atmosphere

MDPI

Article

A Deep Learning Approach to Detecting Atmospheric Rivers in the Arctic

Sinéad McGetrick [1,2,*], Hua Lu [2,*], Grzegorz Muszynski [3], Oscar Martínez-Alvarado [4], Matthew Osman [5], Kyle Mattingly [6] and Daniel Galea [7]

[1] Department of Applied Maths and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK
[2] British Antarctic Survey, Cambridge CB3 0ET, UK
[3] School of Geosciences, University of Edinburgh, Edinburgh EH8 9YL, UK
[4] National Centre for Atmospheric Science, Department of Meteorology, University of Reading, Reading RG6 6BB, UK
[5] Department of Geography, University of Cambridge, Cambridge CB2 3EN, UK
[6] Space Science and Engineering Center, University of Wisconsin–Madison, Madison, WI 53706, USA
[7] Independent Researcher, London HP22 6NJ, UK
* Correspondence: sineadmcgetrick@gmail.com (S.M.); hlu@bas.ac.uk (H.L.)

Abstract

The Arctic is warming rapidly, with atmospheric rivers (ARs) amplifying ice melt, extreme precipitation, and abrupt temperature shifts. Detecting ARs in the Arctic remains challenging, because AR detection algorithms designed for mid-latitudes perform poorly in polar regions. This study introduces a regional deep learning (DL) image segmentation model for Arctic AR detection, leveraging large-ensemble (LE) climate simulations. We analyse historical simulations from the *Climate Change in the Arctic and North Atlantic Region and Impacts on the UK* (CANARI) project, which provides a large, internally consistent sample of AR events at 6-hourly resolution and enables a close comparison of AR climatology across model and reanalysis data. A polar-specific, rule-based AR detection algorithm was adapted to label ARs in simulated data using multiple thresholds, providing training data for the segmentation model and supporting sensitivity analyses. U-Net-based models are trained on integrated water vapour transport, total column water vapour, and 850 hPa wind speed fields. We quantify how AR identification depends on threshold choices in the rule-based algorithm and show how these propagate to the U-Net-based models. This study represents the first use of the CANARI-LE for Arctic AR detection and introduces a unified framework combining rule-based and DL methods to evaluate model sensitivity and detection robustness. Our results demonstrate that DL segmentation achieves robust skill and eliminates the need for threshold tuning, providing a consistent and transferable framework for detecting Arctic ARs. This unified approach advances high-latitude moisture transport assessment and supports improved evaluation of Arctic extremes under climate change.

Keywords: atmospheric rivers; Arctic; integrated water vapour transport; deep learning; image segmentation; U-Net; large-ensemble climate model simulations; ERA5; MERRA-2

Academic Editor: Stephan Havemann

Received: 20 November 2025
Revised: 19 December 2025
Accepted: 30 December 2025
Published: 1 January 2026

1. Introduction

The Arctic is warming nearly four times faster than the global average [1], a phenomenon known as Arctic amplification. Rising temperatures increase atmospheric moisture, driving enhanced precipitation, accelerated sea-ice loss, and more frequent extremes

with global repercussions. Among the key drivers of these changes are atmospheric rivers (ARs), which are long, narrow corridors of strong horizontal water vapour transport typically associated with the low-level jet ahead of the cold front in an extratropical cyclone, as defined by the American Meteorological Society [2].

ARs play a key role in transporting moisture into the Arctic and are frequently associated with episodes of enhanced precipitation, winter warm spells, and rain-on-snow events. Although ARs occur much less frequently in the Arctic than their mid-latitude counterparts [3], they contribute disproportionately to Arctic moistening, accounting for approximately 90% of poleward moisture transport to the Arctic [4]. Lee et al. [5] demonstrated that Arctic amplification during 1989–2009 was primarily driven by enhanced downward longwave radiation linked to increased moisture transport via Arctic ARs (AARs). Luo et al. [6] further showed that North Atlantic sea surface temperature (SST) anomalies and high-latitude blocking facilitate warm, moist air intrusions into the Arctic, particularly in the Barents–Kara Seas, amplifying downward infrared radiation and accelerating sea-ice loss. Studies based on reanalysis and satellite-era observations indicate that AARs have exhibited increasing frequency and intensity in recent decades. For example, Wang et al. [4] show that AARs have become more frequent and intense during summer and have contributed approximately 36% of the observed increase in Arctic atmospheric moisture since 1979. Similarly, Zhang et al. [7] report that early-winter ARs have increased in occurrence and have slowed sea-ice recovery in the Barents–Kara Seas, accounting for roughly one-third of the regional winter sea-ice decline. Consistent with these findings, Ma et al. [8] show that wintertime extreme warming events in the high Arctic occur almost exclusively under atmospheric river conditions and have increased in frequency, duration, and intensity over the past four decades, concurrent with wintertime warming of about 0.8 °C per decade. These findings highlight that the increased moisture import from the lower latitudes plays a pivotal role in Arctic energy balance and cryosphere change.

Beyond their climatological representation, AARs critically influence Arctic sea-ice melt and ice-sheet mass balance, with far-reaching impacts on ecosystems, livelihoods, and infrastructure across polar regions [9,10]. For instance, Gong et al. [11] showed that AARs have emerged as a major driver of sea-ice thinning since 2000, accounting for approximately 44% of total melt through a combination of pre-entry warming, dynamic and thermodynamic impacts during peak intensity, and a prolonged thermodynamic decay phase. Similar to Mattingly et al. [12], their rule-based detection applied a climatological vertically integrated water vapour transport (IVT) percentile threshold to identify anomalous moisture transport. However, to isolate strong, persistent, and spatially coherent events most relevant to sea-ice impacts, Gong et al. [11] adopted a substantially higher absolute IVT threshold of 400 kg m^{-1} s^{-1} and imposed a minimum event duration of 24 h. By contrast, the high-latitude AR algorithm by Mattingly et al. [12] employed a lower raw IVT threshold of 150 kg m^{-1} s^{-1}, allowing for the inclusion of weaker and shorter-lived filaments. Together, these studies underscore the growing role of AARs, the importance of accurate representation given methodological sensitivities, and the need for a robust detection framework.

Despite their importance, AARs remain poorly characterised. It is worth noting that the AMS definition of AR is qualitative rather than quantitative, leading to differences among AR detection algorithms that vary in their interpretation. Furthermore, most AR detection algorithms were developed for mid-latitudes; they perform poorly in polar environments, where atmospheric structure differs, storms are more transient, and observational coverage is limited [7,12,13]. Results from the Atmospheric River Tracking Method Intercomparison Project (ARTMIP) highlight substantial discrepancies among widely used rule-based methods, with particularly large uncertainty in the Arctic [13,14].

For example, comparisons across multiple detection algorithms show that estimated AAR frequency, duration, and intensity can differ by up to a factor of five, underscoring major inconsistencies in AR identification in polar regions [15].

Similar challenges have been documented in the Antarctic, where AR detection is likewise sensitive to algorithm design and threshold choice [16]. Some rule-based approaches attempt to address these issues by adapting IVT thresholds or applying additional geometric constraints for polar regions [12,17,18]. However, even with such adaptations, detected AAR characteristics remain sensitive to methodological choices, particularly for weaker or short-lived events. Together, these findings emphasise the need for coherent, region-specific approaches to AAR identification in order to robustly quantify their role in Arctic moisture transport, surface energy balance, and cryosphere change.

Deep learning (DL) models have recently emerged as a promising alternative for AR detection. Segmentation models such as ARDetect, based on a U-Net architecture, have achieved strong performance in global applications when trained on ARs labelled by rule-based algorithms [19,20]. Galea and Ma [21] compared multiple deep learning architectures for AR prediction, demonstrating the growing potential of DL approaches while also underscoring the importance of model choice and training data.

Recent advances in DL segmentation models highlight both opportunities and challenges for Arctic application. Table 1 summarises representative approaches, illustrating their strengths and limitations. While these models have proven effective for global or mid-latitude ARs, none have yet been designed or evaluated specifically for polar environments.

Table 1. Comparison of DL segmentation models for AR detection, highlighting strengths, limitations, and degree of adaptation.

