

Response to reviewer 2

Review of “Explainable AI shows that a neural network learns extratropical cyclones as predictors of heavy precipitation” by Guillaume-Castel et al.

In this study, the authors train a convolutional neural network (CNN) to predict the occurrence of extreme precipitation over western Norway based on large-scale meteorological variables (specifically Z850 winds and sea level pressure). Predictions are made for the current day and up to 7 days ahead. The emphasis of the study is on analyzing the information used by the network to make extreme precipitation predictions. The authors apply layerwise relevance propagation (LRP), an explainable AI method, to identify pixels in the input domain that are used by the CNN to make predictions in days leading up to extreme precipitation. These pixels are then compared to the location of extratropical cyclones from an existing cyclone database. The results show that many predictions made by the CNN leading up to extreme precipitation are based on the presence of extratropical cyclones.

Overall, I think the study presents an interesting method. However, some of the results are puzzling and I think the scientific contributions are relatively weak unless the authors go deeper into understanding the CNN predictions and the association between cyclones and extreme precipitation predictability. I also think the framing of the study could be improved to better highlight the contributions. The paper was generally easy to follow but could also benefit from some re-organization.

We appreciate the reviewer's comments and thank them for their constructive feedback. One of the main points raised here was also raised by the other reviewer, which made us tighten the framing of the study. We will reframe the study to focus on how using weather features to interpret and analyze XAI results can give us insight into the neural network internal decision process and the underlying physical processes. Hopefully, this framing, along with associated changes, will better highlight the scientific contribution of our study.

Main Comments:

1. General framing: I think the framing that other studies have not used XAI to look at individual predictions is incorrect. LRP specifically produces heat maps for a single prediction at a time, and many studies (including many of those referenced in the paper already) have presented results for individual predictions. Also, in the current study, most of the results are composites or summarized across many predictions (i.e. Fig. 6-9). I think the contributions of this paper are developing methods to link XAI relevance maps with specific meteorological features (cyclones). I think the introduction could be revised to

convey this more clearly and provide a more accurate and nuanced review of the literature.

We agree with the reviewer that the framing was not adequately introduced, and this will be clarified in the revised version (please see the response to point #1 of reviewer 1).

We also agree that other studies have looked at individual predictions. We will edit this part to acknowledge other studies that have looked at individual predictions (like Davenport and Diffenbaugh (2021), or Baño-Medina et al. (2025)) but note that aggregating individual XAI results remains a challenge when the drivers of the predictions are moving and/or transient. Interpreting and aggregating XAI results using tracked weather features, as we show in the study, is one approach to addressing this issue.

To further highlight the benefits of using tracked weather features, we include in the results an additional analysis: we compare our results to a simple composite (simple average) of the relevance maps for all positive events (i.e. days when there is a heavy rainfall event in western Norway). This comparison shows that, while simple composites can highlight that the neural network puts some importance on the North Atlantic storm track, it fails to show that it focuses on specific cyclones. This will be added as a new figure (Draft figure D1):

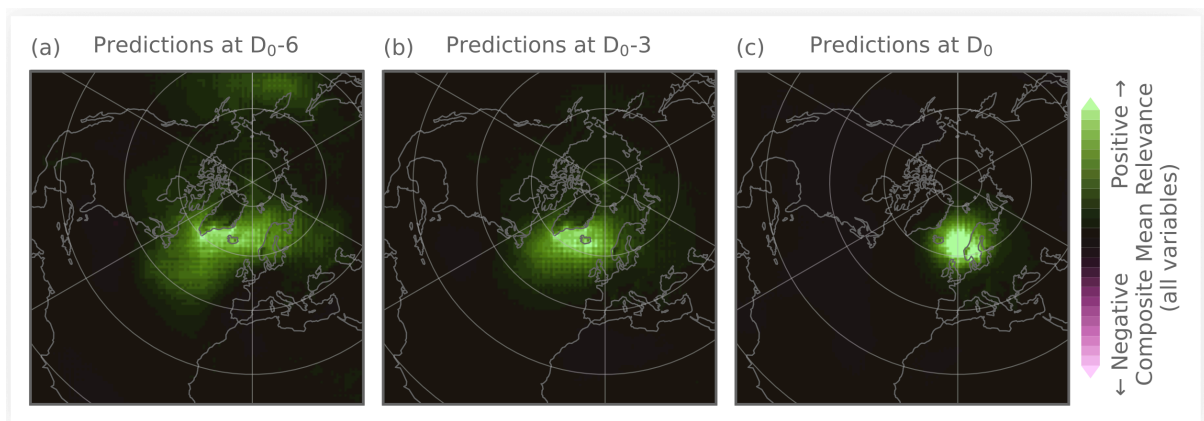


Figure D1: Simple composite of all our relevance maps. This only shows that the neural network puts some importance on the North Atlantic storm track without showing that it focuses on specific cyclones.

2. The paper organization is somewhat confusing. Section 2 is titled “Methods”, but includes results related to the CNN performance. The, most of the novel methods related to comparing LRP maps to cyclone location are presented in Sections 3 and 4. I suggest the authors separate the methods and results into clearly organized and titled sections.

