https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Detection of upper-level troughs and ridges using deep learning – application in the Mediterranean

Ofir Ariel[1], Omer Sela[2], Hadas Saaroni[1], and Baruch Ziv[3]

[1]Department of Environmental and Earth Sciences, Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel
[2]Blavatnik School of Computer Science and AI, Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel
[3]Department of Natural Sciences, The Open University of Israel, Raanana, Israel

Correspondence: Ofir Ariel (ofirariel@gmail.com)

Abstract. Upper-level troughs and ridges are fundamental drivers of mid-latitude weather, organizing cyclogenesis, precipitation, and temperature extremes. Despite their importance, automatic detection remains difficult: existing algorithms rely on rigid rules that struggle to capture the high geometric variability of synoptic features. We introduce a physics-informed deep learning framework for automated axis detection that incorporates physical knowledge into the learning process, by transforming a curvature-based detection algorithm into continuous, differentiable operators embedded within an attention-based network, so geometric criteria are learned from data and axis detection draws on context from the surrounding flow. The network takes ECMWF Reanalysis v5 (ERA5) 500 hPa geopotential height and horizontal wind fields as input, and is trained and evaluated against a new benchmark of expert-labelled Mediterranean trough and ridge scenes. The network output is a confidence map, converted into precise axis lines based on a cyclonic vorticity advection rule. The model significantly outperforms classical baselines (F1 rises from 0.64 to 0.84 for troughs and from 0.54 to 0.75 for ridges). The framework demonstrates generalization capabilities beyond its training data: it qualitatively transfers well to global mid-latitudes without retraining, and the same architecture, initially trained only on troughs, reaches competitive ridge-detection accuracy after fine-tuning on just ten additional labeled ridge scenes. Applying the detector to historical reanalysis, we construct the first expert-calibrated, deep-learning-based climatology of upper-level trough and ridge frequency for the Mediterranean basin, providing a powerful and easily portable tool for understanding large-scale upper-level circulation.

Project page: https://sela-omer.github.io/upper-level-trough-ridge-detection/

1 Introduction

Midlatitude weather variability is fundamentally driven by synoptic-scale Rossby waves, which appear at the 500 hPa level as alternating troughs and ridges. Upper-level troughs (ULTs) are elongated regions of cyclonic curvature extending outward from midlatitude lows. They favour cyclogenesis and storm development downstream of their axes, where positive vorticity advection aloft drives divergence and ascent (e.g. Holton and Hakim, 2012). Upper-level ridges are the complementary anticyclonic phase, extending from the subtropical highs. Persistent ridges can suppress precipitation, favour drought, and set the large-scale conditions for heat extremes (Kautz et al., 2022).

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Not all troughs and ridges have the same effect. Their properties such as orientation, speed, meridional reach, and amplitude vary widely. In the eastern Mediterranean (EM), which sits at the transition between the midlatitude storm track and the subtropical dry zone, these differences matter: for example, a trough that penetrates farther south or tilts differently relative to the coastline can organize a very different rainfall distribution than a shallower or more poleward feature (e.g. Zangvil and Druian, 1990), while a stagnant ridge over the region can divert or suppress lower-level cyclonic activity and favour dry spells (Saaroni et al., 2015, 2019). An objective, reproducible climatology of trough and ridge axes is therefore important for understanding how the upper-level circulation organizes surface weather, and how it may change. Constructing such a climatology requires detectors that can be applied consistently over decades of reanalysis data and that recover axes in a way that would match expert judgment.

Existing detectors vary by the underlying diagnostic: either vorticity- or geometry-based, and the type of feature they extract, i.e., a single centre point, or a continuous axis. Centre-identification methods fall broadly into two groups. The first searches directly for extrema in the geopotential height field. Sanders (1988) produced a global climatology of upper-level trough centres by manually identifying bends in the 552-decametre contour at 500 hPa. Bell and Bosart (1989) automated detection of closed 500 hPa cyclone centres, which by construction excludes the more general open-wave trough. Spanos et al. (2003) studied upper-level cyclones over the Mediterranean during the dry season using a simple minimum-search algorithm, similar to those developed for surface cyclones.

The second group derives centres from a vorticity-related diagnostic. Lefevre and Nielsen-Gammon (1995) detected troughs by multiplying the curvature term of the geostrophic relative vorticity by the magnitude of the geostrophic wind, a quantity they termed Eulerian centripetal acceleration (ECA). Piva et al. (2008) later adopted the ECA method to compile a 500 hPa trough climatology for the Southern Hemisphere. Okajima et al. (2021) used a related quantity, curvature vorticity, which omits the wind-speed weighting, to quantify the contributions of both troughs and ridges to atmospheric energetics. Using local maxima in the Laplacian of the geopotential height field as a vorticity proxy for trough centres, Keable et al. (2002) studied the Southern Hemisphere midlatitude ULT climatology, and Almazroui et al. (2016) examined the linkage between such centres and rainfall over Saudi Arabia. An interesting exception is the work of Funatsu et al. (2008), which utilized the potential vorticity (PV) intrusion signal in satellite measurements to identify trough centres at the 200 to 300 hPa level via maximum brightness temperatures.

All of the above methods identify centres, or more broadly regions of cyclonic curvature. However, the weather driven by a passing trough is not confined to its minimum geopotential height region or maximum vorticity centre. Extracting the precise location and orientation of the entire axis span itself is desired, in order to accurately map the spatial distribution of these impacts.

To detect African easterly waves, Berry et al. (2007) used the curvature vorticity of the 700 hPa streamfunction, a quantity closely related to the curvature vorticity introduced above, defining trough and ridge axes as the contour lines where advection of curvature vorticity by the non-divergent wind is zero. This formulation is consistent with the conceptual framework of positive vorticity advection ahead (behind) and negative vorticity advection behind (ahead) the trough (ridge) axes. Although originally developed for easterly waves, the method is equally applicable to the midlatitude westerlies and to the 500 hPa

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

geopotential height field. Bueh and Xie (2015) termed this quantity curvature vorticity advection (CVA) but instead adopted a geometry-based approach: the minimum bending angle (MBA) method, which computes the geometric slope along geopotential height contours and connects minimum-angle points into trough or ridge axes, depending on their position relative to the slope-node points. Notably, Bueh and Xie (2015) is the only study reviewed here that reports some kind of objective evaluation of their method against an independent test dataset.

Knippertz (2004) identified ULT axes by calculating the zonal geopotential height gradient at each grid point and connecting neighbouring points, an approach that effectively treats troughs as north–south oriented valleys on a topographic landscape. Schemm et al. (2020) instead derived a curvature field from the geopotential height by comparing the tangent direction at each contour point with that at a second point interpolated a short distance along the gradient. A tracing algorithm then followed the geopotential height gradient from a local maximum (minimum) to delineate continuous trough (ridge) axes.

Despite these advances, several important knowledge gaps remain. The first is the relative scarcity of climatological studies of upper-level troughs and ridges compared with the extensive literature on surface cyclones. Most upper-level climatologies are predominantly global in scope. Although they consistently identify the Mediterranean as a region of enhanced wave activity (e.g. Sanders, 1988; Lefevre and Nielsen-Gammon, 1995; Schemm et al., 2020), their broad perspective limits a detailed characterization of regional circulation features. Conversely, studies focusing specifically on the Mediterranean are not designed as general-purpose climatologies. For example, Funatsu et al. (2008) concentrated on heavy precipitation events, whereas Spanos et al. (2003) and Almazroui et al. (2016) limited their analyses to specific seasons.

The second major gap is methodological. Detecting upper-level troughs and ridges is challenging, because their definition is neither unique nor completely objective. Both the slowly evolving planetary-scale Rossby waves and the fast-moving synoptic-scale shortwaves are commonly described by the same term. They can be characterized using different physical diagnostics (e.g., curvature, geopotential height minima or maxima, PV intrusions, and wind field structure) and exhibit substantial geometric diversity, ranging from open to closed systems, narrow to broad features, and a wide spectrum of orientations and amplitudes. The detection methods reviewed above rely on hard, spatially local decision rules - such as fixed geometric and intensity thresholds, or sequential step-by-step tracing — that were probably tuned or designed around a particular class of features. A rule that performs well for one class often fails for another, because trough and ridge geometry varies markedly from one case to another. This limitation is particularly acute over the EM, where even relatively weak upper-level disturbances, which easily might have been filtered out by many global detection schemes, can exert a pronounced influence on surface weather (Fig. 1).

A third major gap, closely related to the previous one, is the lack of benchmark datasets for objective performance evaluation. With the exception of Bueh and Xie (2015), existing ULT detection methods have been assessed primarily through qualitative inspection of a small number of case studies. Although the same challenge has historically affected surface cyclone detection, the much larger body of work on cyclone identification has enabled intercomparison studies that assess robustness through consensus among multiple detection algorithms (e.g. Flaounas et al., 2023). No comparable benchmark currently exists for upper-level troughs and ridges.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

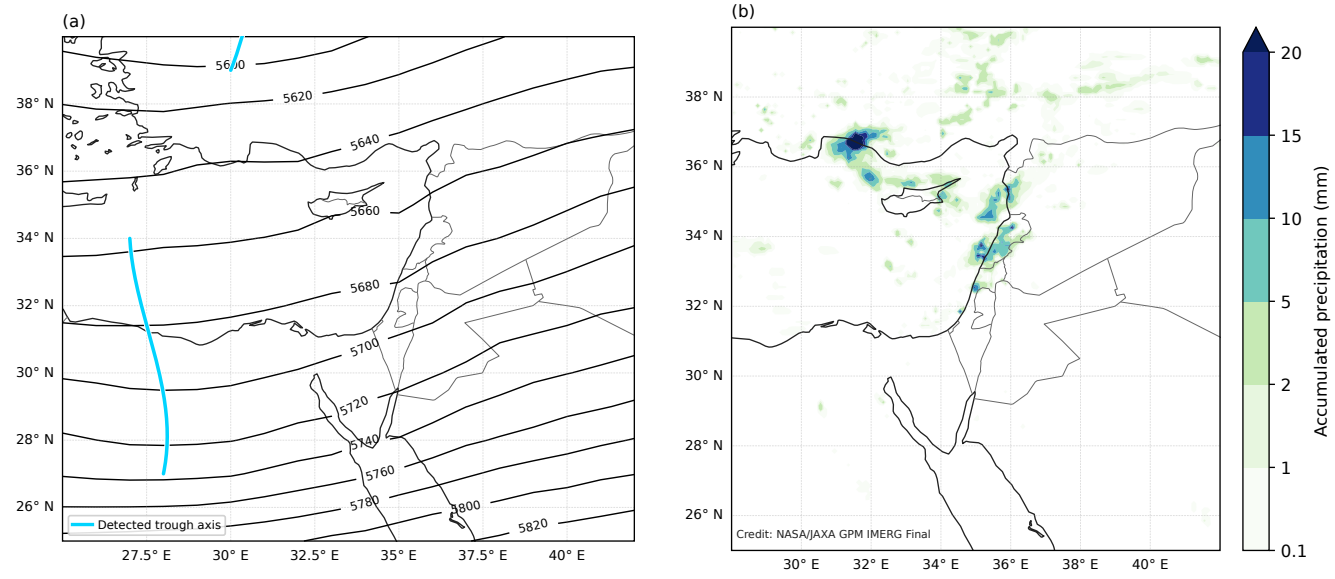

Figure 1. (a) Z_{500} contours and detected upper-level trough axes for 5 December 2024, 00:00 UTC. **(b)** Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) Final accumulated precipitation from 21:00 UTC on 4 December to 03:00 UTC on 5 December 2024. Precipitation in (b) is NASA/JAXA GPM IMERG Final (Huffman et al., 2023) (public-domain data; credit NASA/JAXA).

A data-driven detector, which infers its decision boundary from examples rather than a hand-designed rule, offers a potential solution to these limitations. Cai et al. (2021) developed a UNet-style architecture with an attention-weighted decoder using ECMWF Reanalysis v5 (ERA5) 500 hPa geopotential height, wind, and temperature fields at $0.25°$ resolution over an Asian domain (60 to $160°$ E, 5 to $65°$ N). Their supervised model was trained on 7,122 samples manually labelled by professional meteorologists. However, neither the training dataset nor the model implementation was made publicly available. As a result, reproducing or extending their work requires constructing a new expert-labelled dataset comprising thousands of samples – a truly Sisyphean task – while relying solely on a textual description of the model architecture.

Here we take a different approach. Rather than relying on a purely data-driven "black box" neural network, we incorporate physical knowledge into the learning process by engineering a differentiable geometric module. This design adapts the discrete diagnostic criteria of Schemm et al. (2020) into continuous spatial operators embedded directly within the network architecture, explicitly grounding the learning process with atmospheric physics. The resulting physics-informed model is trained with a high-quality dataset of expert-annotated trough and ridge axes over the Mediterranean basin. We also present qualitative evidence that the detector generalizes beyond its training domain and is applicable globally.

To evaluate our model, we perform a quantitative assessment by comparing our model directly with classical axis-detection methods on the same expert-labelled benchmark dataset, instead of relying on qualitative visual inspection. This enables a fair and robust comparison.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Using the developed detectors, we present an objective climatology of upper-level troughs and ridges over the Mediterranean basin, by quantifying their spatial occurrence. To our knowledge, this is the first expert-calibrated, deep-learning-based climatology of both upper-level troughs and ridges for this domain, and the detectors can be used to curate climatologies for other domains as well.

2 Data

2.1 Dataset and domain

The data used in this study are taken from ERA5 (Hersbach et al., 2020, 2023). We use the 500 hPa geopotential height (Z_{500}) and the horizontal wind components at 500 hPa (u_{500} and v_{500}; the full vector field is denoted $\boldsymbol{U}_{500}$), regridded from their native $0.25°$ resolution to a uniform $1°$ grid, to filter out small-scale variability irrelevant to the synoptic features considered here.

The full 1979-2024 ERA5 archive used in this paper is available from https://huggingface.co/datasets/Omer-Sela/upper-level-trough-ridg

Our study domain covers the Mediterranean basin between 20 to $60°\,\mathrm{N}$, $20°\,\mathrm{W}$ to $50°\,\mathrm{E}$, giving a 41×71 grid.

2.2 Expert-labelled dataset

The ground-truth dataset consists of consensus-based expert annotations, where two professional meteorologists manually labelled trough and ridge axes based on visual analysis of Z_{500} contours and $\boldsymbol{U}_{500}$ barbs, using a purpose-built annotation tool. The curvature and gradient of Z_{500} contours, along with the magnitude and direction of the wind, are the main visual cues used. Each drawn axis is rasterized onto the $1°$ grid as a line one cell wide. A cubic B-spline (degree 3; smoothing parameter $s = 0.1N_{\mathrm{px}}$; 300 evaluation points) is fitted to the raster to obtain a smooth, continuous axis representation for evaluation.