Model	Strengths	Limitations	AR Use
ARDetect [19]	Purpose-built; U-Net style; strong performance	No polar tuning	Global detection
ARCNN (ARCI) [22]	Trained on ARTMIP consensus; reproducible	No polar tuning	Reanalysis-based detection
DeepLabV3+ [23,24]	High accuracy; pretrained weights	Not AR-specific; high memory demand	ClimateNet, general tasks
U-Net++ [25]	Robust to sparse/noisy labels; modular	Complex; limited climate use	Potential for polar segmentation
HRNet [26]	Maintains high-resolution features	Rarely used in climate DL; complex	Potential for polar segmentation

A major challenge in applying DL to detect AARs lies in the scarcity of high-quality observational and labelled training data. Reliable ground-based and satellite observations across the Arctic are also largely limited to the satellite era beginning in 1979 [7,13]. While reanalyses such as ERA5 and MERRA-2 have been widely used to investigate ARs, relatively few studies have examined how features are represented within climate model simulations. Analyses of CMIP5 and CMIP6 ensembles reveal substantial inter-model spread in both the frequency and spatial distribution of AARs, particularly during summer [27,28]. Espinoza et al. [27] showed that although CMIP5 models broadly reproduce the intensification of ARs under a warming climate, they exhibit large regional and seasonal inconsistencies. Building on this, Zhang et al. [28] found that CMIP6 models tend to underestimate AAR frequency over the North Pacific during winter but overestimate it over the North Atlantic during summer, highlighting persistent uncertainties in modelled moisture transport into the Arctic. The limited record and inconsistency among CMIP models restrict both the ability to characterise AAR behaviour and the supply of consistent training samples for supervised DL model development and evaluation.

Single-Model Initial-condition Large Ensemble (SMILE) climate datasets provide an opportunity to address these challenges by offering physically consistent atmosphere–ocean–ice coupled simulations tailored to study extreme events such as AARs. Such datasets enable a more accurate learning of AR features in data-sparse regions, improving both detection skill and uncertainty quantification. For example, Eyring et al. [29] showed that SMILE-based datasets facilitate advanced machine learning for extreme-event detection

and uncertainty estimation, while Higgins et al. [30] used a DL segmentation model to track ARs in a high-resolution large-ensemble simulation, demonstrating the efficiency and adaptability of this approach. In this study, a similar methodology is applied to investigate ARs in the Arctic.

Here, we use the large-ensemble model simulations generated by *the Climate Change in the Arctic and North Atlantic Region and Impacts on the UK* (CANARI) project to build a framework for Arctic-focused AAR detection. Also, for the first time, this CANARI large ensemble (CANARI-LE) is used to examine AARs in terms of the regional distribution, seasonal variation, and sensitivity to threshold values. Specifically, we (i) leverage the CANARI-LE as a regional, ensemble-rich resource for AAR analysis and sensitivity testing; (ii) compare AAR climatology derived from the CANARI-LE with existing ERA5 and MERRA-2 AAR catalogues to assess consistency across model and reanalysis data; and (iii) develop and evaluate a U-Net-based segmentation model that removes the need for threshold tuning while capturing high-latitude moisture transport structures. Together, these contributions establish a unified, reproducible framework for AAR detection and evaluation across models and observations.

2. Data and Methods

This study comprises three main components: preparation of physically consistent predictor fields from a large-ensemble climate model, construction of an AAR label set using a rule-based detector adapted for high latitudes, and development and evaluation of a convolutional encoder–decoder segmentation model that performs grid-cell-wise AAR detection. To explicitly assess how label definition propagates into DL behaviour, we construct two AAR label sets (using the *Default* and *Intermediate* thresholds; Table 2) and train two otherwise identical segmentation models on each set. We refer to these as the *Default* and *Intermediate* models throughout.

Table 2. Threshold configurations evaluated for the rule-based AAR detector. IVT_Thresh is the absolute IVT magnitude threshold (kg m^{-1} s^{-1}); IVT_PR_Thresh is a climatological percentile threshold; min_num_grid_points is the minimum contiguous area; min_length is the minimum object length (km); min_length_width_ratio is the minimum length-to-width ratio required for an object to be considered filamentary; and v_poleward_cutoff_lat defines the latitude beyond which the poleward flow criterion is relaxed (70° N).

Parameter	Strictest	Intermediate	Default	Most Permissive
IVT_Thresh	250	150	150	100
IVT_PR_Thresh (percentile)	95	90	85	85
min_num_grid_points	180	180	150	150
min_length	1500	2000	1500	1500
min_length_width_ratio	1.5	2.0	1.5	1.5
v_poleward_cutoff_lat (°N)	70	70	70	70

2.1. Datasets

We use the CANARI-LE, a newly released large-ensemble climate simulation, to train our DL segmentation models. The CANARI-LE consists of outputs from the global HadGEM3-GC3.1-MM climate model, configured identically to the CMIP6 HighResMIP experiments [31]. Studies have shown that HadGEM3-GC3.1-MM performs well in representing key circulation features, including storm tracks and atmospheric blocking [32]. With forty ensemble members, comparatively high resolution, and six decades of sub-daily output, the CANARI-LE enables the study of high-impact weather systems and captures synoptic-scale variability critical to Arctic moisture transport. This extensive and physically consistent dataset pro-

vides thousands of AAR samples, forming an unprecedented basis for feature detection, DL model training, and sensitivity testing.

More specifically, the dataset that we used is the CANARI-LE historical simulation, which spans 1950–2014 using a 360-day calendar and follows the CMIP6 historical forcing protocol, which prescribes observed changes in greenhouse gas concentrations, aerosols, ozone, solar variability, volcanic eruptions, and land-use changes to ensure consistency with past climate conditions [33]. A macro-initialisation produces five members, which are then stochastically perturbed to generate eight members each for a total of forty members. This method introduces ensemble spread in the ocean state, enabling the robust sampling of internal variability and extreme events, e.g., AARs. The model employs an atmospheric resolution of approximately 60 km and an ocean resolution of 0.25°, sufficient to resolve synoptic-scale moisture transport. A 6-hourly output north of 40° N is used to capture mid-latitude filaments that extend into the Arctic, as previous studies suggest that the majority of poleward moisture transport originates from the eastern North Atlantic, with an uptake maximum poleward of 50° N [34].

The atmospheric fields from the CANARI-LE are defined on the native N216 regular latitude–longitude grid of HadGEM3-GC3.1, with an approximate horizontal resolution of 60 km at mid-latitudes. As with any regular latitude–longitude grid, grid-cell area varies with latitude, decreasing towards the pole. All analyses in this study are performed on this native grid, and grid-cell-based statistics are, therefore, not equal-area measures. However, because all threshold configurations and datasets are evaluated consistently on the same grid, grid-cell counts provide a robust basis for comparing relative differences in AR morphology and detection behaviour.

For independent benchmarking, we use two global reanalyses, ERA5 [35] and MERRA-2 [36]. ERA5, produced by the European Centre for Medium-Range Weather Forecasts, assimilates a wide range of observations using a four-dimensional variational scheme in the Integrated Forecasting System model, providing hourly atmospheric fields at 31 km resolution from 1940 to the present. In this study, only the period since 1979 was used to ensure consistency with the satellite-observation era and the availability of high-quality moisture and wind fields. MERRA-2, from NASA's Global Modeling and Assimilation Office, employs the GEOS-5 satellite model with radiance assimilation to generate a hydrologically consistent global reanalysis from 1980 to the present. Both datasets provide dynamically consistent moisture and wind fields widely used in AR studies.

2.2. Predictor Variables and Preprocessing

Five predictors were derived at each timestep: IVT magnitude, its zonal and meridional components (uIVT, vIVT), total column water vapour (TCWV), and wind speed at 850 hPa. IVT and its components follow the standard vertically integrated formulation [37] using the available pressure levels in the CANARI-LE dataset (925, 850, 700, 600, 500, 300, 250, 200, and 50 hPa). The 850 hPa wind field was included as a predictor because of its strong association with low-level moisture transport and orographic precipitation enhancement [38]. These variables collectively describe the magnitude, direction, and thermodynamic environment of moisture transport relevant to AARs.

All predictors were collocated on the CANARI-LE grid and temporally aligned with the labels. To place variables on comparable scales while preserving spatial gradients, each predictor was min–max-scaled to the interval $[0, 1]$ independently at every timestep, with a small stabilising constant of 10^{-8} added to the denominator. Ensemble members 1–28 were reserved for model training and 29–40 for final testing.

2.3. Rule-Based AAR Labels

AAR labels were generated using the algorithm by Mattingly et al. [12], adapted from reanalysis to the CANARI-LE data structure and 360-day calendar. The procedure follows the standard approach of identifying contiguous regions of high IVT that exceed both an absolute value and a climatological threshold, followed by filtering based on geometric criteria. North of 70° N, the directional constraint (parameter v_poleward_cutoff_lat) was relaxed, meaning that features are not required to have poleward v-wind or vIVT. This ensures that high-latitude filaments with zonal or equatorward moisture transport are not excluded.

