JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2022MS003495

Using Deep Learning for an Analysis of Atmospheric Rivers in a High-Resolution Large Ensemble Climate Data Set

Timothy B. Higgins[1], Aneesh C. Subramanian[1], Andre Graubner[2], Lukas Kapp-Schwoerer[2], Peter A. G. Watson[3], Sarah Sparrow[4], Karthik Kashinath[5], Sol Kim[6], Luca Delle Monache[7], and Will Chapman[8]

[1]Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, CO, USA, [2]ETH Zurich, Zürich, Switzerland, [3]School of Geographical Sciences, Bristol University, Bristol, UK, [4]Atmospheric, Oceanic and Planetary Physics, Oxford University, Oxford, UK, [5]NVIDIA Corporation, Santa Clara, CA, USA, [6]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA, [7]Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA, [8]National Center for Atmospheric Research, Boulder, CO, USA

Key Points:

- CG-Climate is an effective and efficient way to track atmospheric rivers in large datasets
- CG-Climate is flexible with varying spatial resolutions and domains
- Human hand labels are effective in identifying atmospheric river events in the correct location, but have a spatial area bias

Correspondence to:

T. B. Higgins,
timothy.higgins@colorado.edu

Citation:

Higgins, T. B., Subramanian, A. C., Graubner, A., Kapp-Schwoerer, L., Watson, P. A. G., Sparrow, S., et al. (2023). Using deep learning for an analysis of atmospheric rivers in a high-resolution large ensemble climate data set. *Journal of Advances in Modeling Earth Systems*, *15*, e2022MS003495. https://doi.org/10.1029/2022MS003495

Received 26 OCT 2022
Accepted 5 APR 2023

Abstract There is currently large uncertainty over the impacts of climate change on precipitation trends over the US west coast. Atmospheric rivers (ARs) are a significant source of US west coast precipitation and trends in ARs can provide insight into future precipitation trends. There are already a variety of different methods used to identify ARs, but many are used in contexts that are often difficult to apply to large climate datasets due to their computational cost and requirement of integrated vapor transport as an input variable, which can be expensive to compute in climate models at high temporal frequencies. Using deep learning (DL) to track ARs is a unique approach that can alleviate some of the computational challenges that exist in more traditional methods. However, some questions still remain regarding its flexibility and robustness. This research investigates the consistency of a DL methodology of tracking ARs with more established algorithms to demonstrate its high-level performance for future studies.

Plain Language Summary Atmospheric rivers (ARs) are long corridors of water vapor in the lower atmosphere that are associated with a large amount of precipitation on the US west coast. They are important to investigate in future climate change scenarios. To further understand them in climate change scenarios, they must be tracked in large datasets. We demonstrate the efficiency, effectiveness, and flexibility of a machine learning tracking method by comparing it to more established existing tracking methods. This method applies particularly well to large climate datasets and can be useful for future studies.

1. Introduction

Atmospheric rivers (ARs) are elongated corridors of water vapor in the lower troposphere that cause extreme precipitation over many coastal regions around the globe. They play a vital role in the water cycle in the western US, fueling the most extreme west coast precipitation and sometimes accounting for more than 50% of total annual west coast precipitation (Gershunov et al., 2017). Severe ARs are associated with extreme flooding and damages while weak ARs are typically more beneficial to our society as they bring much needed drought relief (Ralph et al., 2019). Future climate projections show an increase in US west coast precipitation variability caused by AR precipitation increasing and non-AR precipitation decreasing (Gershunov et al., 2019). From 2012 to 2016, California experienced a historic drought, which was followed by the state's second wettest year on record. 2020 and 2021 are two of the driest years on record over much of the western US (Williams et al., 2022). The extreme interannual variability in western US precipitation in recent years coinciding with climate change projections of increased precipitation variability is a serious cause for concern over how patterns may change in the coming decades (Polade et al., 2017; Shields & Kiehl, 2016).

A necessary step in understanding changing patterns in ARs as a function of climate change is employing an AR detection method. There are a variety of different algorithms used to track ARs due to their relatively diverse definitions (Shields et al., 2018). The Atmospheric River Tracking Intercomparison Project (ARTMIP) organizes and provides information on all the widely accepted algorithms that exist. Nearly all the algorithms included in ARTMIP rely on absolute and relative numerical thresholds. Absolute thresholds are static constraints that are required for an AR to exist, typically coming in the form of length, width, minimum inte-

grated vapor transport (IVT), or minimum integrated water vapor (IWV), while relative thresholds typically come in the form of percentile based IVT and IWV constraints. The computational expense of running these methods and retrieving their required input variables can be particularly problematic when applied to large climate datasets. The vast majority of algorithms heavily factor in IVT to track ARs, which either needs to be calculated with wind velocity and humidity at multiple vertical model levels (Neiman et al., 2008) or must be output directly. Both methods of acquiring IVT are often unavailable in climate model output at the temporal resolution that ARs occur. When IVT is unavailable in model output, AR detection algorithms that require it as input are unusable.

