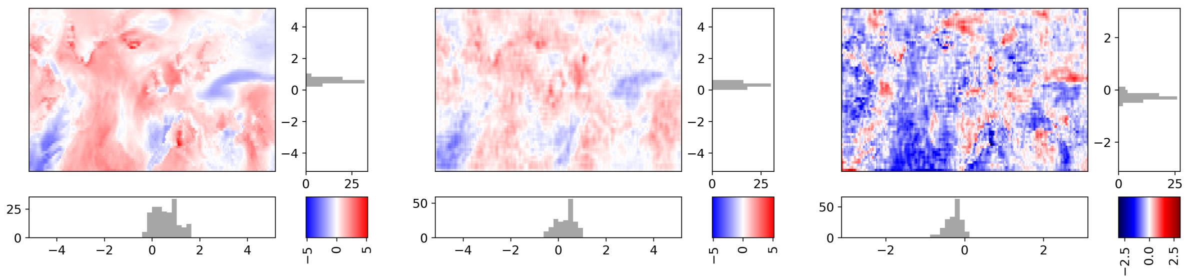
(a) Geopotential height at 500 hPa for a random sample (left), its VAE reconstruction (middle), and the resulting reconstruction bias (right) each averaged over the time dimension. The histograms below and to the right of each panel show the distributions of longitudinal and latitudinal values, respectively.



(b) Geopotential height at 500 hPa for a random sample (left), its VAE reconstruction (middle), and the resulting reconstruction bias (right), each shown at the onset date. The histograms below and to the right of each panel show the distributions of longitudinal and latitudinal values, respectively.



(c) 2-meter temperature for a random sample (left), its VAE reconstruction (middle), and the resulting reconstruction bias (right), each averaged over the time dimension. The histograms below and to the right of each panel show the distributions of longitudinal and latitudinal values, respectively.



(d) 2-meter temperature for a random sample (left), its VAE reconstruction (middle), and the resulting reconstruction bias (right), each shown at the onset date. The histograms below and to the right of each panel show the distributions of longitudinal and latitudinal values, respectively.

Figure 3: Input, reconstruction, and bias maps of geopotential height and temperature, shown as the time mean and at the heatwave onset date.

## 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 revised 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 two dimensions for visualization of the latent space. This two-step approach is recommended by Maaten and Hinton (2008) for t-SNE visualizations of high-dimensional latent space to reduce the number of dimensions and help to ensure efficiency and robustness (Pedregosa et al. 2011). 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.

### 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 4. For each variable, we computed the mean standardized anomaly values over all heatwave samples in a cluster. These mean values represent the typical anomaly structure captured by each cluster. 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 would 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 revised the text to more clearly articulate why a temporal split was chosen in relation to our scientific question and acknowledged this explicitly in the discussion.

> Each heatwave sample consisted of 9 climate variables for 11 days on a $64 \times 192$ spatial grid (North Atlantic, see hatched gray region in Figure 2), resulting in input tensors with the shape $(9, 11, 64, 192)$. The model was trained on heatwave samples from 1941 to 1990, validated on samples from 1991 to 2000, and tested on samples from 2001 to 2022, resulting in 1,408, 320, and 928 samples, respectively. This split enables an out-of-distribution setting where the test years reflect recent climate conditions that differ from the historical baseline because recent decades exhibit non-stationary warming and circulation changes (Chitsaz et al. 2023; Rouges et al. 2023). Because the 1941–1980 baseline precedes the strong post-1980 warming trend (Elguindi, Rauscher, and Giorgi 2013; Reid et al. 2016; IPCC 2021), this choice naturally leads to more detected heatwave events in recent decades. However, our goal is not to quantify frequency changes but to characterize the structural and dynamical patterns of heatwaves. The model focuses on circulation-related variability rather than long-term thermodynamic shifts. To assess robustness against non-stationarity, we also trained a version of the VAE using random train (0.8)/validation(0.05)/test(0.15) splits across the whole 1941–2022 period. While this approach provides a smoother latent space as expected, the resulting latent space (Figure A3) shows consistent seasonal clustering, confirming that the identified heatwave types are not artifacts of temporal trends. By training the model on historical heatwave samples, the VAE learns an internal representation of typical patterns over time, capturing the complex spatial and temporal relationships between the different climate-relevant variables. This allows the model to learn a low-dimensional representation of extreme multivariate events in the latent space.

