https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

Explainable AI shows that a neural network learns extratropical cyclones as predictors of heavy precipitation

Robin Guillaume-Castel [1], Camille Li [1], and Stefan Sobolowski [1]

[1]Geophysical Institute and Bjerknes Centre for Climate Research, University of Bergen, Norway

Correspondence: Robin Guillaume-Castel (robin.guillaume-castel@uib.no)

Abstract. Neural networks are increasingly used in weather and climate science, not only for prediction tasks but also for process understanding and scientific discovery, where model outputs must be linked to physically meaningful processes. Explainable artificial intelligence (XAI) helps establish this link by providing tools to interpret the information a neural network uses to make its predictions. However, most approaches rely on spatially aggregated or composite analyses that do not reveal the physical basis of individual predictions. Here, we present an object-oriented XAI framework that enables such prediction-level evaluation. We use this framework to analyse a simplified prediction task in which a neural network is trained to predict the occurrence of daily heavy precipitation in Western Norway across multiple prediction lead-times, several days in advance. In this study area, heavy precipitation is mainly associated with mid-latitude cyclones, providing a clear criterion: regions of high relevance identified by XAI should correspond to detected cyclones in the input fields. We find that most predictions are indeed associated with cyclones and that relevance patterns match key physical features such as the low-pressure centre and the zone of maximum winds. Furthermore, we show that predictions are based primarily on strong cyclones that travel along the North Atlantic storm track. This study provides a controlled benchmark that demonstrates that neural network predictions of heavy rainfall can align with established physical understanding. More generally, it illustrates how an object-oriented XAI framework can be used to assess physical realism at the level of individual predictions, representing an important step toward building the trust necessary to use these models for research and decision-making applications in weather and climate.

1 Introduction

Neural networks are becoming widely used for climate and weather prediction tasks (de Burgh-Day and Leeuwenburg, 2023). In recent years, they have reached predictive skill that, in some cases, matches or even surpasses traditional numerical weather prediction models (e.g. Kent et al., 2025) and seasonal forecasting models (Unal et al., 2023). Beyond prediction, neural networks are increasingly used for scientific interpretation, through Explainable Artificial Intelligence (XAI), which provides tools to analyse how neural networks make their predictions (Holzinger et al., 2022; Roscher et al., 2020; Toms et al., 2020; Yang et al., 2024). XAI has notably been used to identify the physical drivers of weather events (*e.g.* Alessi et al., 2025;

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

Beobide-Arsuaga et al., 2023) and reveal mechanisms underlying climate change signals (Davenport and Diffenbaugh, 2021; Rugenstein et al., 2025; Toms et al., 2020).

However, using neural networks for scientific understanding implicitly assumes that they learn physically meaningful relationships, which is not guaranteed by predictive skill alone. Since neural networks are trained by minimising a mathematical loss function, they may achieve high skill by exploiting spurious correlations or features that are not causally related to the underlying physical processes. Evaluating the physical consistency of neural network predictions is therefore essential for developing trustworthy AI in climate and weather sciences, both for prediction and for scientific discovery.

XAI methods are increasingly being used to evaluate whether neural network predictions rely on physically meaningful information (Bommer et al., 2024; Mamalakis et al., 2022b; Yang et al., 2024). One widely used approach is Layer-wise Relevance Propagation (LRP Bach et al., 2015; Montavon et al., 2019), which produces relevance maps that highlight the input regions that contribute most strongly to a given prediction (Byrne and O'Gorman, 2013; Davenport and Diffenbaugh, 2021; Mayer and Barnes, 2021; Pegion et al., 2022; Wang et al., 2024). In climate science applications, these maps have been interpreted physically by comparing regions of high relevance with known physical patterns. For example, Davenport and Diffenbaugh (2021) analysed relevance maps from a neural network trained to classify extreme US precipitation events, assessing whether regions of high relevance are collocated with large-scale atmospheric waves. Similarly, Toms et al. (2020) evaluated the physical consistency of neural network predictions of ENSO indices by verifying that they rely on characteristic sea surface temperature patterns in the tropical Pacific. Such assessments demonstrate that neural networks can, on average, extract information from plausible geographical regions. However, they do not establish whether individual predictions are based on the appropriate physical processes or dynamical objects.

Building upon the Toms et al. (2020) benchmark, we propose an object-oriented XAI framework to assess whether neural network predictions align with physical understanding at the level of individual events. Instead of evaluating relevance across broad geographic regions, we link each prediction to tracked physical features in the input fields, enabling object-level interpretation. We apply this framework to a simplified prediction task where the relevant dynamical drivers are well understood: daily heavy precipitation in Western Norway. These events are predominantly caused by processes associated with North Atlantic cyclones. However, unlike ENSO, which exhibits a quasi-stationary spatial structure, cyclones are mobile and transient weather systems. We quantitatively assess whether cyclones in the appropriate location and with the appropriate intensity contribute to the neural network predictions. To evaluate temporal consistency, we evaluate predictions across multiple lead times, allowing us to assess whether they follow individual cyclone trajectories. Overall, demonstrating that neural network predictions are systematically based on relevant cyclones provides an interpretable test of physical grounding and stronger evidence that neural networks can align with physical understanding.

In section 2, we present our prediction task, the data used, and our neural network architecture and training procedure. Section 3 introduces the object-oriented XAI framework, and Section 4 presents the assessment of the physical consistency of the neural network heavy rainfall predictions. Finally, section 5 provides a discussion and conclusions of the current work.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

2 Methods

2.1 Physical benchmark design

2.1.1 Physical motivation for using heavy precipitation in Western Norway as a case study

The occurrence of daily heavy precipitation in Western Norway, defined as accumulated daily precipitation exceeding the annual 95^{th} percentile, is a well-understood phenomenon. Most events are associated with mid-latitude cyclones, mainly originating in the North Atlantic, and related processes such as fronts and atmospheric rivers (Azad and Sorteberg, 2017; Benedict et al., 2019; Konstali et al., 2024; Ødemark et al., 2023; Michel et al., 2021). Cyclones travel eastward in the Atlantic storm track, and upon reaching the Norwegian coast, the moist air they transport collides with the Scandinavian mountains, leading to orographic lifting and heavy precipitation.

This physical case provides an ideal benchmark for testing whether a neural network can associate heavy precipitation events with the correct large-scale dynamical features, *i.e.*, North Atlantic cyclones. The task is conceptually simple, as a single dominant feature is implicated in the majority of events. In addition, these events are associated with the large-scale flow, which is often predictable at several days' lead time: the majority of these cyclones form upstream, for example, off the coast of North America or at the southern tip of Greenland, days before reaching Norway. This case, therefore, allows us to evaluate whether a neural network can identify cyclone trajectories and provide physically consistent predictions across different time horizons.

2.1.2 Prediction task and target definition

We design a classification task to predict the occurrence of daily heavy precipitation in Western Norway from dynamical fields. This is a binary classification task where the target value is 1 for heavy precipitation events and 0 for all other days (including non-heavy precipitation and dry days). To assess whether the predictions align with physical understanding of mid-latitude cyclone trajectories on medium-range forecasting timescales, we predict the occurrence of heavy precipitation up to seven days in advance. In other words, each input sample of dynamical fields on day D_0 is used to make a prediction for the 7 subsequent days: day D_0, D_0+1, D_0+2, … until D_0+6. The output of the neural network is therefore an array of seven probabilities of heavy precipitation occurrence P_H:

$$(P_H(D_0), P_H(D_0+1), P_H(D_0+2), \ldots P_H(D_0+6)) \tag{1}$$

2.2 Data

For consistency between the dynamical fields and the heavy precipitation events, we use the European Centre for Medium-Range Weather Forecasts' reanalysis product ERA5 (Hersbach et al., 2020) for both predictors (input dynamical fields) and prediction targets (precipitation). Limiting the analysis to the satellite era (1979-2024) yields around 12,000 daily samples, which is then split into training (1979-2008), validation (2009-2016) and testing (2017-2024) datasets.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

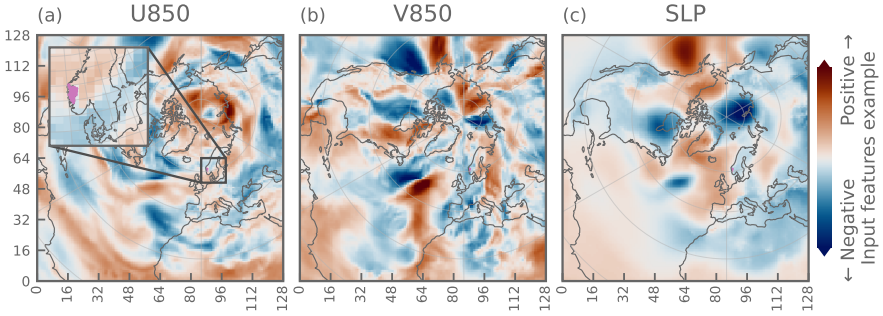

Figure 1. Example input features (August 5th 2016) used for prediction. These fields are anomalies based on local and temporal means and a global standard deviation for each variable. The x and y tick labels indicate the pixel count on each side. a. U-wind at 850 hPa, b. V-wind at 850 hPa, c. Mean sea level pressure. The inset axis in a. shows the location of the Western Norway region used to define the target rain field (in pink).

