



Explaining patterns learned by neural networks for detection of atmospheric surface fronts in gridded atmospheric data

Tim Radke^{1, 2, 3}, Johanna Baehr³, and Marc Rautenhaus^{1, 4}

¹Visual Data Analysis Group, Hub of Computing and Data Science, University of Hamburg, Hamburg, Germany

²IMPRS-ESM, Max Planck Institute for Meteorology, Hamburg, Germany

³Earth and Society Research Hub (ESRAH), University of Hamburg, Hamburg, Germany

⁴Now at: Fulda University of Applied Sciences, Department of Applied Computer Science, Fulda, Germany

Correspondence: Tim Radke (tim.radke@uni-hamburg.de)

Abstract. Atmospheric surface fronts (SFs) strongly influence midlatitude weather through storms, precipitation, and abrupt temperature changes, motivating reliable detection. Here, we reproduce a U-Net, a type of artificial neural network, to detect SFs in ERA5 reanalysis data with labelled datasets over North America and Europe. The goal of this study is to apply Layer-wise Relevance Propagation (LRP) to interpret the learned patterns. To reduce computational cost, we introduce a, compared to previous studies, reduced U-Net variant and a reduced input set of six variables - temperature, specific humidity, and horizontal wind - without reducing detection quality. Relevance analysis shows that U-Net focuses partly on physically meaningful regions near detected SFs, consistent with rule-based methods, but also noisy patterns and spurious correlations far from the detected fronts. Moreover, U-Net employs two nearly inverted relevance patterns for different parts of a front, which we see as not physically plausible. Using a simplified synthetic dataset of a 1-D field with SF labels placed at the minimum of the second derivative, we can link the pattern to the front position relative to the frontal zone. Increasing the label resolution enables the U-Net to learn a more plausible pattern without the split in the synthetic case. When evaluated over a larger region (Pacific to Europe, equator to Arctic), U-Net performs well in midlatitudes but over-detects SFs in the Tropics and Arctic, for example regularly detecting SFs within tropical cyclones. Together with the partially implausible relevance patterns, this indicates that ANN-based SF detection should be used with care outside the training regions. Retraining on the larger domain improves detection skill and the plausibility of relevance patterns, but the split pattern persists. Overall, the results highlight the value of explainable AI for identifying physically implausible shortcuts and the need for a global SF dataset to enable reliable detection worldwide.

1 Introduction

Atmospheric surface fronts (SFs) are defined as boundaries between air masses with differing properties, such as temperature or humidity (e.g., Stull, 2017). They are of importance for midlatitude weather, as they are frequently associated with the development of storms, extreme precipitation events, and abrupt temperature changes (Catto et al., 2012; Catto and Pfahl, 2013; Hope et al., 2014; Naud et al., 2015). Detection of SFs is therefore important for both weather forecasting and climate analysis (Shapiro and Keyser, 1990; Hewson, 1997; Catto et al., 2014). Typically, SFs are identified using either manual expert



analysis (e.g., Saucier, 1955) or rule-based detection methods (e.g., Hewson, 1998; Simmonds et al., 2012; Kern et al., 2019; Beckert et al., 2023). Rule-based approaches often rely on predefined thresholds of one or more variables, such as temperature or humidity gradients. For example, Hewson (1998) detects the warm-air side of zones of horizontal gradients of a thermal variable, a method extended by Kern et al. (2019) and Beckert et al. (2023) to also detect 3-D front structures. In contrast, Simmonds et al. (2012) detect SFs using wind direction shifts and meridional wind change. Other methods use multiple variables for detection. Ribeiro et al. (2016) and Parfitt et al. (2017) use temperature, humidity and wind fields; Foss et al. (2017) detects cold fronts using sea level pressure, temperature and wind fields.

Rule-based methods are used to compute climatologies as well as in operational forecasts. Their dependence on predefined thresholds, however, poses challenges including spurious features that would not be detected by human experts (Beckert et al., 2023) making human expertise in SF detection still indispensable in operational forecasting.

Recent advances in machine learning offer alternatives to rule- or expert-based detection methods in particular offering potential to mimic also detections by human experts and being computationally faster than rule-based methods. Especially convolutional neural networks (CNNs, Goodfellow et al., 2016) — a type of artificial neural networks (ANN) — have demonstrated strong performance in various tasks (Alzubaidi et al., 2021; Zhao et al., 2024). Several studies investigate using CNNs to detect SFs (e.g., Biard and Kunkel, 2019; Lagerquist et al., 2019; Niebler et al., 2022; Justin et al., 2023). They formulate the detection of SFs as a semantic segmentation task, often encountered in computer vision (Long et al., 2015; Liu et al., 2019; Xie et al., 2021; Manakitsa et al., 2024), where each pixel of an image is classified. Similarly for the detection of atmospheric features, each grid point of input data is classified as either background or atmospheric feature.

Biard and Kunkel (2019) and Lagerquist et al. (2019) detect larger "objects" in the area of the SFs and apply a postprocessing algorithm to extract one grid point wide frontal boundaries. In contrast, Niebler et al. (2022) and Justin et al. (2023) formulate the detection task to directly detect frontal boundaries. All four studies use data from weather services with regions including most parts of North America and western Europe as well as the northern Atlantic and Pacific.

While these studies report that CNNs are well suited for the detection of SFs, they are inherently "black box" systems (Zhang et al., 2021), making it difficult to determine whether their predictions are based on physically meaningful patterns or spurious correlations in the training data. This lack of interpretability raises concerns regarding the physical plausibility, generalizability and robustness of CNN-based detection systems, especially when they are trained only on data from localized geographic regions. For instance, Niebler et al. (2022) and Justin et al. (2023) demonstrate that CNNs trained on data from one region perform worse when applied to other regions, suggesting that the CNNs may be learning region-specific features rather than generalizable frontal characteristics.

In this study, we aim to shed light on "black box" CNNs for SF detection by addressing the following objectives:

1. Which meteorological variables are required as input to a CNN detecting SFs (Sect. 5)?
2. Are physically plausible patterns learnt, particularly when using regionally restricted (non-global) data (Sect. 6)?
3. Can a CNN trained on regional data be applied to regions outside the training domain (Sect. 7)?



To address similar concerns for other tasks, explainable artificial intelligence (xAI) techniques have been proposed (Linardatos et al., 2021; Holzinger et al., 2022; Mersha et al., 2024). One such method is Layer-wise Relevance Propagation (LRP, Bach et al., 2015), which was adapted by Radke et al. (2025) to interpret CNNs trained to detect tropical cyclones and atmospheric rivers. They show that their CNNs detect tropical cyclones and atmospheric rivers based on physically plausible patterns. In our study, we adopt this approach to assess whether a CNN trained to detect SFs is learning physically plausible patterns, particularly in the context of training on non-global datasets.

In Sect. 2 the used data is discussed followed by Sect. 3 covering the reproduction of the SF detection by Niebler et al. (2022). We briefly discuss our use of LRP in Sect. 4, which is then applied to analyze which variables are required as input data. In Sect. 6, we discuss whether physically plausible patterns are used to detect SFs, followed by Sect. 7 discussing whether a CNN trained on regional data can be used outside of its training regions.

2 The atmospheric surface front dataset

We utilize data previously discussed by Niebler et al. (2022) and Justin et al. (2023). Geometry data of hand-analyzed SFs is sparsely available; Niebler et al. (2022) presented data by extracting geometry from published surface analysis charts for two regions: North America ([30°N, 75°N], [140°W, 55°W]) and Europe ([30°N, 75°N], [50°W, 40°E]). These hand-analyzed SFs are in the further manuscript referred to as "labels" or "labelled fronts" while the regions are referred to based on their data sources: DWD ("Deutscher Wetterdienst") for Europe (Niebler, 2021) and NWS ("National Weather Service") for North America (National Weather Service, 2019).

In the NWS region, SFs are classified as warm, cold, occluded, and stationary fronts. However, due to Niebler et al. (2022)'s extraction process, stationary fronts are absent in the DWD data (for details, we refer to Niebler et al., 2022). DWD data covers the years 2015 to 2019, whereas NWS data covers 2009 to 2018, both at six-hour intervals (00:00 UTC, 06:00 UTC, 12:00 UTC and 18:00 UTC). To balance the dataset, Niebler et al. (2022) restricted the NWS data to the years 2012–2017.

To evaluate the detection of SFs outside these regions, we use labels from Justin et al. (2023), which cover a different region extending from the Equator up to 80°N and from 130°E eastward to 10°E ([0°N, 75°N], [130°E, 10°E]). This region (NOAA, 2023) is referred to as "unified surface analysis domain" by Justin et al. (2023). In this study we abbreviate this name as "unified" region/domain. The hand-analyzed SFs encompass warm, cold, occluded, and stationary fronts and are available for the years 2008 to 2020.

For CNN input, we use ERA5 reanalysis data (Hersbach et al., 2020) on a regular $0.25^\circ \times 0.25^\circ$ latitude–longitude grid. Scripts to download the ERA5 data used in this study are available at <https://doi.org/10.5281/zenodo.21071307> (Radke et al., 2026). Temperature (t), specific humidity (q), the horizontal wind components (u ; v), and vertical velocity (w) are retrieved from the complete ERA5 global atmospheric reanalysis (Hersbach et al., 2017) using operational analyses (stream: oper, type: an). Surface pressure (p_s) is retrieved from ERA5 hourly data on single levels (Hersbach et al., 2023). The requests are global, and the regional domains are extracted during preprocessing. Following the L137 model-level definition and Niebler et al. (2022), we restrict the input to every fourth model level in the inclusive interval 105–137: 105, 109, 113, 117, 121, 125,



90 129, 133, and 137. These levels cover the atmosphere from the surface to approximately 700 hPa. High-altitude data is excluded as we do not expect meaningful information from these altitudes (also cf. Niebler et al., 2022). Surface pressure is first merged with the model-level variables using Climate Data Operators (CDO; Schulzweida, 2023), after which pressure at each model level is estimated using the corresponding level parameters α and β (ECMWF, 2016, 2026). Hence, the pressure p_k at model level k is calculated as:

95
$$p_k = \frac{(a_{k-1} + b_{k-1}p_s) + (a_{k+1} + b_{k+1}p_s)}{2}. \quad (1)$$

Each variable at a specific model level is referred to as an "input field". The full setup uses temperature, specific humidity, both horizontal wind components, vertical velocity, and derived pressure on all nine model levels. Because the physical zonal distance of a regular latitude–longitude grid decreases toward the poles, following Niebler et al. (2022), we include a 55th input field, *kmPerLon*. At latitude φ , this field represents the distance between adjacent grid points separated by 0.25° in longitude, 100 normalized by the equatorial distance between two grid points of 27.7762km . Under the spherical-Earth approximation,

$$\text{kmPerLon}(\varphi) = \cos \varphi. \quad (2)$$

The resulting field is merged with the model level variables as a single level input field.

To assess the CNNs ability to generalize to unseen data, following Niebler et al. (2022), the data is partitioned into training, validation, and test sets: data for the year 2016 serves as the test set; data for the year 2017 as validation set; and the remaining 105 years (2012–2015, 2018, and 2019) as the training set, totaling 8528, 1460, and 1464 time steps, respectively. Note that, following Niebler et al. (2022), January and February of 2015 are omitted as no DWD front data covering these months was published by Niebler et al. (2022). For each training step for each data sample one region is randomly selected, randomly cropped to a smaller (128×256) subregion and randomly horizontally mirrored. The cropping introduces variability to reduce overfitting (Shorten and Khoshgoftaar, 2019), while horizontal mirroring across the equator creates artificial samples that 110 resemble the Southern Hemisphere, which is otherwise not represented (Niebler et al., 2022).

All data is z-score normalized (LeCun et al., 2012). Each mean and standard deviation is calculated for each variable across all employed model levels using data from 2016. Z-score normalization reduces convergence time (LeCun et al., 2012) and helps a CNN to consider all input fields as "equally important" (Chase et al., 2022). To mitigate the "ignorant to zero" problem in LRP (Mamalakis et al., 2022; Radke et al., 2025), which occurs when normalized data is zero, we add a constant of 10 to all 115 input fields, following Radke et al. (2025). This adjustment ensures all data is positive, thereby avoiding issues associated with zero inputs.

3 Reproduction of Niebler et al. (2022) results

We first reproduce the results presented by Niebler et al. (2022). The CNN setup used by Niebler et al. (2022) transforms the previously described data grids into mutually exclusive binary grids of the same size as the inputs — one for each SF type and 120 one for the background (total of five classes). In these grids, a value of one indicates inclusion within the corresponding class, while zero indicates exclusion.