A recent alternative way of tracking ARs is using machine learning. There are many neural networks that are commonly applied toward identifying objects in a wide range of applications via semantic segmentation. The first of these neural networks that was applied toward detecting ARs is DeepLabv3+ (Prabhat et al., 2021). DeepLabv3+ is a state-of-the-art model that demonstrates one of the highest performances of any present day neural network when tasked with the objective of identifying objects in cityscapes (Wu et al., 2019). In this study, we employ an identical network to (Kapp-Schwoerer & Graubner, 2020), which is a light-weight convolutional neural network (51 convolutional layers and close to 500,000 parameters) adapted from Context Guided Network (CGNet) (Wu et al., 2019) to efficiently track these severe events without using IVT as a predictor variable. Instead of using IVT, this study uses zonal wind at 850 mb, meridional wind at 850 mb, surface pressure, and IWV as input. The selected input features allow the tracking method to capture contextual two-dimensional flow in the lower troposphere while still identifying large plumes of moisture that play a key role in shaping the identity of ARs. By not requiring IVT (which is an integral function of specific humidity and wind that requires additional computations in the vertical dimension from those of IWV, which is only an integral function of specific humidity), this method is immensely flexible with output from a broad variety of different climate simulations.

When applied to cityscapes, CGNet's greatest advantage is its performance relative to its memory footprint (Wu et al., 2019). It has two orders of magnitude less parameters than DeepLabv3+ and is computationally less expensive. This can be especially useful when identifying ARs in large datasets. This will also be the first study to demonstrate the performance of this neural network on a regional domain by providing an objective analysis of its consistency with eight different ARTMIP algorithms. The flexibility of this tracking method allows it to apply particularly well toward large ensemble climate datasets, which continue to increase in size (Chikamoto et al., 2013; Delworth et al., 2012; Kay et al., 2015). Analysis of these datasets is a critical part of quantifying the forced response of our climate system to anthropogenic forcing (Deser et al., 2020; Payne et al., 2020). They are particularly useful for reliably predicting probabilities of rare extreme events, but bring the challenge of requiring additional computational resources. The Weather@Home project addresses this challenge by generating large ensemble and high-resolution simulations of Earth's climate with the assistance of computing power from thousands of volunteers' computers (Guillod et al., 2017; Massey et al., 2015). It creates simulations in various geographic regions and for various warming scenarios by having volunteers run a global model to supply boundary conditions for a higher resolution 1-way nested regional model.

In this work we use output from new Weather@Home simulations using a recently developed configuration of the HadAM4 atmospheric model with 60 km horizontal resolution globally (Watson et al., 2020), which allows for the study of ARs. These simulations include 1,000–2000 winters for each of a set of climate states: recent historical climate and climates with 1.5°C and 2°C global mean warming above pre-industrial temperatures, following the method of Mitchell et al. (2017). This allows us to better quantify the uncertainty of the future of extreme ARs and severe western US precipitation. Precipitation is particularly difficult to predict in traditional climate models. Predicting water vapor is more reliable (Lavers et al., 2016), allowing IVT and possibly ARs to be a favorable method for understanding changing patterns in precipitation (Johnson & Sharma, 2009). Understanding ARs in future climates is a growing area of interest (Espinoza et al., 2018; Warner et al., 2015; Zhao, 2020), which calls for a fast, reliable, and adaptable tracking method for large ensemble climate datasets such as Weather@Home outputs and CMIP6 simulations. Below, we show results from tracking ARs in Weather@Home simulations with a Convolutional Neural Network (CNN) adapted from CGNet, a technique that we refer to as "CG-Climate." In Section 2, we describe the data set that we use for statistical comparisons to other algorithms and the data set that we use to demonstrate its applications to larger scales. In Section 3, we discuss our approach toward estimating the reliability of the algorithm. In Section 4, we show the results of our methods. In Section 5, we give concluding remarks on our findings.

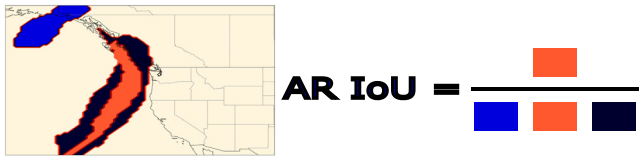

Figure 1. An example of an IoU calculation between CG-Climate and Tempest (Ullrich & Zarzycki, 2017) at a single instant in time. AR IoU is calculated by dividing the orange area (the area where both detection algorithms overlap) by the sum of the orange area, the black area (the area where only one algorithm detected an AR within an event in which both algorithms overlap), and the blue area (the area in which only one algorithm detected an AR that was not within an event in which both algorithms overlap). In AR IoU calculations, the black area and the blue area are weighted equally despite the blue area consisting of grid points that are not associated with an overlapping event and the black area consisting of grid points that are associated with an overlapping event.