To examine the sensitivity of detected AARs to threshold choice, four configurations were evaluated, ranging from permissive to strict (Table 2). The four configurations were designed to represent qualitatively distinct regimes rather than a strictly monotonic progression across all parameters. For DL experiments, two label sets were retained: (i) the *Default* configuration, corresponding to that used by Mattingly et al. [12] for ERA5 and MERRA-2 catalogues, and (ii) an *Intermediate* configuration that applies slightly stricter IVT and geometric thresholds. The *Default* configuration generates frequent AAR detections, providing many positive samples that support model training, but tends to over-identify diffuse or implausible shapes. In contrast, the *Intermediate* configuration yields more coherent and physically realistic AAR filaments but introduces stronger class imbalance, with AAR grid cells comprising less than 3% of the dataset.

Because the CANARI-LE provides 40 members, offering more than 40 times the sampling available from a single reanalysis, this large training base allows for the exploration of stricter thresholds without compromising statistical robustness. Therefore, these two configurations define complementary label regimes to assess how threshold design influences segmentation performance. The corresponding models are referred to as *Default* (trained on the more permissive labels) and *Intermediate* (trained on the stricter "intermediate" labels) in the following sections.

2.4. Segmentation Model

A convolutional encoder–decoder network with U-Net-style skip connections [39] was used for grid-cell-wise segmentation of AARs, following the design principles of ARDetect [19]. The network ingests five predictor channels and produces a binary mask of AAR presence at the native grid resolution. Skip connections transfer fine-scale spatial information from the encoder to the decoder, which reconstructs the output mask by upsampling and merging features across multiple scales.

2.5. Training Procedure

The model was implemented using a custom-built training pipeline. Each input sample comprised five channels, IVT, uIVT, vIVT, TCWV, and 850 hPa wind speed, paired with a binary AAR mask (Figure 1). Data were supplied as 6-hourly NetCDF files and divided into independent development (members 1–28) and test (members 29–40) subsets.

The model was trained to minimise a composite loss function, $\mathcal{L}_{\text{total}}$, consisting of Dice loss [40] and focal loss [41]:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Focal}}, \tag{1}$$

where

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i P_i G_i + \epsilon}{\sum_i P_i + \sum_i G_i + \epsilon} \tag{2}$$

and

$$\mathcal{L}_{\text{Focal}} = -\alpha(1 - P_t)^{\gamma} \log(P_t). \tag{3}$$

where P and G denote the predicted and ground-truth masks, P_t is the predicted probability for the true class, α controls class weighting, and γ adjusts the emphasis on difficult examples. Using both loss functions helps the model capture AAR shapes more accurately and prevents the abundant background grid cells from dominating the optimisation.

Hyperparameters were optimised using Ray Tune with the Asynchronous Successive Halving Algorithm for early stopping [42]. Learning rate, dropout rate, number of convolutional filters per layer, batch size, and focal loss parameters (α, γ) were tuned automatically. Trials were evaluated on a dedicated validation set (a 20% random subset of the training set), and training stopped when the Dice score did not improve for ten consecutive epochs (with each epoch representing one complete pass through the training dataset). Learning rates between 1×10^{-5} and 5×10^{-4}, dropout rates between 0.1 and 0.5, and batch sizes between 8 and 32 were tested. All runs used GPU acceleration on the JASMIN cluster, the UK's collaborative data analysis environment for environmental science [43].

For each epoch, performance was evaluated using grid-cell-wise counts of true positives (TP), false positives (FP), and false negatives (FN). From these, the Dice score, precision, and recall were computed:

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

where the Dice score was the principal optimisation criterion, and precision and recall were used to quantify false-alarm control and detection sensitivity.

All experiments were run with fixed random seeds to ensure reproducibility. The final model configuration, demonstrating consistent validation performance and numerical stability, was selected for evaluation on the CANARI-LE. Training diagnostics were logged with Weights & Biases [44], recording loss evolution, validation metrics, and sample predictions.

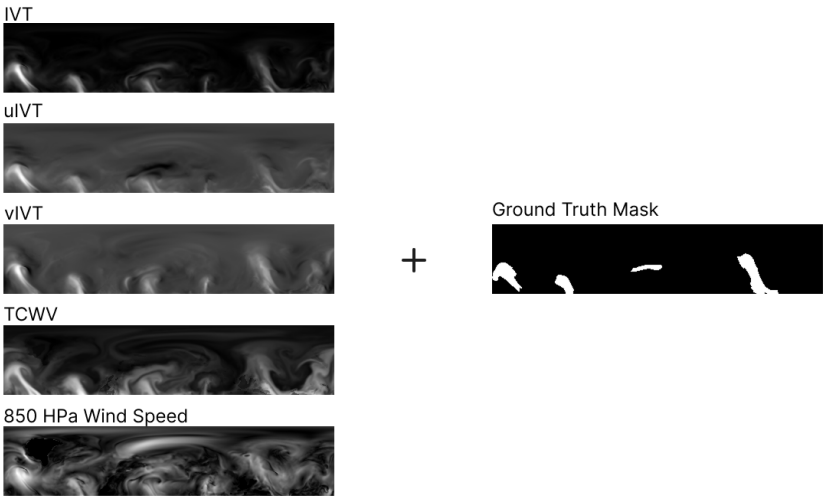

Figure 1. Example model input comprising min–max-scaled IVT, uIVT, vIVT, TCWV and 850 hPa wind speed, paired with a binary AAR mask.

2.6. Evaluation Metrics and Testing

Grid-cell-wise comparisons between predictions and labels provided counts of true positives, false positives, and false negatives. Evaluation metrics included the Dice score for spatial agreement, precision for false-alarm control, and recall for missed detections. An overprediction ratio, defined as the fraction of predicted AAR grid cells absent from the labels, was used as a reliability measure. Spatial frequency maps of predictions and residuals were also examined to identify systematic regional biases. These metrics allow

for a direct comparison between the DL and rule-based detections across the CANARI-LE, ERA5, and MERRA-2.

3. Results

This section presents the main findings from the rule-based and DL detection of AARs in the CANARI-LE. The first part establishes the AAR climatology in the CANARI-LE and compares it with ERA5 and MERRA-2 under identical rule-based logic. The second examines how varying the detection thresholds affects AAR frequency and morphology. The final sections evaluate segmentation model performance and the sensitivity of DL detection to the definition of training labels.

3.1. Integrated Water Vapour Transport Bias in CANARI-LE

Before comparing AAR detections between datasets, it is useful to assess the representation of IVT in the CANARI-LE. The mean IVT from the CANARI-LE was compared with the MERRA-2 and ERA5 reanalyses for the period 2000–2014 (Figure 2). Spatial patterns are broadly consistent across datasets, with maxima over the North Atlantic and North Pacific corresponding to the major storm-track corridors. However, the CANARI-LE systematically underestimates IVT magnitude relative to both reanalyses, particularly north of 70° N and over Greenland and the central Arctic Ocean.

The seasonal climatologies show that this low-moisture bias is the strongest in summer (JJA) and autumn (SON), when poleward moisture transport typically peaks. Modest regional overestimations occur around the Greenland coast and northwest Alaska, which may arise from differences in topographic resolution and the weaker representation of high-latitude atmospheric blocking in the CANARI-LE. In reanalysis datasets, such blocking can suppress meridional transport, whereas its underrepresentation in many climate models can lead to biases in IVT magnitude and structure [45,46].

Recent work by Gao et al. [47] shows that increasing model resolution substantially improves the simulation of atmospheric blocking, particularly for long-lived events exceeding ten days, through improved representation of baroclinic eddies, SST gradients, and orographic effects. Given that the CANARI-LE employs an intermediate resolution (approximately 60 km for the atmosphere and 0.25° for the ocean), it is likely to share some of the blocking-related biases identified in lower-resolution models, which can weaken blocking intensity and associated circulation patterns and, in turn, affect poleward moisture transport and IVT fields.

Overall, the CANARI-LE reproduces the large-scale structure of Arctic moisture pathways but with a general dry bias of 20–40 kg m^{-1} s^{-1}. This systematic offset provides important context for interpreting subsequent differences in AAR frequency among the datasets.

3.2. Rule-Based AARs in the CANARI-LE and Comparison with Reanalyses

Application of the detection algorithm by Mattingly et al. [12] to the CANARI-LE produces spatially coherent AAR structures aligned with the major North Atlantic and North Pacific moisture pathways into the Arctic. The comparison with ERA5 and MERRA-2 in Figure 3 shows that all datasets capture the dominant storm-track corridors and the decline in occurrence toward higher latitudes.

Seasonal patterns are also illustrated in Figure 3, which compares the CANARI-LE with MERRA-2 across the four meteorological seasons, together with their differences and corresponding zonal-mean profiles (60–90° N).