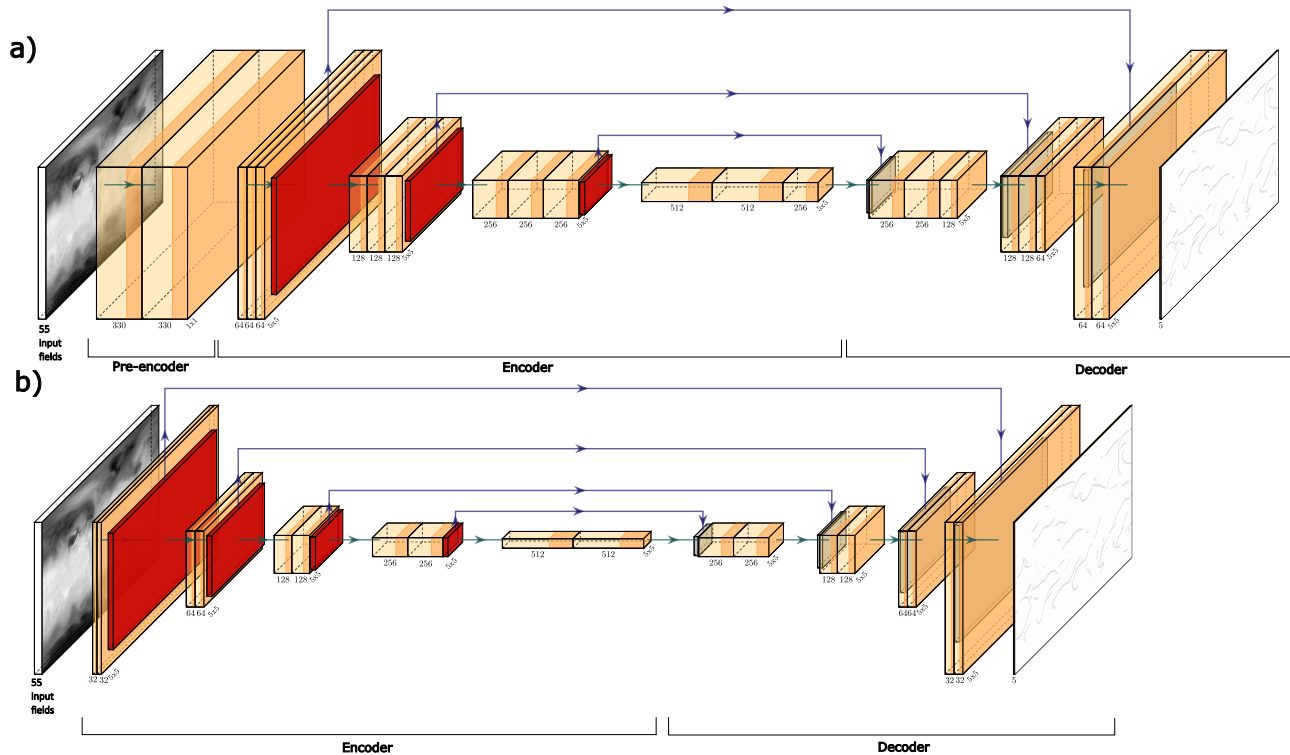


Figure 1. Illustrations of the U-Net architecture used by (a) Niebler et al. (2022) and (b) the simplified, reduced U-Net architecture used in the present study. Orange boxes represent convolutional layers, red boxes pooling layers, and gray boxes transposed convolutional layers. Arrows illustrate connections between layers.

Niebler et al. (2022) adopted a U-Net architecture (Ronneberger et al., 2015). Figure 1 (a) illustrates their architecture. It consists of three main components: a pre-encoder, an encoder, and a decoder. Their pre-encoder includes two convolutional layers that first expand the input grids to 330 "internal" grids (in the machine learning literature referred to as "feature" maps; Chollet, 2021) and subsequently reduce them to 64 "feature" maps. Note that these features are different from the atmospheric features discussed in this study and are abstract concepts like shapes, edges or objects learned by the CNN and used to then detect the atmospheric features. The encoder comprises a series of blocks that extract features from the data, with pooling operations reducing spatial dimensions at each stage to capture features at multiple scales. Following the encoder, the decoder uses expansive blocks that progressively scale the spatial dimensions back to their original extent. Skip connections (in the literature also referred to as residual connections; He et al., 2016) link encoder outputs directly to corresponding decoder layers at the same spatial scale. These connections enhance the network's ability to localize features by providing information from the encoder to the decoder, thereby improving detection accuracy (He et al., 2016).

Niebler et al. (2022) assessed the detection quality of their CNN using the critical success index (CSI), a metric also utilized for SF detection by Lagerquist et al. (2019). The CSI combines two metrics; the success rate (SR) — the proportion of labeled



135 SFs correctly detected — and the probability of detection (POD) — the proportion of CNN-detected SFs that correspond to labeled fronts. Both metrics require a matching process between detected and labeled SFs. We use the same matching algorithm as Niebler et al. (2022). First, eight-connected binary dilation (Serra, 1986) is applied to the binary grids connecting SF segments separated by single grid point wide gaps. Connected components extracted from these dilated grids represent individual SFs. These extracted SFs are filtered with the binary grids to remove the dilation effect. To reduce false positives, 140 SFs with two or fewer grid points are excluded. Matching between SFs is based on the spatial distance between their grid points, computed assuming a spherical Earth with an approximate diameter. Two SFs are considered a match if their average distance falls below a threshold. As threshold we follow Niebler et al. (2022) and use 250 km.

The metric POD is then computed from the total number of fronts detected by weather service experts n_{ws} and the number n_{mws} of these that could be matched against U-Net detected fronts n_d :

$$145 \quad POD = \frac{n_{mws}}{n_{ws}} \quad (3)$$

Similarly the metric SR is computed from the total number of U-Net detected fronts n_d and the number n_{md} of these that could be matched against the weather service fronts n_{ws} .

$$SR = \frac{n_{md}}{n_d}. \quad (4)$$

150 The CSI score, which balances these two metrics to prevent trivial maximization of either metric (e.g., predicting no SFs or all points as SF), is computed as:

$$CSI = \frac{1}{\frac{1}{POD} + \frac{1}{SR} - 1}. \quad (5)$$

Table 1. Evaluation of our U-Net architecture (R26) using 4, 6, 14, and 55 input fields, compared with the Niebler et al. (2022) U-Net architecture (N22) using 55 input fields. Evaluation is performed on the 2016 test dataset for the DWD and NWS regions. Scores are reported for each SF type and for three metrics: POD, SR, and CSI. Because the DWD dataset does not include stationary fronts, these are omitted from the DWD evaluation. Bold values indicate the highest score in each column.

Setup	Warm Front		Cold Front		Occluded		Stationary		Average
	DWD	NWS	DWD	NWS	DWD	NWS	DWD	NWS	
U-Net by N22, 55 input fields	49.2%	36.4%	56.1%	56.8%	52.4%	49.0%	–	43.2%	49.0%
U-Net by R26, 55 input fields	47.0%	36.1%	58.4%	57.9%	55.9%	48.6%	–	47.7%	50.2%
U-Net by R26, 14 input fields	50.0%	36.5%	53.0%	56.6%	56.6%	46.3%	–	45.6%	49.2%
U-Net by R26, 6 input fields	49.3%	36.7%	58.0%	57.1%	56.7%	47.4%	–	45.1%	50.0%
U-Net by R26, 4 input fields	45.0%	36.2%	52.3%	55.7%	55.8%	46.8%	–	45.8%	48.2%

Table 1 compares CSI scores for different SF types across DWD and NWS regions of Niebler et al. (2022) (cf. their Tables 4 and 5) with our reproductions. Niebler et al. (2022) argue that the CNN detects SFs more closely aligned with expert assessments than a selected rule-based system (cf. Niebler et al., 2022, their Table 6), suggesting the CNN's suitability for SF 155 detection similar to human experts.



When applying LRP, we encounter high computational costs especially relating to the required memory, especially the memory units of the used graphic cards (called VRAM). LRP requires all intermediate "feature" maps for calculation leading to large memory footprint. For example, detecting SFs across a global grid requires approximately 37.7 GiB of VRAM, with LRP demanding roughly twice that amount. As a comparison, typical currently available consumer graphic cards have between 8 and 16 GiB of VRAM. Even though one could swap these between RAM and VRAM to reduce VRAM usage, this transfer slows down processing (e.g., with data transfers, an LRP calculation takes approximately 600 ms with our reduced U-Net, compared to about 50 ms without data transfers). Hence, we experiment with a reduction in U-Nets size to allow much faster training and analysis while keeping the same performance scores.

This reduced architecture, shown in Fig. 1 (b), removes the pre-encoder, also not present in the original U-Net Ronneberger et al. (2015), one convolutional layer from each encoder and decoder block, and halves the number of neurons per layer. To compensate for these reductions, one additional encoder and decoder block is added at the smallest spatial scales, where memory demand is the lowest. These low-resolution blocks require approximately 0.01 GiB VRAM per convolutional layer, compared to 1.37 GiB needed by each convolutional layer of the pre-encoder. The changes are straight-forward to reduce the size of U-Net, however due to the large amount of possible changes to the architecture (e.g. number of neurons/number of layers/type of layers etc.) an exhaustive empirical study of all possible U-Net architecture variants is not possible. Hence other variants achieving similar results might exist. As shown in Table 1, detection performance of the reduced U-Net remains comparable with the larger architecture. Importantly, this architecture requires only approximately 8.7 GiB of VRAM for global SF detection with LRP still requiring about twice that amount. Consequently, this reduced architecture will be used throughout the remainder of this study.

4 Layerwise Relevance Propagation (LRP)

In this study, we utilize the LRP adoption developed by Radke et al. (2025), which builds upon the work of Bach et al. (2015) and Montavon et al. (2019). For a singular grid point and a designated class (e.g. cold front), LRP computes the contribution — or relevance — of each grid point of the CNN's input data towards the classification decision for the classified grid point. A relevance value is calculated for each input grid point and input field of the CNN. All relevance values of one input field are referred to as relevance map. To illustrate, Fig. 2 shows ERA5 data on 23 September 2016, 06:00 UTC, the date where extratropical cyclone "Vladiana" was located over the North Atlantic (e.g., Oertel et al., 2019; Choudhary and Voigt, 2022; Neuhauser et al., 2023; Beckert et al., 2023). Fig. 2 (a) shows specific humidity at model level 133 with detected SFs; (b) one of these relevance maps calculated for one selected grid point of the Vladiana cold front marked by an arrow and X.

Positive relevance values indicate a positive influence on the classification (pro argument), negative values indicate a negative influence (contra argument), and values close to zero indicate negligible or no influence. Fig. 2 (b) shows that the CNN considers only information close the classified grid point as important, with the drier air to the northwest as "pro argument" for classifying it as cold front and the moister air to the southeast of it as "contra argument".

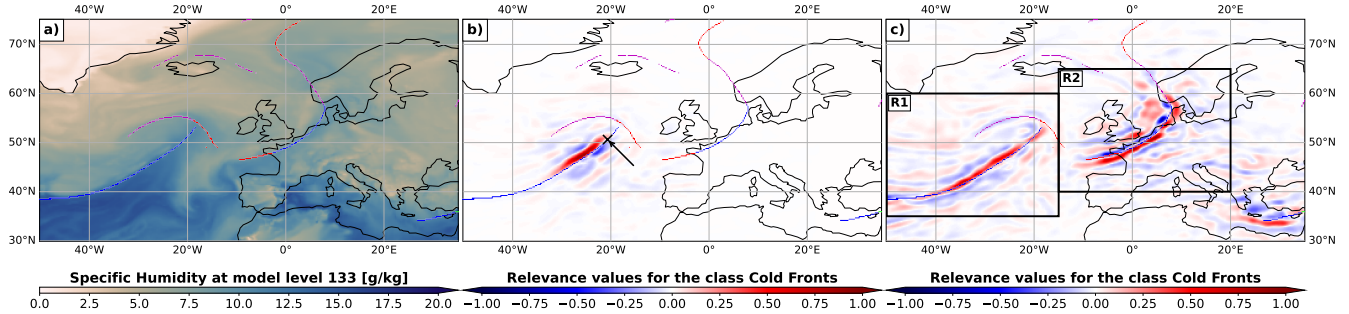


Figure 2. (a) ERA5, specific humidity data at model level 133 for 23 September 2016, 06:00 UTC; (b) corresponding relevance values calculated by LRP for one grid point (marked by X and arrow) of the class "cold fronts"; (c) corresponding relevance values calculated by LRP for the whole class "cold fronts". Red lines show warm, blue cold, purple occluded and green stationary fronts. Rectangles "R1" and "R2" show regions later discussed in Sect. 6 and Fig. 5.

Summing up relevance values for all gridpoints of singular fronts or all features of a selected class allows analysis of learned patterns for complete features. Fig. 2 (c) shows a resulting relevance map for specific humidity at model level 133 and the class cold fronts. To facilitate comparison across different features, Radke et al. (2025) proposed normalizing relevance maps on a per-feature basis; this approach is also adopted in our study.

5 Which meteorological variables are required as input to a CNN detecting SFs?

Analysing relevance patterns as shown in Fig. 2 for all input fields of the Niebler et al. (2022) setup would require analysing 55 times 4 input fields/SF type combinations of relevance maps at each timestep and would thus be very tedious and time consuming. Hence, as our first objective, we apply LRP to determine contribution of each input field and find a reduced set of input fields which achieves similar detection. A similar challenge was discussed by Radke et al. (2025) for the detection of tropical cyclones and atmospheric rivers, albeit for an originally smaller dataset (16 input fields and 2 classes). They find that their CNN performs similarly well when trained with only 4 instead of 16 input fields. To identify this subset, they train their CNN on all available input fields, calculate relevance values for all inputs and grid points and analyze the relevance distribution for the individual grid inputs. Then the inputs are ranked based on the summed relevance of all grid points > 0.1 . The CNN is subsequently retrained using only the most relevant input fields.

We apply the same method to our U-Net. Figure 3 shows the relevance value distributions for each input input field and each detected front type as stacked kernel-density-estimate (KDE)¹. Figure 3 shows that for all SF types most high ($>70\%$) relevance values are distributed over temperature and specific humidity from model levels 113 to 133. Additionally, the horizontal wind speed components are most relevant at model level 133. Another notable input field is the "kmPerLon" ratio, which is relevant

¹ Note that Radke et al. (2025) employs violin plots. However, as violin widths are scaled by density, the high-relevance tails are difficult to compare across 55 separate violins, and direct visual ranking becomes cumbersome. For reference Fig. A1 in the Appendix shows these violin plots for the 55 input fields.