Axis pixels with $|\boldsymbol{U}_{500}|$ below $5\,\mathrm{kn}$ (approximately $2.6\,\mathrm{m\,s^{-1}}$) are considered dynamically insignificant, and their inclusion within the expert annotations is an artifact of manual labelling. Hence, these points are removed from the ground-truth dataset prior to spline-curve construction. Connected remnants that are smaller than four pixels after this edit are also discarded.

The published trough evaluation set contains 600 labelled scenes drawn from 2018–2020 (200 scenes per year), with a spacing of 44 h between successive scenes (three shorter gaps were retained intentionally to have the same amount of scenes per year). In total the benchmark comprises 2,579 individually traced trough axes (mean 4.3 per scene). Axis lengths span roughly 470 to $2{,}250\,\mathrm{km}$ for the central 80% (10th to 90th percentiles) of the distribution (median $\approx 1{,}010\,\mathrm{km}$), reflecting a mix of short Mediterranean disturbances and longer synoptic-scale troughs. This dataset is deliberately year-round, to ensure its utility to synoptic phenomena beyond winter rainfall events: Scenes are distributed evenly across all calendar months (150 December–February (DJF), 150 March–May (MAM), 151 June–August (JJA), and 149 September–November (SON) scenes), and the total axis count varies only modestly by season (601 axes in DJF vs. 640 in JJA).

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

An analogous ridge dataset was created with 200 labelled scenes from 2018. Together these scenes contain 683 expert ridge axes (mean 3.4 per scene), with a median length of $\approx 1{,}050\,\mathrm{km}$ with 80% of the axes ranging between 490 and $2{,}480\,\mathrm{km}$ in length (10th to 90th percentiles). The ridge set is also seasonally balanced (50 scenes per meteorological season).

The full expert-labelled dataset is available to download from https://huggingface.co/datasets/Omer-Sela/upper-level-trough-ridge-detec

Table 1 summarizes both benchmarks. Figure 2 shows representative annotations, axis-length distributions, sample diversity across seasons, and the seasonal balance of labelled axes.

Table 1. Summary of the expert-labelled trough and ridge benchmarks.

Statistic	Troughs	Ridges
Scenes	600	200
Years	2018–2020	2018
Total axes	2,579	683
Axes per scene (mean / median)	4.3 / 4	3.4 / 3
Axis length, median [p10, p90] (km)	1,011 [471, 2,247]	1,047 [494, 2,481]
Scene spacing (median gap, h)	44	44
Axes (DJF / MAM / JJA / SON)	601 / 695 / 640 / 643	151 / 178 / 167 / 187
Mean axis-mask coverage (% of domain)	1.8	4.8

3 Methodology

To overcome the limitations of existing methods when detecting upper-level troughs and ridges, we developed a physics-informed neural network, feeding meteorological features directly into the network architecture. We transformed the geometric algorithm of Schemm et al. (2020) into a continuous, fully differentiable pipeline (Sect. 3.1). By replacing discrete spatial thresholds and morphological operations with differentiable tensor operators (e.g., continuous gradient fields, soft sigmoid thresholding, and continuous curvature operators), gradients can propagate through and allow learning. In the initial stage, this pipeline processes the Z_{500} field to produce a soft axes mask map along with local gradient fields. These features are combined with u_{500} and v_{500} to compose the input fields to an attention-based detector (Sect. 3.2) that outputs two predictions at inference time: a pixel-level axis confidence map, and a complementary normal-sign field. The latter represents CVA structured as a ± 1 step function along the axis normal, used as a consistent orientation rule for axis tracing (Sect. 3.4). The model is trained directly on the curated set of expert annotations (Sect. 3.3). In a post-inference step, a classical algorithm extracts a thin axis along the zero-crossing of the predicted normal-sign field and fits the result to the same smooth cubic B-spline used for the expert-labelled dataset (Sect. 3.4). The complete architecture pipeline is summarized in Fig. 3.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

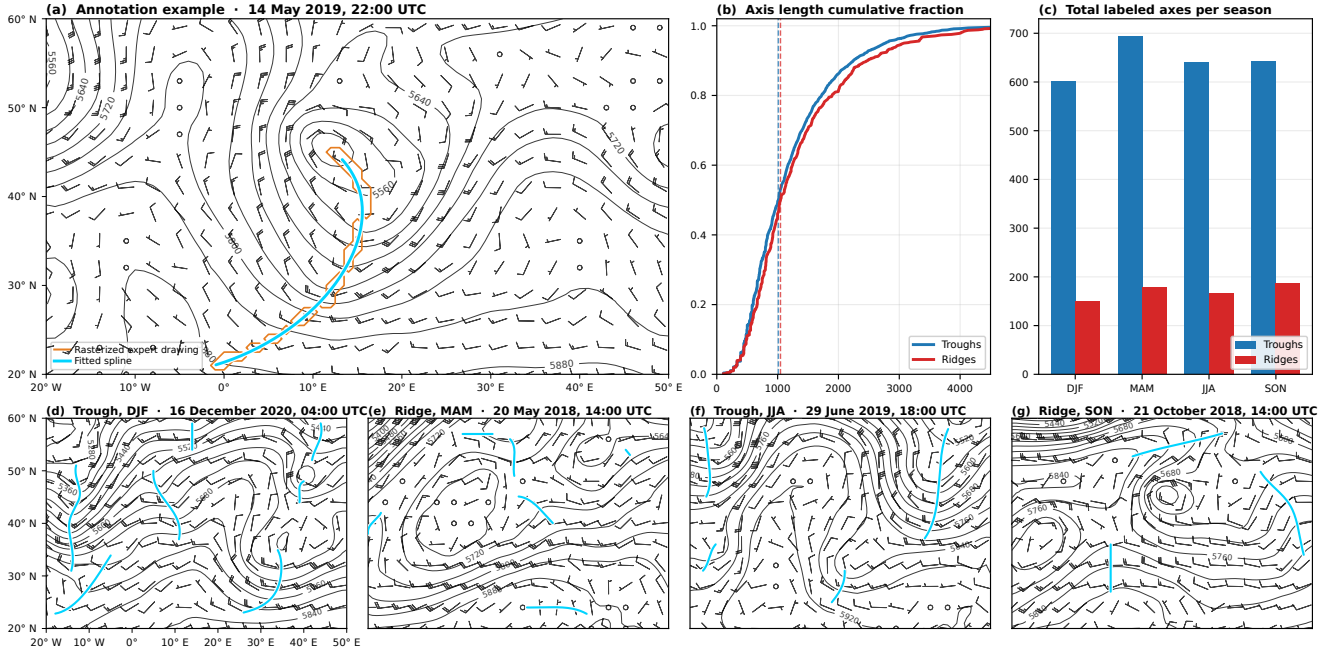

Figure 2. Expert-labelled trough and ridge benchmarks. **(a)** Trough annotation example. **(b)** Empirical cumulative distribution of axis length (trough median $\approx 1{,}010\,\mathrm{km}$; ridge median $\approx 1{,}050\,\mathrm{km}$). **(c)** Seasonal counts of labelled axes. **(d)–(g)** Example labelled scenes.

3.1 Physical feature extraction

To extract physical inputs to the attention-based detector, the Z_{500} field is processed through a sequential series of differentiable geometric operators:

1. **Pre-smoothing.** The Z_{500} field is first smoothed using a conservative diffusion filter. Because the resulting curvature field is highly sensitive to spatial variance, this step ensures the geometric operators extract synoptic-scale features rather than grid-scale noise. We apply the identical filter used during the expert annotation process, guaranteeing that the physical input aligns with the visual structures evaluated by the meteorologists.

2. **Gradients and curvature.** The geopotential gradient $\nabla Z = (\partial Z/\partial x, \partial Z/\partial y)$ is normalized and rotated $90°$ to obtain the direction along the local isoline. The isoline direction vector is compared, via bilinear interpolation, to the same vector a short distance ($20\,\mathrm{km}$) away in the direction of the gradient; the angle between the two vectors, normalized by that distance, defines the local curvature field α, following Schemm et al. (2020).

3. **Soft curvature selection.** The axis mask is the product of two sigmoids: one suppresses flat gradient regions, and the other selects curvature above a threshold for troughs or below its negative for ridges. The curvature and minimum-

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

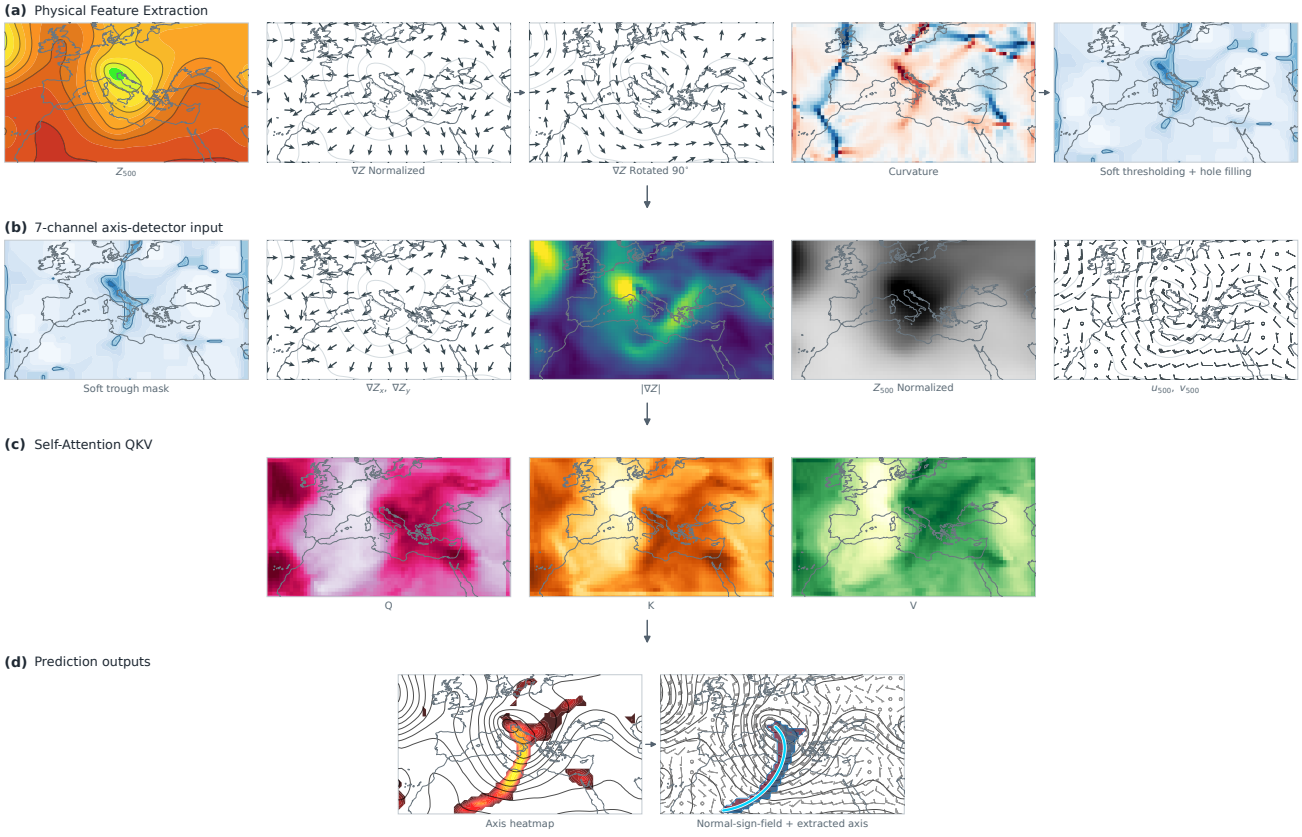

Figure 3. Axis-detection pipeline for one trough scene (14 May 2019, 22:00 UTC). **(a)** Z_{500} input is processed through a series of differentiable operators to output a soft axis mask. **(b)** The soft trough mask together with Z_{500} gradient components, $|\nabla Z|$, normalized Z_{500}, and wind components u_{500} and v_{500} form the seven input channels. **(c)** The detector encodes the stack with two 3×3 convolution blocks and attends within a 7×7 neighbourhood of each grid cell (Sect. 3.2). **(d)** Predicted axis heatmap and normal-sign field, with the extracted axis (Sect. 3.4).

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

gradient thresholds are learnable; a temperature parameter controls the sharpness of the selection. Unlike the hard decisions in the classical detector, these parameters can therefore adapt to the expert labels during training.

4. **Morphological hole filling.** A differentiable dilation-erosion cycle fills small holes in the soft feature mask, analogous to the binary morphology in the algorithm of Schemm et al. (2020).

The output is a soft trough or ridge mask which feeds into the axis detector neural network.

3.2 Attention-based axis detector

To capture both the geometric structure and the dynamic environment, the model concatenates seven physically complementary fields: the soft mask, normalized Z_{500}, its spatial gradients $\boldsymbol{\nabla} Z = (\partial Z/\partial x,\ \partial Z/\partial y)$ and $|\boldsymbol{\nabla} Z|$, and the horizontal wind components u_{500} and v_{500}, normalized by their monthly climatologies. While the geopotential and its gradients dictate the orientation of the trough valley, the wind field provides context on the flow structure and its anomalies, which influence axis placement.

These fields are first processed by a convolutional encoder comprising two 3×3 Conv-BN-GELU blocks, designed to extract local, grid-scale relationships between height geometry and the flow field. Trough and ridge axes are spatially coherent structures rather than isolated grid-point anomalies, so the encoded features are passed to an attention module that aggregates information from a small neighbourhood around each grid cell. Finally, two convolutional heads translate the contextualized features into a continuous axis confidence map and a normal-sign field (Sect. 3.4). Figure 3 summarizes the inference process.

Our model uses local self-attention over a 7×7 grid-cell window: at each location, the network forms query, key, and value representations from the same encoded feature map and computes a weighted average of neighbouring values. This allows nearby synoptic context to inform the axis prediction without prescribing fixed roles for individual input channels. Alternative attention designs and spatial receptive fields are documented in Appendix D, and discussed in Sect. 5.

3.3 Training procedure

Our model is trained directly on the expert-labelled scenes: weights are initialized randomly and optimized jointly on the pixel-level expert axes and pre-computed normal-sign field. To mitigate class imbalance during training – where background pixels heavily outnumber axis pixels – and to account for inherent annotation errors and uncertainty, each ground-truth axis is dilated by one grid cell in both cross-axis directions.