The spatial distribution of AARs is broadly consistent between datasets, with maxima over the North Atlantic and North Pacific in all seasons. The CANARI-LE systematically

underestimates AAR frequency, most notably in summer (JJA) and autumn (SON), when poleward moisture transport is the strongest. The seasonal difference maps highlight a widespread negative bias across the polar cap, while the zonal means confirm that this bias increases towards the pole. These results are consistent with the IVT underestimation identified in Section 3.1 and indicate that the CANARI-LE captures the structure and timing of Arctic moisture transport but with reduced intensity relative to reanalysis.

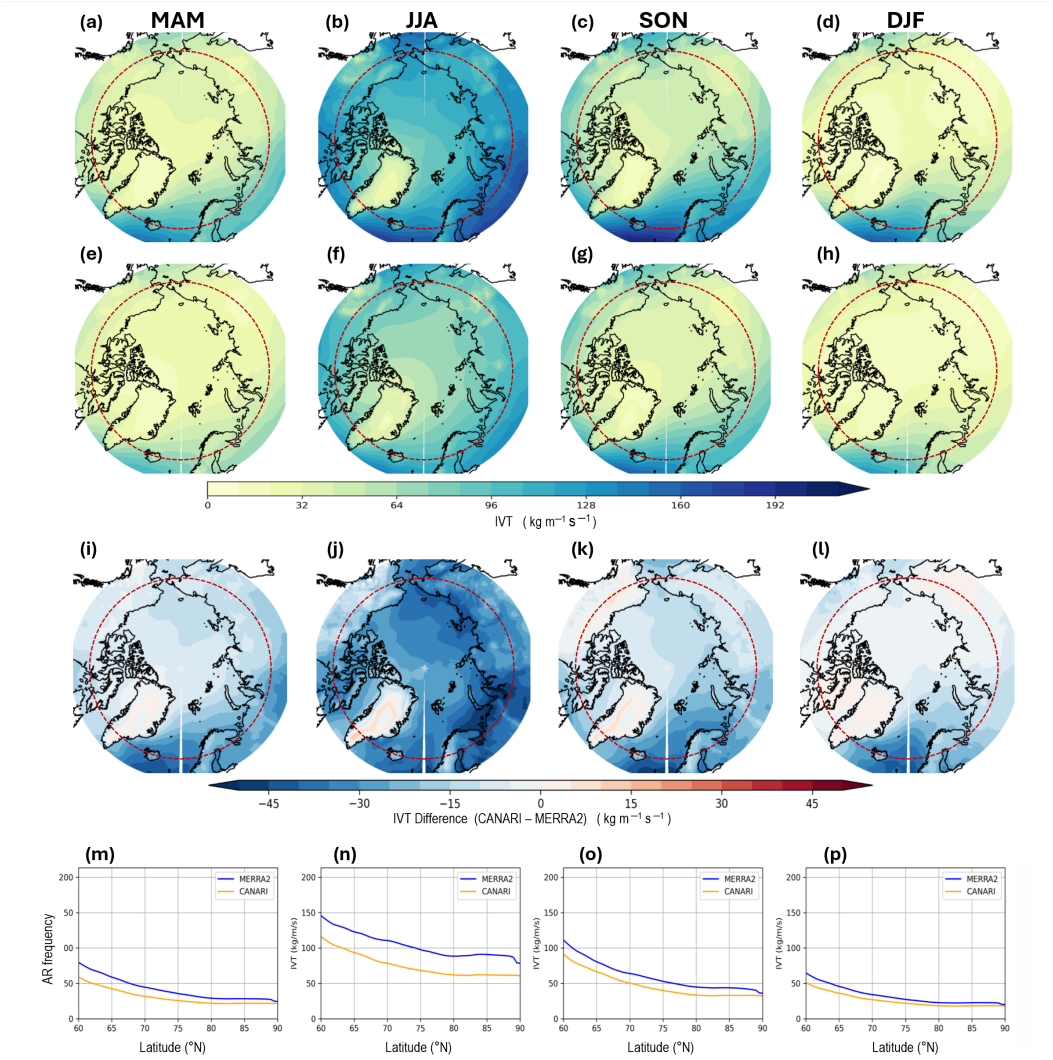

Figure 2. Seasonal mean IVT from MERRA-2 (**a**–**d**) and CANARI-LE (**e**–**h**) averaged over 2000–2014, the IVT difference (CANARI-LE – MERRA-2) (**i**–**l**), and corresponding zonal-mean IVT and AAR frequency at 60°–90° N (**m**–**p**). Blue shading indicates regions where CANARI-LE is drier than MERRA-2. The red-dashed circle marks the Arctic Circle.

3.3. Sensitivity of Rule-Based Detection to Threshold Choice

The rule-based algorithm's behaviour depends on the IVT and geometric thresholds applied. Figure 4 illustrates their influence using a representative timestep on 6 June 2012. Stricter settings retain only narrow, intense filaments, whereas permissive thresholds include broader plumes and diffuse extensions. The zonal-mean response across seasons (Figure 3m–p) shows that sensitivity is the highest in summer and autumn, when moisture transport is the strongest. More permissive configurations converge towards the reanalysis frequencies, while stricter ones underdetect weaker events.

Based on these experiments, two threshold regimes were selected for the DL analysis: the more permissive *Default* configuration and the stricter *Intermediate* configuration.

The *Default* labels provide dense, easily learned training data but may include marginal or fragmented events, whereas the *Intermediate* labels emphasise physical realism and event coherence. Leveraging the extensive sampling of the CANARI-LE makes it feasible to train on both, enabling a controlled test of how label strictness influences segmentation skill and bias.

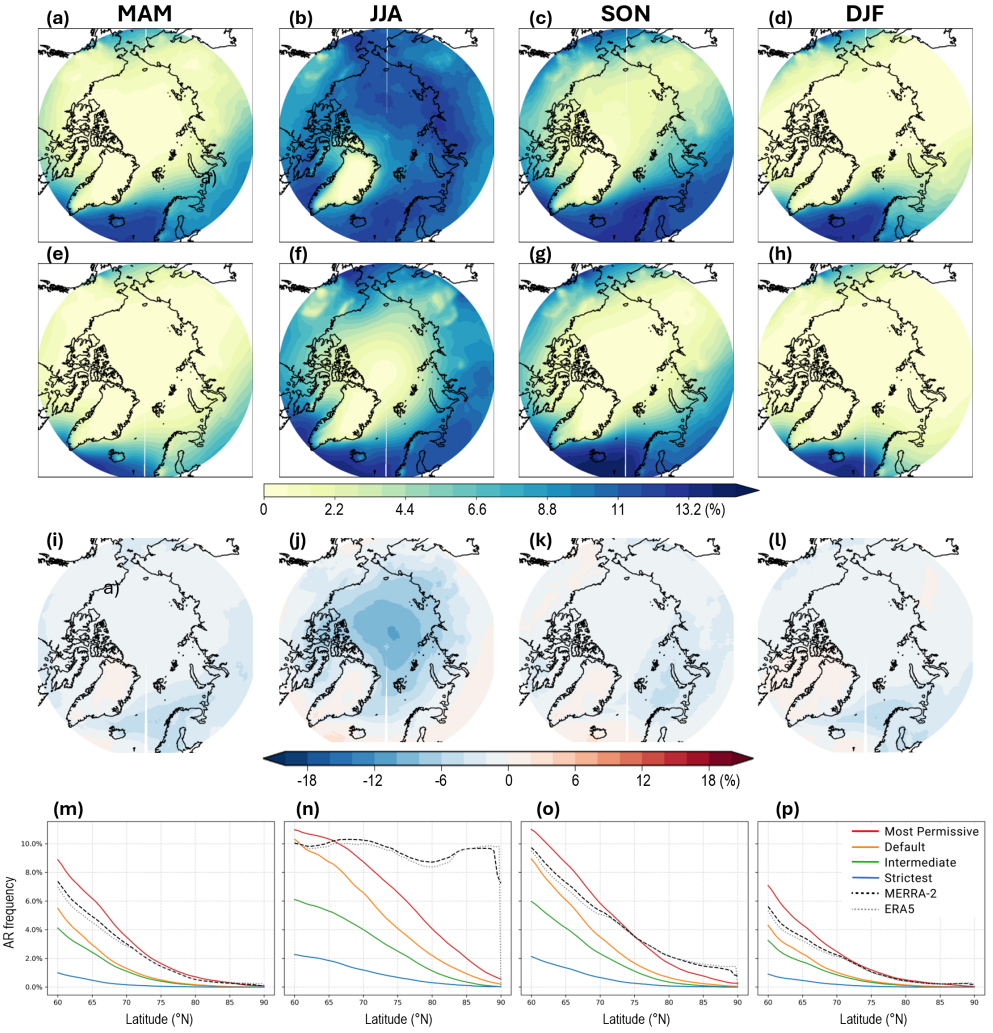

Figure 3. Seasonal AAR frequency (1980–2014) comparing the CANARI-LE using the *Default* thresholds and MERRA-2. Panels (**a**–**d**) show MERRA-2, (**e**–**h**) the CANARI-LE, and (**i**–**l**) the difference (CANARI-LE–MERRA-2) for MAM, JJA, SON, and DJF, respectively. Negative (blue) values indicate fewer AARs in the CANARI-LE. Panels (**m**–**p**) show the corresponding zonal-mean frequencies (60–90° N).