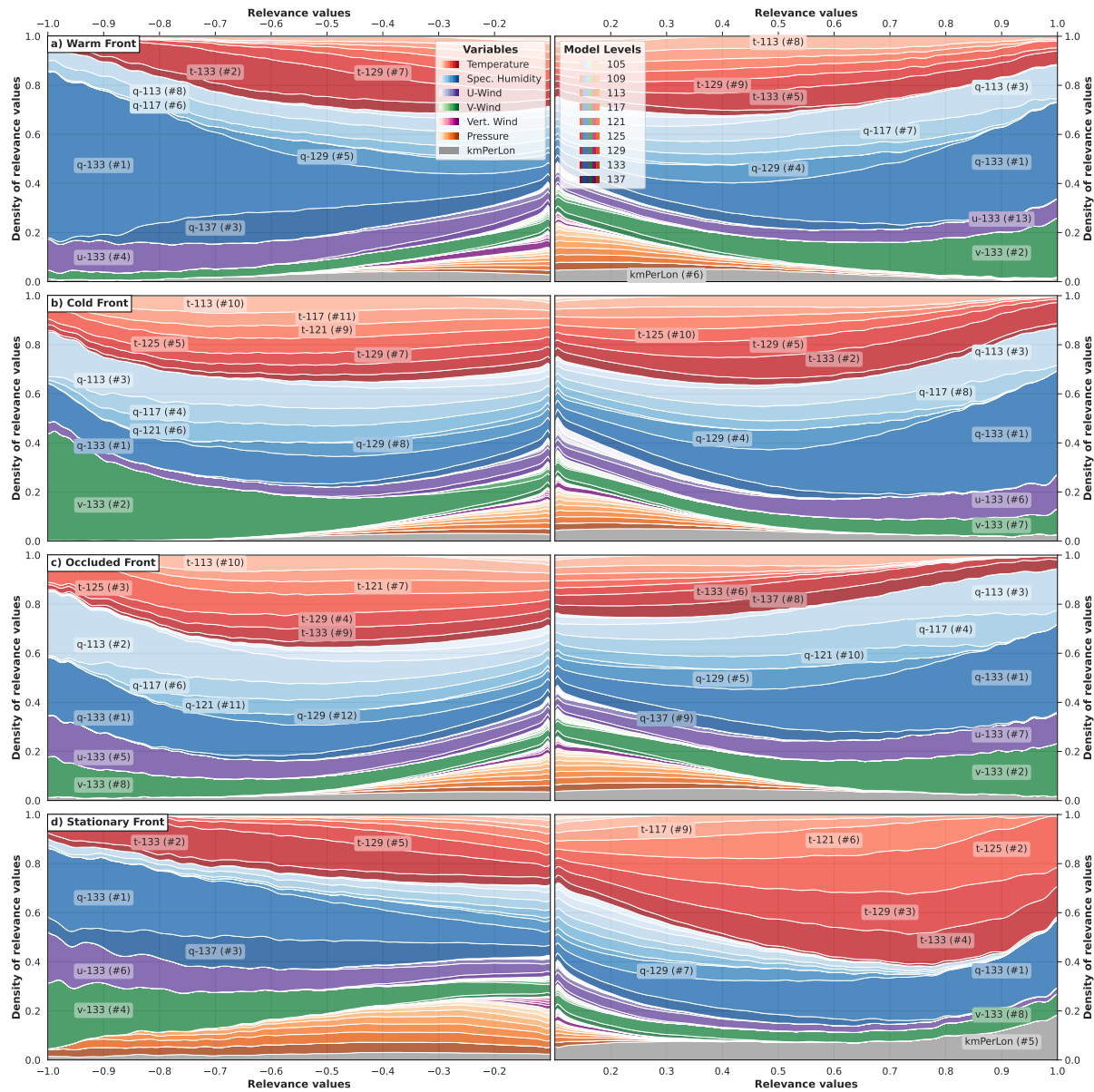


Figure 3. Stacked kernel-density-estimate plots showing the distribution of relevance values split by variable and SF type. Data samples cover the year 2016 in six hour steps taken from the ERA5 dataset. (a) shows distributions for warm, (b) for cold, (c) for occluded and (d) for stationary fronts. Different colors as seen in the legend of (a) show different variables, while shades show model levels. A rank indicating summed relevance is given for the most important variables.

for detecting stationary fronts. As it is less important for other SF types, it is not included in our data subset. Temperature and specific humidity on model level 113 to 133 as well as horizontal wind at model level 133 form a subset of 14 most relevant



input fields. Table 1 shows that this subset of 14 input fields achieves similar detection quality compared to U-Net trained with 55 input fields and is therefore suitable, even though further reduction might be possible.

210 To further reduce the set of input fields we make use of the observation that temperature and humidity at nearby model levels correlate, hence removing some of them seems plausible without losing much information for front detection. Similarly, Radke et al. (2025) discuss that while their CNN shows relevance for wind speeds at different heights using only one height is sufficient. Table 1 shows that training U-Net with temperature and specific humidity only on model level 113 and 133 (6 input fields) achieves similar CSI scores as using 55 input fields. A further reduction of input fields, achieved by removing
 215 temperature and specific humidity on model level 113, decreases the average CSI-score by 1.8% in comparison to U-Net trained on 6-input fields. Hence, we use the subset of six input fields in all further sections.

6 Are physically plausible patterns learnt, particularly when using regionally restricted (non-global) data?

As the second research objective we analyze whether the patterns learned by U-Net reflect the rules used by rule-based detection methods (e.g., Hewson, 1998; Simmonds et al., 2012; Kern et al., 2019; Beckert et al., 2023) or typical SF characteristics used
 220 by human forecasters (e.g., horizontal gradients in temperature, humidity, wind, strong vertical shear of the horizontal wind; and/or a relative minimum in pressure; e.g. Saucier, 1955; Martin, 2013; Stull, 2017).

For example, consider the approach introduced by Hewson (1998) and later extended by Hewson and Titley (2010), Kern et al. (2019) and Beckert et al. (2023). Their method detects SFs as the warm-air side of the frontal zone (cf. Hewson, 1998; Martin, 2013; Beckert et al., 2023) by computing contours where the second derivative of a thermal variable is zero. These
 225 contours are filtered according to the strength of the frontal zone and a thermal front parameter indicating the strength of the gradient region. Hence, here, we are interested whether U-Net also considers gradient regions in one of the input variables as relevant.

Figure 4 shows SFs provided by weather services, detected on 850 hPa by the Hewson (1998) method (here, we use the Beckert et al. (2023) adaptation using equivalent potential temperature as thermal variable and following their filter recom-
 230 mendations), and detected by our U-Net setup using six input fields. U-Net fronts are in most cases closely aligned to the weather service fronts which we consider plausible as those weather services also provided the SFs U-Net was trained on. The rule-based method detects mostly similar SFs with additional SFs being detected implying different patterns being used to detect these fronts compared to the U-Net setup as otherwise the same SFs would be detected by both methods. However, we also note that these rule-based fronts change heavily based on its filter settings.

235 To analyze the patterns learned by U-Net, Fig. 5 shows relevance patterns of Vladiana's cold front (cf. Fig. 2 R1) and the cold front east of it (cf. Fig. 2 R2, in the further manuscript referred to as eastern front). Here, we focus on specific humidity at model level 133. For a single grid point (cf. Fig. 2 (b)) relevance is distributed in an area of around 200km around the grid point with most in absolute terms high relevance values close to the front. Grid points further away from the front show "noise-like" relevance patterns. A similar relevance distribution is shown for the relevance map of all front grid points and also for both
 240 fronts (cf. Fig. 2 (c)). Similar noisy patterns are discussed in computer vision applications of LRP, where they indicate that the

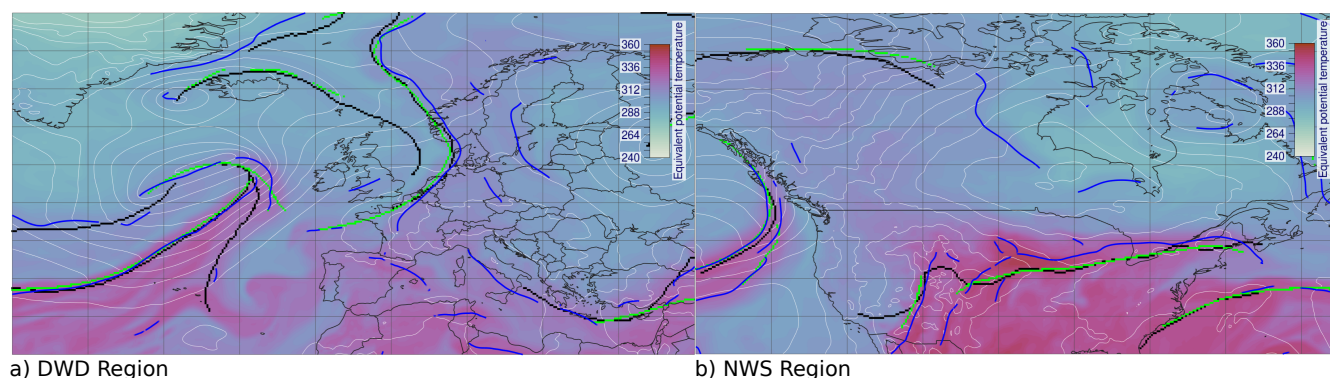


Figure 4. SFs detected in the DWD and NWS region using ERA5 data at 23 September 2016, 06:00 UTC. Black lines show fronts taken from the expert-labeled Niebler et al. (2022) dataset, green lines show fronts detected by U-Net and blue lines show fronts at 850hPa detected by the rule-based method by Beckert et al. (2023). Color shows equivalent potential temperature, white contour lines show sea level pressure in 40 hPa intervals.

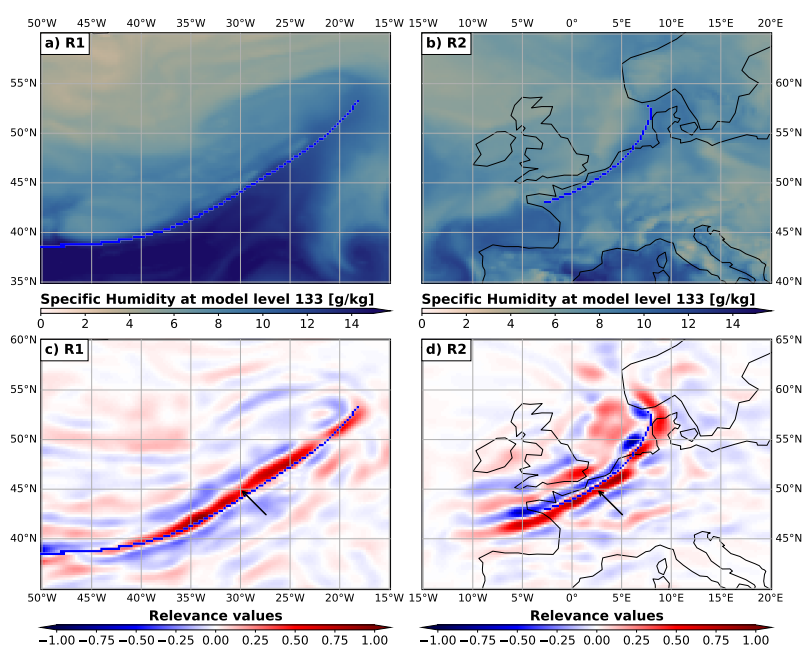


Figure 5. The detected Vladiana and eastern cold front, cf. Fig. 2 (c). (a-b), ERA5 specific humidity at model level 133 on the 23 September 2016, 06:00 UTC. (c-d), the corresponding relevance patterns. Arrows point to the positive relevance band following the front.

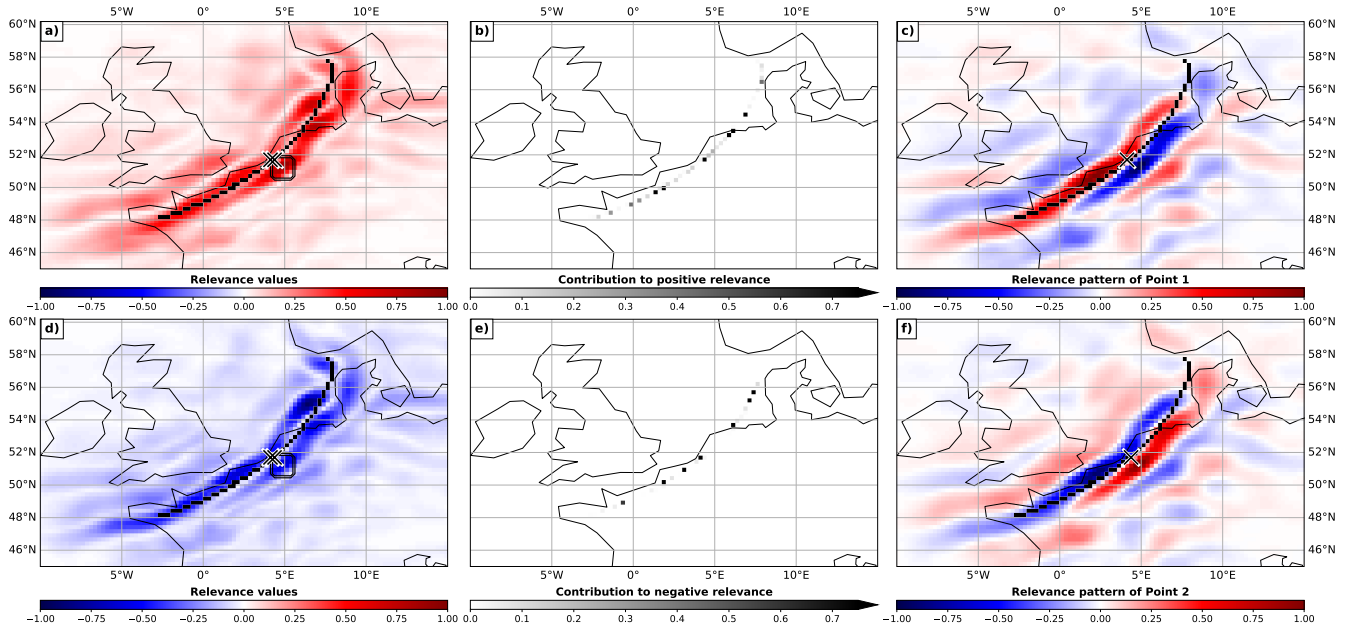


Figure 6. Relevance patterns for ERA5 specific humidity at model level 133 on the 23 September 2016, 06:00 UTC. (a) and (d) show average positive and negative relevance patterns for the cold front. One region (black rectangle) is selected with (b) and (e) visualizing how much each grid point of the front contributes to the positive and negative relevance in this region. Then two grid points of the front causing positive or negative relevance in the region are selected and their relevance patterns shown in (c) and (f).