We understand that the current organization of the paper can be confusing. The original idea behind this organization was based on demonstrating the potential of using XAI for scientific discovery, not on the prediction skill of the (very simple) CNN we chose to help illustrate this. The performance of the CNN is not really a “result” and thus we had thought of including it in the methods. However, given the unclear framing of the original

manuscript, we see that this did not work well. We will ensure that the revised manuscript has a clearer organization, with methods and important results clearly separated. The improved framing should also help in this regard.

3. Performance Metrics: It would be helpful to provide more background on how the PR-AUC is calculated (i.e. based on a graph that plots performance vs. recall at different probability thresholds). Second, it would be helpful to see the actual Precision-Recall curve, and to report statistics like precision and recall in the text (e.g. at the 0.5 probability threshold typically used to distinguish positive and negative predictions).

We thank the reviewer for this comment and agree that PR-AUC should be explained in more detail. This will be added to the revised manuscript. The main figure showing the neural network performance will be updated to include the Precision-Recall curves for each lead time, which will be used to illustrate how to compute PR-AUC (see draft figure below: Fig D2). This new figure shows the precision-recall curves for all prediction lead times. This shows actual skill for the D0 and D0+1, especially for an event present only 5% of the time in the dataset. We can see the skill gradually reducing for longer lead times, while always remaining above the climatology (dashed blue line).

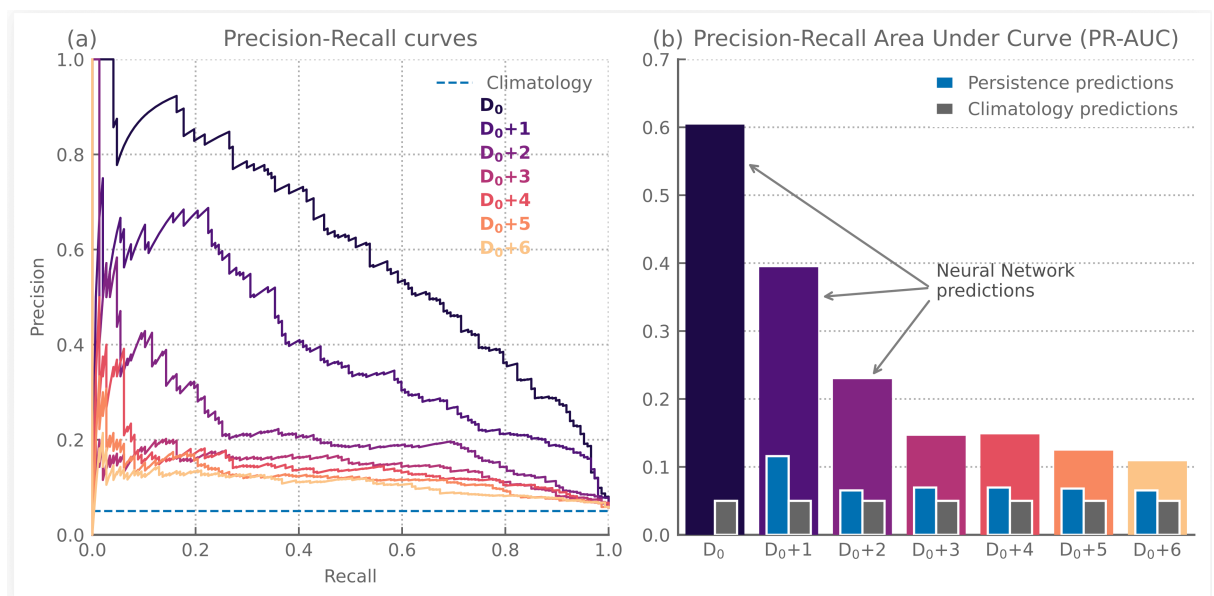


Figure D2: Updated neural network performance metric figure, showing in (a) the Precision-Recall curve, and in (b) the PR-AUC, comparing the results to the climatology and persistence (same as shown in the original Figure 3, but plotted differently such that the colors correspond to the information in panel a)

The following table (draft table D1) reporting True/False Positives/Negatives at the 0.5 threshold will also be reported in the main text, along with corresponding precision (0.08) and recall (0.96) values. Note that better precision would be obtained by tuning the probability threshold for each prediction lead time separately, at the expense of a lower recall.

Table D1: Count of true/false positives/negatives in the test dataset for all 7 prediction lead times, when using a probability cutoff of 0.5.

Ground Truth ↓ / Predictions→	Positive	Negative
Positive	#TP = 983	#FN = 46
Negative	#FP = 10679	#TN = 6212

- I am curious about the persistence baseline for the prediction of extreme precipitation on subsequent days. Why was this chosen as a benchmark? I am also curious what is driving the persistence skill. Is it truly that extreme precipitation is likely to occur in multi-day events? Or is the slightly elevated persistence skill compared to climatology somehow reflecting seasonality (i.e. where extreme precipitation is more likely in a certain season, and thus there is some elevated chance of extreme precipitation happening within the 7-day period following an event)? These are minor points because the persistence model is not the main focus of the paper, but looking into the temporal patterns of extreme precipitation might be useful for understanding other aspects of the CNN prediction.