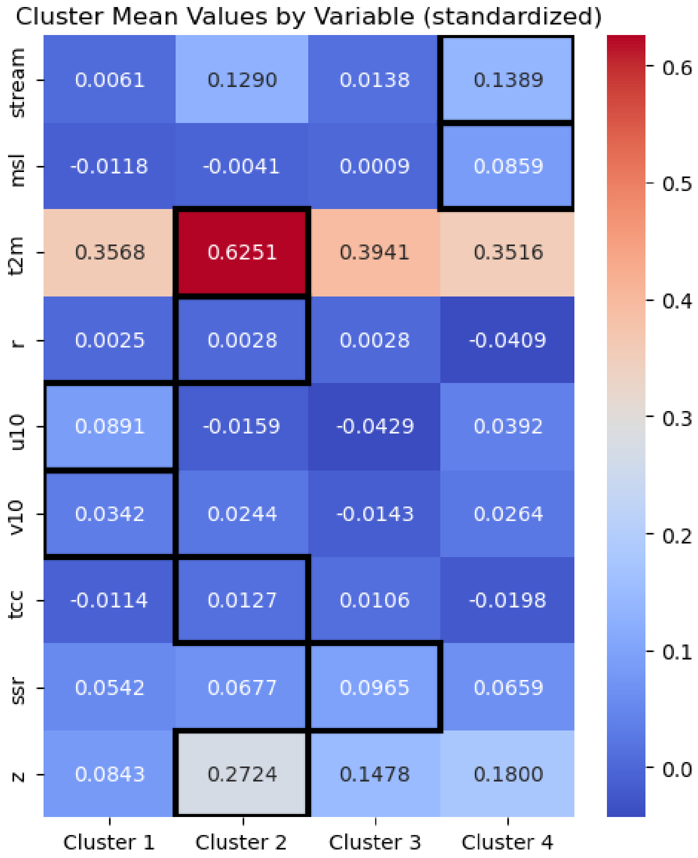



Figure 4: Mean standardized anomaly values for each variable, averaged over all heatwave samples in each latent cluster. The values summarize the typical anomaly patterns represented by the four clusters. Black boxes represent the highest value for each variable.

> **Major Comment 9**
>
> In the Discussion the authors state that the VAE/GMM is sensitive to hyperparameters; 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 complete set of experiments is not feasible, we added comparisons of reconstruction losses for different hidden dimensions and latent sizes to the Supplementary Material.

**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 provided a clearer explanation of the difference and contribution of our study as follows:

> VAEs are primarily developed as generative models, capable of learning complex data distributions and generating realistic synthetic samples (Kingma and Welling 2019). Beyond generative tasks, VAEs have also been widely applied in anomaly detection across various domains, including 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 Happé et al. (2024) to a multivariate and year-round analysis of European heatwaves using reanalysis data. This study aims to understand how heatwaves develop, what contributes to their evolution, and how recent heatwaves are represented in relation to the heatwaves from the historical baseline. In contrast to Happé et al. (2024), who focused on heatwave samples extracted from KNMI-LENTIS dataset over Western Europe, we (i) include nine atmospheric variables from the ERA5 reanalysis dataset to assess their contribution to heatwave evolution, (ii) analyze spatiotemporal samples over a larger window ($\pm 5$ days around the onset, i.e., 11 days in total) to characterize build-up and decay processes better, (iii) train on 11-day multivariate ERA5 heatwave samples that cover the full year with standardized anomalies, and (iv) assess generalization with an explicit historical-to-recent evaluation. To this end, we first train a spatiotemporal VAE on multivariate reanalysis heatwave samples from the historical period of 1941-1980. Then, we cluster and analyze the latent space representation of recent heatwaves (2000-2022) in Western Europe. Section 2 describes the data and methodology followed in this research. Section 3 provides the results for the heatwave cluster analysis. Finally, the main results are summarized and discussed in Section 4.

We used the ERA5 reanalysis dataset (Hersbach et al. 2017; Soci et al. 2024) to detect and understand heatwave dynamics. We focused only on atmospheric variables, as shown in Table 1, including those used in Happé et al. (2024) as well as additional variables relevant to heatwave dynamics identified in previous studies, such as geopotential height at 500 hPa (z500) as it captures mid-tropospheric circulation patterns and has been shown to better reproduce temperature anomalies associated with European heatwaves than sea-level pressure (Jézéquel, Yiou, and Radanovics 2018; Domeisen et al. 2022; Rousi et al. 2022; Barriopedro et al. 2023; Kim and Seo 2023; Tian et al. 2024).

. . .