2.2.1 Dynamical predictor fields

The choice of input variables must include information about the physical process being tested. We select variables that are a priori relevant for mid-latitude circulation: the wind field at 850 hPa (zonal U850 and meridional V850 components), and mean sea level pressure (SLP). Geopotential height at 500 hPa (Z500) was tested as an alternative to SLP and yielded results similar to those of SLP. SLP was favoured as it is used in the cyclone tracking scheme used in this study, as described in section 2.2.3 (Murray and Simmonds, 1991a, b). Studies focused on other aspects of heavy precipitation might wish to choose different predictor variables, for example, column-integrated water vapour or vapour transport for atmospheric rivers, or available potential energy for convection-dominated regions.

To reduce high-latitude distortion, preserve local structure, and maintain approximate isotropy of cyclone geometry across different locations, we regrid the input fields onto a conformal polar projection. Specifically, input fields are regridded to a polar stereographic grid with central longitude of 0°E and a true scale latitude of 70°N, which is a projection commonly used for Arctic machine learning applications (*e.g.* Ali and Wang, 2022), using a bilinear interpolation. The new grid consists of 128×128 pixels with a grid size of 117 km (Fig. 1). The cell size was chosen to be spatially relevant to the scale of cyclones (∼1000 km) and to improve computational efficiency.

Each input field is converted into anomalies by removing the temporal mean from each grid cell, then dividing by the global standard deviation. The global standard deviation was favoured over a local standard deviation to prevent diluting the signal in regions with high variance, which are important for the current task, such as the North Atlantic storm track. To prevent data leakage (inadvertently allowing information from the validation or test period to influence the training process), the mean and standard deviation are computed only from the training period (1979-2008) and are used to compute anomalies for the training, validation, and test datasets.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

2.2.2 Heavy precipitation predictand

The neural network is trained to predict the occurrence of daily heavy precipitation events in Western Norway as captured in ERA5 (see the pink area in Fig. 1.a inset axis). We start by computing a time series of area-weighted mean daily precipitation over the region of interest comprising Vestland, Rogaland and Møre og Romsdal counties. The domain is derived from a definition of European precipitation regions based on high spatial correlation in daily precipitation (Oldham-Dorrington et al., 2025). Heavy precipitation events are defined as days when the accumulated daily precipitation exceeds the 95^{th} percentile of all days, which, for this region in ERA5, is 23.2 mm. The threshold is fixed for all seasons. The choice of the 95^{th} percentile is a trade-off where most of the events can be associated with cyclones, but the events themselves are not too rare such that there are still enough examples in the ERA5 dataset for the neural network to learn from.

While ERA5 generally underestimates heavy precipitation intensities, particularly in regions with complex topography (Lavers et al., 2022), it provides a precipitation product that is dynamically consistent with the large-scale fields used as input fields. In addition, we are primarily interested in the occurrence of heavy precipitation events rather than their intensity, which reduces the limitations associated with using ERA5.

2.2.3 Cyclone tracking dataset

The neural network's predictions are assessed against a cyclone-tracks dataset, also derived from ERA5 reanalysis data (Spensberger and Marcheggiani, 2024). Cyclones are tracked using the Melbourne algorithm (Murray and Simmonds, 1991a, b), which detects maxima in the Laplacian of SLP, that is, local minima in SLP, and is widely used in studies on extratropical cyclones (e.g. Feser et al., 2015; Reid et al., 2025; Marcheggiani et al., 2025; Tao et al., 2025; Madonna et al., 2020). The dataset includes the positions, minimum sea-level pressures, and other characteristics of the tracked cyclones. This dataset consists of around 80,000 low-pressure systems, or 17 per day on average over our study period (1979-2024) and spatial domain (Fig. 1). While several other mid-latitude cyclones tracking schemes exist (see Walker et al., 2020; Neu et al., 2013), the limited level of detail required in our analysis (mainly cyclone position and strength) is not expected to depend strongly on the exact scheme used (Rudeva et al., 2014).

2.3 Convolutional Neural Network

2.3.1 Network architecture

For our task, we design a relatively simple convolutional neural network (CNN) following a LeNet-style architecture (LeCun et al., 2002). CNNs are particularly suited to atmospheric data as they are designed to extract spatial features from input fields. This type of architecture is relatively easy to train and interpret, and is therefore widely used in climate and weather XAI applications (e.g. Davenport and Diffenbaugh, 2021; Wang et al., 2024; Rugenstein et al., 2025; Alessi et al., 2025). While CNNs may not perform as well as more recent architectures such as Vision Transformers, Swin Transformers, or Fourier Neural Operators (e.g. Dosovitskiy et al., 2020; Liu et al., 2021; Kurth et al., 2023; Meo et al., 2024; Li et al., 2021), they

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

are well-suited to our objective, which is not to maximise predictive performance but to develop a benchmark for assessing whether neural network predictions align with the physical understanding of the underlying processes.

The full neural network architecture used in this study is shown in Fig. 2, where information flows from the left (input features) to the right (final predictions). Our network consists of a convolutional encoder that extracts spatial features, followed by a prediction head with simple, fully connected layers that combine these features to make the final prediction. The shape of the data at different stages is shown as orange parallelepipeds in the convolutional encoder (left-hand side in Fig. 2), and as orange dots in the prediction head (right-hand side). In our convolutional encoder, input data go through four sequential convolutional blocks, each of which comprises a series of operations to reduce the size of the input image. Each convolution block (shown by banded grey rectangles) starts with a 3×3 convolution, which learns local spatial structure in the data. The convolution is followed by a Rectified Linear Unit (ReLU) activation layer that introduces non-linearity, then a 2×2 max-pooling layer that retains the most prominent features and downsamples the input image. Finally, a 2D batch normalisation is applied to stabilise training. Additionally, dropout with a probability of 0.3 is applied to the last two convolutional blocks during training, before batch normalisation, to reduce overfitting. Earlier convolutional blocks capture simple local structures such as gradients and edges, and later blocks combine them into more complex features, allowing the neural network to represent increasingly complex and large-scale spatial features. The number of filters used in each convolution layer determines how many distinct patterns the neural network can represent. The four convolutional layers have 16, 32, 64, and 128 filters, respectively. The features extracted by the convolutional encoder are flattened and passed to a prediction head consisting of a single hidden fully connected layer with 128 neurons, followed by ReLU activation, and an output layer with 7 neurons. Dropout with a probability of 0.4 is applied to the first linear layer during training. The optimal hyperparameters for the number of convolution blocks, the number of convolutional filters per convolutional layer, dropout probabilities, batch size and learning rate were determined using a Bayesian hyperparameter optimisation (Wu et al., 2019).

2.3.2 Training strategy and performance evaluation

The data are split chronologically into training (1979-2008, 30 years), validation (2009-2016, 8 years) and testing (2017-2024, 8 years) periods. Training data were used to train the neural network, and validation data were used to tune its hyperparameters. The test data were used only to report the final neural network skill after freezing the architecture, weights, and biases. We used a learning rate of $1.5 \cdot 10^{-4}$ and a batch size of 512. To account for class imbalance (heavy precipitation is rare by construction), we use a binary cross-entropy loss, with the positive class (heavy precipitation events) assigned a weight of 19, approximately equal to its inverse frequency in the training set.

The classification task becomes ambiguous near the threshold, where small differences in input values can lead to different class assignments. To address this, we apply label-smoothing regularisation (Szegedy et al., 2016). Instead of purely binary targets, only values below the 75^{th} percentile are assigned 0 and values between the 75^{th} and 95^{th} percentile transition smoothly from 0 to 1 following a sigmoid function. This reduces the penalty for errors near the threshold, slightly improving performance and helping stabilise training. Similar improvements in training are expected with lower thresholds of the 70^{th}, 80^{th}, and 90^{th} percentiles.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

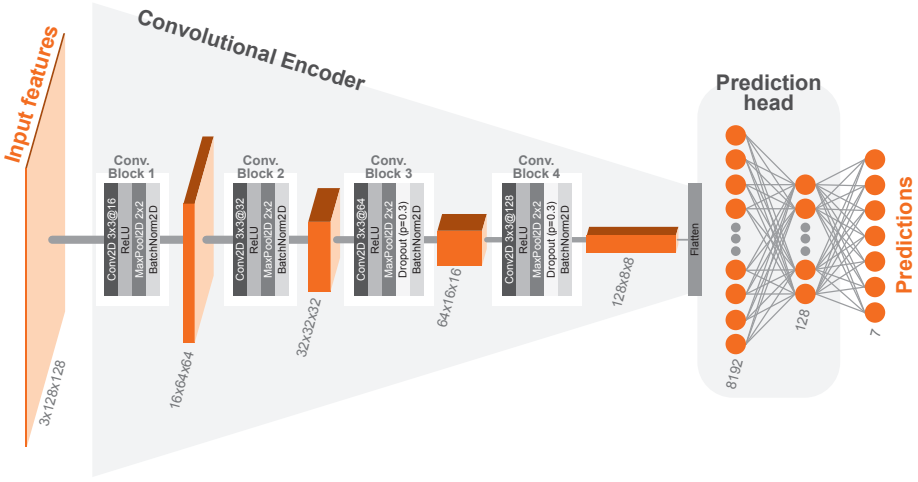

Figure 2. Architecture of the neural network in which information flows from left (input fields) to right (final prediction of seven probabilities) through the different layers. The 3D orange parallelepipeds show the shape of the dataset at different stages of the encoder, starting from the input with 3 channels (U850, V850 and SLP) times 128 x-values times 128 y-values. The shapes are also written at the bottom of the parallelepipeds. The convolutional blocks are represented by 4 or 5 vertical bars, with labels indicating the processing at each step. The fully connected layers of the prediction head are shown on the right-hand side, with neurons represented as red circles and grey lines representing individual connections.