3.4. Training Data Characteristics

Two label regimes were used to train and evaluate the segmentation models. The *Default* label set, based on the thresholds used by Mattingly et al. [12] for MERRA2, captures a large number of ARs and supports learning through abundant positive examples but includes broader and less sharply defined features. The *Intermediate* label set applies stricter IVT and geometric thresholds, improving physical realism while reducing the proportion of AR grid cells to less than 3%. This stronger class imbalance presents a more challenging but physically grounded training target.

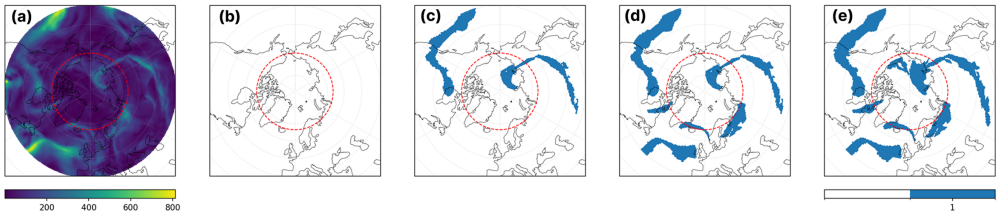

Figure 4. Example from 6 June 2012 showing (**a**) IVT magnitude and (**b**–**e**) AAR detections under the four threshold configurations (Strictest, Intermediate, Default, and Most Permissive). The red-dashed circle marks the Arctic Circle, 0 indicates non-AR grid cells, and 1 indicates AR grid cells.

Figure 5 summarises the statistical characteristics of the *Intermediate* configuration used for model training. Panels (a–c) describe the spatial and categorical properties of detected ARs: (a) shows the distribution of contiguous event sizes, (b) the fractional area of ARs per timestep, and (c) the relative proportions of mid-latitude, Arctic-penetrating, and Arctic-only AR events. Panels (d–f) characterise the dataset composition and input predictors: (d) shows the strong class imbalance between AR and background grid cells, (e) the conditional AR area fraction per timestep, and (f) the normalised mean values of the five predictor channels. Together, these diagnostics show that ARs occupy only a small portion of the Arctic domain but display wide variability in size and spatial extent, emphasising the sparsity and skewness of the training distribution. The corresponding *Default* distributions are provided in the Supplementary Materials.

Under the *Default* thresholds, AR grid cells account for 4.6% of approximately 1.03 billion labelled grid cells. Most timesteps contain at least one AR, with typical coverage per timestep between one and ten percent of the domain. The distribution of contiguous AR sizes is heavy-tailed, with a mean of 431 grid cells and a long upper tail extending beyond 6000 grid cells. About one-quarter of events cross the Arctic Circle, while fewer than three percent occur entirely within the Arctic domain.

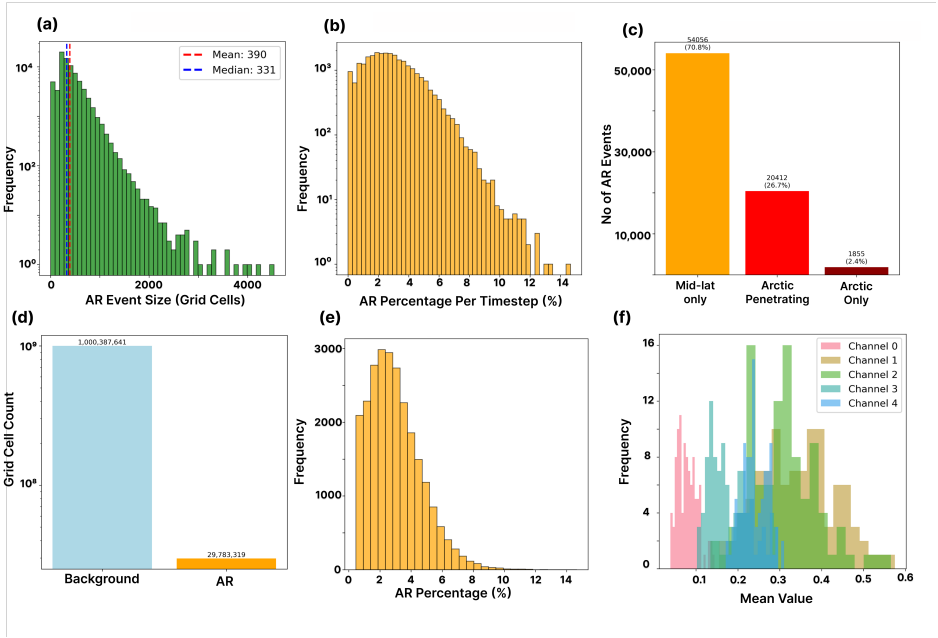

Figure 5. Distribution of AR labels under the *Intermediate* configuration. Panels show (**a**) event size; (**b**) area fraction per timestep; (**c**) proportion of mid-latitude, Arctic-penetrating, and Arctic-only AR events; (**d**) class counts; (**e**) conditional area fraction; and (**f**) normalised predictor distributions.

Applying the *Intermediate* thresholds reduces the AR fraction to 2.9% of the possible grid cells and shifts the distribution toward smaller, narrower filaments. These differences

highlight how threshold choice alters the statistical structure of the training data: the *Default* set favours event inclusivity, whereas the *Intermediate* set prioritises physical coherence. Together, these label sets define the two regimes used for model training and provide the basis for testing model sensitivity to label definition.

3.5. Segmentation Performance and Sensitivity to Label Definition

Two segmentation models were trained on the CANARI-LE using labels from the *Default* and *Intermediate* configurations. Evaluation was performed on held-out ensemble members 29–40. Table 3 summarises the grid-cell-wise results, while Table 4 presents the corresponding confusion matrices aggregated over all test data. The model trained on the *Default* labels achieved a Dice score of 0.76 and a recall of 0.79, indicating high sensitivity to AAR structures but a tendency to overpredict in marginal cases. The *Intermediate* label model attained higher precision (0.80) but lower recall (0.55), reflecting a more conservative learning signal. Both models maintained an accuracy of around 0.98.

Table 3. Segmentation model performance on the held-out ensembles (29–40).

Model	Accuracy	Precision	Recall	Dice
Default	0.98	0.74	0.79	0.76
Intermediate	0.98	0.8	0.56	0.65

The confusion matrices highlight the substantial class imbalance, with more than 94% of grid cells corresponding to non-AAR background. Despite this, both models correctly identify the majority of AAR grid cells. The *Default* model detects a larger proportion of AARs but also produces more false positives, whereas the *Intermediate* model sacrifices recall for improved precision. These contrasting behaviours explain the similar Dice scores yet distinct spatial error patterns seen in Figures 6 and 7. The *Default* model slightly overpredicts AARs along the lower Arctic storm-track latitudes, particularly over the North Atlantic and northwest Pacific, while the *Intermediate* model shows mild underprediction along the same corridors. The frequency fields shown in Figure 7 confirm that differences between the two models occur mainly along the margins of the main filaments rather than introducing spurious detections elsewhere. Both reproduce the overall AAR climatology of the CANARI-LE with high spatial fidelity.

Table 4. Confusion matrices for the segmentation models, showing total counts and percentage of all evaluated grid cells. B stands for Billion.

	Default Model		Intermediate Model	
	Pred. AAR	**Pred. Non-AAR**	**Pred. AAR**	**Pred. Non-AAR**
True AAR	1.32B (3.7%)	0.35B (1.0%)	0.70B (1.6%)	0.57B (1.3%)
True non-AAR	0.47B (1.3%)	33.99B (94.2%)	0.18B (0.4%)	42.36B (96.7%)

The example timesteps in Figure 8 illustrate these behaviours on the event scale. Each row shows one timestep, with columns representing the rule-based ground truth, the default model prediction, and the grid-cell-wise classification outcome. The default model reproduces the main AAR filaments with good spatial alignment, particularly in panels (a–c) and (g–i), where detections follow the observed structure closely. Overprediction occurs mainly when no AARs are present, as in panels (d–f), where diffuse IVT features are misclassified as AARs. These examples demonstrate the inherent precision–recall balance in imbalanced segmentation tasks, where improving sensitivity to rare features often increases

false-positive rates [41,48]. The model effectively captures the core of well-defined events but remains prone to false positives in marginal cases and misses some weaker filaments.