The resulting dataset comprises daily observations from 1940 to 2022 for the North Atlantic region, spanning from 75°W to 60°E and from 30°N to 75°N on a 0.7 ° grid, as shown in Figure 2. ERA5 data was regridded from its native 0.25°resolution to 0.7°to ensure direct comparability with Happé et al. (2024), where this resolution corresponds to the native grid of the EC-Earth3 model. Furthermore, this coarser resolution also reduces the very high dimensionality of the multivariate input while retaining the large-scale circulation patterns. Although ERA5's native resolution is finer, the reconstruction loss of the variational autoencoder emphasizes large-scale circulation patterns while smoothing out small-scale variability. Retaining the native 0.25°grid would therefore not provide additional benefit while substantially increasing computational cost. Since ERA5 data is provided in hourly resolution, we first applied the temporal aggregation operators listed in Table 1 to convert the hourly ERA5 fields to daily values. From these daily fields, we then computed the 1941–1980 climatological mean and standard deviation using a 15-day moving window centered on each calendar day, which we used to standardize the variables and remove seasonal variability. 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 studies on extreme-event detection (Sulikowska and Wypych 2020). We used this period to calculate the climatology, as it precedes the stronger warming trends after 1980 in Europe (Elguindi, Rauscher, and Giorgi 2013; Reid et al. 2016). The standardized daily anomalies were then used for all subsequent analyses.

Our work builds directly on Happé et al. (2024), but extends it in several important ways. While their study focused on summer heatwaves in the KNMI-LENTIS ensemble dataset, our analysis uses ERA5 reanalysis data, encompasses nine variables, spans all seasons, and incorporates a longer heatwave sample window (11 days). We also analyze historical and recent periods separately to reveal how heatwave dynamics have changed under warming.

. . .

While we trained the VAE model with year-round heatwave samples from ERA5, 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. In contrast, those studies focused only on summer events. ... Our results showed that the VAE's latent space effectively captures key atmospheric patterns previously linked to extreme heat events, such as blocking highs, omega blocks, and persistent ridges, similar to those identified by Happé et al. (2024) and Rouges et al. (2023).

### 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.

... 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; Skinner et al. 2025). These trends are projected to continue even at the lowest projected global warming scenario (IPCC 2021). ...

### 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 context for climate extremes, but our scientific question and all analyses are specific to heatwaves over western Europe. We revised the introduction to narrow the framing and state the scope more explicitly.

### 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.

### 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 moved the $r^2$ values to Results, added the $r^2$ for train and validation period and kept the Methods section focused on the training setup. Regarding the resolution, we regridded ERA5 from 0.25° to 0.7° to reduce dimensionality, emphasizing large-scale circulation patterns most relevant to our objective. This regridding also improves training stability under an MSE loss that already smooths small scales, and facilitates comparison with the framework of Happé et al. (2024). We clarified these points in the revised 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)

> We used the ERA5 reanalysis dataset (Hersbach et al. 2017; Soci et al. 2024) to detect and understand heatwave dynamics. We focused only on atmospheric variables, as shown in Table 1, including those used in Happé et al. (2024) as well as additional variables relevant to heatwave dynamics identified in previous studies, such as geopotential height at 500 hPa (z500) as it captures mid-tropospheric circulation patterns and has been shown to better reproduce temperature anomalies associated with European heatwaves than sea-level pressure (Jézéquel, Yiou, and Radanovics 2018; Domeisen et al. 2022; Rousi et al. 2022; Barriopedro et al. 2023; Kim and Seo 2023; Tian et al. 2024).
>
> ...
>
> The resulting dataset comprises daily observations from 1940 to 2022 for the North Atlantic region, spanning from 75°W to 60°E and from 30°N to 75°N on a 0.7 ° grid, as shown in Figure 2. ERA5 data was regridded from its native 0.25°resolution to 0.7°to ensure direct comparability with Happé et al. (2024), where this resolution corresponds to the native grid of the EC-Earth3 model. Furthermore, this coarser resolution also reduces the very high dimensionality of the multivariate input while retaining the large-scale circulation patterns. Although ERA5's native resolution is finer, the reconstruction loss of the variational autoencoder emphasizes large-scale circulation patterns while smoothing out small-scale variability. Retaining the

native 0.25°grid would therefore not provide additional benefit while substantially increasing computational cost. Since ERA5 data is provided in hourly resolution, we first applied the temporal aggregation operators listed in Table 1 to convert the hourly ERA5 fields to daily values. From these daily fields, we then computed the 1941–1980 climatological mean and standard deviation using a 15-day moving window centered on each calendar day, which we used to standardize the variables and remove seasonal variability. 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 studies on extreme-event detection (Sulikowska and Wypych 2020). We used this period to calculate the climatology, as it precedes the stronger warming trends after 1980 in Europe (Elguindi, Rauscher, and Giorgi 2013; Reid et al. 2016). The standardized daily anomalies were then used for all subsequent analyses.