We assess the performance of the neural network with precision-recall metrics. These provide a more informative evaluation of predictive skill for highly imbalanced datasets than other classification metrics, such as the receiver operating characteristic curve (ROC). The Precision-Recall Area Under Curve (PR-AUC) metric (Saito and Rehmsmeier, 2015), which has been used in other heavy precipitation prediction studies (Antoniadou et al., 2023), ranges from 0 to 1, where 1 corresponds to a perfect classification. To evaluate predictive skill, we compare the neural network's PR-AUC against two reference predictors, and consider the model skilful if its PR-AUC exceeds the reference PR-AUC values. The first reference predictor is the climatology, which assigns a constant probability equal to the event frequency. For an event occurring 5 per cent of the time (95^{th} percentile), the PR-AUC is 0.05. The second reference predictor is persistence, which predicts that a heavy precipitation event will occur at future lead times if it occurs on the initial day D_0. Note that persistence-based predictions are not defined at D_0 itself, but yield a PR-AUC of slightly above 0.1 at 1-day lead time, and around 0.06 at longer lead times (Fig. 3).

For same-day predictions, the neural network achieves a PR-AUC of 0.60, which is approximately 12 times higher than climatology. Performance decreases with lead time, dropping to 0.15 at D_0-3, and 0.11 at D_0-6. Despite this decrease, our trained neural network exhibits predictive skill across all lead times (Fig. 3) by consistently outperforming both references (climatology and persistence). Antoniadou et al. (2023) found a best PR-AUC of 0.42 for a similar task, using a neural network to predict the occurrence of same-day heavy precipitation in Denmark from large-scale atmospheric fields (ERA5). While their target events are slightly rarer than ours (13 days per year in their study and 18 days per year in our dataset), the results

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

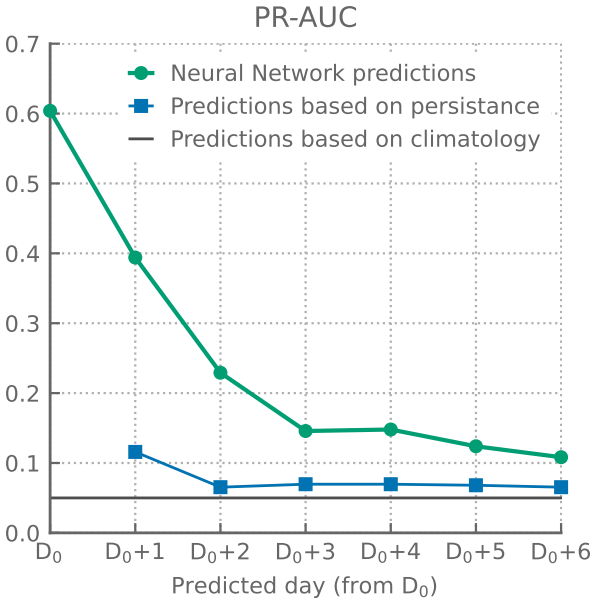

Figure 3. Area Under the Precision-Recall Curve (PR-AUC) computed for our neural network predictions at different lead times (green circles), compared with PR-AUC computed from two reference predictors: persistence-based predictions (blue squares), and climatological predictions (black line).

are broadly comparable. Overall, the assessment of the neural network performance indicates that it is sufficiently skilled to support an investigation of whether its predictions are based on physically relevant features of the input fields.

3 Object-based explainable AI framework

We design an evaluation framework to assess whether our neural network has learned to identify, track and focus on physically relevant weather features when predicting the occurrence of heavy precipitation. To do this, we first use XAI methods to identify which pixels of the gridded input variables contribute to the predicted probability of a heavy precipitation event. We then compare areas of high contribution with North Atlantic cyclone tracks. While it would be interesting to study all predictions, we focus on assessing only the positive cases, *i.e.*, instances when heavy precipitation occurs.

Across the full dataset, including the training, validation and test periods (1979 -2024), there is a total of 840 heavy precipitation events. Our XAI analysis is conducted for all these heavy precipitation events, regardless of the neural network's predicted probability. While predictions are classified as positive only when the probability exceeds a given threshold, relevant cyclones can still contribute to predicted probabilities even when these probabilities remain below the threshold. This could indicate that the neural network detects physically meaningful features but may be insufficiently sensitive to them. The 840 events are predicted by the neural network at seven different lead times, resulting in a total of 5880 predictions analysed and

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

compared with cyclone tracks. The neural network produces predictions for future days (D_0, $D_0 + 1$, ...), but we reorganise the data to focus on how the same event is predicted at different lead times. Thus, for each heavy precipitation event occurring on day D_0, we analyse predictions made on day D_0, $D_0 - 1$, $D_0 - 2$, ... back to $D_0 - 6$.

3.1 Relevance maps computation with Layer-wise Relevance Propagation

To interpret the neural network predictions, we compute relevance maps using Layer-wise Relevance Propagation (LRP, Bach et al., 2015) for each of the 5880 predictions (840 heavy-precipitation events $\times$ 7 lead times). Relevance values quantify the contribution of each input pixel to the final prediction: positive relevance indicates pixels that support the predicted probability of heavy-precipitation occurrence, whereas negative relevance indicates pixels that oppose it. These maps, therefore, indicate where the neural network focuses geographically, which variables it relies on, and how strongly this information influences the predictions.

LRP is one of several XAI techniques commonly used in climate and weather applications. These broadly include backwards-propagation methods, such as LRP and Grad-CAM (Selvaraju et al., 2020), as well as methods based on neural networks' gradients, such as Integrated Gradients (Sundararajan et al., 2017). Although these methods can yield quantitatively different results, previous studies find that large-scale attribution patterns are qualitatively stable (Bommer et al., 2024; Mamalakis et al., 2022a). Here, Integrated Gradients and LRP produce similar spatial structures. However, those from Integrated Gradients are substantially noisier due to gradient-shattering effects in deep neural network architectures (Montavon et al., 2019), while those from LRP are more coherent, allowing for easier physical interpretation. LRP has been used to interpret other heavy precipitation classification studies (e.g. Davenport and Diffenbaugh, 2021; Wang et al., 2024). For these reasons, we adopt LRP as our main XAI method.

LRP works by propagating the prediction backwards through the neural network. At each layer, the prediction score is redistributed to neurons in previous layers according to the neural network weights, thereby quantifying the contribution of individual neurons to the prediction. This process is repeated layer by layer until the input layer is reached, producing a relevance field $R(\text{variable}, x, y)$ with the same dimensions as the input data (see Montavon et al., 2019, for a detailed explanation).

At each back-propagation step, different rules can be applied to constrain the redistribution of information to the previous layer. Following general guidelines to obtain stable and meaningful relevance patterns (Mamalakis et al., 2022a; Montavon et al., 2019), we use a composite LRP rule with:

- $\alpha_1\beta_0$ rule for convolutional layers (as used in Davenport and Diffenbaugh, 2021). In $\alpha_x\beta_y$ rules, positive relevance values are multiplied by x and negative values are multiplied by y before being propagated to the previous layers. Therefore, the $\alpha_1\beta_0$ rule only propagates positive relevance values back. Such a rule helps produce more coherent and smoother relevance patterns.

- ϵ rule for fully connected layers. In this rule, a small constant ϵ is added to the sum of relevance at a given layer before being propagated back. This prevents the amplification of small signals that could happen when dividing by near-zero relevance. Contrary to $\alpha_1\beta_0$, this rule allows for negative relevance.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

To facilitate comparison of relevance across predictions, we normalise the relevance for a given prediction by the sum of all positive relevance values across all input variables and pixels:

$$R_{\text{norm}}(\text{variable}, x, y) = \frac{R(\text{variable}, x, y)}{\sum_{\text{variable},x,y} \max(0, R(\text{variable}, x, y))} \tag{2}$$

This normalisation ensures that the total positive relevance sums to one, allowing the relevance at each pixel to be interpreted as a relative contribution to the prediction. Unlike max-normalisation (used in Toms et al., 2020), our sum-normalisation is less sensitive to outlier pixels dominating the scaling and hiding broader structure in relevance maps. From here onwards, all analyses use these normalised relevance values, which will be referred to as relevance. Figure 4 shows examples of relevance maps for predictions of one heavy precipitation event at different lead times.