2. Data

Due to the diverse definitions of ARs, we apply an AR expert hand-labeled contour data set toward tracking them named ClimateNet. Various AR expert hand-labeling contour campaigns were launched to obtain almost 500 feature-label pairings of AR locations with snapshots of CAM5.1 25 km resolution data on a global domain (Prabhat et al., 2021). This endeavor included campaigns at LBNL, UC Berkeley, Scripps/UCSD, NCAR, the 2019 ARTMIP Workshop, and the 2019 Climate Informatics Workshop, where 80 different weather and climate scientists participated. Labelers were shown geospatial maps of IWV, IVT, surface pressure, and wind velocity at 850 mb at each snapshot in time. Even amongst human experts, there was a significant amount of AR label disagreement. One highly popular metric that is used to estimate performance in semantic image segmentation identification scenarios is mean IoU (intersection over union) (Figure 1). IoU is calculated by dividing the area in which the same class of an object is detected by two different sources by the total area in which that class is detected by either source. The AR IoU was 0.34 when human-labeled ARs were compared to each other. For this work, the test set contains 61 images of hand-labeled contours occurring after 2011 and the training set consists of 400 images of hand-labeled contours occurring before 2011.

Our approach toward verifying the level of reliability of this AR tracking method is comparing its outputs to outputs from various ARTMIP algorithms that tracked ARs in the same data. Here, we use the National Aeronautics and Space Administration (NASA) Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2) data from January and February of 2006–2015 at a 3-hourly temporal resolution as inference data (data that ARs were detected in). The features used were zonal wind at 850 mb, meridional wind at 850 mb, surface pressure, and IWV derived from specific humidity. The algorithm outputs binary arrays containing the locations in which it detects ARs. The resolution was 0.625° longitude × 0.5° latitude. All algorithms were compared to each other on the domain spanning 30°–60°N and 100°–150°W as the features of interest are the ARs making landfall over the Western US. CG-Climate was first run on the MERRA-2 data set and trained on the ClimateNet data set for verification. Various different combinations of the training set domain and resolution were used to inform the user of the best way to apply CG-Climate to their own data. We obtained data from the tier 1 ARTMIP catalog referenced in Shields et al., 2018 of AR masks tracked from eight different ARTMIP algorithms in MERRA-2 as ground "truth" in all verification metrics. Performance was evaluated by calculating mean AR Intersection over Union (IoU), precision of overlapping events, and recall of overlapping events. CG-Climate was also trained on ARTMIP labels to diagnose inconsistencies between it and ARTMIP. A schematic of all comparisons between detection methods and datasets in the study are shown in Figure 2.

The algorithm was also applied to data from the Weather@Home project to gain a better understanding of its speed and flexibility. This data set has a horizontal grid resolution of 5/6° × 5/9° (approximately 60 km in middle latitudes) covering November-February (with November data discarded as spin-up) (Bevacqua et al., 2021). CG-Climate was run on 2 V100 GPUs over 1,244 winters from the historical climate scenario, 1,338 winters from the 1.5° increase scenario, and 1,682 winters from the 2° increase scenario for a total of 4,000 winters. We used variables IWV, zonal wind at 850 mb, meridional wind at 850 mb and surface pressure at 6-hourly intervals from the domain spanning 25°N–90°N and 240°W–35°E, which took up 500 GB after preprocessing was completed. CG-Climate identified ARs on the time scale of hours.

3. Methods

CGNet is a context guided (CG; local segmentation is guided by surrounding contextual information) neural network that was developed by Wu et al. (2019) with the objective of identifying objects in cityscapes with the amount of computing power available on mobile devices. An example of semantic segmentation applied to both cityscapes and ARs is shown in Figure 3. Its relatively light-weight architecture makes this neural network a potentially suitable method of tracking ARs in large climate datasets due to the time and computational resources that is saved from using it. After applying 3 standard convolutional layers, the data is fed into a CGBlock, which