Training minimizes the loss function which is a weighted sum of three terms that supervise distinct parts of the pipeline (Eq. (1); component definitions in Appendix B1, weights in Table B1):

$$\mathcal{L} = \lambda_{\mathrm{h}}\, \mathcal{L}_{\mathrm{axis}}(\mathbf{h}, \mathbf{h}^{\star}) + \lambda_{\mathrm{s}}\, \mathcal{L}_{\mathrm{side}}(\mathbf{z}_{\mathrm{s}}, \mathbf{y}_{\mathrm{s}}; \mathbf{b}^{\star}) + \lambda_{\mathrm{m}}\, \|\mathbf{m}\|_1. \tag{1}$$

Here $\mathbf{m} \in [0,1]^{H \times W}$ is the soft feature mask from the differentiable front end (Sect. 3.1), $\mathbf{h} \in [0,1]^{H \times W}$ is the predicted axis confidence map, and $\mathbf{z}_{\mathrm{s}}$ are the side logits from the complementary normal-sign head (Sect. 3.2). Starred quantities are ground-

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

truth targets derived from the expert set: $\mathbf{h}^\star$ is the dilated axis raster. $\mathbf{b}^\star$ is a supervision band of 5 grid cells around each axis, where the CVA orientation is well defined, and $\mathbf{y}_s \in \{0, 1\}$ marks, within that band, the side of higher CVA.

$\mathcal{L}_{\text{axis}}$ is mean pixel-wise binary cross-entropy between $\mathbf{h}$ and $\mathbf{h}^\star$. It is the primary supervision signal: the network must place high confidence on the expert axis and low confidence elsewhere, despite the strong background/foreground imbalance of thin axes on a synoptic grid. $\mathcal{L}_{\text{side}}$ is masked binary cross-entropy on $\mathbf{z}_s$ inside $\mathbf{b}^\star$, encouraging the predicted normal-sign field to reproduce the CVA-oriented step across the axis, used later for axis tracing (Sect. 3.4). Finally, $\|\mathbf{m}\|_1 = (HW)^{-1} \sum_{i,j} m_{ij}$ penalizes the overall spatial coverage of the soft mask. In the classical detector of Schemm et al. (2020), a grid cell is strictly classified as either inside or outside the trough/ridge region. Here that decision is replaced by soft scores in $[0, 1]$ from learnable sigmoids (Sect. 3.1), which is necessary for gradient-based training but can leave a weak trough-like signal over large areas. The L_1 term counters that tendency and keeps the soft mask localized around the synoptic features of interest.

All reported scores use 10-fold cross-validation over the 600-scene expert set: each scene is predicted only by a model that did not train on it. Training hyperparameter settings are listed in Appendix B. Figure 4 exemplifies the predicted heatmap and normal-sign training.

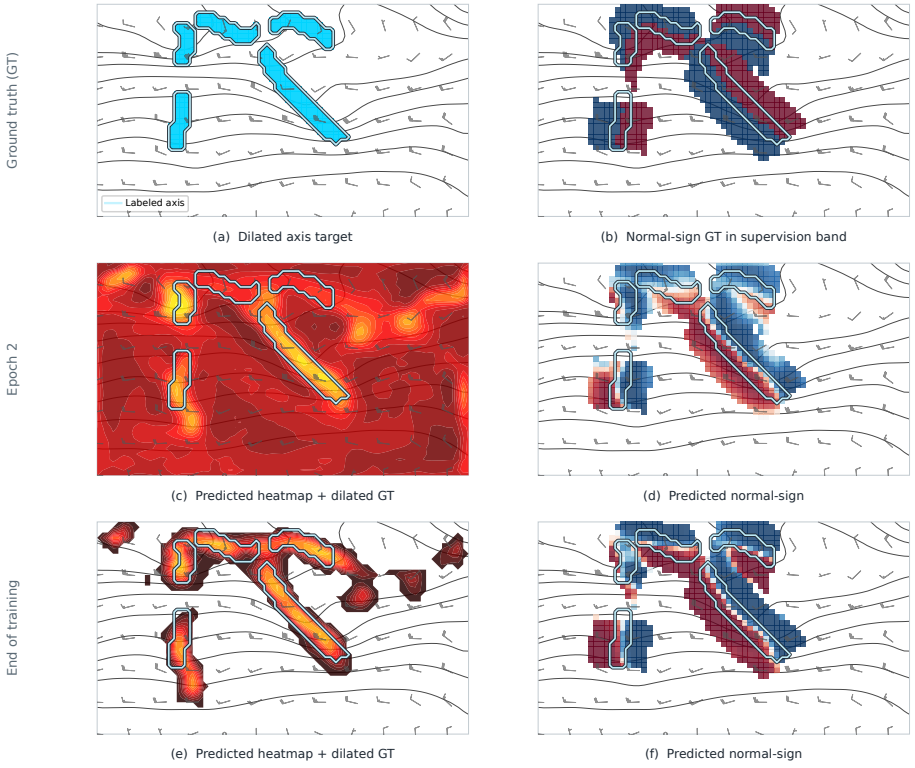

Figure 4. Training procedure on a held-out trough scene. **(a)–(b)** Training targets. **(c)–(d)** Predictions at an early epoch. **(e)–(f)** Predictions after training has converged. Contours show the labelled axes from (a) in all panels.

A training procedure involving self-supervised pretraining with pseudo labels is evaluated as an ablation in Sect. 4.4.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

3.4 Axis extraction

The attention detector identifies a two-dimensional "axis-like" region, but not a precise axis one grid cell wide. We recover thin axes in a separate geometric step that exploits the theoretical idea that along the axis normal, CVA changes sign. Although in practice the vorticity field is noisy, smoothing and averaging the CVA over the cross-axis pixels stabilizes the sign structure. The network's complementary normal-sign field is trained to reproduce that cross-track structure (Sect. 3.2). Within each localized trough or ridge region on the confidence map, the axis is placed where that field crosses zero – a simple, physically motivated rule that does not depend on morphological thinning of the heatmap alone (Fig. 5).

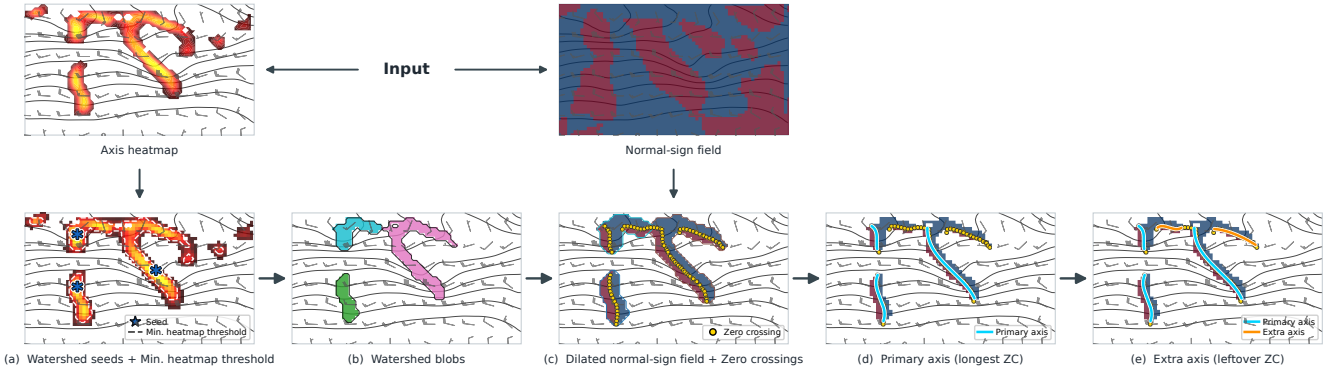

Figure 5. Inference-time axis extraction. The predicted axis heatmap and normal-sign field are the inputs (top row). **(a)** Predicted axis heatmap. **(b)** Watershed algorithm partitions the heatmap into axis-candidate blobs. **(c)** Normal-sign field inside the dilated blobs is used to obtain zero-crossing points. **(d)** Primary axis: the longest valid zero-crossing per blob. **(e)** Additional axes recovered from leftover zero-crossings.

- **Preprocessing**. Before extraction, predicted heatmap cells are zeroed wherever 500 hPa wind speed falls below a weak-wind threshold, matching the ground-truth cleaning rule (Sect. 2.2). The heatmap is also normalized scene by scene so that confidence thresholds can be applied on a comparable scale. Both the wind gate and the confidence thresholds are user-settable: a researcher who wants only the strongest, most unambiguous features can raise them; a more inclusive catalogue can lower them.

- **Separating individual troughs.** The predicted confidence map often contains multiple elongated regions within a single scene, where neighbouring synoptic responses can partially blend. To partition these regions into distinct axis candidates, local peaks in the normalized confidence map are identified as seeds. A standard watershed-style partition then assigns each grid cell to the nearest peak while respecting a minimum-confidence floor. This ignores faint background noise and keeps neighbouring local maxima separated when divided by a saddle in the heatmap. Candidate regions that fall below a minimum size are discarded. The result is one mask per candidate region – a blob around each heatmap maximum – which defines where we search for an axis, but is intentionally wider than one grid cell.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

– **Tracing the primary axis.** Within each candidate blob, the trough axis is traced as the zero level set of the predicted normal-sign field using a marching-squares algorithm. A well-resolved trough or ridge often produces a narrow, high-confidence heatmap that decays rapidly at its margins, so parts of the blob can be too thin to contain the geometric zero-crossing. To expand the normal-sign field context to prevent premature truncation, each blob is dilated by one grid cell before contour extraction. When marching squares yields multiple zero-crossing curves in one blob, the primary axis is the longest valid segment. A segment is valid if its nodes lie within the blob (with a small tolerance near the tips) and if each zero-crossing pixel has both positive and negative side-field values among its immediate neighbours – that is, it sits on a genuine cross-track sign change rather than in a noisy region or where two trough sign patterns interfere. The extracted axis is then clipped to match the initial blob boundaries, before dilation.

– **Secondary axis recovery and segment merging.** A strict "one axis per blob" constraint can fail in more complex or ambiguous synoptic setups. For example, a region with a strong heatmap signal over a broad area can make the saddle between two peaks too high for an extra seeding, so the resultant blob engulfs what a meteorologist would interpret as two distinct axes. While lowering the watershed height parameter is a natural fix, it can easily cause over-fragmentation in other cases. To address this, we evaluate all remaining zero-crossing contours within a region. If remaining segments are of a minimum length of continuous zero-crossing points – they are retained as additional independent axes. In addition, if any secondary segment's endpoints lie in close proximity to an existing axis (typically caused by watershed fragmentation), the pieces are merged as a single continuous axis.

– **Spline fitting**. The ordered zero-crossing polyline is clipped back to the pre-dilated blob and fitted with the same cubic B-spline representation used for the expert labels.

Watershed height, minimum heatmap confidence, and minimum blob size are the main tunable thresholds in this recipe. They are selected separately for trough and ridge evaluation on out-of-fold model predictions, optimizing on the F1 score, Chamfer distance, and completeness (Sect. 3.5). The selected settings are listed in Appendix A.

3.5 Evaluation metrics

Detection skills are assessed by comparing predicted and expert spline curves, each represented by 300 uniformly arc-length-sampled points. All curve-level scores use single spatial tolerance $\tau = 5$ grid cells.

– **Pairing predicted and expert axes.** For each scene, we compare every predicted curve to every ground-truth (GT) curve using the mean nearest-point distance (Chamfer distance, see below). Two curves are considered valid candidate matches if their distance is at most a threshold τ. Each predicted curve is then assigned to its closest GT curve. A GT axis matched to at least one prediction counts as a true positive (TP), while a GT axis with no valid prediction within τ counts as a false negative (FN). Because the extraction algorithm can occasionally split a single continuous synoptic axis into disconnected segments (e.g., if the watershed step divides a single elongated heatmap into two regions), we allow multiple predicted curves to map to a single GT axis. In these cases, the closest prediction registers the TP, while the

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

additional predictions are treated as continuation fragments. These fragments do not inflate the TP count, nor are they penalized as false positives (FPs), provided they lie within τ of the matched GT axis. Any predicted curve that remains completely unpaired is counted as an FP.

- **Curve-level F1.** For scene s, let TP_s, FP_s, and FN_s denote the TP/FP/FN counts from the matching above. Scene-level F1 is

$$\mathrm{F1}_s = \frac{2\,\mathrm{TP}_s}{2\,\mathrm{TP}_s + \mathrm{FP}_s + \mathrm{FN}_s}, \tag{2}$$

and we report the mean $\overline{\mathrm{F1}} = \frac{1}{N}\sum_s \mathrm{F1}_s$ over the N evaluation scenes. F1 evaluates whether the correct number of distinct axes was detected, but it provides no information regarding the geometric quality or exact placement of the matched curves, necessitating complementary spatial metrics.

- **Completeness.** Following Wiedemann et al. (1998), completeness is the fraction of each reference axis length that lies within distance τ of any extracted axis. For expert axis k with arc length L_k, we approximate this by uniform arc-length sampling along the spline (300 points) and define

$$C_k = \frac{1}{L_k}\int_{\gamma_k} \mathbf{1}\left[\min_{\hat{\gamma}\in\hat{\Gamma}} d(\boldsymbol{p},\hat{\gamma}) \le \tau\right] \mathrm{d}\ell, \tag{3}$$

where γ_k is the expert spline, $\hat{\Gamma}$ the set of predicted curves in the scene, and $d(\boldsymbol{p},\hat{\gamma})$ the point-to-curve distance for a point $\boldsymbol{p}$ on γ_k; in practice the integral is evaluated as the fraction of sample points within τ. Scene completeness is the mean of C_k over expert axes in the scene; reported completeness averages this quantity across scenes. Completeness balances F1: a model can achieve reasonable F1 while leaving long unmatched segments on otherwise detected axes, and conversely can raise completeness by adding extra predictions that F1 penalizes as false positives.

- **Chamfer distance.** For each matched expert axis k, let $\hat{\Gamma}_k$ be the union of all predicted curves assigned to that axis. The symmetric mean nearest-point distance is

$$d_k = \frac{1}{2}\left(\frac{1}{|\gamma_k|}\sum_{\boldsymbol{p}\in\gamma_k}\min_{\hat{\boldsymbol{p}}\in\hat{\Gamma}_k}\|\boldsymbol{p}-\hat{\boldsymbol{p}}\| + \frac{1}{|\hat{\Gamma}_k|}\sum_{\hat{\boldsymbol{p}}\in\hat{\Gamma}_k}\min_{\boldsymbol{p}\in\gamma_k}\|\hat{\boldsymbol{p}}-\boldsymbol{p}\|\right), \tag{4}$$

with distances measured in grid cells after arc-length sampling. Reported Chamfer is the mean of d_k over matched expert axes, averaged across scenes. Together, F1 and completeness separate "how many troughs" from "how much of each trough was recovered"; Chamfer then quantifies geometric accuracy on the recovered subset.