We used persistence as it is a standard comparison metric for prediction studies, and specifically for precipitation (see for example gmd.copernicus.org/articles/16/2737/2023/). Both of the reviewer's suggestions are likely true. Firstly, cyclones leading to heavy precipitation in Western Norway can cause heavy rainfall (based on our definition of 95th percentile) over multiple days. To validate this, we computed the duration of heavy rainfall events (consecutive days of heavy rainfall) in our dataset (Figure D3). This shows that 1-2 day events are most common (blue and orange bars), and account for 94% of all events. This is consistent with the highest persistence skill being on day D0+1. Moreover, autumn and winter (September to February) see more total heavy precipitation days with longer duration of each heavy precipitation event, meaning that there is also a seasonality component of predictive skill of persistence.

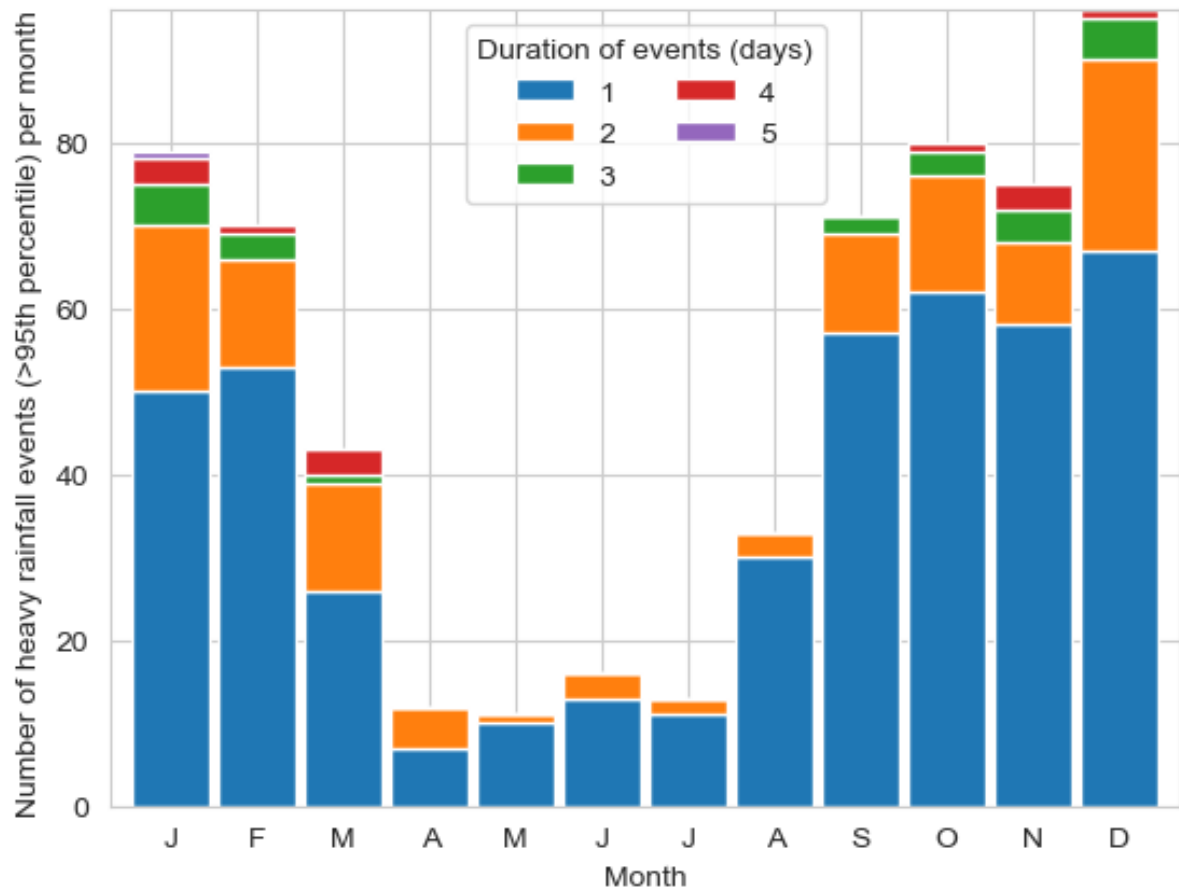


Figure D3: Number of heavy rainfall events (above the 95th percentile) in Western Norway per month, and their associated durations in days. A single event is considered when the precipitation is above the 95th percentile for multiple consecutive days.

- When performing the cyclone-based LRP comparison, the authors only focus on positive cases (I take this to mean cases where extreme precipitation actually occurred regardless of whether the CNN prediction was correct or not). I don't think this choice makes sense, nor is it clear why the authors make this choice. When interpreting the results, there is no context about whether the CNN predictions are correct. For example, are CNN predictions that were based on cyclones more likely to be correct or vice-versa? Do false positive predictions occur where there are cyclones that don't result in extreme precipitation? It would be helpful to see the results of the object-based XAI analysis split up for true positive, false positive, true negatives, and false negatives.