3.2 Detection of relevance patches and association with cyclones

Once relevance maps are computed for all predictions, we quantify to what extent positive relevance pixels are spatially clustered around specific cyclones, as seen in Fig. 4. Our method is presented with an example in Fig. 5. 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% (Fig 5.a. to b.). For comparison, if relevance were uniformly distributed over all pixels, each pixel would have a relevance of 0.006% ($= 100\%/(128 \times 128\text{pixels})$), which is 16$\times$ smaller than the threshold chosen. Most relevant patches exhibit maximum values well above this threshold, making the results mostly insensitive to the exact choice of threshold. Positions of cyclone centres on the day of the prediction are overlaid with the relevance patches (Fig 5.b), and a given patch is associated with a cyclone if any pixel of the patch is within 500 km of the centre (Fig 5.b. to c.).

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. If several cyclones are still associated with the same sub-patch, we select the cyclone for which a circle of 500 km radius around its centre has the most overlap with the sub-patch. Finally, if multiple sub-patches are associated with the same cyclone, then they are combined into a single relevance patch. We then compute the total relevance per patch, i.e., the sum of the relevance over all variables and pixels within the patch. Patches with a total relevance of less than 10% are discarded. Overall, each prediction is either associated with the cyclone corresponding to the strongest patch, or not associated with any cyclone if there is no coherent relevant patch or if none of the coherent patches is close enough to a cyclone.

4 Assessing the physical consistency of predictions

Based on the association of cyclones with our neural network's predictions in Section 3.2, we now assess whether the predictions of our neural network align with physical understanding. Specifically, we first show that most predictions are based on cyclones, regardless of their location or intensity. We then test whether the associated cyclones have reasonable intensities and occur in appropriate locations across successive lead times.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

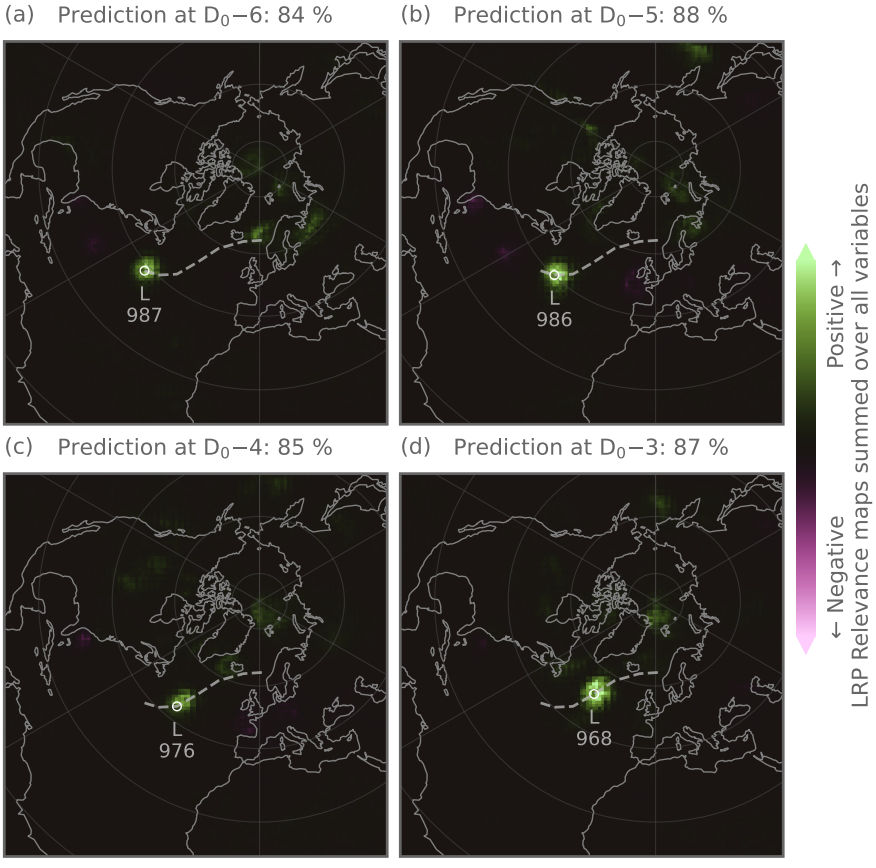

Figure 4. Example of relevance maps summed over all input variables. These four maps show the relevance for the prediction of a heavy precipitation event occurring on August, 14^{th} 2005 at six (a), five (b), four (c) and three (d) days' lead time. The title of each subplot shows the predicted probability of heavy precipitation. The trajectory of one specific cyclone is shown (dashed line), along with the position of its centre and its minimum sea level pressure value at every prediction lead time. This highlights that high relevance is concentrated around the cyclone trajectory.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

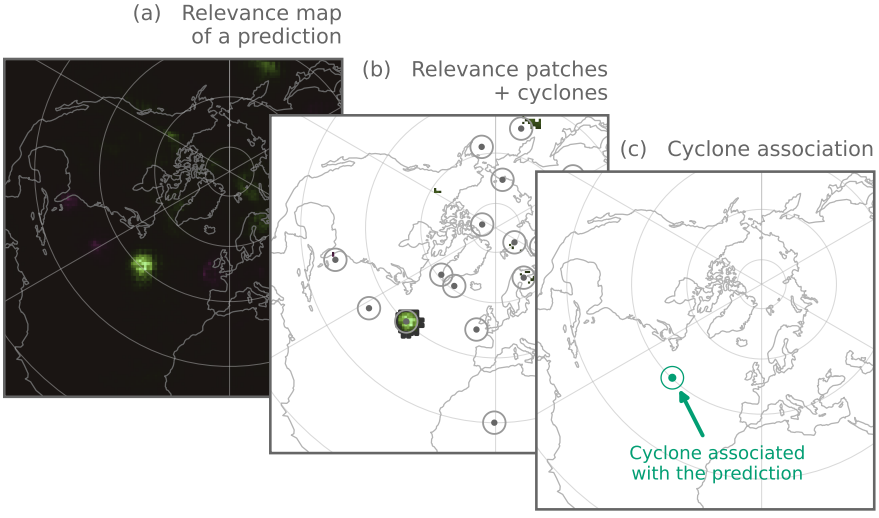

Figure 5. Example of the method to associate a cyclone with a prediction, highlighting the different steps of our analysis. a. Relevance map summed over the three input variables. This example shows the relevance for the prediction of an event occurring on August, 14^{th} 2005 with a 5-day lead time (corresponding to Fig. 5.b). b. Segmentation of the relevance map into contiguous relevance patches. In this example, one main patch is found, but other smaller patches exist. Relevance patches are overlaid with all cyclones in the Melbourne cyclone tracking dataset occurring on the date of prediction. The dark dots highlight the centres and the grey circles a 500 km radius around the cyclone centres. The main relevance patch is collocated with a cyclone. c. Map showing the cyclone associated with the prediction (centre and 500 km radius in red).

4.1 Are the neural network predictions based on cyclones?

We evaluate whether cyclones contribute to the neural network's predictions by testing whether high-relevance patches are systematically collocated with cyclones and whether these patches correspond to known structural characteristics of cyclones.

We start by quantifying the fraction of predictions that is associated with cyclones (Fig. 6, orange line). A higher fraction indicates that more relevance patches are collocated with cyclones. We find that, for same-day predictions (D_0), 93% of predictions are associated with a cyclone. This fraction decreases with increasing lead-time and stabilises at 30-35% by D_0-4. The decline is expected, given that many cyclones associated with heavy precipitation form only a few days before the event and are therefore not present at earlier lead times. Furthermore, if the neural network is temporally consistent, relevance patches at increasing lead times should correspond to the same cyclone as it propagates along its track. To test this, we track the cyclones associated with D_0 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).

However, the number of predictions associated with cyclones remains slightly higher than expected from this constraint alone. One possible explanation is that the neural network may assign relevance to cyclones that are dynamically related to the

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

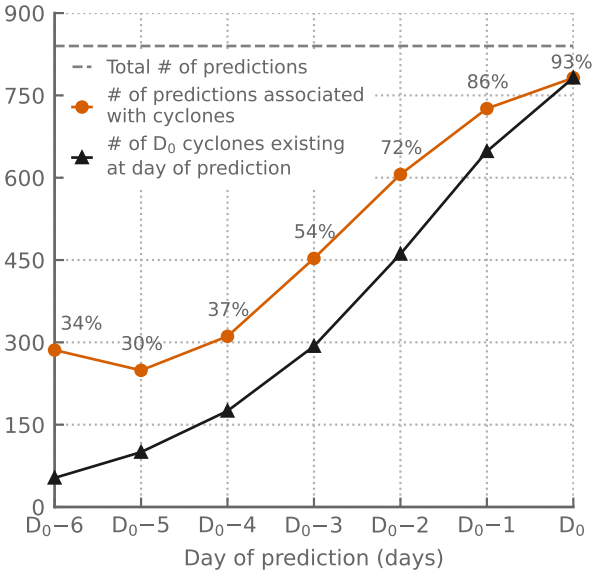

Figure 6. Number of predictions associated with cyclones at different lead times (orange line with dots). The cyclone association rate at each lead time is shown above the dot, and the total number of predictions is shown by the dashed grey line (840 occurrences of heavy precipitation events in the full dataset, including training, testing and validation). For example, at a 2-day lead time, 72% of the 840 predictions were associated with a cyclone. The black line with triangles shows the number of cyclones associated with D_0 predictions that existed at a given lead time. This gives an expected number of predictions that should be associated with cyclones if the neural network were consistently using the same cyclones across prediction lead times.

event-producing cyclone, even if the tracking algorithm labels them as separate systems or /if the exact precipitation outcome depends on subsequent atmospheric evolution.