- **Pixel average precision (AP).** To evaluate the raw axis heatmap independently of post-processing, we compute the area under the pixel-level precision-recall curve across all heatmap thresholds,

$$\mathrm{AP}_s = \int_0^1 P(R)\,\mathrm{d}R, \tag{5}$$

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

where $P(R)$ is precision at recall R for scene s against the dilated axis target. AP is averaged across scenes and used principally for architecture and input-channel ablations.

Taken together, the metrics answer these complementary questions: F1 – "did we find the right number of axes?"; Completeness – "did we recover the full length of each labelled axis?"; Chamfer – "among matched axes, how accurately are they placed?"; AP – "does the model assign consistently high confidence to the right pixels?"

3.6 Statistical protocol

All scores are computed out of fold. Methods are compared in a paired-by-scene design: for each synoptic map we form scene-level differences, then report the mean paired effect and a 95% confidence interval (CI) from 10,000 scene-level bootstrap resamples. Bootstrap resampling is used because it summarizes uncertainty on the mean paired difference without assuming a particular distribution for per-scene scores. A sign test on F1 differences is reported as supporting evidence where scene-level F1 is discrete and often tied; Wilcoxon signed-rank tests on Chamfer and completeness provide analogous paired checks for these continuous scene-level scores.

When several pre-specified ablation contrasts are evaluated together, p-values are adjusted with Holm–Bonferroni to limit false positives from multiple comparisons. Final configuration choices among close variants additionally consider effect size, and training or operational simplicity.

4 Results

4.1 Classical baselines

We re-implement three published trough and two ridge axes detectors for quantitative comparison: for both troughs and ridges, we use the curvature-based method of Schemm et al. (2020) and a midlatitude adaptation of Berry et al. (2007) CVA zero-contour method. The geopotential-height-gradient method of Knippertz (2004) is used for troughs only. We do not include the minimum-bending-angle method of Bueh and Xie (2015) among the compared baselines: we lack a reliable implementation of their algorithm. For Schemm et al. (2020) We use the reported algorithm parameters from the paper, and a set of parameters tuned by Optuna (**?**) for Berry et al. (2007) and **?**, because we use our adaption of those algorithms. Tuning protocols and parameter settings for are listed in Appendix A3. The methods are applied to all 600 expert-labelled scenes for the paired comparison with the model's out-of-fold predictions, providing quantitative benchmark scores for the classical methods.

4.2 Trough detection

The model's out-of-fold score on the 600 trough scenes is F1 = 0.84, Chamfer = 1.04, and completeness = 0.87 (Table 2).

CVA emerges as the strongest classical baseline for trough detection, achieving an F1 of 0.64, a completeness of 0.78, and a Chamfer distance of 1.64, though its completeness comes at the cost of high FP count. Against CVA, our model yields a

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

significant F1 gain of 0.20, completeness gain of 0.09 and a Chamfer reduction of 0.59. Figure 6 shows a matched out-of-fold scene with Schemm, CVA, Knippertz, and our detector.

Table 2. Matched comparison on the 600-scene trough benchmark. Values are scene means; bracketed intervals are 95% bootstrap CIs on the mean (10,000 resamples). TP/FP/FN are pooled counts.

Method	F1 ↑		Compl. ↑		Chamfer ↓		TP	FP	FN
Ours	**0.84**	**[0.83, 0.85]**	**0.87**	**[0.86, 0.88]**	**1.04**	**[1.01, 1.07]**	**2171**	**404**	**408**
Schemm	0.63	[0.61, 0.65]	0.62	[0.61, 0.64]	2.06	[2.00, 2.12]	1567	724	1012
CVA / Berry	0.64	[0.63, 0.66]	0.78	[0.77, 0.80]	1.63	[1.59, 1.68]	1977	1587	602
Knippertz	0.62	[0.60, 0.64]	0.53	[0.51, 0.55]	2.16	[2.10, 2.22]	1417	466	1162

4.3 Ridge detection

On the 200-scene dataset, the ridge detector's out-of-fold score is F1 = 0.75, Chamfer = 1.20, and completeness = 0.84 (Table 3).
325 Similar to the trough detection, CVA again emerges as the strongest classical baseline, achieving an F1 of 0.54, completeness of 0.60 and a Chamfer distance of 1.67. Against CVA, our model yields a significant F1 gain of 0.21, completeness gain of 0.24 and a Chamfer reduction of 0.47. Figure 7 shows a matched out-of-fold scene with Schemm, CVA, and our detector.

Table 3. As Table 2 but for the 200-scene ridge benchmark.

Method	F1 ↑		Compl. ↑		Chamfer ↓		TP	FP	FN
Ours	**0.75**	**[0.73, 0.78]**	**0.84**	**[0.82, 0.87]**	**1.20**	**[1.13, 1.28]**	**562**	**241**	**121**
Schemm	0.47	[0.43, 0.50]	0.52	[0.49, 0.56]	2.07	[1.92, 2.22]	341	400	342
CVA / Berry	0.54	[0.52, 0.57]	0.60	[0.57, 0.64]	1.67	[1.56, 1.78]	411	403	272

4.4 Ablation study

Table 4 condenses the model-selection ablation experiments. Among the three attention variants evaluated (local self-attention,
330 cross-attention, and unified-channel attention), local self-attention clearly outperforms cross-attention, and is preferred over unified-channel attention. Pretrain yields no gain over random initialization. We therefore retain random-init training for our selected model as it simplifies the production recipe.

Relative to a wind-plus-CVA configuration, omitting CVA while keeping the wind has does not move the metrics in a significant way; omitting the wind while keeping CVA significantly harm all the metrics. A minimal-input variant without wind or
335 CVA channels matches F1 but has worse Chamfer and AP; we therefore keep wind and drop CVA in the selected configuration (wind-only random-init self-attention). Finally, the architecture proves robust to variations in attention hyperparameters. Most

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

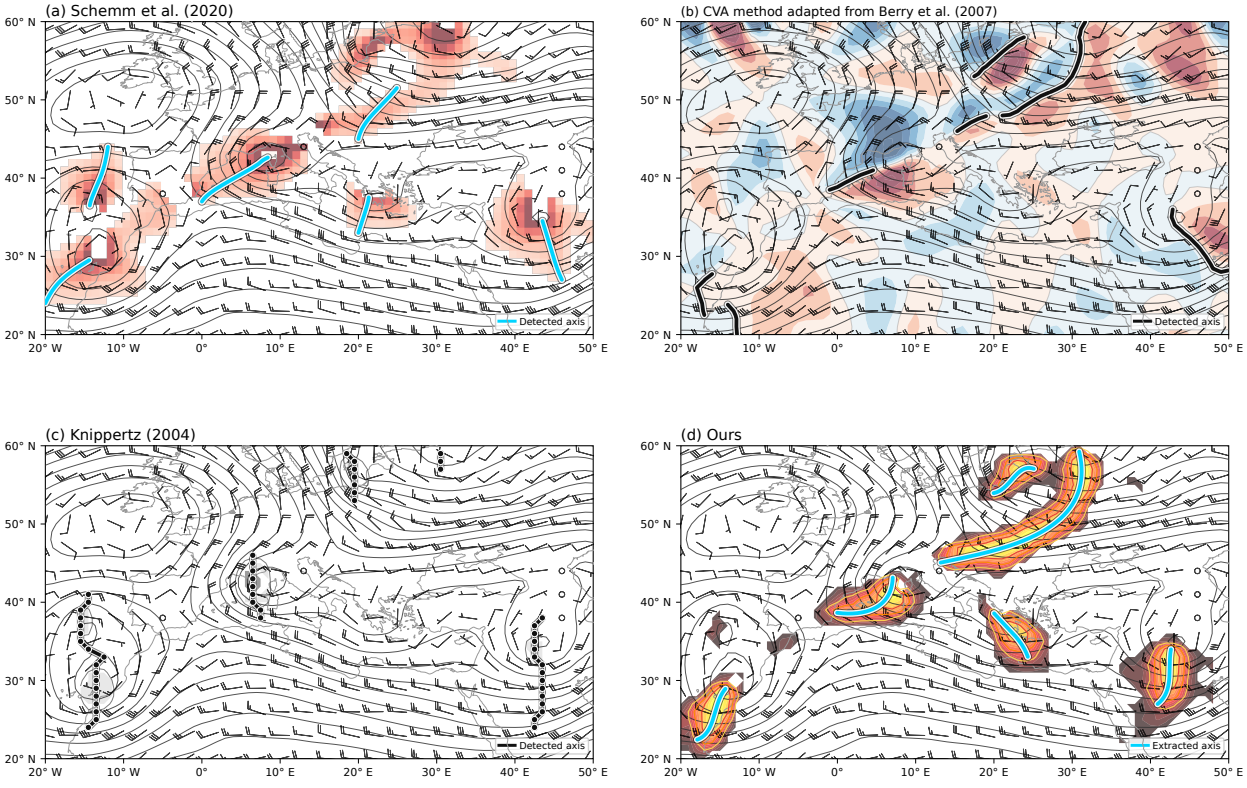

Figure 6. Matched comparison of our model and classical baselines on an out-of-fold trough scene. Each panel shows Z_{500} contours and detected axes for the same date. **(a)** Schemm et al. (2020). **(b)** CVA method adapted from Berry et al. (2007). **(c)** Knippertz (2004). **(d)** This study.

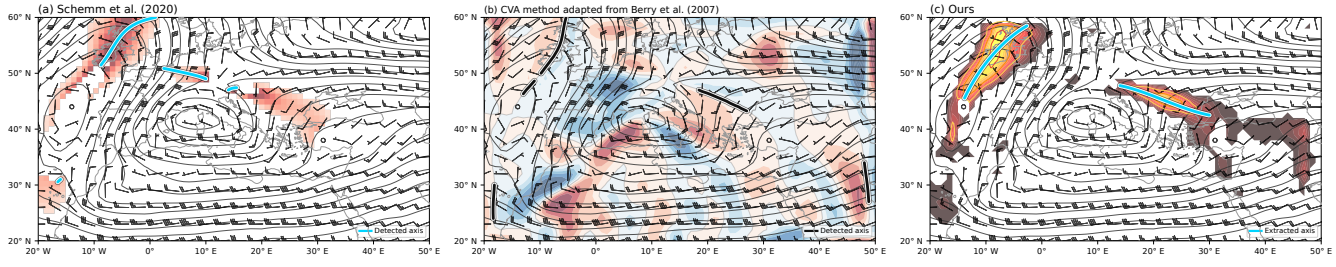

Figure 7. As Fig. 6 but for a ridge scene. **(a)** Schemm et al. (2020). **(b)** CVA method adapted from Berry et al. (2007). **(c)** This study.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

notably, expanding the self-attention receptive field from a local 7×7 synoptic window to the full spatial domain yields no improvement in detection skill. Window-size and receptive-field ablations are reported in Appendix D.

All ablations were conducted on the trough detector only. The ridge detector inherits the same random-init architecture and training recipe (Sect. 3.3) with ridge-specific post-processing settings (Appendix A).

Table 4. Ablation summary on the 600-scene trough benchmark. All listed variants use random initialization except pretrain.

Variant	F1 ↑		Compl. ↑		Chamfer ↓		AP ↑	
Self-attention (local 7×7)	0.84	[0.83, 0.85]	0.87	[0.86, 0.88]	1.04	[1.01, 1.07]	0.67	[0.66, 0.68]
Cross-attention	0.80	[0.79, 0.81]	0.80	[0.78, 0.81]	1.18	[1.14, 1.22]	0.61	[0.60, 0.62]
Unified-channel attention	0.83	[0.81, 0.84]	0.84	[0.82, 0.85]	1.08	[1.04, 1.12]	0.63	[0.62, 0.64]
Pretrain	0.84	[0.83, 0.85]	0.88	[0.87, 0.89]	1.03	[1.00, 1.06]	0.67	[0.66, 0.68]
Random init	0.84	[0.83, 0.85]	0.87	[0.86, 0.88]	1.04	[1.01, 1.07]	0.67	[0.66, 0.68]
Wind + CVA	0.84	[0.83, 0.85]	0.87	[0.85, 0.88]	1.03	[1.00, 1.06]	0.67	[0.66, 0.68]
Wind only (no CVA)	0.84	[0.83, 0.85]	0.87	[0.86, 0.88]	1.04	[1.01, 1.07]	0.67	[0.66, 0.68]
CVA only (no wind)	0.81	[0.80, 0.82]	0.78	[0.77, 0.80]	1.21	[1.17, 1.25]	0.60	[0.59, 0.61]
Neither (no wind, no CVA)	0.84	[0.83, 0.85]	0.87	[0.85, 0.88]	1.09	[1.05, 1.12]	0.65	[0.64, 0.66]

4.5 Seasonal climatology

For climatological inference, we apply the selected model to ERA5 Z_{500}, u_{500}, and v_{500} at 00:00 and 12:00 UTC from 1979 through 2025 on the same 41×71 grid used for training. Per-grid-cell axis frequency is the fraction of analyzed times at which an extracted axis intersects that cell. The annually averaged normalized trough-frequency map (Fig. 8) reveals the most prominent maximum over the eastern Mediterranean, with additional centres west of the Moroccan coast and over the Black Sea. During winter (DJF), trough frequency over the EM is lower, and troughs are distributed across a broader area of the Mediterranean basin. Nevertheless, the main frequency maxima remain over the eastern and central Mediterranean, with troughs occurring more frequently over maritime areas than over the continental regions north and east of the basin. It should be emphasized that trough frequency is not necessarily related to trough depth, which is reflected in Z_{500} values that reach their greatest magnitude during winter. In summer (JJA), the maximum trough frequency is observed over the EM. This maximum is associated with the relative trough situated between two upper-level ridges: one over the Arabian Peninsula to the east and the other over North Africa to the west. These troughs occur within a high geopotential-height environment. Despite the persistence of this troughing pattern, summer is virtually rainless in the EM, as discussed below. During the transition seasons – spring (MAM) and autumn (SON) – the spatial pattern resembles that observed in summer, with relatively high trough frequency over the EM. However, the frequency is lower than in summer, and the troughs are distributed over a broader area. A secondary maximum is evident over the Black Sea in spring but disappears in autumn. The spatial distribution of upper-

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

level ridge frequency largely mirrors that of upper-level trough frequency (Fig. 9), consistent with the wave-like structure of low-wavenumber Rossby waves. At lower latitudes, however, high-pressure systems are broad and extend over large areas; consequently, the algorithm does not necessarily identify them as distinct ridge axes.