We thank the reviewer for raising this important point. Our current XAI analysis is indeed restricted to the 840 actual heavy-precipitation events, regardless of predicted probability. The main reason for doing this is that, because of the class-weighted loss mentioned earlier, restricting to actual events yields a sample that is almost entirely

true positives, with negligible false negatives (see results in table D1), making it conceptually straightforward to interpret relevance maps without further stratifying by prediction outcome.

We acknowledge that the lower precision at longer lead times implies a high rate of false positives: our PR curves (fig. D3) show precision of only around 0.15 at moderate recall for lead times D0+3 through D0+6. Examining whether cyclones are also implicated in these false positives would be a valuable extension of this work, and this would help clarify whether the neural network's use of cyclones as a predictive feature is specific to actual heavy-precipitation events, or if cyclones in relevant areas generally increase the predicted probability. Extending our full relevance and object-matching pipeline to a comparably sized sample of false-positive cases would require substantial additional computation, but we will note this as a direction for future work in the discussion.

6. Figure 6: it would be helpful if the authors could elaborate on the results in this figure with some further analysis. There is a relatively large gap between the orange and black lines. If I understand correctly, this indicates that, at 1- to 6-day lead times, the CNN is making predictions based on different cyclones than the cyclone this is associated with extreme precipitation on Day 0. Are the CNNs making correct predictions at these lead times? Or is it just coincidence that the CNN identifies a cyclone at those lead times, and the prediction isn't necessarily, correct? Splitting these results into true/false positives/negatives could help show what is going on.

We thank the reviewer for this comment. We will provide an updated figure with more details (see draft Figure D4), allowing us to ask whether the neural network follows the heavy precipitation-causing cyclones from D0 or if it focuses on different, less relevant, cyclones. Of all the predictions included (all days when there is a heavy rainfall event), 96% are true positives, as mentioned in the response above, meaning that this analysis is mainly conducted on true positive results. The figure D4 shows that one day before the event, around 60% of positive predictions are still associated with the same cyclone as at D0. Further back, a smaller proportion of predictions are associated with the same cyclone, but the neural network is still mainly making predictions based on other cyclones. It is probable that the CNN is sometimes basing the predictions on the strongest cyclone, or another strong cyclone as they are strong features in the input fields, which may not be the relevant cyclone more than 2 days before the event.

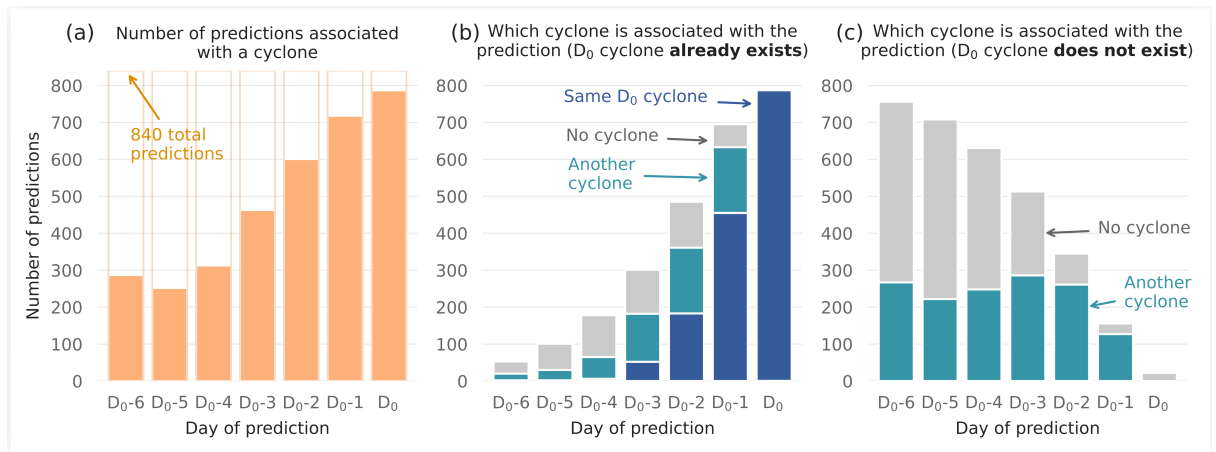


Figure D4: updated figure 6 showing additional information. (a): number of predictions associated with a cyclone. As mentioned before these are only predictions of events that actually happened, and virtually all of them are positive. (b) When the cyclones that are associated with the prediction at D₀ already exist before the event, this figure shows whether the neural network is using the same D₀ cyclone (dark blue) another cyclone (lighter blue) or no cyclone (grey). (c) same as (b) but when the cyclone associated with D₀ prediction does not yet exist.

That said, we will also provide another updated figure to highlight the location of the other cyclones used by the neural network (see draft Figure D5). Now we specifically ask the question, “Is the cyclone used by the NN the same or in the same region as the one it should be using”. Now we can answer that in some cases the neural network tracks cyclones very well from one timestep to another. And when it does not track the exact same cyclone, it still mostly uses a cyclone in the same sector.