While our primary focus is to provide physical interpretability of the predictions, these results also inform us on the reduction in predictive skill at longer lead times. With increasing lead-time, fewer relevant cyclones are present, leaving the neural network with less information to establish strong statistical links to heavy precipitation. Under these conditions, the neural network must rely on dynamical precursors to cyclogenesis and cyclone development. These precursors are more diverse and less spatially coherent than mature cyclones, making them harder for the neural network to learn consistently.

Thus far, a cyclone associated with a prediction is one that is collocated with a strong relevance patch. To assess whether this association reflects more than simple spatial collocation, we construct cyclone-centred composites of relevance for the associated cyclones. This allows us to examine whether areas of high relevance correspond to known structural features of cyclones, such as the location of strongest winds or the sea-level pressure minimum. Composites of wind speed and SLP are plotted along with the corresponding relevance maps in Fig. 7, all rotated according to the direction of propagation of individual cyclones (following Catto et al., 2010; McErlich et al., 2023; Yu et al., 2026). Cyclone-centred composites of relevance indicate that predictions are based on features of the input variables that are physically interpretable in terms of cyclone structure (Fig.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

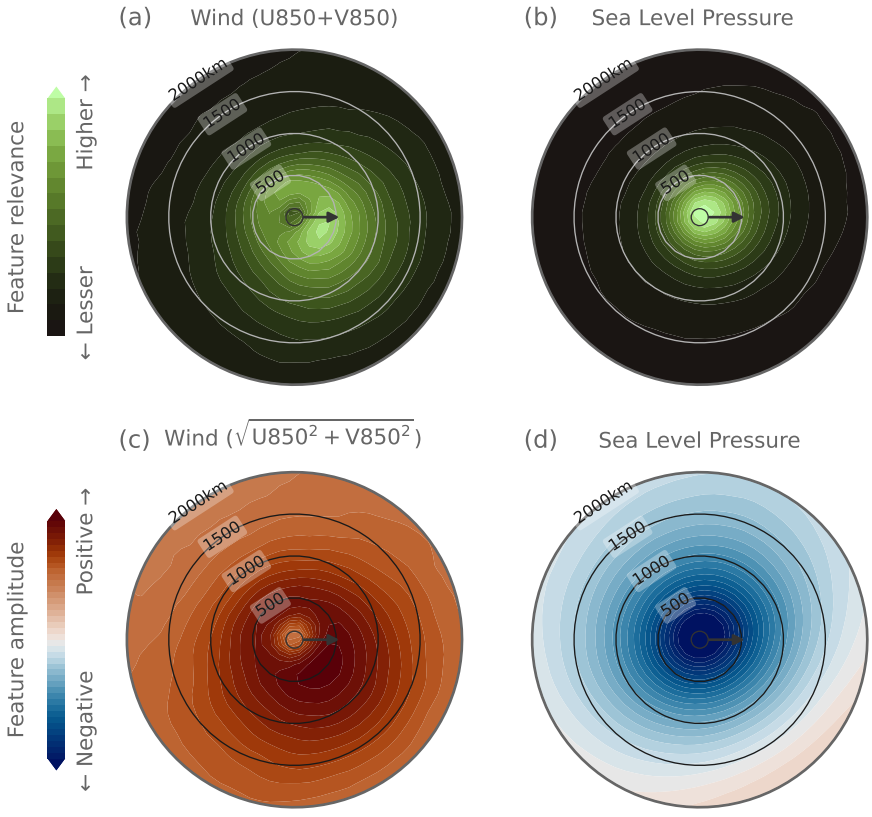

Figure 7. (a-b) Cyclone-centred composites of relevance for all predictions associated with a cyclone for 850 hPa wind (U850 + V850, a), and sea level pressure (SLP, b). (c-d) Corresponding cyclone-centred composites of input fields of wind speed ($\sqrt{U850^2 + V850^2}$, c) and sea level pressure (d). All composites are centred on the cyclone's centre (small circle in the middle) and rotated such that the direction of propagation (black arrow) is aligned with the positive x-axis. The concentric circles show the distance to cyclone centres every 500 km.

7). For SLP (Fig. 7.a), the composite relevance peaks at the cyclone centre where the SLP minimum occurs, and relevance decreases sharply away from the centre on length scales matching the typical horizontal extent of mid-latitude cyclones. This is consistent with the SLP field composite (Fig. 7.c). For wind, relevant pixels are concentrated on the outer edge of the cyclone, around 500 km from the centre and predominantly in the forward-right quadrant relative to the direction of motion (Fig. 7.b). This is consistent with where winds are often strongest due to the presence of a fast-moving cold front (Fig. 7.d). These results highlight that regions of high relevance are not merely collocated with cyclones but are also associated with their dynamical characteristics.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

4.2 Which cyclones contribute to the neural network predictions?

Showing that cyclones contribute to the neural network predictions is a first step towards establishing physical consistency, but we still need to assess whether the predictions are based on the appropriate cyclones. We do this by studying the intensity and location of the associated cyclones.

We first assess whether predictions are based on stronger cyclones, which are more likely to generate heavy precipitation in Western Norway. We evaluate the distribution of the minimum sea level pressure (SLP) of cyclones associated with neural network predictions compared to the distribution of minimum SLP of all low-pressure systems in the cyclone tracks dataset (Fig. 8.a), as well as the extent to which intense cyclones are associated with predictions (Fig. 8.a). We find that predictions are systematically based on systems with lower central SLP (i.e., stronger cyclones) – those with a median of 969 hPa and a 5-95th percentile range of 945-991 hPa, compared with a median of 997 hPa and range of 966-1015 hPa for all tracked cyclones. Additionally, more intense cyclones have a higher probability of being associated with a prediction (Fig. 8.b). 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. This intensity dependence is physically plausible and reassuring, but does not provide the full picture. The neural network may partially rely on cyclone strength as a proxy for heavy precipitation potential, but cyclone trajectory also matters for whether a system will produce heavy rainfall in Western Norway.

To examine the role of cyclone location, we next analyse the geographical locations of cyclones associated with predictions at different lead times. We divide our domain into geographical sectors (Fig. 9.a), including three relevant domains along the North Atlantic storm track (West Atlantic, East Atlantic and Europe), the Tropics (which includes the tropical Atlantic, North Africa and the Mediterranean), the Pacific, and Asia. Each cyclone associated with a prediction is assigned to one of these domains, and we count how many cyclones are in each domain at each lead time (red lines in Fig. 9). We find that cyclones contributing to the neural network predictions are mainly found in the three North Atlantic sectors. At D_0, most of the associated cyclones are in the European sector, primarily just offshore of Norway (∼650, Fig. 9.a exact location not shown), with a smaller number in the East Atlantic sector, close to Iceland (∼150 Fig. 9.b). At D_0-1 and D_0-2, the associated cyclones are mainly in the North Atlantic, with up to 55% of predictions (∼450 cases) based on a cyclone in the East Atlantic (Fig. 9.b), while the number of predictions associated with cyclones in the European sector is smaller, with 30% of all predictions at D_0-1 (∼ 250 cases) and 6% at D_0-2 (∼ 50 cases, Fig. 9.a). By D_0-3, the East Atlantic and Europe remain the main contributors, but with fewer cyclones than at shorter lead times. The West Atlantic's contribution peaks at this lead time, with just under half the number of cyclones as in the East Atlantic (Fig. 9.c). Beyond D_0-3, the predictions are roughly evenly split between European and East Atlantic cyclones (around 100 each, or 10% of predictions), with some contribution from West Atlantic cyclones (around 50, or 5%). Contributions from the three other sectors (Asia, Pacific and Tropics) are negligible (Fig. 9.d-f), as expected, because these regions are mostly disconnected dynamically from heavy precipitation in Western Norway, except for rare cases of tropical to extratropical cyclone transitions. These results highlight that the locations of cyclones associated

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

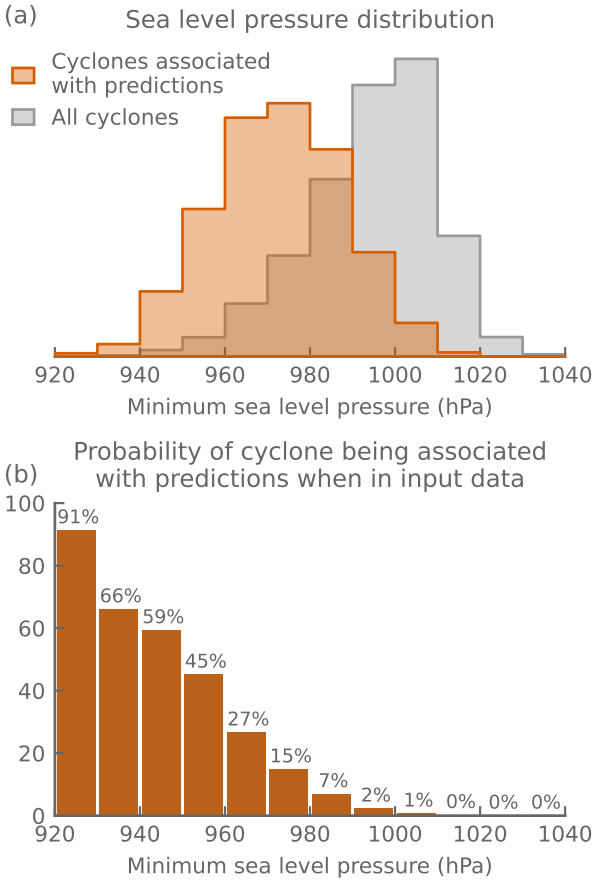

Figure 8. a. Histogram of the minimum sea level pressure of all cyclones in the Melbourne tracking dataset (grey) and of the subset of cyclones associated with predictions of the neural network (orange). b. The probability that a cyclone seen by the neural network is associated with a prediction of a heavy precipitation event, depending on its minimum sea level pressure. Probabilities are computed in bins of 10 hPa.

with predictions shift upstream as lead time increases, consistent with the expected evolution of weather systems along the North Atlantic storm track.