4.6 Global generalization

The detector is fully convolutional and its attention is local, so inference does not require the geographic dimensions or longitudinal extent used for EM training. Application over the full Northern and Southern Hemispheres recovers qualitatively plausible trough and ridge axes without retraining (Fig. 10); the Southern Hemisphere application requires flipping the latitude order and the meridional wind sign. These examples demonstrate qualitative portability, not benchmarked out-of-domain skill, because expert labels were not made globally.

4.7 Few-shot ridge transfer

We test whether the trough detector can be repurposed to ridge detection from a small labelled set. The trough checkpoint is applied in ridge mode – negated Z_{500} gradients and reversed curvature sign in the physical pipeline – and fine-tuned on ten labelled ridge scenes from the 200-scene expert ridge set (10 / 10 / 180 train / validation / test split; Fig. 11).

Two initializations of the same self-attention architecture (500 hPa wind on, no CVA input) are compared. Ours is the selected trough detector trained from random weights on expert labels only. Ours + pretraining adds the optional Schemm pseudo-label stage described in Appendix C before expert-trough fine-tuning. Table 5 reports zero-shot and few-shot scores on the held-out test split.

Zero-shot ridge transfer is weak (Ours F1 = 0.23; Ours + pretraining F1 = 0.28). After ten-scene fine-tuning, test F1 rises to 0.53 and 0.63, respectively, and mean Chamfer falls from 2.24 to 2.05 (Ours) and from 1.76 to 1.67 (Ours + pretraining), displaying a marked improvement.

Table 5. Few-shot ridge transfer on the held-out 180-scene test split.

Initialization	Stage	F1 ↑	Completeness ↑	Chamfer ↓
Ours	Zero-shot	0.23	0.24	2.24
Ours	Few-shot	0.53	0.66	2.05
Ours + pretraining	Zero-shot	0.28	0.23	1.76
Ours + pretraining	Few-shot	0.63	0.70	1.67

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

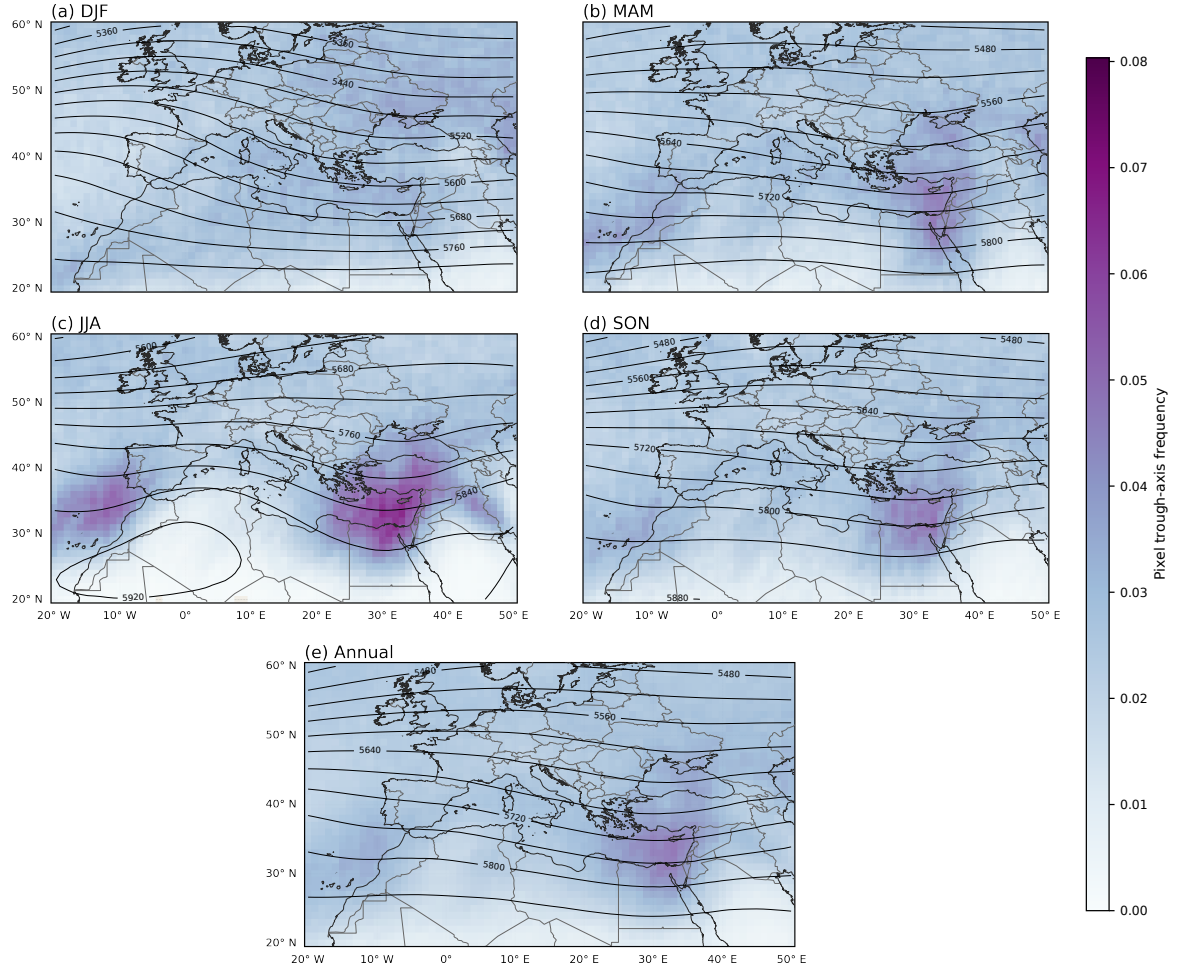

Figure 8. Per-grid-cell trough-axis frequency for 1979–2025. **(a)–(d)** Seasonal maps (DJF, MAM, JJA, SON). **(e)** Annual mean. Frequency is the fraction of analyzed 00:00 and 12:00 UTC times in each period at which a detected trough axis intersects a cell. Black contours are the corresponding mean Z_{500}. The colour scale is shared with Fig. 9.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

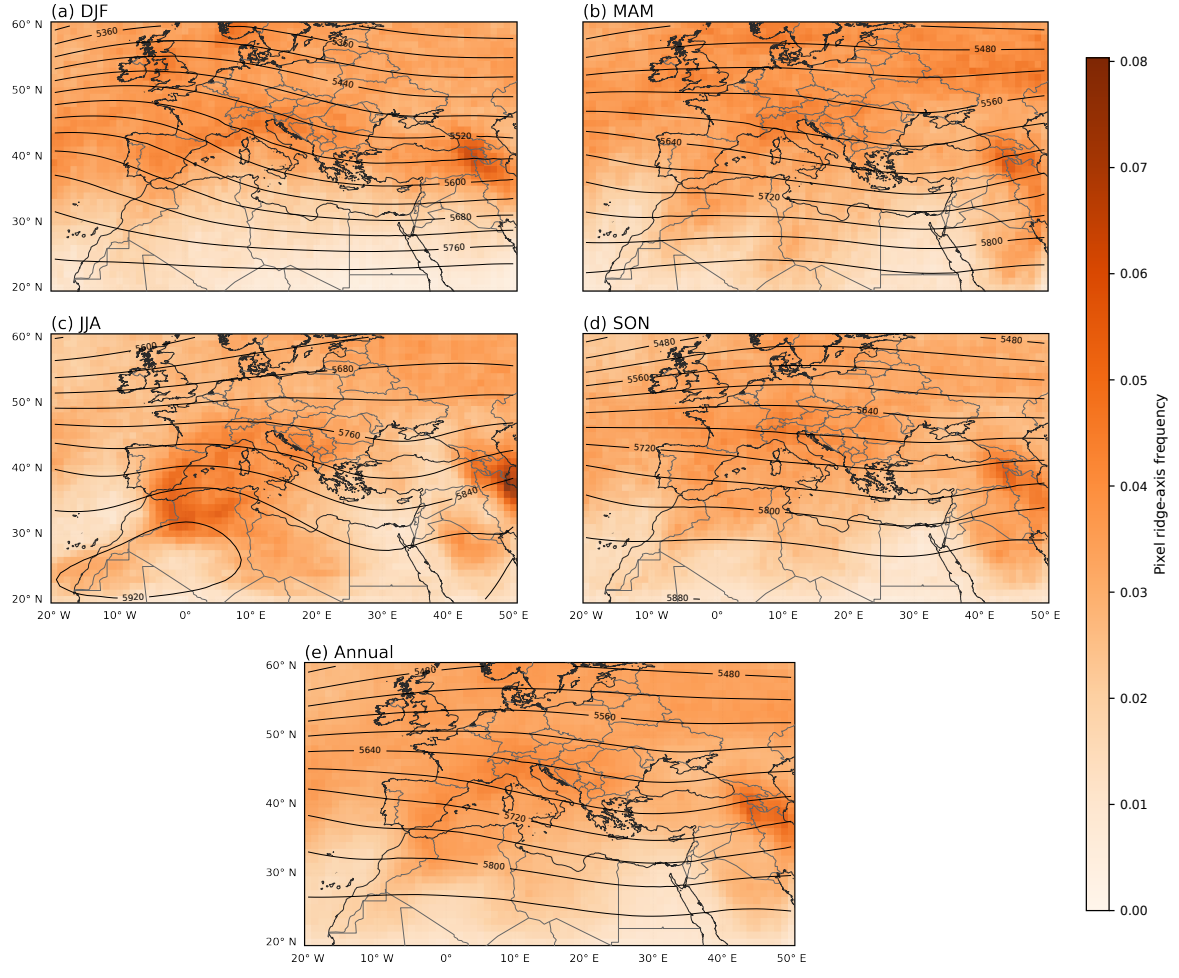

Figure 9. Same as Fig. 8, but for ridges.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

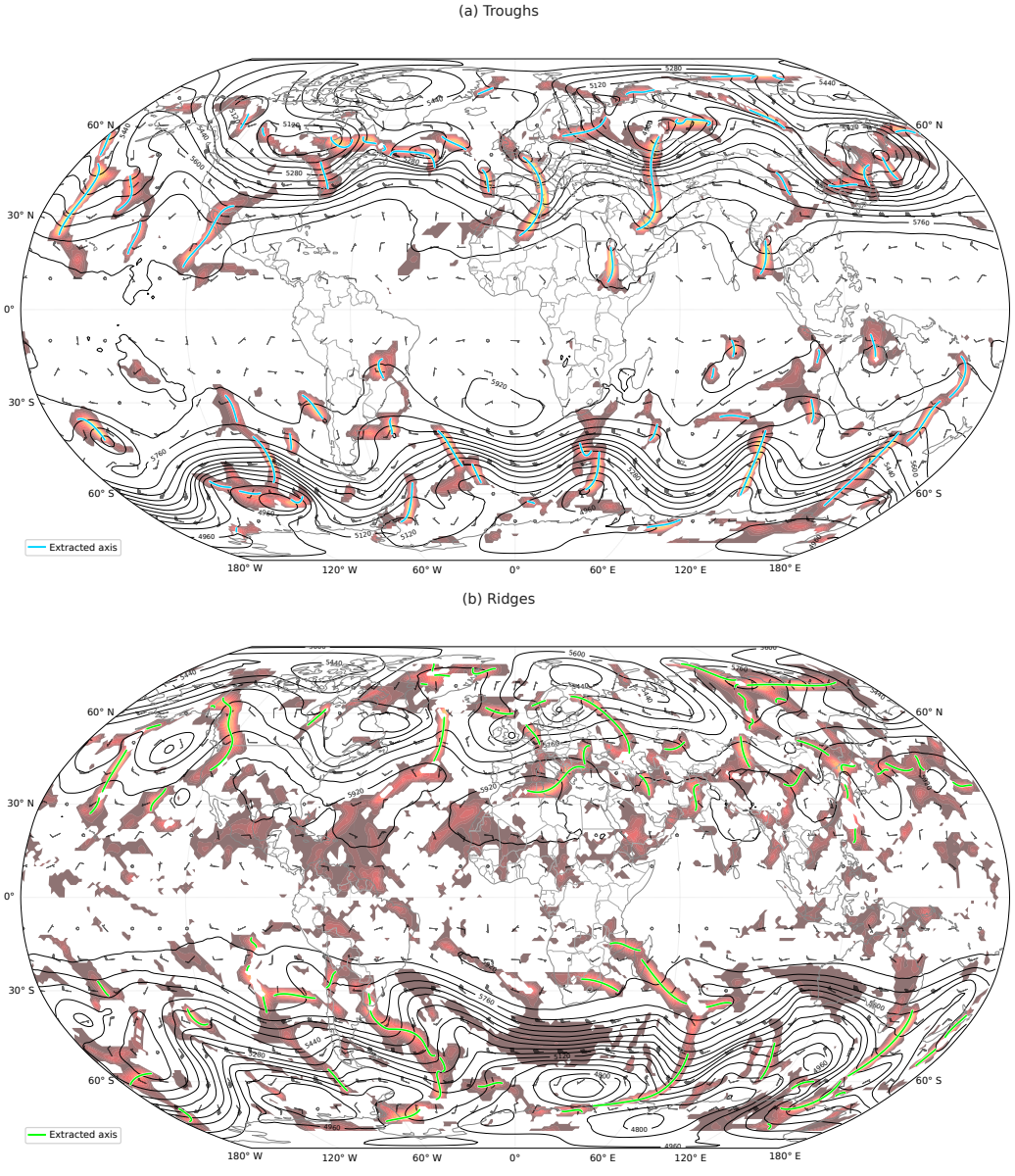

Figure 10. Global zero-shot detection without retraining. **(a)** Troughs. **(b)** Ridges. Southern Hemisphere applications use flipped latitude and reversed meridional wind. Extracted axes are overlaid on the predicted heatmap.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

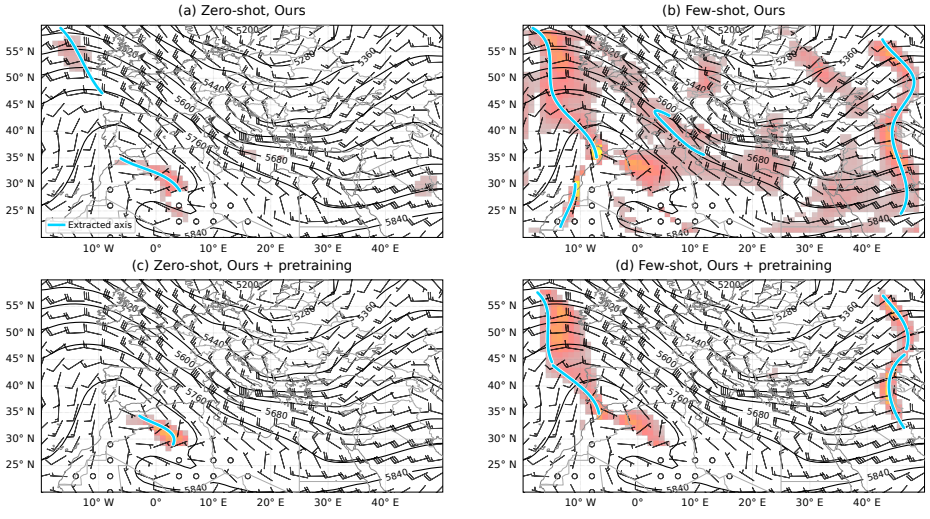

Figure 11. Zero-shot versus few-shot ridge detection on one held-out scene. **(a)** Zero-shot transfer from Ours in ridge mode. **(b)** Ours, after fine-tuning on ten labelled ridge scenes. **(c)** Zero-shot transfer from Ours + pretraining. **(d)** Ours + pretraining after ten-scene fine-tuning. Shading is the predicted ridge heatmap; extracted axes are shown in the legend.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

5 Discussion

5.1 Comparison to classical baselines

On both the 600-scene expert-labelled troughs and 200-scene ridges benchmark, the learned model improves both curve detection and localization relative to the published methods retained for quantitative comparison.