CNN is learning spurious correlation between the background data and the features indicating low generalization (e.g. Bassi et al., 2024).

Close to Vladiana's cold front a band of positive relevance (marked by the arrow in Fig. 5 (c)) follows the front. The positive band is located at the cold/dry side of the front and embedded into two in absolute terms lower negative relevance bands. The relevance map for the R2 front shows the positive relevance band at the warm/humid side (marked by the arrow in Fig. 5 (d)). We consider it surprising that U-Net "looks at" different sides of the fronts as important for detection.

To further investigate this behaviour, we follow Radke et al. (2025) and split the relevance map into one only positive and one only negative relevance map. Figures 6 (a) and (d) show that both positive and negative relevance form bands on either side of the front. Since each input grid point can only contribute either positively or negatively to the detection of a given output grid point, U-Net must treat each band as "pro-argument" for some detected grid points and "contra-argument" for others. We examine this by selecting one sub-region, marked by the black rectangle in Fig. 6, and visualizing how much grid points contribute positive or negative relevance within this sub-region. Figures 6 (b) and (e) show that distinct parts of the SF account for negative and positive relevance, respectively. Moreover, Figures 6 (c) and (f) compare relevance patterns for two neighbouring grid points; one follows a pattern similar to the whole SF, whereas the other is nearly inverted. Thus, U-Net considers the same spatial region for detection but uses two distinct relevance patterns. We consider this implausible, as neither



a human expert nor a rule-based detection system would rely on two different patterns to classify neighbouring grid points as part of the same SF.

To gain a better understanding of this behaviour, we create a simplified synthetic 1-D dataset of a single variable mimicing the rising temperature or humidity between two air masses, typical for SFs. Figure 7 (a) shows a data sample of the synthetic dataset. The variable is defined as a function going from a unit-less scalar value of 5 ("cold/dry") to 15 ("warm/humid") following a sigmoid function (black line in Fig. 7 (a)). These values were chosen as they are close to the z-score normalized values + 10 used for the SF detection. The function is sampled at 96 points creating a 1-D data vector. As our U-Net setup has been designed for 2-D input fields, 48 copies of this sampled 1-D vector are stacked to create the input of U-Net. Following Hewson (1998), SF features are placed at the 0-contour of the 2nd derivative.

U-Net is trained in epochs. One epoch consists of 20000 training steps and 2000 evaluation steps, each using a newly generated data sample. To reduce overfitting, cold and warm side are varied between 4-6 (cold side) and 14-16 (warm side). Additionally, the data is randomly flipped vertically, and both position and steepness of the sigmoid function are randomly selected. Training is stopped once the valuation loss stagnates for 5 epochs. For evaluation, the Intersection over Union metric (IoU, Everingham et al., 2010) is used. IoU is a commonly used metric for semantic segmentation tasks (Cordts et al., 2016; Zhou et al., 2017; Abu Alhaija et al., 2018; Prabhat et al., 2021) which divides the intersection of labeled SFs and detected SFs by the union of both and is therefore in the range of zero to one, one indicating perfect detection and zero indicating no detection.

After training, U-Net reaches an IoU score of 93.4% indicating well detected SFs. Fig. 7 shows two nearly identical data samples (a) and (c) which differ only in the position of the sigmoid function; one is placed at 49.75 the other at 50.25. The corresponding relevance patterns, shown in Fig. 7 (b) and (d), resemble the previously discussed ones. One shows a positive relevance band on the warm side of the front with a negative band on the cold side, while the other is mostly inverted. Varying the data samples, indicates that the relevance pattern changes only when the position of the SF changes. As an example, Fig. 7 (e) shows a scatter plot where each point represents one data sample and the points are colored depending on the pattern used to detect the corresponding front. The relevance patterns are classified based on the distribution of relevance values from left to right; if the relevance minimum is further to the left of the relevance maximum it is classified as pattern 1, if it is further to the right it is clasified as pattern 2. The points are placed based on the center of the sigmoid function (x-axis), which are values between 47.5 and 52.5 with a step size of 0.01 and the location of the SF (y-axis), which can only be a natural number.

A stair case pattern emerges; moving the center of the sigmoid function also moves the front further to the right. The relevance patterns follow this "stair case", where the left side of each "step" shows mostly pattern 1 and the right side pattern 2. We hypothesize that the cause of these two patterns is the sampling of the 1-D function into a discrete space leading to U-Net learning two patterns as "easiest" solution to detect SFs. However, we cannot directly prove this hypothesis or determine the underlying reason why U-Net learned this pattern, because LRP reveals what a CNN has learned but does not explain the learning process that led to it. To the best of our knowledge, there is no straightforward method to understand the reason a CNN has learned a pattern.

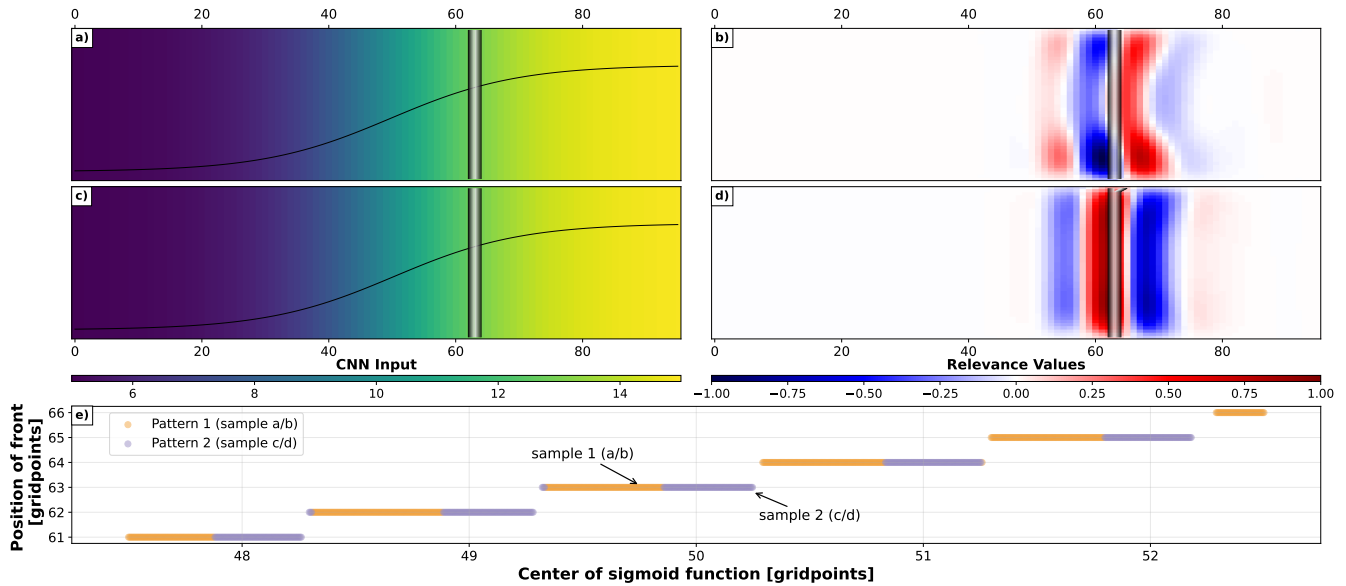


Figure 7. (a) and (c) show two samples of the synthetic dataset with the labelled front in gray. (b) and (d) show the corresponding SF detected by U-Net and the corresponding relevance patterns. In e) each point represents one data sample of the synthetic dataset, where we move the center of the sigmoid function from 47.5 to 52.5 in 0.01 steps as shown by the x-axis. As we move the center, the position of the SF also changes shown only the y-axis. The color of each point visualizes the relevance patterns used by U-Net. Purple corresponds to the pattern in (a) and (b) and orange to pattern the pattern in (c) and (d).

290 However, as a potential solution leading to a more plausible pattern, we propose to increase the resolution of the output data to allow a more precise placement of SFs in relation to the 1-D function. The input data stays the same, as in our real use case, using input data with a higher resolution is not feasible. As in our U-Net setup, the network expects its output to have the same spatial resolution as the output, its architecture is slightly changed by adding a final module containing an interpolation layer that upsamples the output to five times the resolution and two subsequent convolutional layers. Without these convolutional

295 layers, U-Net would not be able to learn different patterns, as it would just upsample the already known output. Everything else regarding the architecture or training stays the same. With this setup, U-Net achieves an IoU score of 87.8%. As shown in Fig. 8, only one pattern exists showing positive relevance close to the front and negative relevance further away. While the amount of negative relevance at both sides varies between different samples, the general pattern stays the same indicating that this method is suited to train U-Net to detect SFs in a much more plausible way on the synthetic dataset.

300 Now, it would be desirable to apply the same method to the to the dataset and training method of Niebler et al. (2022). However, placing the weather service SFs more precise in higher resolution is not easily possible. Assume the underlying data (e.g., the temperature field) to behave like the 1D-function of the simplified dataset. By placing fronts by hand, as done by the weather services, slight variations in placement between SFs will exist, even if the underlying data stays the same. Hence, even

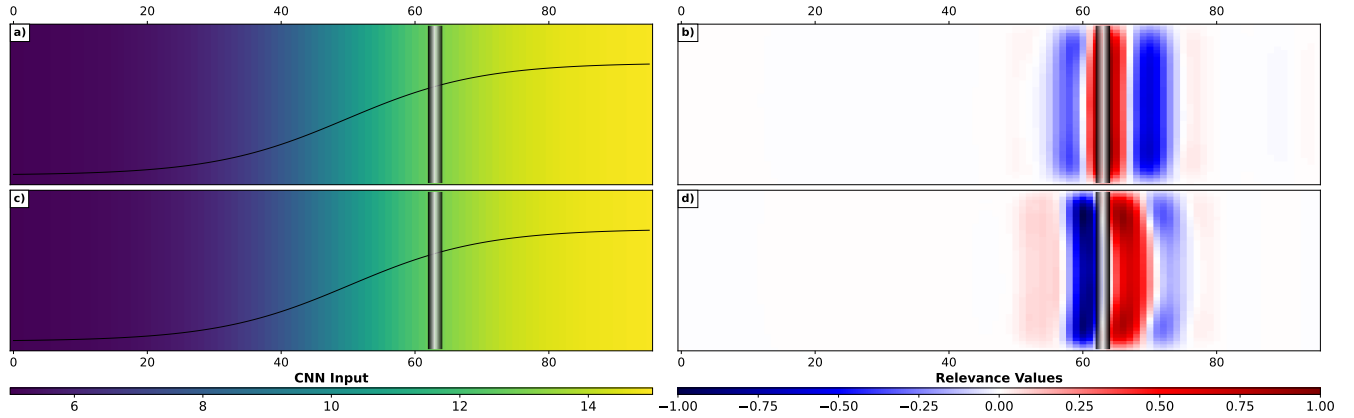


Figure 8. Same as Fig. 7 (a-d) but with U-Net trained on a five times larger output resolution. (a) and (c) show two samples of the synthetic dataset with the labelled front in gray. (b) and (d) the corresponding SF detected by U-Net and the corresponding relevance patterns.

in higher resolution U-Net learns two patterns to detect different parts of the SFs, depending on the position in relation to the frontal zone.

We outline another possible solution, which is out of scope for this study. Instead of directly using weather service provided SFs for training one could use SFs detected by a rule-based detection method, as these methods place SFs always in the same way without human variation. However, to also incorporate knowledge of human experts, we propose to filter the SFs detected by a rule-based method based with the weather service provided fronts to create a new training dataset. We hypothesize that training U-Net with this dataset will lead to U-Net learning a more plausible single pattern.

To assess how frequently this split relevance pattern occurs across the full dataset, we compute composites of relevance maps. Composite relevance maps were also used by Radke et al. (2025), who averaged relevance patterns for a CNN detecting tropical cyclones and atmospheric rivers. In their case, the relevance patterns were sufficiently similar to be averaged directly. However, as shown in Fig. 4, SFs differ in length and curvature, making a direct average over all relevance maps unsuitable for analysis.

To make the SFs comparable, we transform them into straight north–south oriented lines with a consistent direction of movement. This leads to both their relevance patterns and the underlying input fields being comparable across SFs. Fig. 9 illustrates the method.