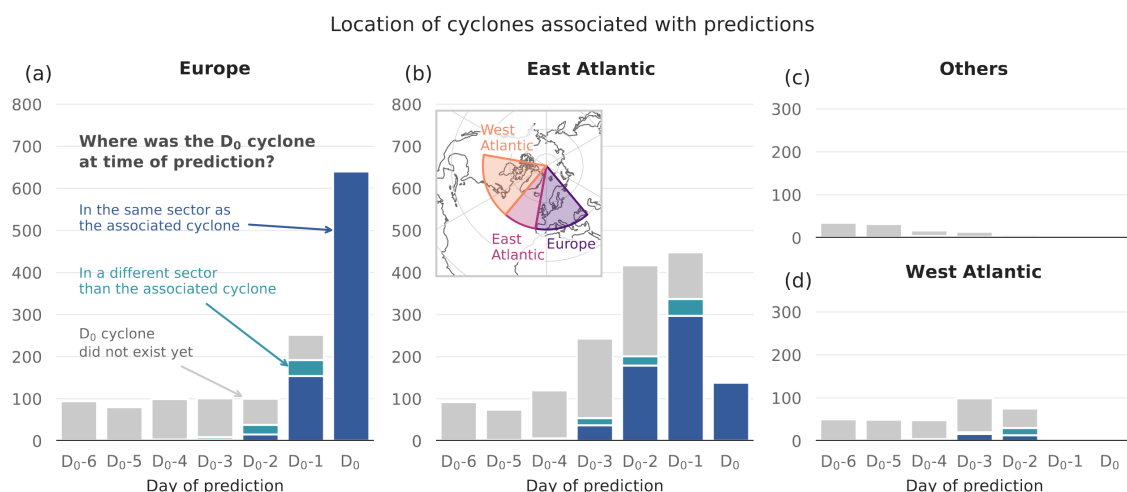


Figure D5: Updated figure showing the location of the cyclones associated with predictions (total bar height) compared to the positions of the cyclones associated with D₀ predictions at earlier lead time.

Line-by-line comments:

7. Line 23: “XAI has notably been used to identify the physical drivers of weather events (e.g. Alessi et al., 2025;...”. I think this reference applies to mean precipitation and not weather events and may not be appropriate here.

Thanks, this reference will be removed.

8. Line 101: “Each input field is converted into anomalies by removing the temporal mean from each grid cell, then dividing by the global standard deviation.” Does global standard deviation refer to the entire globe or the entire input domain?

We thank the reviewer for pointing this out. It is indeed a domain-wide standard deviation and not a global one. This will be clarified in the text.

9. Section 2.2.3 Cyclone tracking dataset: In this section, it would be helpful to clarify that cyclones are tracked from one day to the next, so cyclones can be “backtracked” from D0.

Thank you for the comment, a mention to this will be added.

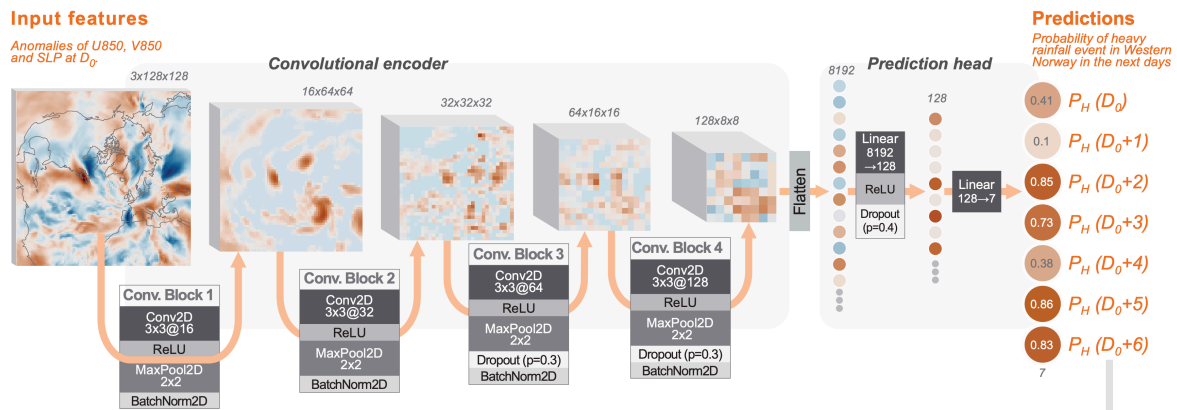
10. Line 70: “Similar improvements in training are expected with lower thresholds of the 70th, 80th, and 90th Percentiles”. I don't understand what this sentence is trying to say. Is this referring to different thresholds for the label smoothing regularization, or for the extreme threshold?

This refers to the lower threshold of the label smoothing regularisation. This will be clarified in the text.

11. Figure 2: I find the labels for the convolutional blocks too small/squished to read.

An updated figure showing the neural network architecture will be provided, with bigger labels for the convolution blocks (see below, draft figure D6).