It is worth emphasizing that the classical detectors are not weak. In a large fraction of scenes they place axes that an analyst would accept, and the climatologies built on them remain informative. Their aggregate scores are limited less by an inability to locate the axes than by a number of recurrent, systematic failure modes, each traceable to a hard decision taken somewhere in the algorithm. Figure 6 illustrates these failures, and they largely motivated the development of the present model.

Each of the three baselines fails in its own characteristic way. The tracing step of Schemm et al. (2020) is sequential and greedy: starting from an extremum inside the curvature object, it follows the geopotential height gradient one step at a time, so a single unfavourable step, typically where curvature is locally weak, the gradient is noisy or where two curvature objects touch, can terminate an axis prematurely. This is illustrated in Fig. 6, where the northernmost trough is cut short due to the gradient being noisy in an arm of weak curvature. Because the decision at each step depends only on the local gradient and on the path already traced, there is no mechanism for later evidence to correct an earlier mistake. The attention module reaches its decision differently: every grid cell aggregates evidence from its neighbourhood before an axis probability is assigned, so a segment is supported by the surrounding flow configuration rather than by the history of a walk, and the truncations characteristic of greedy tracing disappear. The scheme of Knippertz (2004) isolates regions of zonal geopotential height gradient and is, as Knippertz notes in the paper, best suited to troughs with a north–south orientation; strongly tilted or nearly zonally oriented axes are correspondingly truncated or missed, as is visually evident in Fig. 6. It should be noted that the original algorithm use a 2.5° grid so the algorithm had been adapted. The CVA zero-contour approach adapted from Berry et al. (2007) produces, by construction, a dense set of zero lines, only some of which correspond to synoptically meaningful axes. While CVA actually performs well on the scene in Fig. 6, it still has a spurious, extra, meaningless axis sticking out of the southwestern low. Separating real from spurious axes relies on a cyclonic curvature threshold, and a fixed choice doesn't performs well across the range of trough and ridge amplitudes present in the benchmark: permissive settings retain spurious lines, whereas stricter settings begin to remove real axes.

5.2 Robustness to architectural choices

The ablations help define a practical final configuration. The three attention variants differ only in how query, key, and value are formed.

In cross-attention, roles are prescribed: the query is built from normalized Z_{500}, keys from the geopotential gradient and $|\nabla Z|$, and values from an encoded mix of all fields. Neighbour weights therefore answer a fixed question – which nearby cells have gradient structure consistent with this height? – rather than letting the network learn how each diagnostic should participate in matching.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

In unified-channel attention, every input channel contributes symmetrically to query, key, and value before the local aggregation step. No channel is designated as "asking" vs. "being matched"; wind, height, and geometry enter attention on equal footing.

The selected self-attention design first fuses the input stack with a convolutional encoder, then forms query, key, and value from that shared representation. Channel roles are learned rather than hard-coded.

Self-attention clearly outperforms cross-attention, suggesting that the prescribed height to geometry factorization is too restrictive. Unified-channel attention recovers most of the detection skill (F1 0.83 vs 0.84) but with worse Chamfer (1.08 vs 1.04), consistent with axes being found but less accurately placed when attention operates on raw channels rather than on a learned fused representation.

Within the ranges we explored, detection skill is largely insensitive to the attention window size (Table D3). The most informative of these negative results concerns the receptive field. Replacing the 7×7 local window with global attention, in which every grid cell attends to every other cell in the domain, does not improve skill on the matched pretrain comparison (Table D2); the resulting heatmaps are nearly indistinguishable from those of the local model (Fig. E2). This suggests that the evidence required to decide whether a given cell lies on a trough axis is essentially local, independent of the longer-range structure of the wave. It also has a practical consequence: the selected configuration is inexpensive to train and to apply to multi-decadal archives, and its computational cost grows only linearly with domain size.

Pseudo-label pretraining gives no significant edge over random initialization and is omitted from the main model, yielding a simplified training procedure.

5.3 Input-channel interpretation

The channel ablations are best read as a statement about information content rather than about the physical relevance of the discarded diagnostics. At 500 hPa the flow is close to geostrophic balance, so the horizontal wind is, to a good approximation, a derivative of the geopotential height field, and curvature vorticity and its advection are higher-order derivatives of the same field. Most of the information carried by these diagnostics is therefore already present in Z_{500}, and a network that is free to compute its own spatial derivatives can recover much of it. However, the retention of wind is not only due to a small metric improvement. It also helps to distinguish the model from a curvature detector with softened thresholds. The ageostrophic component, although small in the mean at this level, is not distributed randomly: it is largest precisely where the flow turns sharply, and an abrupt directional shift in the wind is one of the main cues an analyst uses to place a trough axis. Access to u_{500} and v_{500} allows the network to exploit that cue directly, which is consistent with the localization-only character of the improvement, seen in Chamfer distance and pixel AP rather than in detection F1.

Two further considerations bear on the CVA result, whose lack of improvement seems contradictory to its use as the algorithm guidance to extract the axis. First, the annotators worked from geopotential height contours and wind barbs and did not consult the CVA field, both because of its noisiness and because expert axes do not always coincide with its zero line in the manner assumed by Berry et al. (2007). The labels do nonetheless follow the curvature vorticity field itself closely (Fig. E3), so the absence of a measurable benefit is not evidence that curvature vorticity is unrelated to the target.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Second, supplying CVA as an input channel visibly changes the heatmap: it sometimes tightens the axis onto the expert label, but it also tends to shorten axes so that they terminate around the local maxima of curvature vorticity. Gains and losses of this kind partly cancel, which is a plausible explanation for the small, non-significant net effect. We therefore omit CVA, without claiming that it carries no useful signal.

5.4 Ridge detection

The only substantive difference of the ridge detector is the size of the expert dataset: 200 scenes rather than 600. Ridge skill is accordingly lower than trough skill (F1 = 0.75 versus 0.84), which we attribute partly to the smaller labelled set and partly to the nature of the target: ridges are broader and flatter than troughs, their curvature signal is weaker relative to the background, and the axis position is correspondingly less tightly constrained, for the annotators as well as for the model. This is also evident in the score drop experienced by the classical algorithms, from F1 = 0.63 (Schemm) and 0.64 (CVA) in trough detection to 0.47 and 0.54 in ridge detection. Despite the lower detection score, our ridge detector still reaches skill that is useful for climatological work.

5.5 Subjectivity of the labels

Inspecting a large number of predicted heatmaps led to an observation that nuances the reported metrics. The model frequently indicates a trough or ridge at a location that carries no expert label, and on re-examination of the corresponding synoptic situation the suggestion is usually defensible. Some of these cases reflect simple oversight, which is unsurprising given how demanding manual labelling is. Others reflect something more fundamental: annotating synoptic features is inherently subjective and open to meteorological interpretation. A substantial fraction of the events counted as false positives are thus not errors in a meteorological sense, hence F1 should be regarded as a lower bound on the agreement an analyst would perceive. The heatmap has, in practice, become a reference we are willing to consult alongside the labels themselves.

This ambiguity is a property of the object, not of our protocol, and it does not diminish the value of a common evaluation set. Until now, upper-level trough and ridge detectors have been assessed almost exclusively through visual inspection of a handful of cases, and there has been no way to state how much better one method is than another. Providing an expert-labelled benchmark together with a fixed matching rule and a paired statistical protocol allows such statements to be made.

5.6 Generalization of the model

Labelling troughs and ridges over the entire globe for long enough to obtain a meaningful sample of interannual and intra-annual variance is a highly tedious and error-prone task. Our labelling was carried out over a domain the annotators know well, which made the work faster and the judgments more consistent. Nevertheless, the model proves to be more general than its training set. Applied without retraining to the whole Northern Hemisphere and, after flipping the latitude axis and reversing the sign of the meridional wind, to the Southern Hemisphere, it recovers axes that are visually meaningful (Fig. 10). Over regions whose flow regime differs more strongly from the Mediterranean sector the result might be less robust, and

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

this is precisely the situation the few-shot pathway is designed for: a researcher can label a modest number of local cases where the model underperforms and fine-tune the released model with the supplied code (model checkpoints are available at https://huggingface.co/Omer-Sela/upper-level-trough-ridge-detection-models).

The few-shot ridge experiment shows what such an adaptation can achieve. Large-scale pretraining has been shown to improve downstream transfer and sample efficiency, including when only a few labelled target examples are available (Kolesnikov et al., 2020). In weather and climate applications, pretrained representations have also been successfully adapted across tasks, variables, spatial scales, and geographic regions (Nguyen et al., 2023), while pretraining on climate-model simulations has been shown to reduce overfitting and improve subsequent forecast skill on reanalysis data (Rasp and Thuerey, 2021). For a researcher targeting another region or another axis type, the released trough detector can serve as an initialization: only a small set of local maps needs to be labelled with the supplied annotation tool, after which the model can be fine-tuned while retaining most of its learned representation.

The experiment deliberately starts from the trough detector's weights and fine-tunes on only ten labelled ridge scenes. Ridges are not merely sign-flipped troughs – their flow characteristics differ, and zero-shot transfer remains weak. Yet the ten-scene adaptation lifts performance to a level that is useful in practice (Table 5). Ours + pretraining yields a higher gain. Representation quality affects adaptation, but it is not required for a substantial few-shot improvement: even the selected-model initializer more than doubles zero-shot detection skill. Because the target here differs from the source in kind and not only in geography, this is a deliberately demanding test, and the improvement it demonstrates should be at least as attainable for a researcher adapting the released model to a new geographic domain.

Flexibility also extends to how the output is used. Because the detector emits a continuous heatmap before any thresholding, a researcher can adapt the definition of the objects retained without retraining. For example, we deliberately labelled small and weak troughs, since in the eastern Mediterranean even shallow disturbances can organize substantial rainfall (Fig. 1); in a domain where such features are noise, they can be removed by raising the heatmap gate or the minimum connected-component size, and analyses restricted to deep, long, or strongly tilted axes can be constructed by filtering the extracted splines on the corresponding geometric attributes.

Finally, the pipeline depends on the input fields only through Z_{500}, u_{500}, and v_{500} on a 1° grid, so it can be applied to other reanalyses and to climate-model output easily, provided they have been regridded accordingly.

5.7 Climatology

The annual and seasonal frequency maps of upper-level troughs and ridges correspond closely to the mean geopotential height field. However, closer examination reveals patterns that are not evident from the mean geopotential field alone and provide information on the underlying synoptic activity. One such pattern emerges from a comparison between the Mediterranean basin and the European landmass. Trough density is consistently higher over the Mediterranean than over most of Europe, even though Europe experiences considerably more rainy days throughout the year. One possible interpretation is that the precipitation-producing systems over Europe are embedded in persistent zonal westerly flow and tend to exhibit less meridional amplification. Consequently, although such systems occur frequently, they less often develop the type of trough identified by

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

our detection method. By contrast, systems over the Mediterranean more frequently take the form of meridionally elongated troughs, which may explain the higher frequency of detected troughs in this region. The decoupling between trough frequency and rainfall frequency is itself a noteworthy result and points to a broader conclusion: a trough identified by our algorithm does not necessarily have a uniform physical meaning across regions and seasons. This point is most clearly illustrated over the EM, where trough frequency exhibits its strongest and most pronounced maximum in summer, despite the region being almost entirely rainless during this season. Ziv et al. (2004) showed that this region coincides with one of the strongest upper-level subsidence maxima, and the most significant minimum of the upper tropospheric moisture content for the Northern Hemisphere covers the EM (Stephens et al., 1996). Thus, despite the high frequency of upper-level troughs, large-scale subsidence strongly suppresses precipitation development, and the summertime trough signal over the EM is not associated with rainfall. The trough, whose location remains nearly stationary during summer, is a relative feature situated between the ridging regions to its east and west. Trough occurrence and precipitation are related but distinct characteristics of the atmospheric flow, and the impact of an approaching upper-level trough on the surface weather depends on both region and season. The frequency maps also reveal a methodological aspect that warrants explicit consideration. Because the detection algorithm was designed to identify wave-like, elongated troughs and ridges, it is less effective at detecting broad, flat systems. For the same reason, polar lows that occasionally reach the northern boundary of the study domain, as well as the persistent high-pressure systems at lower latitudes, are generally not identified. These systems tend to be too broad, rounded, and zonally oriented to be classified as troughs or ridges. We didn't designated a separate label for closed lows or highs. They are represented in our climatology primarily indirectly, through the troughs extending from them. For example, a cut-off low is generally identified as a trough extending equatorward from its centre rather than as a distinct closed entity. Finally, it is important to clarify what a frequency-based climatology of upper-level troughs and ridges adds beyond the information provided by the mean geopotential height field. Rather than showing only the mean locations of troughs and ridges, these maps depict the spatial distribution of the locations through which individual trough and ridge axes actually pass. This information is directly relevant to the degree of precision with which weather phenomena associated with troughs and ridges can be anticipated in a given region. Beyond the climatology presented here, the frequency of occurrence at each grid point makes it possible to identify preferred trough locations and their seasonal displacement. However, this approach alone cannot describe trough life cycles or establish causal relationships with surface impacts. The natural continuation of the present study, and the focus of our future work, is therefore to track trough axes through time and characterize their properties over the EM, including their propagation speed, tilt, amplitude, and meridional extent. Object-based statistics of this kind make it possible to relate atmospheric circulation features to the phenomena they organize, using composite analyses of precipitation, cyclogenesis, dry spells, and heat extremes relative to the positions of trough and ridge axes. Because the same detection algorithm can also be applied to climate-model simulations, the resulting diagnostics may be used both to examine trends in the observational record and to assess projected changes in upper-level flow over the region.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

6 Conclusions

We have presented a differentiable, physically constrained detector of upper-level trough axes for the Mediterranean sector. The model uses Z_{500} input from ERA5, transforms it into a soft-trough mask, and alongside $\boldsymbol{U}_{500}$ it serves as an input to a self-attention model with local receptive field, trained directly on an expert-labelled dataset. On the full 600-scene benchmark, our model significantly outperforms the classical baselines.