In step (a), U-Net detects SFs, which are extracted as polylines in step (b). Each connected component is treated as one SF. The grid points, together with information on their directly neighbouring grid points, define the polyline. Two grid points are considered neighbours if they are connected vertically, horizontally, or diagonally.

In step (c), for each grid point $P(x_i, y_i)$ with two neighbouring grid points $Q(x_{i-1}, y_{i-1})$ and $R(x_{i+1}, y_{i+1})$, we approximate the local SF direction by the straight line connecting Q and R . The corresponding unit vector $\hat{i}$ is computed as

$$\hat{i} = \left(\frac{x_{i+1} - x_{i-1}}{\sqrt{(x_{i+1} - x_{i-1})^2 + (y_{i+1} - y_{i-1})^2}}, \frac{y_{i+1} - y_{i-1}}{\sqrt{(x_{i+1} - x_{i-1})^2 + (y_{i+1} - y_{i-1})^2}} \right) = (x_i, y_i). \quad (6)$$

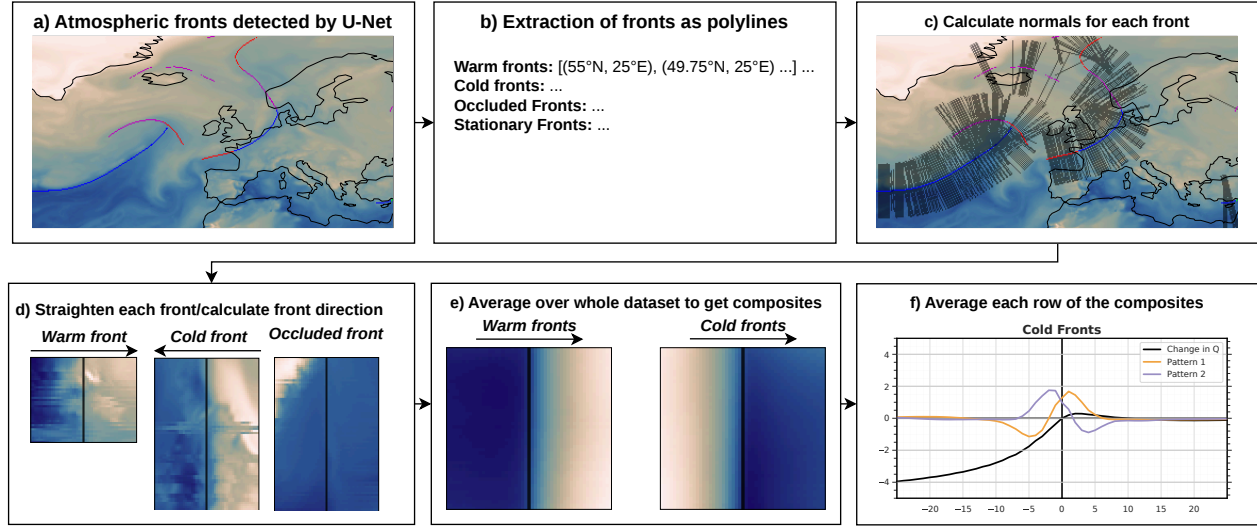


Figure 9. Illustration of our method to calculate composites of both relevance maps and input fields. The SFs (a) are transformed into polylines (b) which are used to calculate normals in both perpendicular directions (c). Then, relevance and input field values are used to "straighten" each front in (d) and calculate the direction of movement. These straightened SFs are averaged to produce composite maps (e). Finally, these composites are additionally averaged over all front grid points and along their corresponding normal directions leading to one-dimensional representations of relevance values and input values (f).

325 This vector is then rotated by 90° and -90° to obtain the two normal vectors $N_1(x'_1, y'_1)$ and $N_2(x'_2, y'_2)$:

$$N_1(x'_1, y'_1) = N_1(x_i \cdot \cos(90^\circ) - y_i \cdot \sin(90^\circ), x_i \cdot \sin(90^\circ) + y_i \cdot \cos(90^\circ)) = N_1(-y_i, x_i) \quad (7)$$

and

$$N_2(x'_2, y'_2) = N_2(x_i \cdot \cos(-90^\circ) - y_i \cdot \sin(-90^\circ), x_i \cdot \sin(-90^\circ) + y_i \cdot \cos(-90^\circ)) = N_2(y_i, -x_i). \quad (8)$$

Normals are then defined as lines extending from each front grid point in both normal-vector directions. We use a length of 25 grid points for each normal, as this covers the relevance patterns around each front.

Step (d) shows the resulting SFs transformed into straight lines. For the straightening, each front grid point and its corresponding normals are stacked from the start to the end of the front. Input fields and relevance maps are transformed according to the values along these normals. For the input fields, we use anomalies relative to the value at the front rather than absolute values. Otherwise, fronts in generally colder and warmer regions would cancel each other out in the composites.

335 Similarly, to prevent the warm/humid and cold/dry sides of different fronts from cancelling each other out, all straightened fronts are oriented consistently before averaging. We define the warm/humid side as the reference side. For each straightened front, the mean temperature and specific humidity on both sides are compared. If the warm/humid side is opposite to the reference side, the front is flipped.



In step (e), composites are computed by averaging all detected fronts in the full test dataset. Since the input fields and
 340 relevance patterns are, on average, assumed to be similar for every front grid point independent of position along the front,
 the composites are additionally averaged along the front (step (f)). This yields one-dimensional representations of both input
 variables and relevance values. As in the previous analysis, the composites are further separated by relevance pattern, based on
 the locations of the relevance maximum and minimum.

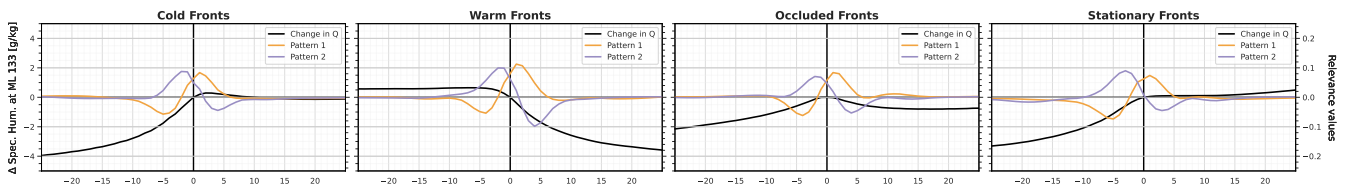


Figure 10. Composites calculated for the year 2016 at 6-hourly intervals, corresponding to the test dataset. Columns distinguish between front types. Black lines show the change in ERA5 specific humidity at model level 133 relative to the specific humidity at the front, while orange and purple lines show the relevance composites.

Figures 10 shows the composites for all four SF types for specific humidity at model level 133 and the two corresponding
 345 relevance patterns. One pattern has its positive relevance maximum on the warm/humid side of the frontal zone, similar to
 pattern 1 in the synthetic dataset (Fig. 7 (b)). The second pattern has its relevance maximum within the frontal zone and
 towards the cold/dry side. Both patterns occur with similar frequency, each accounting for approximately 50% of the cases. This
 is consistent with the synthetic dataset, where the two patterns show a similar ratio. Although both patterns assign relevance to
 parts of the SF that are also important for rule-based detection systems, namely the frontal zone and the warm side, they should
 350 not appear as separate patterns. In particular, the pattern that primarily focuses on the warm side of the SF could, on its own,
 lead to false detections, as weak gradients might be classified as SFs when the frontal zone itself is not considered important.

For completeness we also cover horizontal wind. Fig. 11 shows that in absolute terms lower relevance values indicate that
 U-Net uses wind to a lower extend than the other variables, consistent with the input field analysis in Sect. 5. Similar to before,
 "noisy" patterns are located around the front with a positive relevance band following the front. The positive band is found at
 355 a location of a strong wind shear, a characteristic of a SF Stull (2017). Hence, we consider this pattern as plausible.

Returning to the research question whether U-Net in our setup learns physically plausible patterns; LRP showed that gener-
 ally U-Net looks at changing humidity and temperature close to the SF which is plausible as this pattern characterizes SFs.
 Similarly, the wind shear in horizontal wind, another characteristic of SFs, is used by U-Net. But it was also shown that this
 relevance pattern is split into two different patterns, with one seeing the frontal zone as important to detect SFs and the other
 360 seeing the warm side as important. We hypothesise that especially the second pattern can lead to the overdetection of SFs in
 regions with only weak changes in humidity/temperature, which is discussed further in the following section. Furthermore,
 noisy relevance patterns are found in a distance of up to 200 km around the SFs. This phenomenon, also described in computer
 vision literature (e.g. Bassi et al., 2024), corresponds to overfitting and spurious correlations with the background data U-Net
 learned which we consider as not plausible.

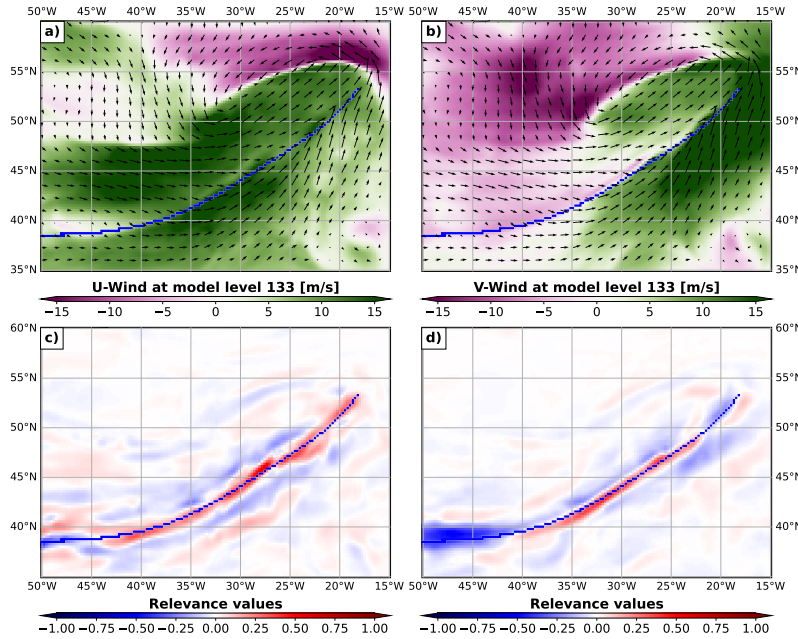


Figure 11. (a) and (b) ERA 5, u- and v-wind at model level 133 on the 23 September 2016, 06:00 UTC; (c) and (d) the corresponding relevance patterns used to detect the Vladiana cold front. Arrows show the horizontal wind.

365 7 Can a CNN trained on regional data be applied to regions outside the training domain?

Niebler et al. (2022) show that training U-Net on only one of their regions leads to lower CSI scores on the other when compared to training U-Net on both regions. Additionally, in Sect. 6, we show that U-Net learns relevance patterns which are only in part physically plausible. Hence, as our third research objective, U-Net is evaluated on a different region than the DWD and NWS region to investigate how well U-Net detects SFs outside of its training regions. This region, as discussed in Sect. 2, contains labeled SFs presented by Justin et al. (2023).

Due to potential differences between the labels used by Niebler et al. (2022) and Justin et al. (2023), U-Net might detect differently well depending on the label source. However, as shown in Table 2, evaluating our U-Net on the NWS region which is part of both the labels used by Niebler et al. (2022) and Justin et al. (2023) shows that independent of the label source CSI-scores are similar (average of 46.3% CSI score for both sources). Therefore, the labels of Justin et al. (2023) are also suitable to evaluate U-Net trained on the data by Niebler et al. (2022).

Evaluating U-Net on the whole "Unified" region shows a drop in both average CSI-score from 46.3% to 34.8% and CSI-scores for all SF classes. The biggest difference being the stationary fronts dropping from 46% to 21.4%. Different subregions of the unified region, split by latitude into "Arctic" (60°N to 80°N), "Midlatitude" (30°N to 60°N) and "Tropics" (0°N to 30°N), show largely differing detection quality based on the region. Table 2 shows these CSI-scores. While for the "Midlatitude" region, which also contains the two training regions, CSI-scores are even better than only for the NA region (average of



Table 2. CSI scores for all front types are given in percentages. In the DWD region of Niebler et al. (2022) stationary fronts are not explicitly labeled. Hence, they are not evaluated for the DWD region. The author names indicate the different label sources. In the first column we indicate the regions used for training. Additionally, the number of detected and labeled atmospheric fronts is shown in the two rightmost columns.

Region		CSI-Scores					Number of fronts	
Training	Evaluation	Warm	Cold	Occluded	Stationary	Average	Detected	Labeled
DWD and NWS	DWD (Niebler et al., 2022)	49.8	57.8	56.3	–	–	16,397	17,580
	NWS (Niebler et al., 2022)	34.3	56.9	45.7	47.2	46.0	19,333	17,411
	NWS (Justin et al., 2023)	36.2	57.6	45.4	46.0	46.3	–	–
	Unified (Justin et al., 2023)	32.4	43.7	41.8	21.4	34.8	58,617	37,442
	Arctic (Justin et al., 2023)	22.7	30.7	39.2	24.4	29.25	8,631	7,621
	Midlatitude (Justin et al., 2023)	44.8	65.7	55.2	40.1	51.5	35,229	30,873
	Tropics (Justin et al., 2023)	4.7	35.6	2.7	12.6	13.9	17,077	3,329
Unified	DWD (Niebler et al., 2022)	26.0	47.3	43.3	–	–	11,548	17,580
	NWS (Niebler et al., 2022)	36.4	56.2	46.9	45.8	46.3	19,301	17,411
	NWS (Justin et al., 2023)	–	–	–	–	–	–	–
	Unified (Justin et al., 2023)	39.5	56.3	52.4	37.1	46.3	47,298	37,442
	Arctic (Justin et al., 2023)	24.0	32.6	44.3	38.5	34.9	8,884	7,621
	Midlatitude (Justin et al., 2023)	47.5	68.3	60.4	48.6	56.2	35,398	30,873
	Tropics (Justin et al., 2023)	23.6	80.3	12.1	45.0	40.25	4,433	3,329

49.2%), both Arctic and Tropics regions show much lower CSI scores. In the Tropics this reduction is especially severe with warm and occluded fronts reaching only 3.3% and 1.5% respectively.