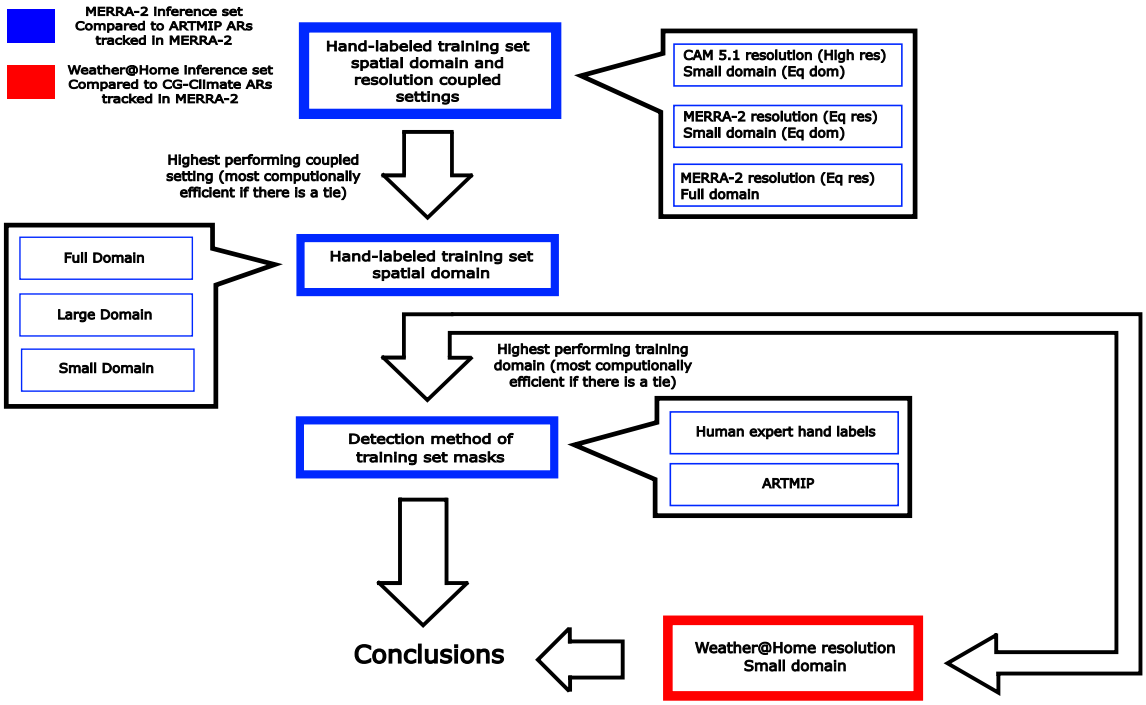

Figure 2. Schematic diagram showing the flow of comparisons and datasets used in the study.

is the core building block of the CNN. The CNN in this study used 24 total CGBlocks. The CGBlock first applies a 1×1 convolution to apply a dimensionality reduction across the respective input feature maps (i.e., physical variables in the first layer); this reduces the number of resulting convolutional fields to 1. Two 3×3 convolutional

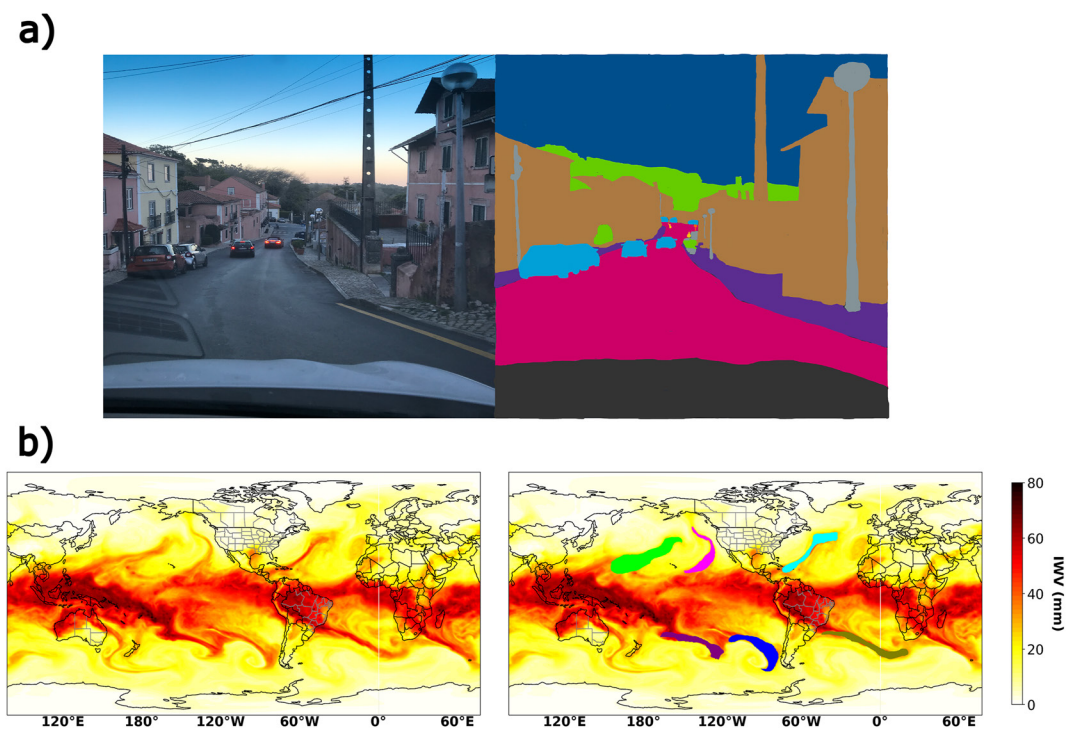

Figure 3. An example of two different applications of semantic segmentation. (a) A feature-label pair of objects in a cityscape. (b) A feature label pair of ARs in an IWV field. AR contours are shown in purple, blue, green, and pink.

kernels are then applied. One is dilated to represent surrounding context and one is not dilated to represent local context. The outputs from both the standard and dilated convolutions are concatenated into a joint feature map. A batch normalization is performed to normalize the mean and variance from the output and a PRelU activation function is applied (Xu et al., 2015). 1×1 average pooling is applied to the output from the PRelU function followed by a linear layer, the RelU activation, and another linear layer. Finally, a sigmoid function is applied to the output of the second linear layer before element-wise multiplication is used between it and the output from the PRelU activation function. The Jaccard Loss function is used for training. The loss function uses gradient descent, which minimizes the loss over 15 epochs with a batch size of 4 and a learning rate of 0.001. Stopping is applied when the gradient approaches zero.