### 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 two dimensions for visualization of the latent space. This two-step approach is recommended by Maaten and Hinton (2008) for t-SNE visualizations of high-dimensional latent space to reduce the number of dimensions and help to ensure efficiency and robustness (Pedregosa et al. 2011). 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.

### 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 the reviewer for this suggestion. We modified this section to better explain the patterns we observed in Cluster #3 ans #4 and compare with findings from the literature.

> Although both occur mainly during MAM and SON, they represent different dynamical regimes. Cluster *#3* is dominated by enhanced solar radiation, while Cluster *#4* is linked to stronger large-scale pressure anomalies. Similarly, autumn heatwaves are associated with high-pressure anomalies over western and southern Europe, leading to persistent warm conditions (Boboc et al. 2025). (Träger-Chatterjee, Müller, and Bendix 2013) found that enhanced solar irradiation during late winter and spring is associated with the development of warm conditions in the coming months.
>
> ...
>
> For instance, in Figure 15, cluster *#3* displays a positive geopotential height at 500 hPa (*z500*) anomaly over Greenland and negative anomalies over the North Atlantic extending into Scandinavia, a pattern consistent with Greenland blocking events which are known to cause European heatwaves (Kornhuber et al. 2017). Conversely, cluster *#4* shows the opposite, i.e., negative anomalies over Greenland and positive anomalies over the North Atlantic–Scandinavia sector.

### 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. . . .

# References


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

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.

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.

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.

Boboc, Loredana et al. (Sept. 2025). "Trends and variability of heat waves in Europe and the association with large-scale circulation patterns". In: *Weather and Climate Extremes* 49, p. 100794. DOI: 10.1016/j.wace.2025.100794.

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.

Chitsaz, Fatemeh et al. (2023). "Heatwave Duration and Heating Rate in a Non-Stationary Climate: Spatiotemporal Pattern and Key Drivers". en. In: *Earth's Future* 11.12. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2023EF003995, e2023EF003995. DOI: 10.1029/2023EF003995.

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.

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.

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.

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.

Fischer, E. M., S. Sippel, and R. Knutti (2021). “Increasing probability of record-shattering climate extremes”. en. In: *Nature Climate Change* 11.8, pp. 689–695. DOI: 10.1038/s41558-021-01092-9.

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].

Hersbach, H. et al. (2017). *Complete ERA5 from 1940: Fifth generation of ECMWF atmospheric reanalyses of the global climate*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Accessed on 15-JUN-2023. Data distribution by the German Climate Computing Center (DKRZ). DOI: 10.24381/cds.143582cf.

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.

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).

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.

Kharin, V. V. et al. (2018). “Risks from Climate Extremes Change Differently from 1.5°C to 2.0°C Depending on Rarity”. en. In: *Earth's Future* 6.5, pp. 704–715. DOI: 10.1002/2018EF000813.

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.

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.

Kornhuber, K. et al. (Aug. 2017). “Summertime Planetary Wave Resonance in the Northern and Southern Hemispheres”. en. In: *Journal of Climate*. DOI: 10.1175/JCLI-D-16-0703.1.

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.

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.

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

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.

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

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.

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

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.

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.

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.

Skinner, Christopher B. et al. (Aug. 2025). "The spatial extent of heat waves has changed over the past four decades". en. In: *Communications Earth & Environment* 6.1. Publisher: Nature Publishing Group, p. 662. DOI: 10.1038/s43247-025-02661-y.

Soci, Cornel et al. (2024). "The ERA5 global reanalysis from 1940 to 2022". In: *Quarterly Journal of the Royal Meteorological Society*. DOI: 10.1002/qj.4803.

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.

Sulikowska, Agnieszka and Agnieszka Wypych (July 2020). "Summer temperature extremes in Europe: how does the definition affect the results?" en. In: *Theoretical and Applied Climatology* 141.1, pp. 19–30. DOI: 10.1007/s00704-020-03166-8.

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.

Träger-Chatterjee, C., R. W. Müller, and J. Bendix (May 2013). "Analysis of extreme summers and prior late winter/spring conditions in central Europe". English. In: *Natural Hazards and Earth System Sciences* 13.5. Publisher: Copernicus GmbH, pp. 1243–1257. DOI: 10.5194/nhess-13-1243-2013.