(a) Neural network architecture and inference example



(b) Example of obtention of relevance maps using Layer-wise Relevance Propagation

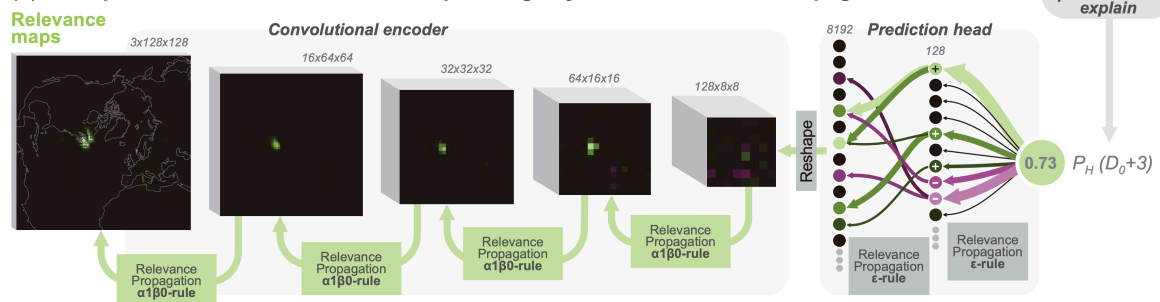


Figure D6: a) Updated figure showing the neural network architecture, along with inputs/outputs concrete example. b) new subfigure showing how relevance maps are obtained from the prediction and the model architecture.

12. Line 245: “We start by defining a relevance patch as a contiguous area of high positive relevance within which every pixel must account for a relevance greater than 0.1%”. Are relevance values summed across all input variables?

Yes, relevance values are summed for all variables before applying the 0.1% mask. This will be mentioned in the text.

13. Line 252: “If a large patch spans multiple cyclones, we first split the patch into individual patches by applying watershed segmentation to the distance-transformed masked relevance field, then assign cyclones to the resulting sub-patches.” I have two comments on this. First, how often does a large patch span multiple cyclones? It would help to have some context about this. Second, the statement on watershed segmentation is confusing. What is the watershed segmentation procedure? What is the distance-transformed masked relevance field?

We agree with the reviewer that this is not clear in the main text and will add a paragraph to detail what the distance transform is, and what the watershed algorithm is. Briefly, the distance transform assigns to a given pixel its distance to a border (here the outer border of a given patch) (see draft figure D7.d). The watershed algorithm then separates

the patches by detecting local peaks (i.e. pixels further away from the border than their neighboring pixels) and separating the patch pixels using slopes from the peaks.

We will also include a comment stating how many cases are affected by the splitting of large patches. Overall, fewer than 10 percent of the larger patches required splitting. An updated method figure will also be provided (see draft figure D7 below)