The wide variety of accepted detection algorithms in the AR research community creates a dilemma in determining the truth to compare this algorithm to for verification. Due to the lack of agreement between ARTMIP algorithms and the types of algorithms used to calculate ARs (Inda-Díaz et al., 2021; Zhang et al., 2021), it is necessary to compare CG-Climate to additional common tracking methods to best understand its reliability. We compare the output of CG-Climate to eight different ARTMIP algorithms (Table 1). We also compare each ARTMIP algorithm to the remaining seven ARTMIP algorithms. We do not claim our evaluation metrics capture the detection skill of any of the ARTMIP algorithms, largely due to the fact that different algorithms were created to identify ARs in different contexts. All ARTMIP algorithms are already considered to be reputable. The goal of our approach is to acquire a reference point for the consistency of CG-Climate relative to a variety of other algorithms and show that the inconsistencies that do exist within its detection results are within a threshold that is common when reputable algorithms are compared to each other.

Once events were calculated in MERRA-2 data, we filtered out events that existed for less than 12 hr and did not exceed an area greater than 150,000 km^2 to exclude events that were unable to sustain moderate development and to stay consistent with (Kapp-Schwoerer & Graubner, 2020). Various ARTMIP algorithms require the length of the AR to be greater than 1,500–2,000 km and do not have a width requirement (e.g., Brands et al., 2017; Gershunov et al., 2017; Goldenson et al., 2018; Guan & Waliser, 2015; Reid et al., 2020; Rutz et al., 2014). We chose a cluster area of 150,000 km^2 because that is roughly the product of two grid points in width and 1,500 km in length. Beyond filtering out events that did not meet the 12 hr threshold, changes in temporal resolution had no effect on the results, as masks are first calculated at individual time steps.

AR IoU between two detection algorithms is calculated by dividing the number of grid points in which both algorithms detected an AR by the total number of grid points in which at least one algorithm detected an AR. While IoU is a useful metric to measure consistency of algorithms with each other, it is not perfect for AR detection algorithms because AR detection algorithms only attempt to identify one type of object. Some detection algorithms may disagree on what spatial area an AR occupies, but agree that there is an AR in a particular location. This is a common theme given that all AR detection algorithms have their own unique spatial biases. The training set used in this study used AR masks that were particularly larger than ARs detected from other studies (Inda-Díaz et al., 2021). An extreme example of this happened on 3 January 2006 at 18Z (Figure 4) and serves as a useful case study to demonstrate this phenomenon. All eight algorithms agree that an AR did exist in the eastern Pacific and did impact the west coast, so we can confidently say there was an AR there. While CG-Climate correctly decides that there was an AR at that location, the IoU with Mundhenk is only 0.307, which is less than the average IoU between those two algorithms. The main difference between these two detection results is the horizontal spatial extent of the AR rather than the central location in which it occurred or the existence of an AR existed within the domain at that time. The backbone of the AR identified from CG-Climate remained the same as it was in all eight ARTMIP algorithms. The discrepancy in spatial extent is accounted for by grid points that do not play as strong of a role in defining the main characteristics of the AR as grid points closer to the center despite having the same amount of relevance in AR IoU calculations. While we still consider mean AR IoU to be a useful and valid verification method, its limitations warrant the use of an additional verification method that focuses on the performance of detecting overlapping AR events regardless of discrepancies in shape or size.

In our AR event precision and recall calculations, we define individual AR events as clusters of AR grid points that do not have any space separating them. We classify events to overlap when they share at least one AR grid-point from the same event. We calculate precision by dividing the true positives (instances in which specified algorithm detects an AR that overlaps with given threshold of ARTMIP algorithms) by the sum of the true positives and false positives (instances in which the specified algorithm detects an AR that does not overlap with the

Journal of Advances in Modeling Earth Systems

10.1029/2022MS003495

Table 1
The Eight Different ARTMIP Algorithms We Compare CG-Climate to for Verification