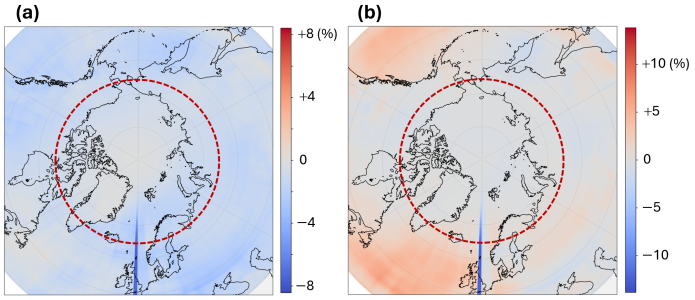

Figure 6. Spatial error (prediction–truth) in AAR frequency for the (**a**) *Intermediate* model and (**b**) *Default* model. Red indicates overprediction, and blue indicates underprediction. The red-dashed circle marks the Arctic Circle.

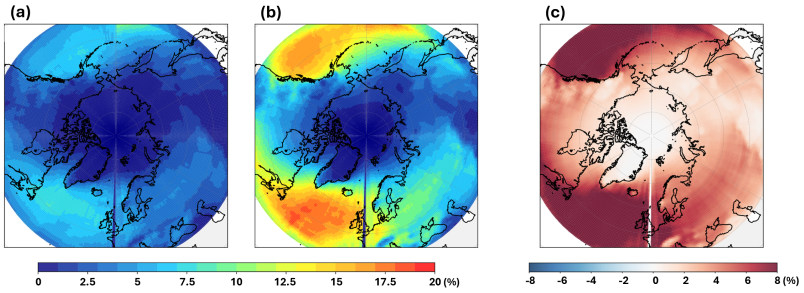

Figure 7. Predicted AAR frequency from the segmentation models: (**a**) *Intermediate* model, (**b**) *Default* model, and (**c**) difference (**a**,**b**).

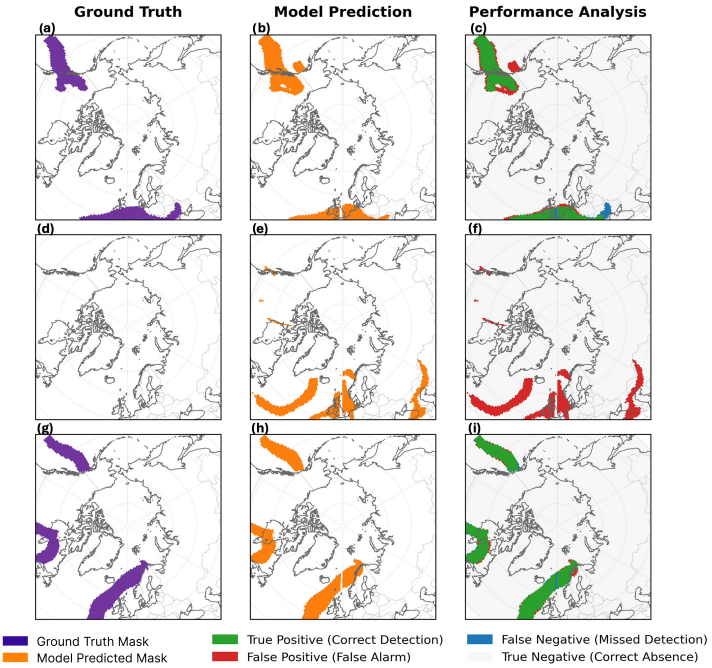

Figure 8. Representative model predictions for three timesteps. Each row corresponds to a different case: (**a**–**c**) a strong AAR event, (**d**–**f**) a non-AAR case with overprediction, and (**g**–**i**) a weaker but well-defined event. Columns show (left to right) rule-based ground truth (**a**,**d**,**g**) with mask shown in purple, model prediction (**b**,**e**,**h**) with prediction shown in orange, and grid-cell-wise classification outcome (**c**,**f**,**i**). Green indicates true positives (correct detections), red false positives (false alarms), blue false negatives (missed detections), and grey true negatives (correct absences).

4. Discussion and Conclusions

This work provides the first assessment of AARs in the CANARI-LE using a unified framework combining rule-based detection with machine learning. Three main contributions stand out. First, the CANARI-LE serves as a regional, ensemble-rich resource for AAR detection and sensitivity analysis, offering a large, internally consistent event sample at 6-hourly resolution. Second, a polar-specific detection logic is applied to the CANARI-LE, ERA5, and MERRA-2, enabling close comparison of AAR climatology across model and reanalysis data. Third, this study examines how threshold settings in the rule-based algorithm shape training labels and, consequently, the learning behaviour of the segmentation model. Although the DL approach removes the need for manual threshold tuning, understanding this dependency remains essential to interpreting performance and ensuring robustness.

Applying the algorithm by Mattingly et al. [12] to the CANARI-LE yields coherent AAR corridors aligned with principal North Atlantic and North Pacific moisture pathways. Relative to ERA5 and MERRA-2, the CANARI-LE shows lower AAR frequency, particularly north of 75° N and during summer and autumn. This difference reflects reduced IVT magnitude in the CANARI-LE rather than inconsistencies in the detection logic. Because DL targets derive from these fields, supervised models inherit each dataset's climatological imprint. Segmentation metrics, therefore, quantify agreement with a given algorithm–dataset pair rather than an absolute truth, consistent with the algorithm and data dependence highlighted by ARTMIP [13,14].

These dataset-dependent features align with earlier multi-model studies. Consistent with Zhang et al. [28], the CANARI-LE underestimates winter AAR frequency, especially north of 75° N, and lacks the summer enhancement seen in several CMIP6 models. This behaviour reflects persistent biases in simulating large-scale circulation and blocking over the North Atlantic [45]. Reduced Greenland blocking and weaker meridional moisture flux likely contribute to the underestimation of poleward moisture transport into the Arctic. Similar discrepancies were reported by Espinoza et al. [27], who found that most CMIP5 models capture overall AR intensification under warming but miss regional and seasonal variability. The moisture-source analysis by Papritz et al. [34] supports our interpretation that the North Atlantic sector dominates poleward moisture flux into the Arctic, consistent with the CANARI-LE maximum in AAR frequency. These parallels underline both the value of the CANARI-LE for physically consistent regional analysis and the need to improve representation of blocking dynamics and moisture transport in climate models.

Beyond circulation biases, algorithm design also shapes detection outcomes. Threshold sensitivity tests show that permissive IVT and geometry settings increase event counts and recall but broaden plume margins, while stricter settings reduce event counts, favour precision, and slightly underpredict along main transport paths. The segmentation models reproduce this trade-off: training on default labels increases recall and Dice, whereas stricter labels raise precision but miss edge detections. These results indicate that model behaviour mirrors label design rather than introducing artefacts, highlighting the CANARI-LE's value for systematic testing.

The CANARI-LE enables robust sensitivity testing because it supplies thousands of events across members and decades. Performance improves with the size and diversity of the training sample, and the DL pipeline offers a substantial speed increase for ensemble-scale inference compared with the rule-based algorithm. These properties make the approach well suited to large-member evaluation and scenario analysis, provided that the label provenance and threshold regime are treated as part of the experimental design.

Despite this, several limitations remain. Labels are generated from the CANARI-LE, so both training and evaluation inherit its physics, resolution, and moisture biases relative to

reanalysis. There is no independent human-annotated Arctic benchmark, and expert labels can diverge [24]. Limited pressure-level availability constrains the IVT integral and may influence high-latitude magnitude differences. The models operate on single timesteps, which limits sensitivity to synoptic continuity. Although dataset size affects segmentation skill, we did not vary training set size due to GPU and time constraints, focusing instead on model skill. Finally, although the architecture is reproducible and effective, it is not the focus of novelty; model behaviour is governed primarily by label definition and data characteristics [19].

Overall, rule-based AAR detection in the CANARI-LE is physically credible but yields lower frequencies than reanalyses because of background IVT differences. Label thresholds govern the precision–recall balance and spatial error structure in both rule-based and DL detections. An ensemble-rich dataset such as the CANARI-LE is, therefore, valuable for stress-testing these methodological choices and training models for polar environments.