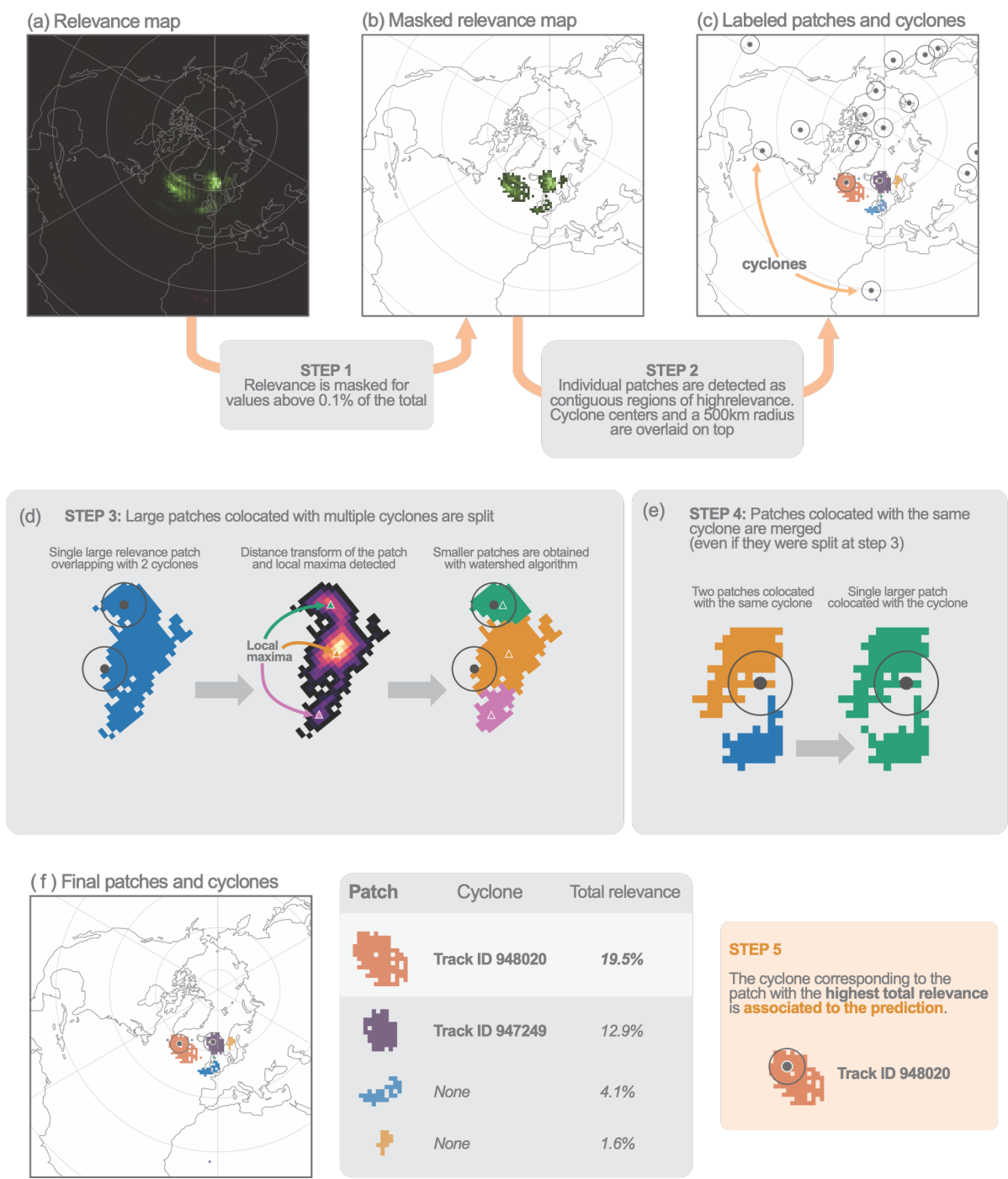


Figure D7: Updated methodology figure detailing our workflow. From one relevance map, we extract contiguous areas of high relevance and associate them with cyclones.

14. Line 257: “Patches with a total relevance of less than 10% are discarded.” How often does this occur?

In total, we detected more than 1 000 000 individual patches, and out of these 4181 had more than 10% relevance. In fact, the vast majority of patches detected consisted of a very few number of pixels, with 60% of patches having fewer than 3 pixels, 80% fewer than 20 and 95% less than 100. Most of these patches are discarded by the 10% total relevance threshold. A mention of this will be added in the text.

15. Figure 4: Why was this particular example chosen? It might also be helpful to see details of other individual predictions, for example in cases where relevance was not associated with a cyclone, or associated with a cyclone outside the Atlantic/Europe regions.

There was no specific reason to choose this example, apart from the fact that it only included a single patch that nicely corresponded to a cyclone that matched across multiple lead times. We will provide a new figure with more different examples of relevance maps (see below, draft figure D8). These include cases where a physically reasonable cyclone is associated to the prediction (a-c), a case where the NN focuses on a cyclone in the Pacific (d), a case where the neural network is not strongly focusing on a specific area (e), and a case where a cyclone reduces the predicted probability (f).