Algorithm	Geometry	Threshold	Time req.	Region
Brands et al. (2017)	Length > 1,500 km	Relative and Absolute: 90th percentile IVT at point of detection, 85th percentile IVT along the AR structure, all months considered for threshold calculation, 250 kg m^{-1} s^{-1}, Spatial tracking guided by vector IVT	None	150°W to 30°E 30°N to 62°N
Tempest (Ullrich & Zarzycki, 2017)	Laplacian IVT thresholds most effective for widths >1,000 km, cluster size >120,000 km	IVT ≥ 250 kg m^{-1} s^{-1}	2 Days	Latitude >15°N
Gershunov et al. (2017)	Length ≥ 1,500 km, intersects with coastline	IVT ≥ 250 kg m^{-1} s^{-1}, IWV ≥ 15 mm	12 hr	Western US
Goldenson et al. (2018)	Length > 2,000 km, Width < 1,000 km, Object recognition, intersects with coastline	IWV > 20 mm	None	Western US
Guan and Waliser (2015)	Length > 2,000 km, Length-width ratio > 2, Coherent IVT direction within 45° of AR shape orientation and with a poleward componenet	85th percentile IVT, IVT > 100 kg m^{-1} s^{-1} in polar locations	None	Global
Reid et al. (2020)	Length > 2,000 km, Length-width ratio > 2, Orientation angle > 10°	IVT > 250 kg m^{-1} s^{-1}	None	Global
Mundhenk et al. (2016)	Length > 1,400 km Aspect ratio 1:4 lat limit >16°N/S Axis orientation based on IVT	Relative IVT percentiles and anomalies both temporal and spatial	None	Global
Rutz et al. (2014)	Length ≥ 2,000 km	IVT > 250 kg m^{-1} s^{-1}	None	Global
CG-Climate	None	None	12 hr	Global

Note. All eight except for Goldenson require wind velocity at multiple vertical levels, which requires additional computational resources. The information provided in this table was taken from https://www.cgd.ucar.edu/projects/artmip/algorithms.html

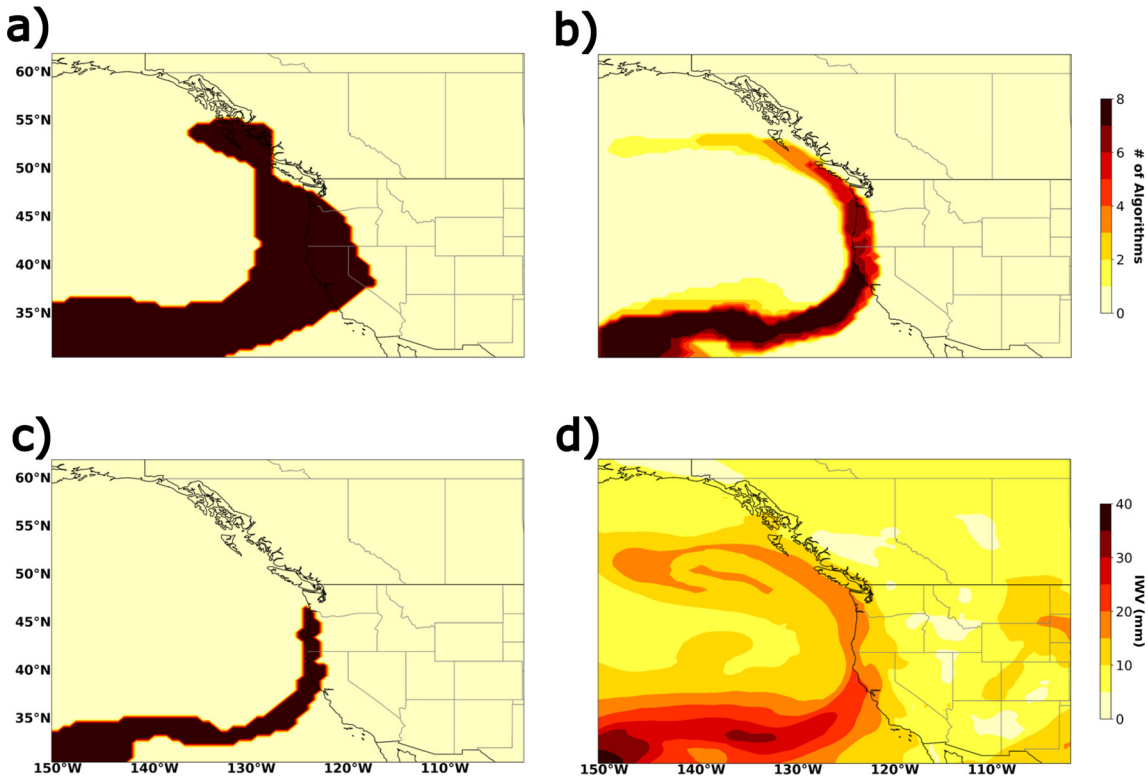

Figure 4. A display of an AR that occurred on 3 January 2006. (a) The area of the AR when identified with CG-Climate. (b) The area of the AR when identified by Mundhenk et al. (2016). (c) A composite of the areas of the AR when identified by eight different ARTMIP algorithms. (d) A spatial plot of IWV.

given threshold of ARTMIP algorithms). We calculate recall by dividing the true positives by the sum of the true positives and false negatives (instances in which the specified algorithm does not detect an AR i.e., detected by the given threshold of ARTMIP algorithms). The thresholds used are the ranges from each possible number of ARTMIP algorithms to eight.