We further test whether the spatial distribution of cyclones associated with predictions at different lead times is consistent with trajectories from the cyclone tracks dataset. To do so, we follow the cyclones identified at zero lead (D_0) backwards in time and record the geographical regions they occupied at earlier lead times (black lines in Fig. 9). Our results show that cyclones associated with predictions at longer lead-times (orange lines in Fig. 9) largely correspond to the earlier positions of cyclones associated with D_0 predictions (black lines in Fig. 9). This indicates that the predictions are aligned with the expected trajectories of mid-latitude cyclones along the North Atlantic storm track.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

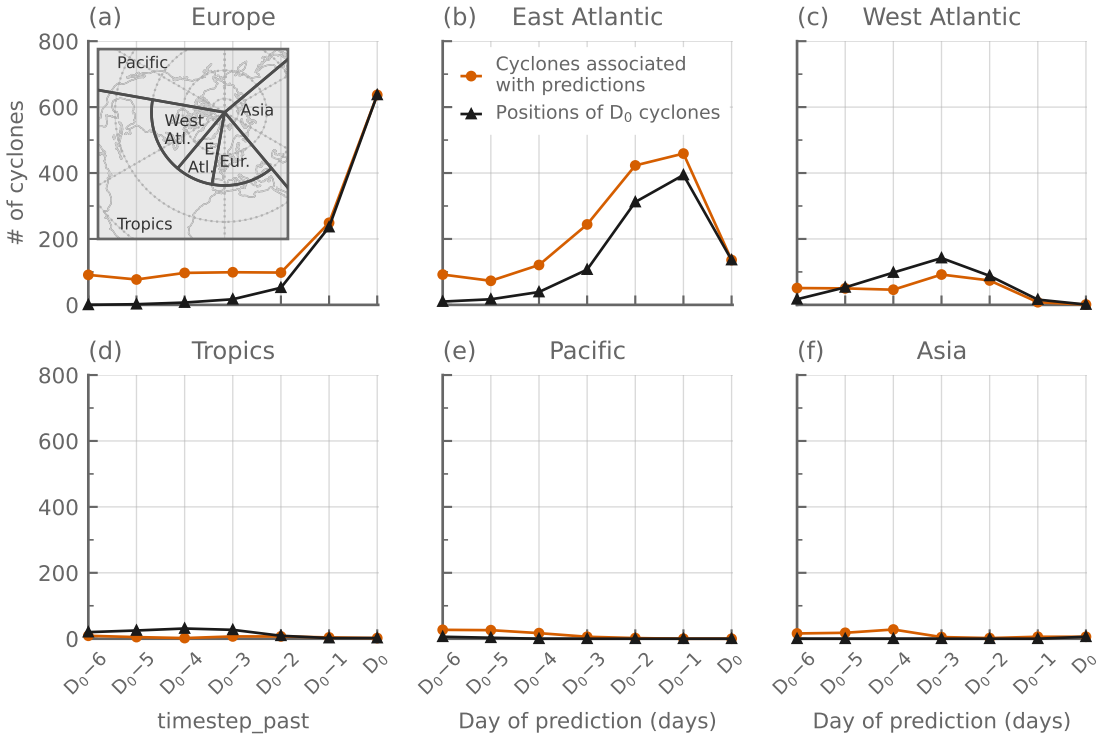

Figure 9. Location of the cyclones associated with predictions at different lead times, split by geographical sectors (red lines with circles). These are compared to the past positions of the cyclones associated with predictions at D_0 tracked back in time (black lines with triangles). The map in a. shows the different sectors studied: Europe (Eur.), the East Atlantic (E. Atl.) sector, the West Atlantic (West Atl.) sector, Asia, the Pacific and the Tropics.

4.3 Limitations

While we provide evidence that the neural network predictions align with physical understanding, our framework relies on methodological and data choices designed to help the neural network learn the appropriate cyclone-related features, which can introduce an inductive bias. Notably, the neural network can only predict from a partial representation of cyclones existing in the input data, which does not include higher-level winds and vertical gradients, for example. Similarly, other processes not included in input variables, such as moisture availability and transport (Azad and Sorteberg, 2017; Benedict et al., 2019), convection (Xie et al., 2025) and orographic enhancement (Barstad and Caroletti, 2013), cannot contribute to predictions even though they influence heavy precipitation in the western Norway. Furthermore, our analysis depends on the ERA5 reanalysis for dynamical fields and precipitation. While the ERA5 reanalysis has known limitations in the representation of heavy precipitation (e.g. McErlich et al., 2023), we argue that it is suitable for our objective of assessing whether the neural network's

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

predictions are physically reasonable, specifically since mid-latitude cyclones have been shown to be the main cause of heavy precipitation in Western Norway in ERA5 (Konstali et al., 2024).

Finally, we note that LRP is a post hoc explainability method (i.e., not built directly into the network architecture) that only provides an approximate explanation of the network behaviour (Montavon et al., 2019) and does not fully capture its internal decision-making process. In that sense, the results from any such method should be taken more as an indication than a definitive answer. In addition, similar post hoc explainability methods, such as integrated gradients or LRP with different rules, may yield quantitatively different results, particularly at finer spatial scales (Mamalakis et al., 2022a). While large-scale features such as cyclones are relatively robust due to their extent in the input space, smaller-scale features may be more sensitive to the choice of XAI method. Fundamentally, our method identifies only statistical alignment between the neural network relevance fields and the tracked physical features, not causal links, despite the apparent correspondence between relevance maps and key physical characteristics of cyclones. This distinction is inherent to all post hoc explainability approaches and remains an open challenge in XAI for geoscientific applications. Interpretable AI methods, which involve building predictive models that are explicitly understandable and interpretable, may be more promising for understanding or discovering physical causality (Yang et al., 2024; Lipton, 2018).

5 Concluding remarks

5.1 Main findings

In this work, we present an object-oriented XAI framework that allows us to assess whether neural network outputs align with physical understanding. Specifically, we combine XAI relevance analysis with trajectories of tracked dynamical objects to associate individual predictions with specific weather features. This approach moves beyond composite spatial analysis toward a more quantitative and process-oriented interpretation of neural network predictions, enabling physical assessment at the level of individual predictions.

Our application focuses on predicting heavy precipitation events in Western Norway. We used our framework to show that (1) the neural network predictions are systematically associated with mid-latitude cyclones, and (2) these cyclones exhibit physically reasonable intensity, location, and evolution along the North Atlantic storm track. These findings hold for lead times up to 3 to 4 days. Beyond this range, predictions become more uncertain, and interpretability decreases, consistent with the fact that many cyclones causing heavy precipitation have not yet formed at longer lead times. By linking predictions to individual cyclones, our object-based XAI framework provides evidence that the neural network predictions are based on the correct weather features, not just the correct geographical regions. Overall, our results strengthen the conclusion of Toms et al. (2020), indicating that neural networks can make physically grounded predictions when applied to weather and climate science.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

5.2 Generalisation

Our object-oriented XAI framework could be extended to more complex applications. In the current study, the predicted events (i.e. heavy rainfall in Western Norway) are mainly driven by a single, well-identified physical feature: mid-latitude cyclones. This makes the problem particularly suitable for our object-based XAI framework. Applications where the predictions are driven by multiple interacting physical features should be tractable, provided that the features can be tracked in space and time. For example, in regions like West Africa, heavy precipitation is driven by a combination of meso-scale convective systems, easterly waves and large-scale circulation patterns, notably associated with monsoon systems (Kamara, 1986). Such a prediction task would require sufficient training data for the neural network to learn to separate the multiple interacting features.

While we designed a relatively simple neural network architecture by choice, our framework could also be applied to evaluate the physical consistency of state-of-the-art AI weather prediction models such as Pangu (Bi et al., 2022), FourCastNet (Kurth et al., 2023) or AIFS (Lang et al., 2024). This would likely require adapting the XAI method used, as LRP is not compatible with all neural network architectures. Other methods, such as Grad-CAM or gradient-based approaches, may be more suitable for these AI weather models (Baño-Medina et al., 2025; Rampal et al., 2022). Combining such a framework with existing benchmarks for new AI weather models would provide a complete assessment of both predictive skill and physical consistency, therefore improving trust in operational AI forecasting.