Future work should evaluate model transfer across domains by training on the CANARI-LE and testing on reanalysis data, and vice versa, to distinguish algorithmic from data-model influences. Investigating how model performance varies when trained on smaller subsets of the dataset could also provide insight into data efficiency and generalisation. Incorporating temporal context, for example, by analysing sequences of timesteps, may enhance sensitivity to weaker but continuous AAR filaments. Extending the framework to produce probabilistic outputs and calibrated uncertainty estimates would enable threshold-independent evaluation and support downstream climate applications. As the model architecture is fully compatible with the CANARI-LE output, it can also be applied to future simulations (2015–2100) to examine projected changes in AAR frequency, intensity, and pathways under warming scenarios.

In summary, this study demonstrates that a regional, polar-focused framework combining rule-based and machine-learning methods can generate consistent AAR detections across a large ensemble while revealing how detection outcomes depend on label design and data characteristics. These results establish the CANARI-LE as a valuable resource for Arctic AR research and provide a foundation for reproducible, scalable approaches to high-latitude moisture transport analysis.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/atmos17010061/s1, Figure S1: Annual mean AAR frequency (1980–2014) for (a) CANARI-LE, (b) MERRA-2, and (c) ERA5. Frequencies are expressed as the percentage of 6-hourly timesteps classified as AARs. Figure S2: Distribution of AAR labels under the *Default* configuration. Panels show (a) event size, (b) area fraction per timestep, (c) proportion of midlatitude, Arctic-penetrating, and Arctic-only events, (d) class counts, (e) conditional area fraction, and (f) normalised predictor distributions.

Author Contributions: S.M.: formal analysis; writing—original draft, review, and editing; methodology; software; investigation; validation. H.L.: Conceptualisation; methodology; validation; funding acquisition; supervision; writing—review and editing; project administration; resources. G.M.: Conceptualisation; methodology; validation; supervision; writing—review and editing. O.M.-A.: Writing—review and editing; software; data curation; validation; supervision. M.O.: Writing—review and editing; supervision. K.M.: Software; data curation; writing—review and editing. D.G.: software; validation; writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: H.L. and O.M.A. were supported by CANARI, a National Capability Multi-Centre Science programme of the Natural Environment Research Council (NE/W004984/1), which provided the moisture and wind fields for model training and comparison.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The CANARI large-ensemble historical simulations are available via the Joint Advanced Supercomputing Infrastructure for the Natural Environment (JASMIN). AR categorisations based on MERRA-2 and IVT source files based on ERA5 and MERRA-2 were accessed through the ARTMIP database [49,50]. Additional derived datasets and analysis scripts are available from the corresponding author upon request. The U-Net model was implemented in PyTorch v2.5.0.

Acknowledgments: This work used JASMIN, the UK's collaborative data analysis environment for storage and analysis (www.jasmin.ac.uk, accessed on 29 December 2025). We thank Tony Phillips for data management and computational support, Christine Shields for assistance with ERA5-based AR catalogues from ARTMIP, and Michelle MacLennan for valuable discussions. We also wish to thank the anonymous reviewers for their constructive comments, which helped to improve this manuscript. This work was carried out at the British Antarctic Survey as part of S.M.'s MPhil thesis at the University of Cambridge.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARs	Atmospheric rivers
AARs	Arctic atmospheric rivers
SST	Sea surface temperature
ARTMIP	Atmospheric River Tracking Method Intercomparison Project
CANARI-LE	Climate change in the Arctic and North Atlantic Region and Impacts on the UK Large Ensemble
ERA5	Reanalysis version 5 by the European Centre for Medium-Range Weather Forecasts
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications, Version 2
IVT	Integrated water vapour transport
uIVT	Zonal component of IVT
vIVT	Meridional component of IVT
TCWV	Total column water vapour
DL	Deep learning
U-Net	Convolutional Networks for Image Segmentation

References

1. Rantanen, M.; Karpechko, A.Y.; Lipponen, A.; Nordling, K.; Hyvärinen, O.; Ruosteenoja, K.; Vihma, T.; Laaksonen, A. The Arctic has warmed nearly four times faster than the globe since 1979. *Commun. Earth Environ.* **2022**, *3*, 168. [CrossRef]
2. American Meteorological Society. Atmospheric River. Available online: https://glossary.ametsoc.org/wiki/Atmospheric_river (accessed on 25 June 2025).
3. Guan, B.; Waliser, D.E. Tracking Atmospheric Rivers Globally: Spatial Distributions and Temporal Evolution of Life Cycle Characteristics. *J. Geophys. Res. Atmos.* **2019**, *124*, 12549–12568. [CrossRef]
4. Wang, Z.; Ding, Q.; Wu, R.; Ballinger, T.J.; Guan, B.; Bozkurt, D.; Chen, Z. Role of atmospheric rivers in shaping long-term Arctic moisture variability. *Nat. Commun.* **2024**, *15*, 5505. [CrossRef]
5. Lee, S.; Gong, T.; Feldstein, S.B.; Screen, J.A.; Simmonds, I. Revisiting the cause of the 1989–2009 Arctic surface warming using the surface energy budget: Downward infrared radiation dominates the surface fluxes. *Geophys. Res. Lett.* **2017**, *44*, 10654–10661. [CrossRef]
6. Luo, B.; Wu, L.; Luo, D.; Dai, A.; Simmonds, I. The winter midlatitude—Arctic interaction: Effects of North Atlantic SST and high-latitude blocking on Arctic sea ice and Eurasian cooling. *Clim. Dyn.* **2019**, *52*, 2981–3004. [CrossRef]
7. Zhang, P.; Chen, G.; Ting, M.; Yu, Z.; Ma, J.; Li, P.; Wang, S. More frequent atmospheric rivers slow the seasonal recovery of Arctic sea ice. *Nat. Clim. Change* **2023**, *13*, 266–273. [CrossRef]
8. Ma, W.; Wang, H.; Chen, G.; Qian, Y.; Baxter, I.; Huo, Y.; Seefeldt, M.W. Wintertime Extreme Warming Events in the High Arctic: Characteristics, Drivers, Trends, and the Role of Atmospheric Rivers. *Atmos. Chem. Phys.* **2024**, *24*, 4451–4472. [CrossRef]
9. Li, H.; Ke, C.-Q.; Shen, X.; Zhu, Q.; Cai, Y.; Luo, L. The Varied Role of Atmospheric Rivers in Arctic Snow Depth Variations. *Geophys. Res. Lett.* **2024**, *51*, e2024GL110163. [CrossRef]