CG-Climate can be trained on varying domains and resolutions that do not equal those of the inference set. Training the model on the largest available domain allows the user to use the maximum amount of available training data. However, training on a domain that is different from the inference data domain also creates a degree of inconsistency between training and inference data. Since the original training data set resolution was higher than the inference set resolution, a similar dilemma occurs when the model is trained on a resolution that is higher than that of the inference set. To understand the impact of training the model with domains and resolutions that are different from those of the inference set, we first train CG-Climate on three different datasets. The inference set remains the same for all three runs and uses the small domain (Figure 5b). The first using an equal domain and equal resolution to the inference set (Eq Dom/Eq Res), the second using an equal domain and the native 25 km CAM5.1 resolution rather than the re-gridded MERRA-2 0.625° longitude × 0.5° latitude resolution (Eq Dom/High Res), and the third using the full domain (Figure 5a) and equal resolution to the training set (Full Dom/Eq Res). The performance of each output is evaluated by the AR IoU with all eight other algorithms that tracked ARs in the same data set over the same time period within the evaluation domain (Figure 5d). All ARs were tracked in MERRA-2 data with both CG-Climate and ARTMIP methods.

In Figure 6, box-whisker plots are calculated from all AR IoU values between a given algorithm and all remaining ARTMIP algorithms. Based on this data, CG-Climate is the most consistent with the ARTMIP algorithms when the training data domain and resolution is equal to that of the inference set. When CG-Climate is trained on the same domain and same resolution as that of the inference set, AR IoU is similar to that of various ARTMIP algorithms. It also has more consistency than Goldenson, which understandably has the lowest IoU values because it is the only heuristic based ARTMIP algorithm shown here that does not require wind velocity at multiple vertical levels as an input variable. These high AR IoU values exist in spite of a significant positive areal extent

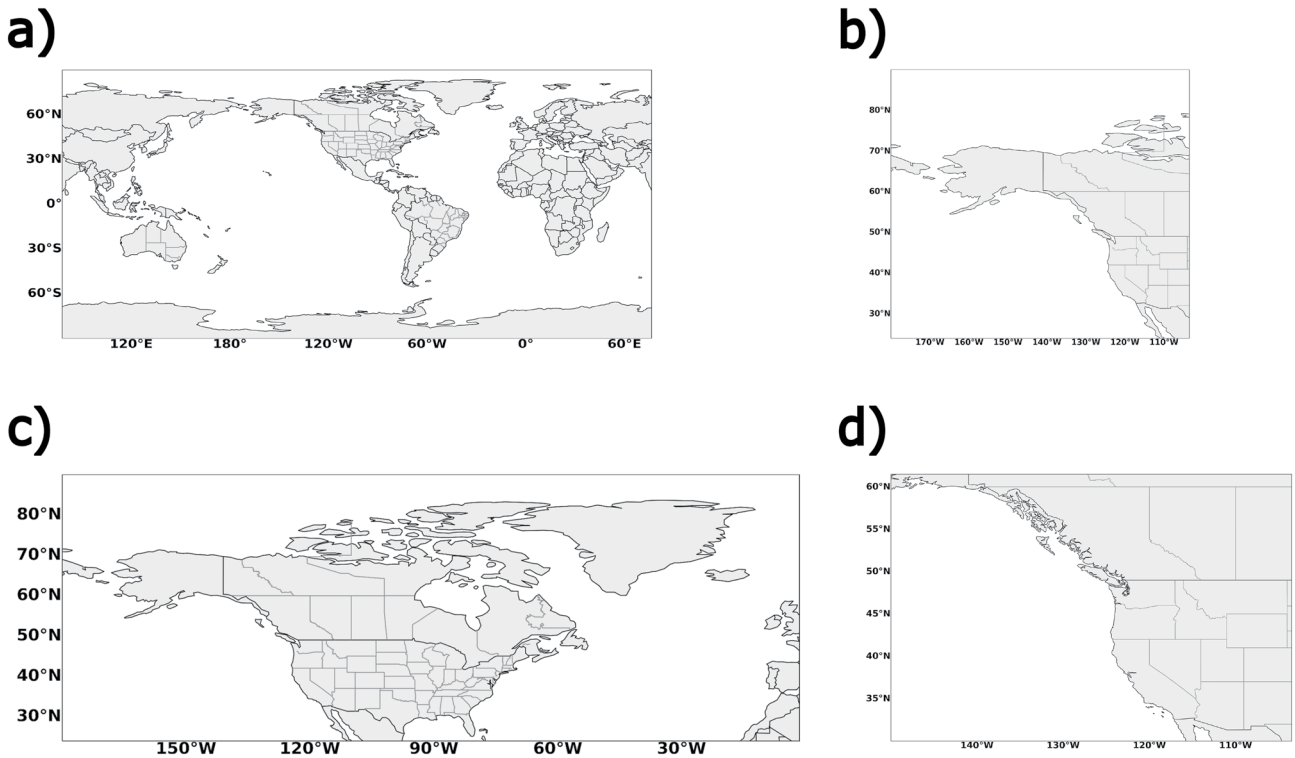

Figure 5. All four domains used for training and analysis. (a) The original training domain (Full). (b) A domain only containing the northeastern Pacific (Small). (c) A domain containing two active regions of AR activity in output from the Weather@Home data set (North Atlantic and northeastern Pacific) (Large). (d) The domain used for comparisons between ARTMIP algorithms (Evaluation).

bias mentioned in Kapp-Schwoerer and Graubner (2020). For the rest of our analysis we will use the output of CG-Climate when trained on the same domain and resolution as that of the inference set to demonstrate its performance.