The central lesson is that detection of objects as variable as upper-level troughs and ridges is underserved by hard, local, rule-based decisions. Classical detectors succeed in most situations but fail in systematic ways when the geometry departs from the configuration their parameters were tuned for. A neural network generalizes across that variability as it learns directly from expert labels, and embedding a geometric algorithm in the architecture allows this to be achieved without abandoning a physically motivated definition of the object. The result retains the classical method's good predictions and closes much of the gap where it fails.

The detector is not restricted to the domain it was trained on. It carries over qualitatively to mid-latitudes across the globe. It can be adapted to a new region or a new axis type from a handful of additional labels, and its continuous heatmap output allows users to impose their own criteria for which features to retain.

Identifying a trough in a single synoptic situation requires only a brief look by a trained meteorologist; the value of an automated detector lies in the thousands of cases that a reliable climatology demands. We release the model alongside the expert-labelled benchmark and annotation tool, so this work provides a strong foundational basis for a wide range of applications that are available with this tool.

Our next steps are to track axes through time and characterize their properties, and to apply the same framework to climate projections in order to assess how the upper-level flow over the region may change.

Appendix A: Implementation details

A1 Post-processing threshold tuning

Watershed height h, minimum heatmap confidence, and minimum blob size are tuned separately for the trough and ridge detectors. In each case, candidate settings are scored on out-of-fold predictions from the selected model only (10-fold cross-validation: each scene is evaluated by a model that did not train on it). The objective is the same weighted F1–Chamfer composite at $\tau = 5$ grid cells used for classical baseline tuning (Sect. 4.1): two-thirds mean scene F1 plus one-third Chamfer utility (lower Chamfer is better). Once selected, thresholds are frozen and applied to the full trough (600-scene) or ridge (200-scene) evaluation pool. All other extraction settings – weak-wind gating, blob dilation, secondary-axis recovery, and endpoint merging – are shared and were not re-tuned between morphologies.

The selected values are: trough $h = 0.4$, minimum confidence 0.3, minimum blob size 10 pixels; ridge $h = 0.45$, minimum confidence 0.25, minimum blob size 15 pixels. Table B5 lists the full locked recipes.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

A2 Locked axis-extraction settings

All reported neural-model scores use the axis extractor of Sect. 3.4 with a $5\,\mathrm{km}$ weak-wind gate on the heatmap, scene-wise heatmap normalization, one-pixel blob dilation before zero-crossing tracing, secondary-axis recovery, and endpoint merging within 2 grid cells. Trough-specific and ridge-specific watershed and confidence thresholds are given in Table B5.

A3 Classical baseline tuning

For Schemm et al. (2020) we use the curvature, gradient, and morphology thresholds reported in that paper, with pad-then-crop tracing for troughs and a Z-minimum/uphill walk for ridges. The expert-labelled fields store geopotential height in metres, not raw geopotential. The published weak-gradient gate $|\boldsymbol{\nabla} Z|_{\min} = 5 \times 10^{-5}$ applied to those height fields is the setting used for Table 2. A tenfold smaller gate (5×10^{-6}, or $5 \times 10^{-5}/g$), which would be the height-unit equivalent if the original code ingested $\Phi = gZ$, is substantially worse on the 600-scene pool (F1 0.51 vs 0.63). Free parameters for the CVA / Berry and Knippertz re-implementations are tuned with Optuna (Akiba et al., 2019) on a held-out validation subset of the expert labels, optimizing a composite of mean scene F1 at $\tau = 5$ grid cells and mean Chamfer, then frozen before the full 600-scene (trough) or 200-scene (ridge) evaluation. Ridge-specific CVA settings are tuned on the ridge validation subset under the same composite objective. Table A1 lists the values used for the scores in Tables 2 and 3.

Table A1. Baseline detector settings used for the headline comparisons. "Published" means the values reported by the original authors; "tuned" means Optuna on a held-out expert-label validation subset under the same F1–Chamfer composite used for axis-extraction thresholds (Appendix A1).

Method	Origin	Settings used here
Schemm (trough, ridge)	published	Curvature threshold $0.05^\circ\,\mathrm{km}^{-1}$; $\lvert\boldsymbol{\nabla} Z\rvert_{\min} = 5 \times 10^{-5}$; integration distance $5\,\mathrm{km}$; diffusion $n_{\mathrm{filt}} = 50$, $f = 0.2$; minimum object size 20 grid cells. Trough tracing is pad-then-crop; ridge tracing is a Z-minimum/uphill walk.
CVA / Berry (trough)	tuned	Diffusion $n_{\mathrm{filt}} = 80$, $f = 0.2$; cyclonic-curvature mask $K_{1\mathrm{T}}$ at the 64th percentile of positive ζ_{curv}; $K_{2\mathrm{T}} = K_{3\mathrm{T}} = 0$; westerly geostrophic-wind mask; minimum contour 5 points.
CVA / Berry (ridge)	tuned	As trough, except $n_{\mathrm{filt}} = 78$ and $K_{1\mathrm{T}}$ at the 77th percentile of $\lvert\zeta_{\mathrm{curv}}\rvert$ on anticyclonic cells (sign-flipped A1/A2).
Knippertz (trough)	tuned	Native 1° grid (the original method used a 2.5° working grid); latitude window 3 cells; longitude windows 4+3+5 cells; TP threshold $14.7\,\mathrm{gpm}$ (published default $25\,\mathrm{gpm}$); linking 4.1°; minimum chain 2 points.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Appendix B: Hyperparameter and training details

B1 Training loss definitions

Equation (1) in Sect. 3.3 combines three component losses. Let $\mathbf{h}, \mathbf{h}^\star \in [0,1]^{H\times W}$ denote the predicted axis confidence map and its dilated raster target, $\mathbf{z}_\mathrm{s} \in \mathbb{R}^{H\times W}$ the side logits, $\mathbf{y}_\mathrm{s} \in \{0,1\}^{H\times W}$ the CVA-oriented side labels, and $\mathbf{b}^\star \in \{0,1\}^{H\times W}$ the supervision band (radius 5 grid cells around each expert spline). The axis term is mean pixel-wise binary cross-entropy,

$$\mathcal{L}_\mathrm{axis}(\mathbf{h}, \mathbf{h}^\star) = -\frac{1}{HW}\sum_{i,j}\Big[h^\star_{ij}\log h_{ij} + \big(1-h^\star_{ij}\big)\log\big(1-h_{ij}\big)\Big]. \tag{B1}$$

The side term is masked binary cross-entropy with logits, averaged only over band pixels:

$$\mathcal{L}_\mathrm{side}(\mathbf{z}_\mathrm{s}, \mathbf{y}_\mathrm{s}; \mathbf{b}^\star) = \frac{\sum_{i,j} b^\star_{ij}\,\ell_{ij}}{\sum_{i,j} b^\star_{ij} + \varepsilon}, \qquad \ell_{ij} = -y_{\mathrm{s},ij}\log\sigma(z_{\mathrm{s},ij}) - \big(1-y_{\mathrm{s},ij}\big)\log\big(1-\sigma(z_{\mathrm{s},ij})\big), \tag{B2}$$

where $\sigma(\cdot)$ is the logistic function and $\varepsilon = 10^{-7}$. The mask regularizer is the spatial mean of the soft feature mask,

$$\|\mathbf{m}\|_1 = \frac{1}{HW}\sum_{i,j} m_{ij}. \tag{B3}$$

Loss weights for the selected scratch model are listed in Table B1.

When training on Schemm pseudo-labels (pretraining ablation only), the dataset additionally provides region masks and reference curvature fields, enabling two further terms: soft Dice overlap on $\mathbf{m}$ and mean-squared error on α masked by the target region. These terms are inactive under expert spline supervision and are not used in the selected pipeline.

Table B1. Loss weights for the selected scratch model (Eq. (1)).

Term	Weight
$\mathcal{L}_\mathrm{axis}$	$\lambda_\mathrm{h} = 1.0$
$\mathcal{L}_\mathrm{side}$	$\lambda_\mathrm{s} = 1.0$
$\|\mathbf{m}\|_1$	$\lambda_\mathrm{m} = 0.01$

Appendix C: Schemm pseudo-label pretraining

The selected trough and ridge detectors in the main text are trained from scratch on expert spline labels (Sect. 3.3): weights are initialized randomly and optimized end-to-end on the curated annotation pool, without an intermediate training stage. For ablation purposes we also evaluated an optional **short pseudo-label pretraining** stage that is *not* part of the selected production recipe.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Table B2. Differentiable physical front-end constants (Sect. 3.1). The soft minimum-gradient gate is a learnable parameter initialized at 5×10^{-6} on geopotential height. This is not the classical Schemm hard threshold: Table A1 uses the published $|\nabla Z|_{\min} = 5 \times 10^{-5}$ on the same height fields.

Parameter	Value	Learnable
Curvature threshold	$0.05^\circ\,\mathrm{km}^{-1}$	yes
Minimum gradient magnitude	5×10^{-6}	yes
Integration distance	20 km	no
Sigmoid temperature	10	yes
Hole-filling kernel / iterations	5×5 / 5	no
Input smoothing iterations / coefficient	20 / 0.2	no
Additional learnable diffusion	disabled	no

Table B3. Self-attention axis detector hyperparameters (Sect. 3.2).

Parameter	Value
Hidden dimension	64
Attention window size	7×7
Number of attention heads	1
Number of attention blocks	1
Input channels (selected trough model)	7
External predictor fields	Z_{500}, u_{500}, v_{500}
Side supervision enabled	yes

Table B4. Training hyperparameters for the selected scratch model (Sect. 3.3).

Parameter	Value
Training data	expert labels (600 scenes)
Cross-validation	10-fold, pooled; each scene scored out of fold
Initialization	random weights (no Schemm pretrain)
Batch size	16
Learning rate	3×10^{-4}
Max. epochs	100
Early stopping patience	15 (validation loss)
Optimizer	AdamW, cosine annealing
Axis target widening	1 px normal, baked into ground truth
Side supervision band radius	5 grid cells

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Table B5. Locked axis-extraction settings (Sect. 3.4; threshold tuning in Appendix A1). Trough settings apply to the 600-scene benchmark and all trough ablations; ridge settings apply to the 200-scene ridge benchmark.

Parameter	Trough	Ridge
Watershed height h	0.4	0.45
Minimum heatmap confidence	0.3	0.25
Minimum blob size (pixels)	10	15
Shared settings (both morphologies)		
Weak-wind gate	5 kn	
Blob dilation before zero-crossing	1 pixel	
Secondary-axis heatmap floor	= minimum confidence	
Endpoint merge tolerance	2 grid cells	
Climatology period	1979–2025	
Analysis times	00:00 and 12:00 UTC	

C1 Protocol

Pseudo-labels are generated by applying the classical trough-ridge detector of Schemm et al. (2020) to ERA5 Z_{500} fields. For each synoptic time step in a multi-year archive (2005–2015 over the same Mediterranean domain and 1° grid as the expert labels), the algorithm produces axis masks and side-orientation targets. These algorithmic labels supervise the same differentiable Schemm front end and attention-based detector used at inference time, with the full pseudo-label loss including mask-Dice and curvature mean-squared-error terms (Appendix B1). Training runs for a fixed epoch budget on this larger, noisier dataset. The resulting weights are then fine-tuned on the 600-scene expert trough pool under the usual cross-validation protocol. A **scratch** control skips pseudo-label pretraining and initializes weights randomly before expert-label training only.

C2 Main benchmark

On the full 600-scene trough evaluation under locked axis extraction, short pseudo-label pretraining yields a small, non-significant F1 gain over scratch (Table 4); we therefore retain scratch training for the selected model because it matches pretrain skill within confidence intervals and avoids an extra pipeline stage.

The few-shot experiment in Sect. 4.7 tests whether pretraining affects *cross-morphology* adaptation rather than in-domain trough skill. Both the **Ours** and **Ours + pretraining** initializers use the same self-attention architecture (500 hPa wind on, no CVA channel) and the same fold-4 trough checkpoint from the 600-scene ablation before ridge-mode fine-tuning on ten labelled scenes.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Appendix D: Attention and receptive-field ablations

Table D1 compares three random-init attention variants on the full 600-scene pool under identical channels (500 hPa wind on, no CVA input), cross-validation folds, and locked axis extraction. Cross-attention, in which queries are formed from encoded features but keys and values come from a parallel projection, reduces detection and completeness relative to local self-attention. Unified-channel attention (`full_field` mode) lets all seven input channels participate symmetrically in query, key, and value while retaining the same 7×7 neighbourhood; it underperforms self-attention on every reported curve metric.

Table D2 reports a separate receptive-field ablation on the pretrain initialization (the only weights available for both window sizes at the time of the experiment). Replacing the 7×7 window with global $HW \times HW$ self-attention does not improve F1 or Chamfer; Fig. E2 shows that the predicted heatmaps are visually almost identical.

Table D3 reports a window-size sweep on a self-attention configuration without wind or CVA input channels (fold-0, 61 held-out scenes). Local windows of 5×5, 7×7 and 9×9 differ by at most 0.02 in F1; enlarging to 13×13 or replacing the window with global attention does not help. A separate 10-fold scratch sweep that varied the number of attention blocks (1 vs. 2) and attention heads (1, 2 or 4) likewise left F1 within 0.79 to 0.80.

Table D1. Random-init attention ablation at $\tau = 5$ grid cells with locked axis extraction. All variants use random initialization, wind input, and no CVA channel.

Attention mode	F1 ↑	Compl. ↑	Chamfer ↓	AP ↑
Self (local 7×7)	0.84	0.87	1.04	0.67
Cross	0.80	0.80	1.18	0.61
Unified-channel	0.83	0.84	1.08	0.63

Table D2. Local vs. global self-attention receptive field at $\tau = 5$ grid cells (pretrain initialization; otherwise matched recipe).