5.3 Implications for scientific discovery

Beyond evaluating neural networks, object-based XAI offers new opportunities for scientific discovery and data-driven hypothesis generation. With this framework, individual predictions can be linked to physical features. This could enable the identification of potential physical drivers, the quantification of the relative contribution of different processes for a given prediction, and the analysis of long-term changes in driver importance (as in Davenport and Diffenbaugh, 2021). While causal inference remains limited, consistent attribution patterns across many events can provide multiple lines of evidence to support such hypotheses.

5.4 Achieving credibility through physical consistency

Building trust in AI weather models remains an important challenge due to their complexity, but it is critical if these tools are to be used in operational and decision-making contexts. As argued by O'Loughlin et al. (2025), establishing trust requires first explaining the neural network predictions, which can be done using XAI methods. Such explainability provides a foundation for establishing the credibility of predictions, eventually enabling trust. One way to achieve credibility is through component-level understanding of the neural network, in which individual components (i.e., specific layers, blocks, or groups of neurons) are linked to neural network performance (O'Loughlin et al., 2025). Here, we propose an alternative: assessing the physical consistency of individual predictions. We show that neural network predictions align with physical understanding by relying on meaningful physical features, thereby increasing their credibility. Ultimately, trust requires predictions to be accurate, but

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

also credible through a grounding in physical interpretability. Strengthening this trust supports the use of neural networks for scientific discovery and process understanding (Mamalakis et al., 2022b; Roscher et al., 2020; Yang et al., 2024).

Code and data availability. Code to make all processing and figures will be shared with a Zenodo archive. All data used in this study are publicly available.

Author contributions. RGC, CL and SS designed the study. RGC conducted the analysis, produced the results and the figures. RGC wrote the first draft of the manuscript and all three authors contributed to writing the final manuscript.

Competing interests. CL is a member of the editorial board of Weather and Climate Dynamics.

Acknowledgements. We thank Joshua Dorrington, Linus Ericsson and Qidi Yu for helpful discussions and insightful comments on this work. The project was supported in part by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101126560. This paper makes use of the scientific colormaps Vik and Vanimo by Crameri (2023).

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

References

Alessi, M. J., Connolly, C. J., Barnes, E. A., and Rugenstein, M.: Southern Hemisphere Surface Warming Drives Southwestern U.S. Precipitation according to AI-Informed Climate Model Simulations, Journal of Climate, 38, 7655–7667, https://doi.org/10.1175/JCLI-D-25-0176.1, 2025.

Ali, S. and Wang, J.: Mt-icenet-a spatial and multi-temporal deep learning model for arctic sea ice forecasting, in: 2022 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), pp. 1–10, IEEE, https://doi.org/10.1109/BDCAT56447.2022.00009, 2022.

Antoniadou, N., Sørup, H. J. D., Pedersen, J. W., Gregersen, I. B., Schmith, T., and Arnbjerg-Nielsen, K.: Comparison of data-driven methods for linking extreme precipitation events to local and large-scale meteorological variables, Stochastic Environmental Research and Risk Assessment, 37, 4337–4357, https://doi.org/10.1007/s00477-023-02511-3, 2023.

Azad, R. and Sorteberg, A.: Extreme daily precipitation in coastal western Norway and the link to atmospheric rivers, Journal of Geophysical Research: Atmospheres, 122, https://doi.org/10.1002/2016JD025615, 2017.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLOS ONE, 10, e0130 140, https://doi.org/10.1371/journal.pone.0130140, 2015.

Barstad, I. and Caroletti, G. N.: Orographic precipitation across an island in southern Norway: model evaluation of time-step precipitation, Quarterly Journal of the Royal Meteorological Society, 139, 1555–1565, https://doi.org/10.1002/qj.2067, 2013.

Baño-Medina, J., Sengupta, A., Doyle, J. D., Reynolds, C. A., Watson-Parris, D., and Monache, L. D.: Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia, npj Climate and Atmospheric Science, 8, 92, https://doi.org/10.1038/s41612-025-00949-6, 2025.

Benedict, I., Ødemark, K., Nipen, T., and Moore, R.: Large-Scale Flow Patterns Associated with Extreme Precipitation and Atmospheric Rivers over Norway, Monthly Weather Review, 147, 1415–1428, https://doi.org/10.1175/MWR-D-18-0362.1, 2019.

Beobide-Arsuaga, G., Düsterhus, A., Müller, W. A., Barnes, E. A., and Baehr, J.: Spring Regional Sea Surface Temperatures as a Precursor of European Summer Heatwaves, Geophysical Research Letters, 50, e2022GL100 727, https://doi.org/10.1029/2022GL100727, 2023.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast, https://doi.org/10.48550/arXiv.2211.02556, 2022.

Bommer, P. L., Kretschmer, M., Hedström, A., Bareeva, D., and Höhne, M. M.-C.: Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science, Artificial Intelligence for the Earth Systems, 3, https://doi.org/10.1175/AIES-D-23-0074.1, 2024.

Byrne, M. P. and O'Gorman, P. A.: Land–Ocean Warming Contrast over a Wide Range of Climates: Convective Quasi-Equilibrium Theory and Idealized Simulations, Journal of Climate, 26, https://doi.org/10.1175/JCLI-D-12-00262.1, 2013.

Catto, J. L., Shaffrey, L. C., and Hodges, K. I.: Can climate models capture the structure of extratropical cyclones?, Journal of Climate, 23, https://doi.org/10.1175/2009jcli3318.1, 2010.

Crameri, F.: Scientific colour maps, https://doi.org/10.5281/zenodo.8409685, 2023.

Davenport, F. V. and Diffenbaugh, N. S.: Using Machine Learning to Analyze Physical Causes of Climate Change: A Case Study of U.S. Midwest Extreme Precipitation, Geophysical Research Letters, 48, https://doi.org/10.1029/2021GL093787, 2021.

de Burgh-Day, C. O. and Leeuwenburg, T.: Machine learning for numerical weather and climate modelling: a review, Geoscientific Model Development, 16, https://doi.org/10.5194/gmd-16-6433-2023, 2023.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, https://doi.org/10.2139/ssrn.5180447, 2020.

Feser, F., Barcikowska, M., Krueger, O., Schenk, F., Weisse, R., and Xia, L.: Storminess over the North Atlantic and northwestern Europe—A review, Quarterly Journal of the Royal Meteorological Society, 141, 350–382, https://doi.org/10.1002/qj.2364, 2015.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., De Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, https://doi.org/10.1002/qj.3803, 2020.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W.: Explainable AI Methods - A Brief Overview, in: xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers, edited by Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., and Samek, W., pp. 13–38, Springer International Publishing, Cham, ISBN 978-3-031-04083-2, https://doi.org/10.1007/978-3-031-04083-2_2, 2022.

Kamara, I. R.: The origins and types of rainfall in West Africa, Weather, https://doi.org/10.1002/j.1477-8696.1986.tb03787.x, 1986.

Kent, C., Scaife, A. A., Dunstone, N. J., Smith, D., Hardiman, S. C., Dunstan, T., and Watt-Meyer, O.: Skilful global seasonal predictions from a machine learning weather model trained on reanalysis data, npj Climate and Atmospheric Science, 8, 314, https://doi.org/10.1038/s41612-025-01198-3, 2025.

Konstali, K., Spensberger, C., Spengler, T., and Sorteberg, A.: Global attribution of precipitation to weather features, Journal of Climate, 37, https://doi.org/10.1175/JCLI-D-23-0293.1, 2024.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., and Anandkumar, A.: FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators, in: Proceedings of the Platform for Advanced Scientific Computing Conference, pp. 1–11, ACM, Davos Switzerland, ISBN 979-8-4007-0190-0, https://doi.org/10.1145/3592979.3593412, 2023.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z. B., Nemesio, A. P., Dueben, P. D., Brown, A., Pappenberger, F., and Rabier, F.: AIFS – ECMWF's data-driven forecasting system, https://doi.org/10.48550/arXiv.2406.01465, 2024.

Lavers, D. A., Simmons, A., Vamborg, F., and Rodwell, M. J.: An evaluation of ERA5 precipitation for climate monitoring, Quarterly Journal of the Royal Meteorological Society, 148, 3152–3165, https://doi.org/10.1002/qj.4351, 2022.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86, https://doi.org/10.1109/5.726791, 2002.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A.: Fourier Neural Operator for Parametric Partial Differential Equations, https://doi.org/10.48550/arXiv.2010.08895, 2021.

Lipton, Z. C.: The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue, 16, 31–57, https://doi.org/10.1145/3236386.3241340, 2018.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, https://doi.org/10.48550/arXiv.2103.14030, 2021.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

Madonna, E., Hes, G., Li, C., Michel, C., and Siew, P. Y. F.: Control of Barents Sea Wintertime Cyclone Variability by Large-Scale Atmospheric Flow, Geophysical Research Letters, 47, https://doi.org/10.1029/2020GL090322, 2020.

Mamalakis, A., Barnes, E. A., and Ebert-Uphoff, I.: Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience, Artificial Intelligence for the Earth Systems, 1, https://doi.org/10.1175/AIES-D-22-0012.1, 2022a.

Mamalakis, A., Ebert-Uphoff, I., and Barnes, E.: Explainable Artificial Intelligence in Meteorology and Climate Science: Model Fine-Tuning, Calibrating Trust and Learning New Science, in: xxAI - Beyond Explainable AI, edited by Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., and Samek, W., vol. 13200, pp. 315–339, Springer International Publishing, Cham, ISBN 978-3-031-04082-5 978-3-031-04083-2, https://doi.org/10.1007/978-3-031-04083-2_16, 2022b.