10. Serreze, M.C.; Gustafson, J.; Barrett, A.P.; Druckenmiller, M.L.; Fox, S.; Voveris, J.; Stroeve, J.; Sheffield, B.; Forbes, B.C.; Rasmus, S.; et al. Arctic rain on snow events: Bridging observations to understand environmental and livelihood impacts. *Environ. Res. Lett.* **2021**, *16*, 105009. [CrossRef]
11. Gong, Z.; Zhong, L.; Hua, L.; Feng, J. Dynamic and thermodynamic impacts of atmospheric rivers on sea ice thickness in the Arctic since 2000. *J. Clim.* **2025**, *38*, 2873–2888. [CrossRef]
12. Mattingly, K.S.; Mote, T.L.; Fettweis, X. Atmospheric River Impacts on Greenland Ice Sheet Surface Mass Balance. *J. Geophys. Res. Atmos.* **2018**, *123*, 7584–7604. [CrossRef]
13. Shields, C.A.; Rutz, J.J.; Leung, L.-Y.; Ralph, F.M.; Wehner, M.; Kawzenuk, B.; Lora, J.M.; McClenny, E.; Osborne, T.; Payne, A.E.; et al. Atmospheric River Tracking Method Intercomparison Project (ARTMIP): Project goals and experimental design. *Geosci. Model Dev.* **2018**, *11*, 2455–2474. [CrossRef]
14. Shields, C.A.; Payne, A.E.; Shearer, E.J.; Wehner, M.F.; O'Brien, T.A.; Rutz, J.J.; Leung, L.-Y.R.; Ralph, F.M.; Marquardt Collow, A.B.; Ullrich, P.A.; et al. Future atmospheric rivers and impacts on precipitation: Overview of the ARTMIP Tier 2 high-resolution global warming experiment. *Geophys. Res. Lett.* **2023**, *50*, e2022GL102091. [CrossRef]
15. Zhou, Y.; O'Brien, T.A.; Ullrich, P.A.; Collins, W.D.; Patricola, C.M.; Rhoades, A.M. Uncertainties in atmospheric river lifecycles by detection algorithms: Climatology and variability. *J. Geophys. Res. Atmos.* **2021**, *126*, e2020JD033711. [CrossRef]
16. Wille, J.D.; Favier, V.; Gorodetskaya, I.V.; Agosta, C.; Baiman, R.; Barrett, J.E.; Barthelemy, L.; Boza, B.; Bozkurt, D.; Casado, M.; et al. Atmospheric rivers in Antarctica. *Nat. Rev. Earth Environ.* **2025**, *6*, 178–192. [CrossRef]
17. Lauer, M.; Mech, M.; Guan, B. *Global Atmospheric Rivers Catalog for ERA5 Reanalysis [Dataset]*; PANGAEA: Bremen, Germany, 2023. [CrossRef]
18. Wille, J.D.; Favier, V.; Gorodetskaya, I.V.; Codron, F.; Kittel, C.; Agosta, C.; Lenaerts, J.T.M. Antarctic Atmospheric River Climatology and Precipitation Impacts. *J. Geophys. Res. Atmos.* **2021**, *126*, e2020JD033788. [CrossRef]
19. Galea, D.; Ma, H.; Wu, W.; Kobayashi, D. Deep Learning Image Segmentation for Atmospheric Rivers. *Artif. Intell. Earth Syst.* **2024**, *3*, e230048. [CrossRef]
20. Ullrich, P.A.; Zarzycki, C.M.; McClenny, E.E.; Mullendore, G.; Rhoades, A.M.; Ullrich, R.; Lauritzen, P.H. TempestExtremes v2.1: A community framework for feature detection, tracking, and analysis in large datasets. *Geosci. Model Dev.* **2021**, *14*, 5023–5048. [CrossRef]
21. Galea, D.; Ma, H. Intercomparison of Deep Learning Model Architectures for Atmospheric River Prediction. *Artif. Intell. Earth Syst.* **2025**, *4*, 240057. [CrossRef]
22. Marquardt Collow, A.B.; Shields, C.A.; Guan, B.; O'Brien, T.A.; Wilson, A.B.; Rutz, J.J.; Mahoney, K.; Wick, G.A.; Ralph, F.M.; Leung, L.-Y.R. An Overview of ARTMIP's Tier 2 Reanalysis Intercomparison: Uncertainty in the Detection of Atmospheric Rivers and Their Associated Precipitation. *J. Geophys. Res. Atmos.* **2022**, *127*, e2021JD035158.
23. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder–Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1802.02611.
24. Prabhat; Kashinath, K.; Mudigonda, M.; Kim, S.; Kapp-Schwoerer, L.; Graubner, A.; Karaismailoglu, E.; von Kleist, L.; Kurth, T.; Greiner, A.; et al. ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geosci. Model Dev.* **2021**, *14*, 107–124. [CrossRef]
25. Zhou, Z.; Rahman Siddiquee, M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *arXiv* **2018**, arXiv:1807.10165. [CrossRef]
26. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef]
27. Espinoza, V.; Waliser, D.E.; Guan, B.; Lavers, D.A.; Ralph, F.M. Global analysis of climate change projection effects on atmospheric rivers. *Geophys. Res. Lett.* **2018**, *45*, 4299–4308. [CrossRef]
28. Zhang, L.; Zhao, Y.; Cheng, T.F.; Lu, M. Future changes in global atmospheric rivers projected by CMIP6 models. *J. Geophys. Res. Atmos.* **2024**, *129*, e2023JD039359. [CrossRef]
29. Eyring, V.; Collins, W.D.; Gentine, P.; Barnes, E.A.; Barreiro, M.; Beucler, T.; Zanna, L. Pushing the Frontiers in Climate Modelling and Analysis with Machine Learning. *Nat. Clim. Chang.* **2024**, *14*, 916–928. [CrossRef]
30. Higgins, T.B.; Subramanian, A.C.; Graubner, A.; Kapp-Schwoerer, L.; Watson, P.A.G.; Sparrow, S.; Kashinath, K.; Kim, S.; Delle Monache, L.; Chapman, W. Using Deep Learning for an Analysis of Atmospheric Rivers in a High-Resolution Large Ensemble Climate Data Set. *J. Adv. Model. Earth Syst.* **2023**, *15*, e2022MS003495. [CrossRef]
31. Roberts, M.J.; Baker, A.; Blockley, E.W.; Calvert, D.; Coward, A.; Hewitt, H.T.; Jackson, L.C.; Kuhlbrodt, T.; Mathiot, P.; Roberts, C.D.; et al. Description of the resolution hierarchy of the global coupled HadGEM3-GC3.1 model as used in CMIP6 HighResMIP experiments. *Geosci. Model Dev.* **2019**, *12*, 4999–5028. [CrossRef]
32. Palmer, T.E.; McSweeney, C.F.; Booth, B.B.B.; Priestley, M.D.K.; Davini, P.; Brunner, L.; Borchert, L.; Menary, M.B. Performance-based sub-selection of CMIP6 models for impact assessments in Europe. *Earth Syst. Dyn.* **2023**, *14*, 457–483. [CrossRef]

33. CANARI Project Team. CANARI: Climate Change in the Arctic-North Atlantic Region and Impacts on the UK. Available online: https://canari.ac.uk/ (accessed on 27 June 2025).
34. Papritz, L.; Hauswirth, D.; Hartmuth, K. Moisture origin, transport pathways, and driving processes of intense wintertime moisture transport into the Arctic. *Weather. Clim. Dyn.* **2022**, *3*, 1–20. [CrossRef]
35. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
36. Gelaro, R.; McCarty, W.; Suárez, M.J.; Todling, R.; Molod, A.; Takacs, L.; Randles, C.A.; Darmenov, A.; Bosilovich, M.G.; Reichle, R.; et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **2017**, *30*, 5419–5454. [CrossRef]
37. Zhu, Y.; Newell, R.E. A Proposed Algorithm for Moisture Fluxes from Atmospheric Rivers. *Mon. Weather. Rev.* **1998**, *126*, 725–735. [CrossRef]
38. Ralph, F.M.; Neiman, P.J.; Rotunno, R. Dropsonde Observations in Low-Level Jets over the Northeastern Pacific Ocean from CALJET-1998 and PACJET-2001: Mean Vertical-Profile and Atmospheric-River Characteristics. *Mon. Weather. Rev.* **2005**, *133*, 889–910. [CrossRef]
39. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
40. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
41. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
42. Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J.E.; Stoica, I. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv* **2018**, arXiv:1807.05118. [CrossRef]
43. Lawrence, B.N.; Bennett, V.L.; Churchill, J.; Juckes, M.; Kershaw, P.; Pascoe, S.; Pepler, S.; Pritchard, M.; Stephens, A. Storing and manipulating environmental big data with JASMIN. In Proceedings of the IEEE International Conference on Big Data, San Francisco, CA, USA, 6–9 October 2013; IEEE: New York, NY, USA, 2013. [CrossRef]
44. Biewald, L. Experiment Tracking with Weights and Biases. Available online: https://www.wandb.com/ (accessed on 27 June 2025).
45. Kleiner, N.; Chan, P.W.; Wang, L.; Ma, D.; Kuang, Z. Effects of Climate Model Mean-State Bias on Blocking Underestimation. *Geophys. Res. Lett.* **2021**, *48*, e2021GL094129. [CrossRef]
46. Reynolds, C.A.; Crawford, W.; Huang, A.; Barton, N.; Janiga, M.A.; McLay, J.; Flatau, M.; Frolov, S.; Rowley, C. Analysis of Integrated Vapor Transport Biases. *Mon. Weather. Rev.* **2022**, *150*, 1097–1113. [CrossRef]
47. Gao, Y.; Guo, X.; Lu, J.; Woolings, T.; Chen, D.; Guo, X.; Wu, L. Enhanced simulation of atmospheric blocking in a high-resolution Earth system model: Projected changes and implications for extreme weather events. *J. Geophys. Res. Atmos.* **2025**, *130*, e2024JD042045. [CrossRef]
48. Saito, T.; Rehmsmeier, M. The precision–Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
49. Marquardt Collow, A.B.; Guan, B.; Kim, S.; Lora, J.; McClenny, E.; Nardi, K.; Wehner, M. Atmospheric River Tracking Method Intercomparison Project Tier 2 Reanalysis Source Data and Catalogues. *Clim. Glob. Dyn. Div.* **2024**, *127*, 20 . [CrossRef]
50. Jonathan, J.; Rutz, C.A.; Lora, J.M.; Payne, A.E.; Guan, B.; Ullrich, P.A. Atmospheric River Tracking Method Intercomparison Project Tier 1 Source Data and Catalogues. *Nsf. Natl. Cent. Atmos. Res.* **2024**, *124*, 13777–13802. [CrossRef]