The amount of detected and labeled SFs for both Arctic and Tropics show that especially in the Tropics much more SFs are detected than labelled (3329 labelled compared to 17077 detected). To understand where these fronts are detected, Fig. 12 shows climatologies of the Tropics with both labels and detected SFs. Warm fronts are detected mostly in the North Pacific and Atlantic tropical cyclone basins. Additionally, both cold and stationary fronts are detected in a long band parallel to the equator and a lot of stationary fronts are detected at the western coast of Mexico.

Figure 13 shows detected fronts on 21 September 2016, 12:00 UTC, with these three anomalies visible together with the specific humidity field at model level 133. Temperature at model level 133 is shown in the Appendix Fig. A2. Additionally, tropical cyclones detected by Radke et al. (2025) are shown as orange contours. In comparison to the labelled fronts, neither anomaly corresponds to an expert labelled front. A part of the incorrectly detected SFs are detected inside of a tropical cyclones (Fig. 13, ○), which for U-Net might "look" similar to extratropical cyclones as both share rotating wind fields. However, as tropical cyclones are not connected to front systems, these fronts should not be detected. Similarly, the SFs parallel to the equator (Fig. 13, ★) are detected even though temperature and humidity do not show strong gradients hence would typically not be classified as SFs. This matches with the first pattern described in Sect. 6, where weak rising temperature/specific humidity close to the front is used for detection instead of the whole frontal zone.

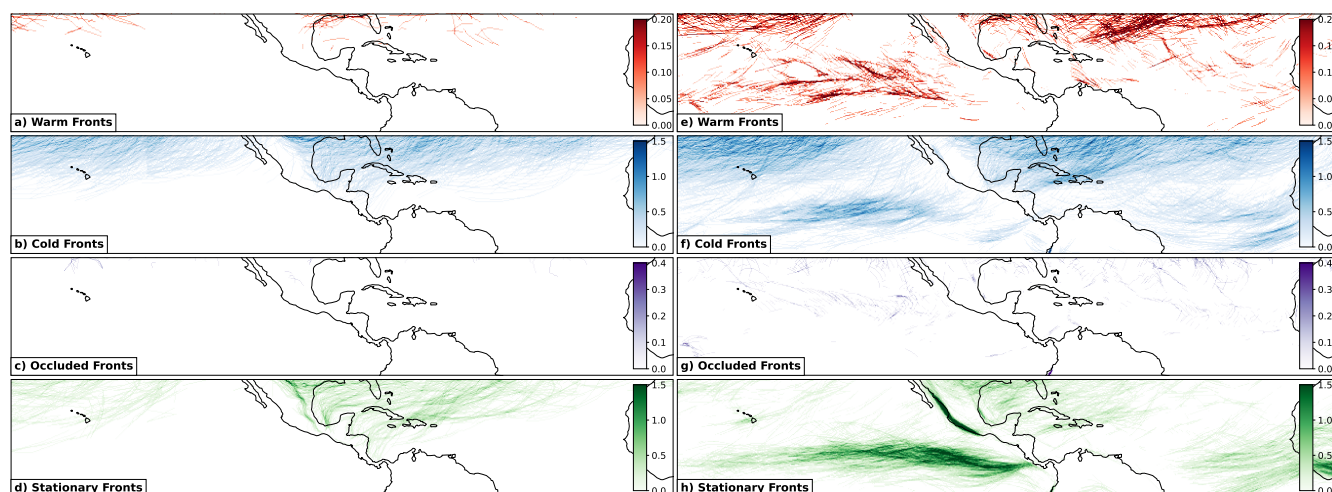


Figure 12. Climatologies of the four SF types for the "tropic" region. From a-d, they are calculated for the expert-labeled SFs, from e-h for the detected SFs. Climatologies are calculated for the year 2016 in six hour intervals and the colorbars show frequency of SF occurrence as a percentage of timesteps.

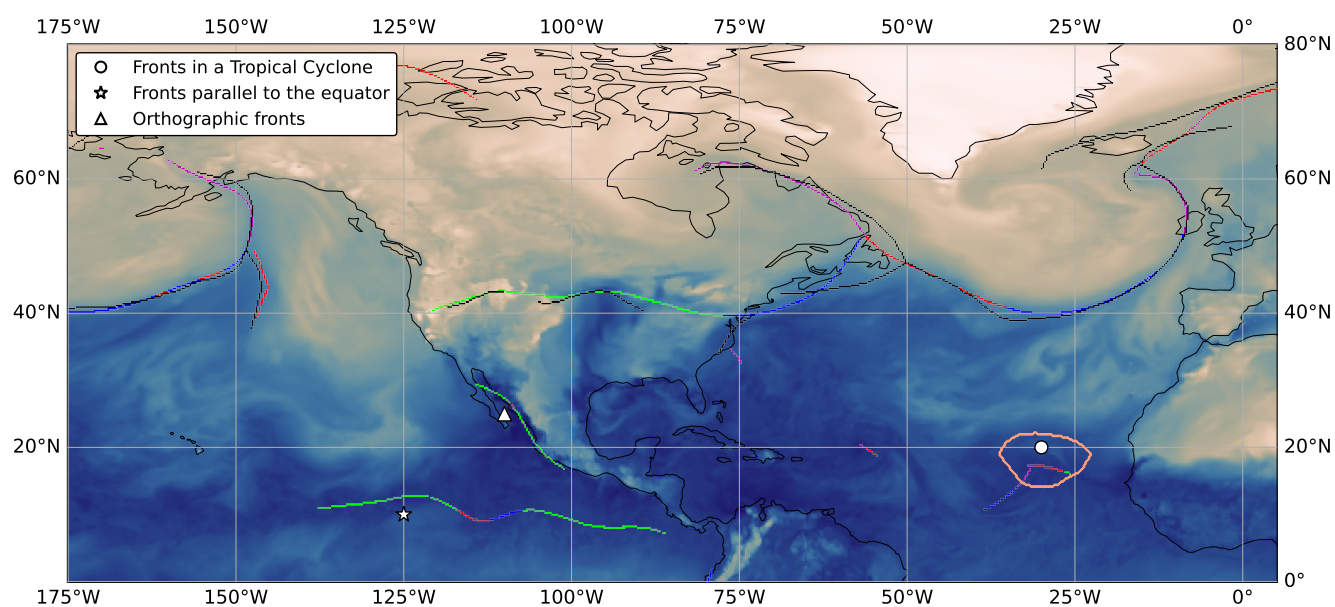


Figure 13. ERA5 specific humidity at model level 133 on the 21 September 2016, 12:00 UTC. Additionally tropical cyclones (orange outlines), SFs detected by U-Net (red - warm fronts, blue - cold fronts, purple - occluded fronts, green - stationary fronts) and SF detected by experts as provided by Justin et al. (2023) (black lines) are shown.

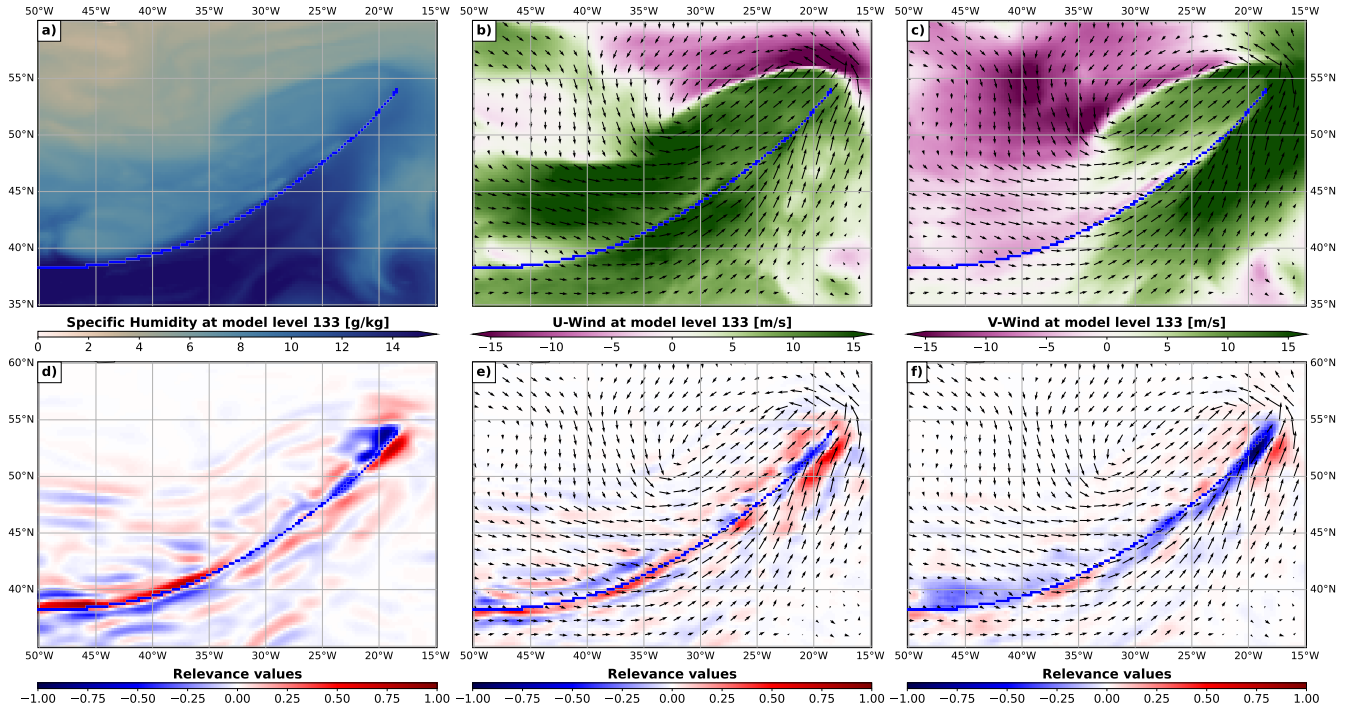


Figure 14. Same as Fig. 5 but with U-Net trained on the "Unified" region. The cold front of extratropical cyclone Vladiana is shown. Subplots (a-c) show ERA5 specific humidity, u and v wind, at model level 133 on the 23 September 2016, 06:00 UTC, while subfigures (d-f) show the corresponding relevance patterns used to detect this cold front. Arrows in (b/c/e/f) show horizontal wind.

The SFs close to the West coast of Mexico (Fig. 13, $\triangle$) show a steep change in humidity, so detecting them as SFs seems not far fetched. However, these are geographic fronts: in ERA5 model level data the model levels follow the terrain, leading to dry and cold air in mountainous regions in contrast to the moist, warm air over the Pacific directly neighbouring each other in the data. Hence, in these areas SFs are frequently detected by U-Net.

To improve the performance of U-Net especially regarding the implausibly detected SFs, we propose to keep the U-Net architecture with the training setup of Niebler et al. (2022) but train it with the labels and region of Justin et al. (2023). Table 2 shows resulting CSI scores for the test dataset. They show similar or improved CSI-scores across all regions, with the exception of the DWD region. For the DWD region we assume that SFs are detected worse, because the labels of the DWD do not contain stationary fronts, while the larger regions does contain them. Furthermore not all parts of the DWD region are covered by the "Unified" region. When not training U-Net with the DWD data it will detect stationary fronts over Europe not contained in the labels, hence leading to lower scores. We cannot argue whether this is correct or not, as for a direct comparison a European SF dataset containing all four types of SFs would be required, which to the best of our knowledge is not readily available.

As the detection performance has changed, relevance patterns have to have also changed. Fig. 14 shows specific humidity at model level 133 and both horizontal wind components together with the relevance patterns for detecting the shown cold



front. In general, relevance patterns are more concentrated on the region close to the SF, with less noise, indicating that U-Net learned fewer spurious correlations with the background data. Otherwise, relevance patterns for specific humidity and wind look similar. The same two patterns as discussed in Sect. 6 exist, which is not surprising as hand labeled data is still used even though from a different source.

415 To summarize, U-Net detects SFs poorly outside of the training regions especially in the Tropics and the Arctic, where it detect much more SFs than an expert would label. We attribute these erroneously detected SFs to the spurious correlations with the background data learned by U-Net as well as one of the two general patterns with U-Net looking at weak gradients in the data without considering the frontal zone much. Training U-Net on a larger domain reduces the spurious correlations and improves the detection quality outside of the smaller training regions used by Niebler et al. (2022). However training on the
420 larger domain cannot fix the previously discussed issue of the two patterns, as the labels are still placed by hand.