Marcheggiani, A., Dacre, H., Spensberger, C., and Spengler, T.: Weather features drive free-tropospheric baroclinicity variability in the North Atlantic storm track, Quarterly Journal of the Royal Meteorological Society, 151, e5061, https://doi.org/10.1002/qj.5061, 2025.

Mayer, K. J. and Barnes, E. A.: Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network, Geophysical Research Letters, 48, e2020GL092 092, https://doi.org/10.1029/2020GL092092, 2021.

McErlich, C., McDonald, A., Renwick, J., and Schuddeboom, A.: An Assessment of Southern Hemisphere Extratropical Cyclones in ERA5 Using WindSat, Journal of Geophysical Research: Atmospheres, 128, https://doi.org/10.1029/2023JD038554, 2023.

Meo, C., Roy, A., Lică, M., Yin, J., Che, Z. B., Wang, Y., Imhoff, R., Uijlenhoet, R., and Dauwels, J.: Extreme Precipitation Nowcasting using Transformer-based Generative Models, https://doi.org/10.48550/arXiv.2403.03929, 2024.

Michel, C., Sorteberg, A., Eckhardt, S., Weijenborg, C., Stohl, A., and Cassiani, M.: Characterization of the atmospheric environment during extreme precipitation events associated with atmospheric rivers in Norway - Seasonal and regional aspects, Weather and Climate Extremes, 34, 100 370, https://doi.org/10.1016/j.wace.2021.100370, 2021.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., and Müller, K.-R.: Layer-Wise Relevance Propagation: An Overview, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, edited by Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., pp. 193–209, Springer International Publishing, Cham, ISBN 978-3-030-28954-6, https://doi.org/10.1007/978-3-030-28954-6_10, 2019.

Murray, R. J. and Simmonds, I.: A numerical scheme for tracking cyclone centres from digital data. Part II: Application to January and July general circulation model simulations, Australian Meteorological Magazine, 39, https://doi.org/10.1071/es91021, 1991a.

Murray, R. J. and Simmonds, I.: A numerical scheme for tracking cyclone centres from digital data Part I: development and operation of the scheme, Australian meteorological magazine, 39, 155–166, https://doi.org/10.1071/es91020, 1991b.

Neu, U., Akperov, M. G., Bellenbaum, N., Benestad, R., Blender, R., Caballero, R., Cocozza, A., Dacre, H. F., Feng, Y., Fraedrich, K., Grieger, J., Gulev, S., Hanley, J., Hewson, T., Inatsu, M., Keay, K., Kew, S. F., Kindem, I., Leckebusch, G. C., Liberato, M. L. R., Lionello, P., Mokhov, I. I., Pinto, J. G., Raible, C. C., Reale, M., Rudeva, I., Schuster, M., Simmonds, I., Sinclair, M., Sprenger, M., Tilinina, N. D., Trigo, I. F., Ulbrich, S., Ulbrich, U., Wang, X. L., and Wernli, H.: IMILAST: A Community Effort to Intercompare Extratropical Cyclone Detection and Tracking Algorithms, Bulletin of the American Meteorological Society, 94, 529–547, https://doi.org/10.1175/BAMS-D-11-00154.1, 2013.

Oldham-Dorrington, J., Li, C., Sobolowski, S., and Guillaume-Castel, R.: Understanding biases and changes in European heavy precipitation using dynamical flow precursors, EGUsphere, pp. 1–37, https://doi.org/10.5194/egusphere-2025-4977, 2025.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

O'Loughlin, R. J., Li, D., Neale, R., and O'Brien, T. A.: Moving beyond post hoc explainable artificial intelligence: a perspective paper on lessons learned from dynamical climate modeling, Geoscientific Model Development, 18, 787–802, https://doi.org/10.5194/gmd-18-787-2025, 2025.

Pegion, K., Becker, E. J., and Kirtman, B. P.: Understanding predictability of daily southeast US precipitation using explainable machine learning, Artificial Intelligence for the Earth Systems, 1, https://doi.org/10.1175/AIES-D-22-0011.1, 2022.

Rampal, N., Gibson, P. B., Sood, A., Stuart, S., Fauchereau, N. C., Brandolino, C., Noll, B., and Meyers, T.: High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand, Weather and Climate Extremes, 38, https://doi.org/10.1016/j.wace.2022.100525, 2022.

Reid, K. J., Barnes, M. A., Gillett, Z. E., Parker, T., Udy, D. G., Ayat, H., Boschat, G., Bowden, A., Grosfeld, N. H., King, A. D., Richardson, D., Shao, Y., Teckentrup, L., Trewin, B., Hope, P., Zhou, L., Borowiak, A. R., Holgate, C. M., and Isphording, R. N.: A Multiscale Evaluation of the Wet 2022 in Eastern Australia, Journal of Climate, 38, 909–929, https://doi.org/10.1175/JCLI-D-24-0224.1, 2025.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J.: Explainable machine learning for scientific insights and discoveries, Ieee Access, 8, https://doi.org/10.1109/ACCESS.2020.2976199, 2020.

Rudeva, I., Gulev, S. K., Simmonds, I., and Tilinina, N.: The sensitivity of characteristics of cyclone activity to identification procedures in tracking algorithms, Tellus A: Dynamic Meteorology and Oceanography, 66, 24 961, https://doi.org/10.3402/tellusa.v66.24961, 2014.

Rugenstein, M., Van Loon, S., and Barnes, E. A.: Convolutional Neural Networks Trained on Internal Variability Predict Forced Response of TOA Radiation by Learning the Pattern Effect, Geophysical Research Letters, 52, https://doi.org/10.1029/2024GL109581, 2025.

Saito, T. and Rehmsmeier, M.: The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets, PLOS ONE, 10, e0118 432, https://doi.org/10.1371/journal.pone.0118432, 2015.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, International Journal of Computer Vision, 128, https://doi.org/10.1007/s11263-019-01228-7, 2020.

Spensberger, C. and Marcheggiani, A.: ERA5 cyclone tracks, https://doi.org/10.11582/2024.00023, 2024.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, p. 3319–3328, JMLR.org, https://doi.org/10.5555/3305890.3306024, 2017.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z.: Rethinking the Inception Architecture for Computer Vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, ISBN 978-1-4673-8851-1, https://doi.org/10.1109/CVPR.2016.308, 2016.

Tao, D., Li, C., Davy, R., He, S., Spengler, T., Michel, C., and Rosendahl, A.: Arctic-Atlantic Cyclones: Variability in Thermodynamic Characteristics, Large-Scale Flow, and Local Impacts, Geophysical Research Letters, 52, e2024GL111 769, https://doi.org/10.1029/2024GL111769, 2025.

Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I.: Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability, Journal of Advances in Modeling Earth Systems, 12, e2019MS002 002, https://doi.org/10.1029/2019MS002002, 2020.

Unal, A., Asan, B., Sezen, I., Yesilkaynak, B., Aydin, Y., Ilicak, M., and Unal, G.: Climate model-driven seasonal forecasting approach with deep learning, Environmental Data Science, 2, https://doi.org/10.1017/eds.2023.24, 2023.

Walker, E., Mitchell, D., and Seviour, W.: The numerous approaches to tracking extratropical cyclones and the challenges they present, Weather, 75, https://doi.org/10.1002/wea.3861, 2020.

Wang, B., Wei, W., Yin, Z., and Xu, L.: Using machine learning to analyze the changes in extreme precipitation in southern China, Atmospheric Research, 302, 107 307, https://doi.org/10.1016/j.atmosres.2024.107307, 2024.

https://doi.org/10.5194/egusphere-2026-3362
Preprint. Discussion started: 29 June 2026

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., and Deng, S.-H.: Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb, Journal of Electronic Science and Technology, 17, 26–40, https://doi.org/10.11989/JEST.1674-862X.80904120, 2019.

Xie, K., Li, L., Chen, H., Mayer, S., Dobler, A., Xu, C.-Y., and Göktürk, O. M.: Enhanced evaluation of hourly and daily extreme precipitation in Norway from convection-permitting models at regional and local scales, Hydrology and Earth System Sciences, 29, 2133–2152, https://doi.org/10.5194/hess-29-2133-2025, 2025.

Yang, R., Hu, J., Li, Z., Mu, J., Yu, T., Xia, J., Li, X., Dasgupta, A., and Xiong, H.: Interpretable machine learning for weather and climate prediction: A review, Atmospheric Environment, 338, 120 797, https://doi.org/10.1016/j.atmosenv.2024.120797, 2024.

Yu, Q., Spensberger, C., Magnusson, L., and Spengler, T.: Distinct bias structures for extratropical cyclones with strong or weak diabatic heating, EGUsphere, 2026, 1–22, https://doi.org/10.5194/egusphere-2026-257, 2026.

Ødemark, K., Müller, M., Palerme, C., and Tveito, O. E.: Recent changes in circulation patterns and their opposing impact on extreme precipitation at the west coast of Norway, Weather and Climate Extremes, 39, 100 530, https://doi.org/10.1016/j.wace.2022.100530, 2023.