Receptive field	F1 ↑	Compl. ↑	Chamfer ↓	AP ↑
Local 7×7 window	0.84	0.88	1.01	0.67
Global $HW \times HW$	0.84	0.88	1.04	0.66

Appendix E: Supplementary figures

Code and data availability. This project is fully reproducible and expandable with the code available at https://doi.org/10.5281/zenodo.22388962.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

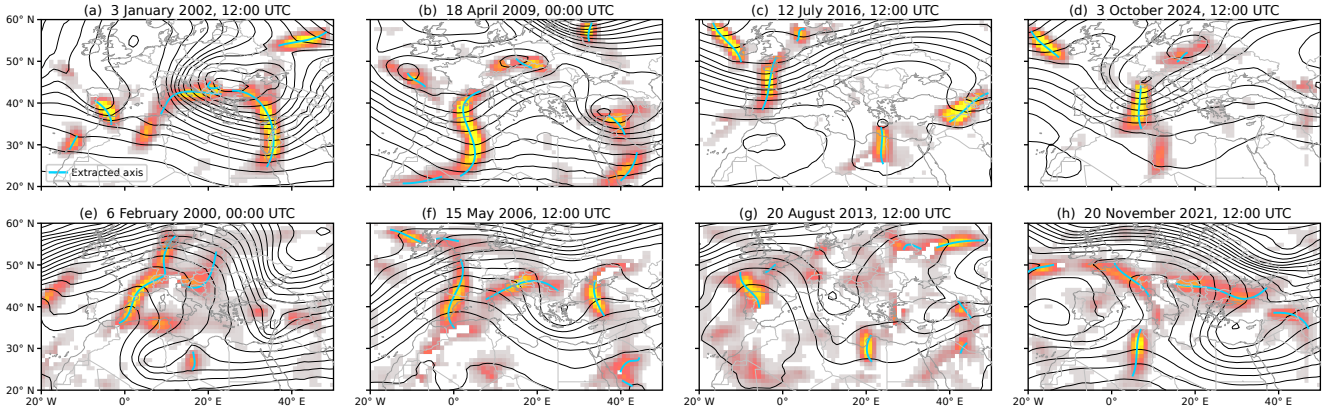

Figure E1. Additional predictions from the selected detector on randomly selected 00:00 and 12:00 UTC dates spanning 2000–2024. Top row: troughs. Bottom row: ridges. Extracted axes are overlaid on the calm-zeroed heatmap; contours are Z_{500}.

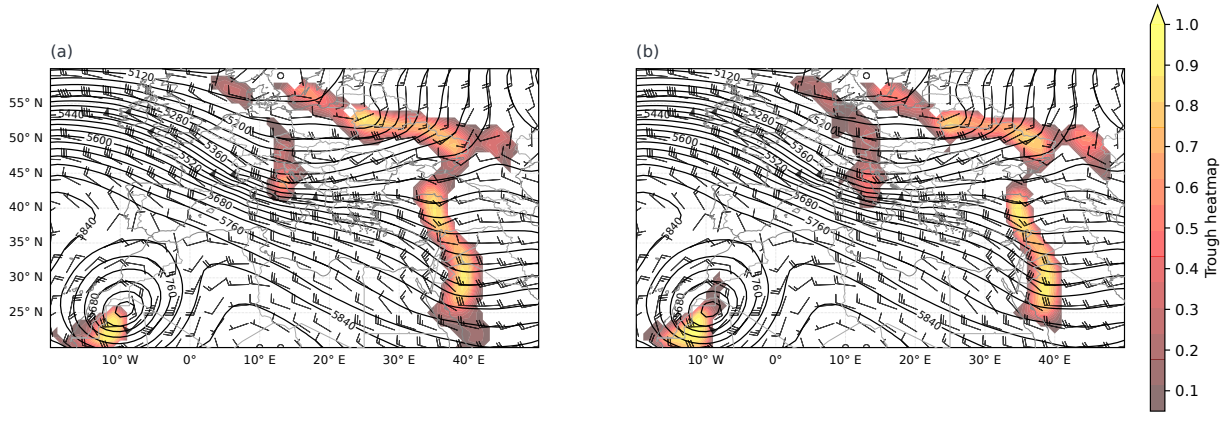

Figure E2. Local 7×7 attention versus global attention on a representative scene. The predicted heatmaps are nearly indistinguishable, consistent with the finding that enlarging the receptive field does not improve skill.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Table D3. Attention-window sweep at $\tau = 5$ grid cells on a self-attention configuration without wind or CVA input channels (fold-0, 61 held-out scenes). Extraction settings follow the in-job monitor recipe, not the locked paper extractor; the comparison is therefore relative within this sweep.

Attention window	F1 ↑	Chamfer ↓
5×5	0.82	1.56
7×7 (selected)	0.83	1.43
9×9	0.82	1.60
13×13	0.81	1.76
Global $HW \times HW$	0.83	1.54

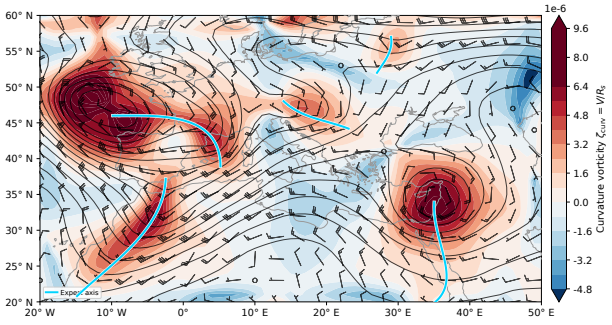

Figure E3. Expert-labelled axes overlaid on the curvature-vorticity field. Labels follow curvature vorticity closely even though annotators did not consult CVA during digitization. Expert axes are identified in the legend.

The new benchmark, consists of 600 expert-labelled trough scenes, 200 expert-labelled ridge scenes, and the derived B-spline ground truth are available from https://doi.org/10.57967/hf/10303.

Model checkpoints are available from https://doi.org/10.57967/hf/10305

Author contributions. O.A. designed and implemented the model and training pipeline, conducted the experiments, produced the figures, and wrote the paper. O.S. conceptualized the AI approach to trough and ridge detection, designed the model and supervised its implementation, and designed the statistical protocol. H.S. and B.Z. supervised the research project, labelled the troughs and ridges, provided meteorological interpretation, and supervised the writing the climatology discussion.

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

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Acknowledgements. This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904.

We acknowledge the use of ERA5 reanalysis data provided by the Copernicus Climate Change Service (C3S) at ECMWF.

We thank Prof. Ori Adam for his helpful advice.

During preparation of this manuscript, the authors used an AI-based language assistants to help refine and proofread the manuscript text and to create the figures. The scientific content, methodology, analysis, and conclusions are entirely the authors' own; the tool was not used to generate results, interpret findings, or write the scientific substance of the paper. The authors reviewed and take full responsibility for all text in the manuscript.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: a next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019, 2623–2631, https://doi.org/10.1145/3292500.3330701, 2019.

Almazroui, M., Kamil, S., Ammar, K., Keay, K., and Alamoudi, A. O.: Climatology of the 500-hPa mediterranean storms associated with Saudi Arabia wet season precipitation, Clim. Dynam., 47, 3029–3042, https://doi.org/10.1007/s00382-016-3011-0, 2016.

Bell, G. D. and Bosart, L. F.: Climatology of Northern Hemisphere 500 mb closed cyclone and anticyclone centers, Mon. Weather Rev., 117, 2142–2163, https://doi.org/10.1175/1520-0493(1989)117<2142:CONHMC>2.0.CO;2, 1989.

Berry, G., Thorncroft, C., and Hewson, T.: African easterly waves during 2004 – analysis using objective techniques, Mon. Weather Rev., 135, 1251–1267, https://doi.org/10.1175/MWR3343.1, 2007.

Bueh, C. and Xie, Z.: An objective technique for detecting large-scale tilted ridges and troughs and its application to an East Asian cold event, Mon. Weather Rev., 143, 4765–4783, https://doi.org/10.1175/MWR-D-14-00238.1, 2015.

Cai, Y., Li, Q., Yin, F., Zhang, L., Huang, H., and Ding, X.: An automatic trough line identification method based on improved UNet, Atmos. Res., 264, 105839, https://doi.org/10.1016/j.atmosres.2021.105839, 2021.

Flaounas, E., Aragão, L., Bernini, L., Dafis, S., Doiteau, B., Flocas, H., Gray, S. L., Karwat, A., Kouroutzoglou, J., Lionello, P., Miglietta, M. M., Pantillon, F., Pasquero, C., Patlakas, P., Picornell, M. Á., Porcù, F., Priestley, M. D. K., Reale, M., Roberts, M. J., Saaroni, H., Sandler, D., Scoccimarro, E., Sprenger, M., and Ziv, B.: A composite approach to produce reference datasets for extratropical cyclone tracks: application to Mediterranean cyclones, Weather Clim. Dynam., 4, 639–661, https://doi.org/10.5194/wcd-4-639-2023, 2023.

Funatsu, B. M., Claud, C., and Chaboureau, J.-P.: A 6-year AMSU-based climatology of upper-level troughs and associated precipitation distribution in the Mediterranean region, J. Geophys. Res.-Atmos., 113, D15120, https://doi.org/10.1029/2008JD009918, 2008.

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.-N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146, 1999–2049, https://doi.org/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 pressure levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/cds.bd0915c6, 2023.

Holton, J. R. and Hakim, G. J.: An Introduction to Dynamic Meteorology, 5th edn., International Geophysics Series, vol. 88, Academic Press, Waltham, MA, USA, 552 pp., ISBN 9780123848666, 2012.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V07, NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], https://doi.org/10.5067/GPM/IMERG/3B-HH/07, 2023.

Kautz, L.-A., Martius, O., Pfahl, S., Pinto, J. G., Ramos, A. M., Sousa, P. M., and Woollings, T.: Atmospheric blocking and weather extremes over the Euro-Atlantic sector – a review, Weather Clim. Dynam., 3, 305–336, https://doi.org/10.5194/wcd-3-305-2022, 2022.

Keable, M., Simmonds, I., and Keay, K.: Distribution and temporal variability of 500 hPa cyclone characteristics in the Southern Hemisphere, Int. J. Climatol., 22, 131–150, https://doi.org/10.1002/joc.728, 2002.

https://doi.org/10.5194/egusphere-2026-5334
Preprint. Discussion started: 23 September 2026

Knippertz, P.: A simple identification scheme for upper-level troughs and its application to winter precipitation variability in northwest Africa, J. Climate, 17, 1411–1418, https://doi.org/10.1175/1520-0442(2004)017<1411:ASISFU>2.0.CO;2, 2004.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N.: Big Transfer (BiT): general visual representation learning, in: Computer Vision – ECCV 2020, Glasgow, UK, 23–28 August 2020, Lecture Notes in Computer Science, vol. 12350, 491–507, https://doi.org/10.1007/978-3-030-58558-$7_2 9, 2020.

Lefevre, R. J. and Nielsen-Gammon, J. W.: An objective climatology of mobile troughs in the Northern Hemisphere, Tellus A, 47, 638–655, https://doi.org/10.3402/tellusa.v47i5.11558, 1995.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A.: ClimaX: a foundation model for weather and climate, in: Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023, Proceedings of Machine Learning Research, 202, 25904–25938, https://proceedings.mlr.press/v202/nguyen23a.html, 2023.

Okajima, S., Nakamura, H., and Kaspi, Y.: Cyclonic and anticyclonic contributions to atmospheric energetics, Sci. Rep., 11, 13202, https://doi.org/10.1038/s41598-021-92548-7, 2021.

Piva, E. D., Gan, M. A., and Rao, V. B.: An objective study of 500-hPa moving troughs in the Southern Hemisphere, Mon. Weather Rev., 136, 2186–2200, https://doi.org/10.1175/2007MWR2135.1, 2008.

Rasp, S. and Thuerey, N.: Data-driven medium-range weather prediction with a Resnet pretrained on climate simulations: a new model for WeatherBench, J. Adv. Model. Earth Syst., 13, e2020MS002405, https://doi.org/10.1029/2020MS002405, 2021.

Saaroni, H., Ziv, B., Lempert, J., Gazit, Y., and Morin, E.: Prolonged dry spells in the Levant region: climatologic-synoptic analysis, Int. J. Climatol., 35, 2223–2236, https://doi.org/10.1002/joc.4143, 2015.

Saaroni, H., Ziv, B., Harpaz, T., and Lempert, J.: Dry events in the winter in Israel and its linkage to synoptic and large-scale circulations, Int. J. Climatol., 39, 1054–1071, https://doi.org/10.1002/joc.5862, 2019.

Sanders, F.: Life history of mobile troughs in the upper westerlies, Mon. Weather Rev., 116, 2629–2648, https://doi.org/10.1175/1520-0493(1988)116<2629:LHOMTI>2.0.CO;2, 1988.

Schemm, S., Rüdisühli, S., and Sprenger, M.: The life cycle of upper-level troughs and ridges: a novel detection method, climatologies and Lagrangian characteristics, Weather Clim. Dynam., 1, 459–479, https://doi.org/10.5194/wcd-1-459-2020, 2020.

Spanos, S., Maheras, P., Karacostas, T., and Pennas, P.: Objective climatology of 500-hPa cyclones in central and east mediterranean region during warm-dry period of the year, Theor. Appl. Climatol., 75, 167–178, https://doi.org/10.1007/s00704-003-0726-8, 2003.

Stephens, G. L., Jackson, D. L., and Wittmeyer, I.: Global observations of upper-tropospheric water vapor derived from TOVS radiance data, J. Climate, 9, 305–326, https://doi.org/10.1175/1520-0442(1996)009<0305:GOOUTW>2.0.CO;2, 1996.

Wiedemann, C., Heipke, C., Mayer, H., and Jamet, O.: Empirical evaluation of automatically extracted road axes, in: Empirical Evaluation Techniques in Computer Vision, edited by: Bowyer, K. W. and Phillips, P. J., IEEE Computer Society Press, Los Alamitos, CA, USA, 172–187, ISBN 0818684011, 1998.

Zangvil, A. and Druian, P.: Upper air trough axis orientation and the spatial distribution of rainfall over Israel, Int. J. Climatol., 10, 57–62, https://doi.org/10.1002/joc.3370100107, 1990.

Ziv, B., Saaroni, H., and Alpert, P.: The factors governing the summer regime of the eastern Mediterranean, Int. J. Climatol., 24, 1859–1871, https://doi.org/10.1002/joc.1113, 2004.