8 Conclusions

In this study, we applied the explainable artificial intelligence (xAI) method layer-wise relevance propagation (LRP) to the detection of atmospheric surface fronts (SFs). Building on previous studies by Niebler et al. (2022) and Justin et al. (2023), a reduced U-Net variant and a reduced set of input variables to detect SFs was evaluated. We found that a reduction in U-Nets size
425 as well as a reduction of the dataset from 55 to 6 input fields is possible without decreasing detection performance. With this architecture and dataset, the relevance patterns learned by U-Net were analyzed with both plausible and implausible patterns visible, leading to the creation and evaluation of a synthetic front dataset. Additionally, we evaluated U-Net on a larger region covering parts of the Tropics and Arctic. We then compared relevance patterns between training U-Net on two smaller regions covering the midlatitudes of northern America and western Europe and a larger one stretching from the Pacific over northern
430 America to western Europe. The main findings of our study are as follows:

- The detection of SFs as discussed by Niebler et al. (2022) was reproducible and a reduction of the size of U-Net is possible without decreasing the detection performance while reducing computational costs (cf. Sect. 3).
- In Sect. 5, using LRP, we identified a minimal subset of six input fields, temperature (model level 113 and 133), specific humidity (model level 113 and 133), and horizontal wind (model level 133) that maintained a similar detection
435 performance compared to the 55 input fields used by Niebler et al. (2022).
- LRP-Analysis of one selected SF case showed that U-Net uses both physically plausible and implausible patterns to detect SFs. We expected LRP to show U-Net using the frontal zone as positive criteria to detect the front, with each part of the front using similar local information of either rising temperature/humidity or wind shear. While most information is drawn from the temperature and humidity field in the frontal zone, which is similar to rule-based detection systems,
440 different grid points of the same SF showed two distinct patterns. One pattern shows U-Net "looking at" the frontal zone which we consider plausible while the other "looks" more at the warm side of the SFs, which we consider as less



plausible as it uses only weak gradients for detection. Additionally, noisy patterns in a large area around the front suggest the presence of spurious correlations with the background data limiting generalization.

– To investigate the two patterns, we proposed a synthetic dataset to shed light on these two patterns. In this synthetic dataset, U-Net also detects SFs using these two patterns, even though it changes only between different samples, not between different pixels of the same front. Further investigation showed depending on the underlying function and hence the underlying data, the relevance patterns change, even if the front is detected at the same location. We argue that U-Net learns these two patterns to "easily" distinguish fronts that are at the same location even though the background data changes, which applies to the real world data.

– Increasing the resolution of the output reduces the difference between fronts placed at the same location, as the fronts are now placed more discretely. Training U-Net on this higher resolution, we found it to learn only one pattern, which we consider as more plausible than two different patterns. Applying this method of increasing the output resolution on the dataset by Niebler et al. (2022) is challenging, as the handlabeled SFs are not placed in a consistent distance to the frontal zone (as we did in the synthetic data). Hence, we found that SFs will still be detected using two patterns.

– When applied outside the training regions, U-Net's detection performance declined, especially in the Arctic and Tropics where it often misidentified SFs. SFs are detected inside of tropical cyclones or in parallel to the equator where the implausible detection patterns cause U-Net to detect erroneous fronts.

– Retraining U-Net on a larger region improved detection skill. Relevance patterns show reduced spurious correlations.

For the future, we outline methods to improve U-Net to detect SFs more reliably and physically plausible. As training U-Net on a larger region already improved detection quality on this region, we expect similar improvements for other regions. Hence, extending the dataset should be the first priority, however the creation of curated handlabeled data sets is challenging. Instead one could use rule-based detection systems to create a global dataset to train U-Net. To not fully lose the advantages of hand-labelled data, one could also combine both, by for example filtering the rule-based detected fronts using hand-labeled fronts, in regions where hand-labelled data is available. A dataset created by a rule-based system might also be able to avoid the two patterns discussed earlier and allow U-Net to only learn one plausible pattern.

We conclude that the U-Net setup used in this study, trained on two regions, detects SFs with only partially physically plausible patterns and should therefore be used with caution. Outside the training regions, particularly in the Tropics, the model substantially over-detects SFs compared to expert analyses. Similar over-detection occurs at coasts and mountain ranges outside the training domain, indicating that U-Net partly relies on region-specific or spurious correlations rather than generally valid frontal characteristics. Consequently, applications beyond the training regions should account for an increased false-detection rate, and depending on the intended use, may require additional filtering or retraining with more representative data. Nevertheless, within the training regions, we consider the investigated U-Net setup a useful tool for rapid SF detection. With further (global) training data available, the setup can be further improved.



475 *Code and data availability.* The code used to generate the results presented in this paper is available at <https://doi.org/10.5281/zenodo.21071307> (Radke et al., 2026). This archive also contains the ERA5 retrieval scripts. ERA5 model-level data are available from the Complete ERA5 global atmospheric reanalysis at <https://doi.org/10.24381/cds.143582cf> (Hersbach et al., 2017) and surface pressure from ERA5 hourly data on single levels at <https://doi.org/10.24381/cds.adbb2d47> (Hersbach et al., 2023). The used NWS SF data are available at <https://doi.org/10.5281/zenodo.2642801> (National Weather Service, 2019). The polylines of the DWD data are available at <https://doi.org/10.5281/zenodo.5785816> (Niebler, 2021). SFs for the "Unified" domain are available at <https://doi.org/10.5281/zenodo.7505022> (NOAA, 2023).

480 *Author contributions.* TR worked on conceptualization, data curation, formal analysis, and investigation; applied the LRP code; performed the CNN training; conducted the relevance analysis; and wrote all sections. JB co-supervised the study, contributed to the general discussion, conceptualization and administration of the study. MR primarily supervised the study, lead the general discussion, conceptualization and administration of this study.

Competing interests. The authors declare that they have no conflict of interest.

485 *Acknowledgements.* Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2037 'CLICCS - Climate, Climatic Change, and Society' – Project Number: 390683824. Contribution to the 'Earth and Society Research Hub' (ESRAH) of University of Hamburg.

Appendix A

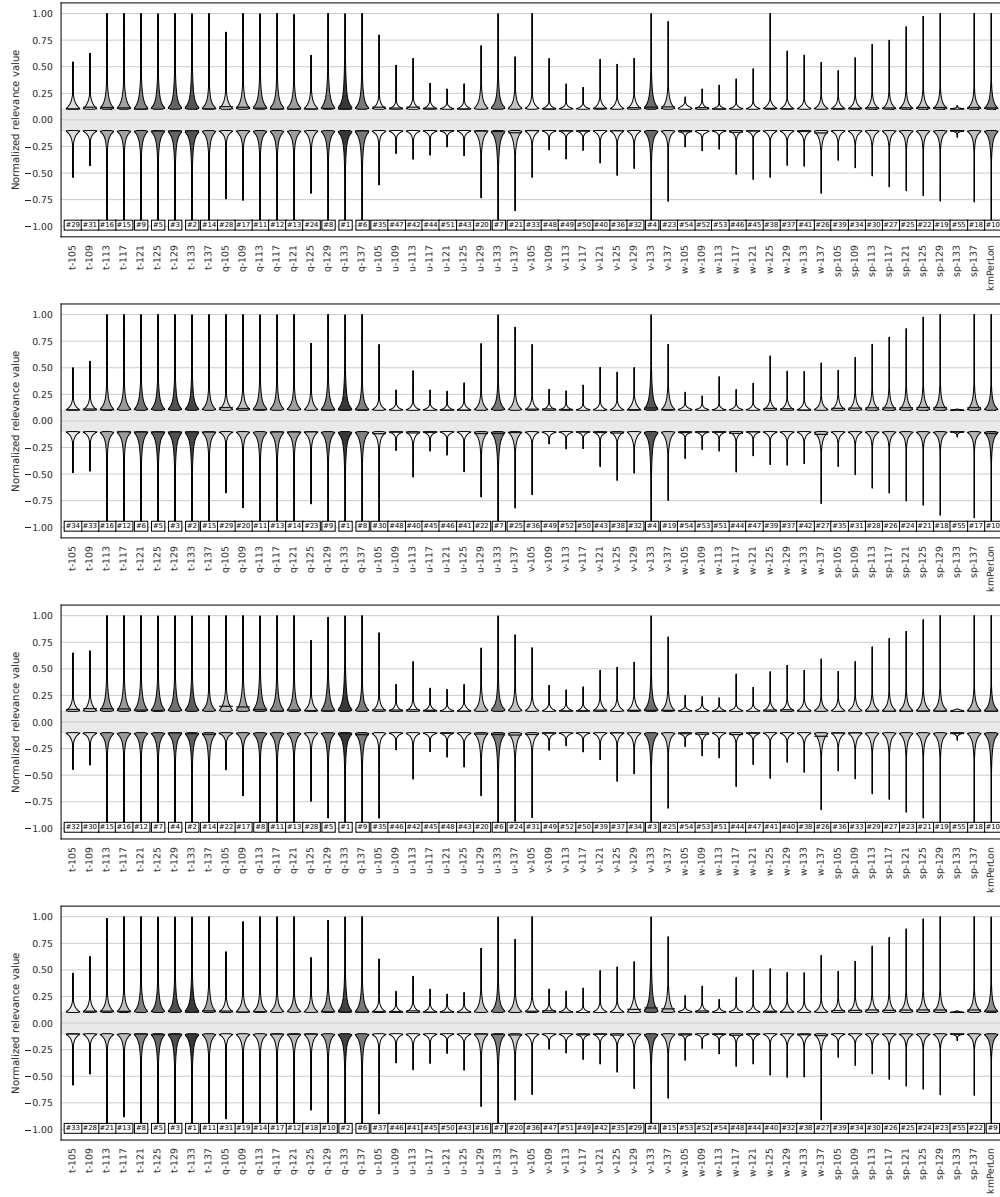


Figure A1. Cf. Radke et al. (2025) Fig. 10. Distributions of relevance values (computed for data taken from the ERA5 dataset from 2016 in six hour intervals) for U-Net trained with all 55 variables contained in the dataset. Distribution are split for warm, cold, occluded and stationary fronts. Width of violin plots is differently scaled between front types but consistent for all variables within each subplot. Relevance values in the range $[-0.1...0.1]$ are omitted. Numbers at the bottom as well as grey shade indicate ranking in terms of numbers of grid points with absolute relevance > 0.1 .

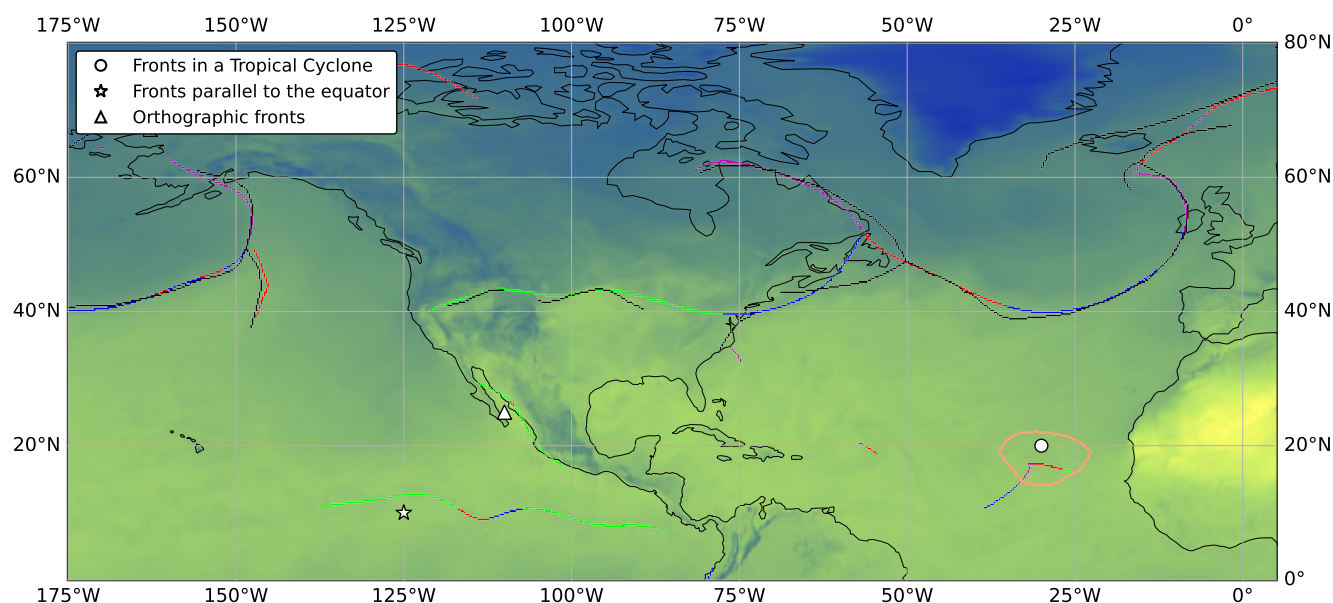


Figure A2. Same as Fig. A2 but for temperature at model level 133. ERA5 data on the 21 September 2016, 12:00 UTC is used. Additionally tropical cyclones (orange outlines), SFs detected by U-Net (red - warm fronts, blue - cold fronts, purple - occluded fronts, green - stationary fronts) and by experts as provided by Justin et al. (2023) (black lines) are shown.