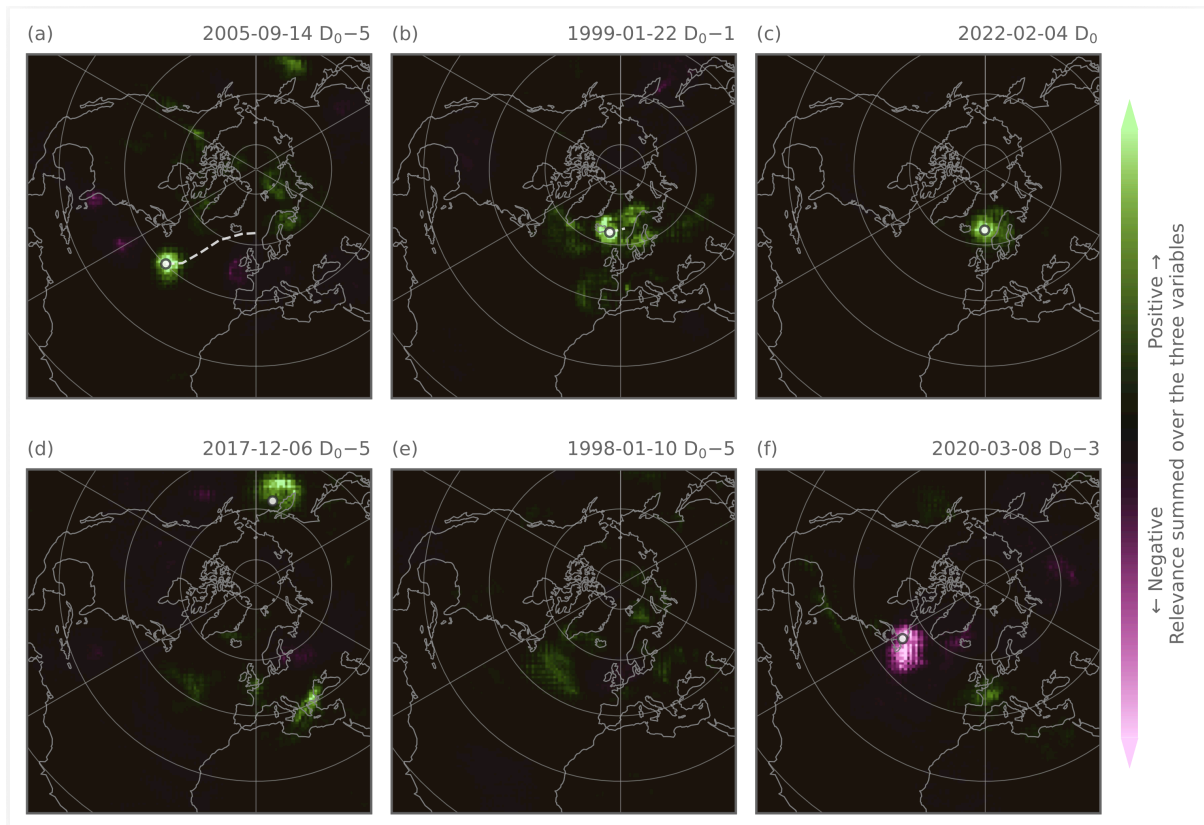


Figure D8: Updated figure showing more diverse examples of relevance maps. In (a,b,c,d,f), the white dots show the location of specific cyclones collocated with high relevance.

16. Figure 5: Caption refers to 5b but should be 4b. Also, caption for panel c indicated the cyclone is shown in red, but appears to be green.

Figure 5 will be fully updated to include more details about the segmentation method (including the segmentation of large patches using the distance transform), and the caption will be adapted accordingly (see draft figure D7.).

17. Line 273: “To test this, we track the cyclones associated with D0 predictions backwards in time and find that the decrease in the number of associated cyclones is broadly consistent with this physical constraint: the neural network generally does not assign relevance to cyclones that have not yet formed (Fig. 6, black line).” It took me a minute to understand what this is saying. It would be helpful to describe the methods and results from Figure 6 more in the text.

We thank the reviewer for the comment. The methodology text will be updated, as well as figures showing results of the backtracking along with the text describing them (see draft figures D4 and D5).

18. Line 310: “For example, moderately strong cyclones with a minimum SLP of 960 to 970 hPa are associated with predictions about 18% of the time they exist in the input fields, while strong cyclones deeper than 940 hPa are associated with predictions more than 50% of the time.” It might make more sense to perform

this analysis (and that in Figure 8) for cyclones close to western Norway, since one wouldn't expect cyclones in certain regions (e.g. Pacific or Tropics) to be associated with extreme P in western Norway.

This figure is specifically made to focus only on the intensity of the cyclone, which is why we computed this on all cyclones and not just the local ones. Computing this only for cyclones closer to Western Norway would assume part of the neural network behaviour *a priori*. Here, we make no assumptions about cyclone location and assess first intensity (Figure 8 in the paper) and location (Figure 9 in the paper) afterwards as a physical consistency check. The location information in Figure 9 helps support the selection of physically reasonable cyclones by the neural network. We will add a short discussion of these points to the manuscript.

19. Figure 8 (and related text): The wording “associated with predictions” is confusing. Does this mean associated with a correct, positive prediction? Or any prediction? I also found this terminology to be confusing throughout the text.

We agree this terminology was not clear, and specifically whether “predictions” implied a correct positive prediction or any output of the neural network. We will clarify in the text that a “prediction” refers to any output of the neural network for a given lead time of the 840 events we focus on, regardless of whether the predicted probability is above a given threshold or not. Because of our focus on positive events, most of the cases reported consist of true positive predictions, and a minority are false negatives.

This definition will be clarified at the beginning of the result section when introducing the first results regarding how cyclones are associated with predictions. Additionally, we will update the method figure (draft Figure D7) to clarify the terminology.

20. Figure 9: I find it interesting in panel c that there are more cyclones in the West Atlantic than are used by the neural network to make predictions for D-3 and D-4. This suggest that the CNN is not as accurate in tracking the cyclones from day to day leading up to extreme precipitation? Perhaps the authors can comment on this.

As mentioned above, we will clarify the temporal consistency of the neural network by updating two figures (see draft figures D4 and D5 above). Overall, we find that in some cases the neural network tracks cyclones very well from one timestep to another, and when it does not track the exact same cyclones, it still mostly uses a cyclone in the same geographical sector.

21. Sections 5.2, 5.3, and 5.4: These sections are pretty broad and make sweeping claims about the potential novelty or impact of the work. I would encourage the

authors to reframe the conclusion section to focus on summarizing the key results, and discussing more specific contributions of the study rather than sweeping generalizations.

The concluding section will be revised according to the new framing of the work. Now our discussion points will focus on the following:

- Main findings and contributions:
 - We use an object-based method for the first time to systematically interpret relevance maps obtained with XAI by matching them with tracked weather features.
 - This provides a deeper understanding of the neural network internal decision process than simple composites: we find evidence that our neural network uses mid latitude cyclones of the appropriate magnitude and location, and the same individual cyclone can be tracked across lead times.
- Limitations of our approach
 - We essentially focus on true positive cases and applying our pipeline to false positive cases would be a valuable extension of this work
 - While we provide a better way to interpret relevance maps that include moving drivers, the relevance maps are themselves limited, as they only show an approximation of the neural network behavior.
- An object-based interpretation of relevance maps could be adapted to more complex setups with more physical features and using state of the art models (e.g. AIFS, GraphCast, ...), allowing us to do a meaningful exploration of true/false negative and positive cases.