References

- 490 Abu Alhaija, H., Mustikovela, S. K., Mescheder, L., Geiger, A., and Rother, C.: Augmented Reality Meets Computer Vision: Efficient Data Generation for Urban Driving Scenes, *International Journal of Computer Vision*, 126, 961–972, <https://doi.org/10.1007/s11263-018-1070-x>, 2018.
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L.: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, *Journal of Big Data*, 8, 53, <https://doi.org/10.1186/s40537-021-00444-8>, 2021.
- 495 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, e0130140, <https://doi.org/10.1371/journal.pone.0130140>, 2015.
- Bassi, P. R. A. S., Dertkigil, S. S. J., and Cavalli, A.: Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization, *Nature Communications*, 15, 291, <https://doi.org/10.1038/s41467-023-44371-z>, publisher: Nature Publishing Group, 2024.
- 500 Beckert, A. A., Eisenstein, L., Oertel, A., Hewson, T., Craig, G. C., and Rautenhaus, M.: The three-dimensional structure of fronts in mid-latitude weather systems in numerical weather prediction models, *Geoscientific Model Development*, 16, 4427–4450, <https://doi.org/10.5194/gmd-16-4427-2023>, 2023.
- Biard, J. C. and Kunkel, K. E.: Automated detection of weather fronts using a deep learning neural network, *Advances in Statistical Climatology, Meteorology and Oceanography*, 5, 147–160, <https://doi.org/10.5194/ascmo-5-147-2019>, 2019.
- 505



- Catto, J. L. and Pfahl, S.: The importance of fronts for extreme precipitation, *Journal of Geophysical Research: Atmospheres*, 118, 10,791–10,801, <https://doi.org/10.1002/jgrd.50852>, _eprint: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/jgrd.50852>, 2013.
- Catto, J. L., Jakob, C., Berry, G., and Nicholls, N.: Relating global precipitation to atmospheric fronts, *Geophysical Research Letters*, 39, <https://doi.org/10.1029/2012GL051736>, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2012GL051736>, 2012.
- 510 Catto, J. L., Nicholls, N., Jakob, C., and Shelton, K. L.: Atmospheric fronts in current and future climates, *Geophysical Research Letters*, 41, 7642–7650, <https://doi.org/10.1002/2014GL061943>, _eprint: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014GL061943>, 2014.
- Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., and McGovern, A.: A Machine Learning Tutorial for Operational Meteorology. Part I: Traditional Machine Learning, *Weather and Forecasting*, 37, 1509–1529, <https://doi.org/10.1175/WAF-D-22-0070.1>, 2022.
- 515 Chollet, F.: *Deep Learning with Python*, Manning, 3 edn., ISBN 978-1-61729-686-4, 2021.
- Choudhary, A. and Voigt, A.: Impact of grid spacing, convective parameterization and cloud microphysics in ICON simulations of a warm conveyor belt, *Weather and Climate Dynamics*, 3, 1199–1214, <https://doi.org/10.5194/wcd-3-1199-2022>, 2022.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223, <https://doi.org/10.1109/CVPR.2016.350>, iSSN: 1063-6919, 2016.
- 520 ECMWF: IFS Documentation CY41R2, Part III: Dynamics and Numerical Procedures, European Centre for Medium-Range Weather Forecasts, <https://doi.org/10.21957/83wouv80>, 2016.
- ECMWF: L137 model level definitions, ECMWF Confluence Wiki, <https://confluence.ecmwf.int/spaces/UDOC/pages/108117123/L137+model+level+definitions>, last updated 20 April 2026; accessed 20 October 2026, 2026.
- 525 Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge, *International Journal of Computer Vision*, 88, 303–338, <https://doi.org/10.1007/s11263-009-0275-4>, 2010.
- Foss, M., Chou, S. C., and Seluchi, M. E.: Interaction of cold fronts with the Brazilian Plateau: a climatological analysis, *International Journal of Climatology*, 37, 3644–3659, <https://doi.org/10.1002/joc.4945>, _eprint: <https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.4945>, 2017.
- 530 Goodfellow, I., Bengio, Y., and Courville, A.: *Deep Learning*, MIT Press, 2016.
- He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, IEEE, Las Vegas, NV, USA, ISBN 978-1-4673-8851-1, <https://doi.org/10.1109/CVPR.2016.90>, 2016.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simons, 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.-N.: Complete ERA5 from 1940: Fifth generation of ECMWF atmospheric reanalyses of the global climate, <https://doi.org/10.24381/cds.143582cf>, data set, 2017.
- 535 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simons, 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., Vil-



- laume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803>, 2020.
- Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, <https://doi.org/10.24381/cds.adbb2d47>, data set, 2023.
- Hewson, T. D.: Objective identification of frontal wave cyclones, *Meteorological Applications*, 4, 311–315, <https://doi.org/10.1017/S135048279700073X>, _eprint: <https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1017/S135048279700073X>, 1997.
- Hewson, T. D.: Objective fronts, *Meteorological Applications*, 5, 37–65, <https://doi.org/10.1017/S1350482798000553>, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1017/S1350482798000553>, 1998.
- Hewson, T. D. and Titley, H. A.: Objective identification, typing and tracking of the complete life-cycles of cyclonic features at high spatial resolution, *Meteorological Applications*, 17, 355–381, <https://doi.org/10.1002/met.204>, _eprint: <https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.204>, 2010.
- 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., *Lecture Notes in Computer Science*, 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.
- Hope, P., Keay, K., Pook, M., Catto, J., Simmonds, I., Mills, G., McIntosh, P., Risbey, J., and Berry, G.: A Comparison of Automated Methods of Front Recognition for Climate Studies: A Case Study in Southwest Western Australia, *Monthly Weather Review*, 142, 343–363, <https://doi.org/10.1175/MWR-D-12-00252.1>, 2014.
- Justin, A. D., Willingham, C., McGovern, A., and Allen, J. T.: Toward Operational Real-Time Identification of Frontal Boundaries Using Machine Learning, *Artificial Intelligence for the Earth Systems*, 2, <https://doi.org/10.1175/AIES-D-22-0052.1>, 2023.
- Kern, M., Hewson, T., Schätler, A., Westermann, R., and Rautenhaus, M.: Interactive 3D Visual Analysis of Atmospheric Fronts, *IEEE Transactions on Visualization and Computer Graphics*, 25, 1080–1090, <https://doi.org/10.1109/TVCG.2018.2864806>, 2019.
- Lagerquist, R., McGovern, A., and Ii, D. J. G.: Deep Learning for Spatially Explicit Prediction of Synoptic-Scale Fronts, *Weather and Forecasting*, 34, 1137–1160, <https://doi.org/10.1175/WAF-D-18-0183.1>, 2019.
- LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R.: Efficient BackProp, in: *Neural Networks: Tricks of the Trade: Second Edition*, edited by Montavon, G., Orr, G. B., and Müller, K.-R., *Lecture Notes in Computer Science*, pp. 9–48, Springer, Berlin, Heidelberg, ISBN 978-3-642-35289-8, https://doi.org/10.1007/978-3-642-35289-8_3, 2012.
- Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S.: Explainable AI: A Review of Machine Learning Interpretability Methods, *Entropy*, 23, 18, <https://doi.org/10.3390/e23010018>, number: 1 Publisher: Multidisciplinary Digital Publishing Institute, 2021.
- Liu, X., Deng, Z., and Yang, Y.: Recent progress in semantic image segmentation, *Artificial Intelligence Review*, 52, 1089–1106, <https://doi.org/10.1007/s10462-018-9641-3>, 2019.
- Long, J., Shelhamer, E., and Darrell, T.: Fully convolutional networks for semantic segmentation, in: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, <https://doi.org/10.1109/CVPR.2015.7298965>, ISSN: 1063-6919, 2015.
- 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>, 2022.



- Manakitsa, N., Maraslidis, G. S., Moysis, L., and Fragulis, G. F.: A Review of Machine Learning and Deep Learning for Object Detection, Semantic Segmentation, and Human Action Recognition in Machine and Robotic Vision, *Technologies*, 12, 15, <https://doi.org/10.3390/technologies12020015>, 2024.
- 585 Martin, J. E.: *Mid-Latitude Atmospheric Dynamics: A First Course*, John Wiley & Sons, ISBN 978-1-118-68789-5, google-Books-ID: g_aHATw2d1cC, 2013.
- Mersha, M., Lam, K., Wood, J., AlShami, A. K., and Kalita, J.: Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction, *Neurocomputing*, 599, 128 111, <https://doi.org/10.1016/j.neucom.2024.128111>, 2024.
- 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., *Lecture Notes in Computer Science*, 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.
- 590 National Weather Service: National Weather Service Coded Surface Bulletins, 2003-, <https://doi.org/10.5281/zenodo.2642801>, 2019.
- Naud, C. M., Posselt, D. J., and Heever, S. C. v. d.: A CloudSat–CALIPSO View of Cloud and Precipitation Properties across Cold Fronts over the Global Oceans, *Journal of Climate*, 28, 6743–6762, <https://doi.org/10.1175/JCLI-D-15-0052.1>, 2015.
- 595 Neuhauser, C., Hieronymus, M., Kern, M., Rautenhaus, M., Oertel, A., and Westermann, R.: Visual analysis of model parameter sensitivities along warm conveyor belt trajectories using Met.3D (1.6.0-multivar1), *Geoscientific Model Development*, 16, 4617–4638, <https://doi.org/10.5194/gmd-16-4617-2023>, 2023.
- Niebler, S.: Front polylines extracted from DWD Maps, <https://doi.org/10.5281/zenodo.5785817>, 2021.
- 600 Niebler, S., Miltenberger, A., Schmidt, B., and Spichtinger, P.: Automated detection and classification of synoptic-scale fronts from atmospheric data grids, *Weather and Climate Dynamics*, 3, 113–137, <https://doi.org/10.5194/wcd-3-113-2022>, 2022.
- NOAA: NOAA Unified Surface Analysis Fronts, <https://doi.org/10.5281/zenodo.7505022>, 2023.
- Oertel, A., Boettcher, M., Joos, H., Sprenger, M., Konow, H., Hagen, M., and Wernli, H.: Convective activity in an extratropical cyclone and its warm conveyor belt – a case-study combining observations and a convection-permitting model simulation, *Quarterly Journal of the Royal Meteorological Society*, 145, 1406–1426, <https://doi.org/10.1002/qj.3500>, 2019.
- 605 Parfitt, R., Czaja, A., and Seo, H.: A simple diagnostic for the detection of atmospheric fronts, *Geophysical Research Letters*, 44, 4351–4358, <https://doi.org/10.1002/2017GL073662>, eprint: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017GL073662>, 2017.
- Prabhat, Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., Karaismailoglu, E., von Kleist, L., Kurth, T., Greiner, A., Mahesh, A., Yang, K., Lewis, C., Chen, J., Lou, A., Chandran, S., Toms, B., Chapman, W., Dagon, K., Shields, C. A., O'Brien, T., Wehner, M., and Collins, W.: ClimateNet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather, *Geoscientific Model Development*, 14, 107–124, <https://doi.org/10.5194/gmd-14-107-2021>, publisher: Copernicus GmbH, 2021.
- 610 Radke, T., Fuchs, S., Wilms, C., Polkova, I., and Rautenhaus, M.: Explaining neural networks for detection of tropical cyclones and atmospheric rivers in gridded atmospheric simulation data, *Geoscientific Model Development*, 18, 1017–1039, <https://doi.org/10.5194/gmd-18-1017-2025>, 2025.
- Radke, T., Baehr, J., and Rautenhaus, M.: Code for: "Explaining patterns learned by neural networks for detection of atmospheric surface fronts in gridded atmospheric data", <https://doi.org/10.5281/zenodo.21071307>, 2026.



- Ribeiro, B. Z., Seluchi, M. E., and Chou, S. C.: Synoptic climatology of warm fronts in Southeastern South America, *International Journal of Climatology*, 36, 644–655, <https://doi.org/10.1002/joc.4373>, [_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.4373](https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.4373),
620 2016.
- Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., *Lecture Notes in Computer Science*, pp. 234–241, Springer International Publishing, Cham, ISBN 978-3-319-24574-4, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.
- 625 Saucier, W. J.: *Principles of Meteorological Analysis*, http://archive.org/details/principlesofmete0000walt_t6v2, 1955.
- Schulzweida, U.: CDO User Guide, <https://doi.org/10.5281/zenodo.10020800>, 2023.
- Serra, J.: Introduction to mathematical morphology, *Computer Vision, Graphics, and Image Processing*, 35, 283–305, [https://doi.org/10.1016/0734-189X\(86\)90002-2](https://doi.org/10.1016/0734-189X(86)90002-2), 1986.
- Shapiro, M. A. and Keyser, D.: Fronts, Jet Streams and the Tropopause, in: *Extratropical Cyclones: The Erik Palmén Memorial Volume*,
630 edited by Newton, C. W. and Holopainen, E. O., pp. 167–191, American Meteorological Society, Boston, MA, ISBN 978-1-944970-33-8, https://doi.org/10.1007/978-1-944970-33-8_10, 1990.
- Shorten, C. and Khoshgoftaar, T. M.: A survey on Image Data Augmentation for Deep Learning, *Journal of Big Data*, 6, 60, <https://doi.org/10.1186/s40537-019-0197-0>, 2019.
- Simmonds, I., Keay, K., and Tristram Bye, J. A.: Identification and Climatology of Southern Hemisphere Mobile Fronts in a Modern Re-analysis, *Journal of Climate*, 25, 1945–1962, <https://doi.org/10.1175/JCLI-D-11-00100.1>, 2012.
- 635 Stull, R.: *Practical Meteorology: An Algebra-based Survey of Atmospheric Science*, Univ. of British Columbia, ISBN 978-0-88865-283-6, 2017.
- Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, in: *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, Curran Associates, Inc., <https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html>, 2021.
- 640 Zhang, Y., Tiño, P., Leonardis, A., and Tang, K.: A Survey on Neural Network Interpretability, *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5, 726–742, <https://doi.org/10.1109/TETCI.2021.3100641>, 2021.
- Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., and Parmar, M.: A review of convolutional neural networks in computer vision, *Artificial Intelligence Review*, 57, 99, <https://doi.org/10.1007/s10462-024-10721-6>, 2024.
- 645 Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A.: Scene Parsing through ADE20K Dataset, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, IEEE, Honolulu, HI, ISBN 978-1-5386-0457-1, <https://doi.org/10.1109/CVPR.2017.544>, 2017.