Weather@Home simulations output data in regional domains to conserve computational resources. In some cases, the regional output contains multiple regions that experience high levels of AR activity and have varying AR climatologies. Training the model on a domain specific to the climatology the user is investigating requires less computational resources and enables the model to be trained and run at faster speeds compared to training on the full domain. ARs that occur in areas with different climatologies also have varying characteristics regardless of detection algorithm (Inda-Díaz et al., 2021). However, cropping the domain also reduces the amount of training data. To better understand this relationship, we trained and ran CG-Climate on three different domains (Figure 5).

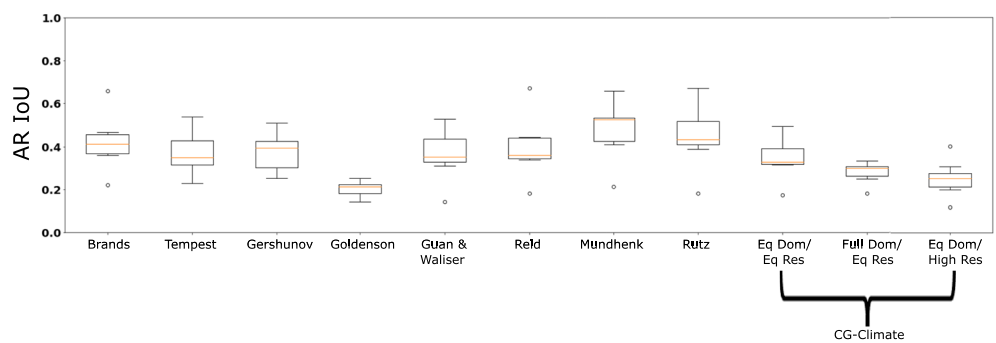

Figure 6. Box plots consisting of all values of AR IoU between each given detection method and all remaining ARTMIP algorithms used for evaluation. Every box that represents an ARTMIP algorithm does not include mean AR IoU values between itself and any of the variations of CG-Climate.

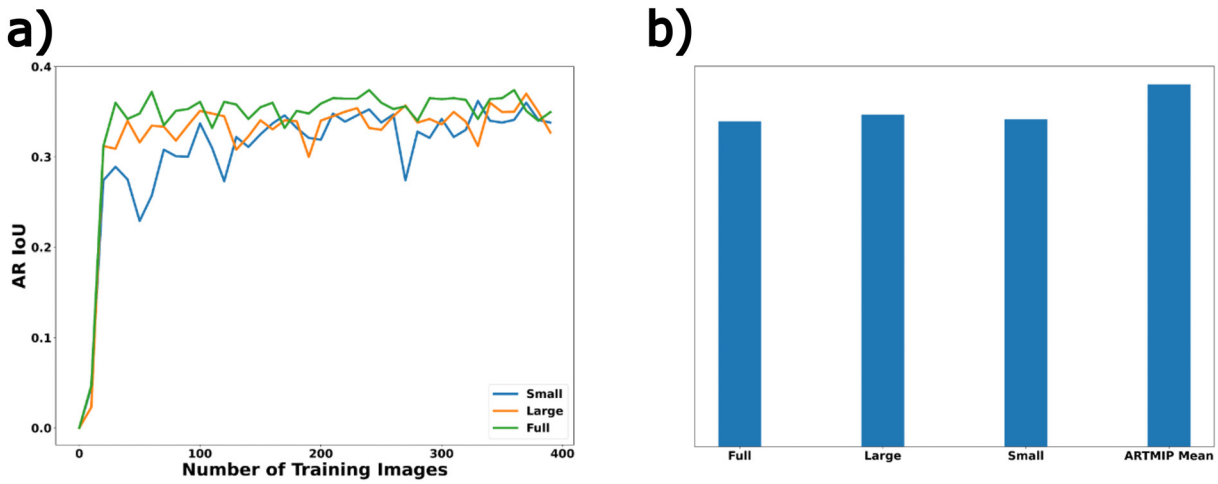

Figure 7. AR IoU of the output of CG-Climate with each of the eight ARTMIP algorithms over the evaluation domain when trained on each of the three domains referred to in (a–c). (a) The AR IoU for each domain when trained on varying amounts of training images. (b) The AR IoU for each domain when trained on 400 training images. The ARTMIP mean represents the AR IoU value of all eight ARTMIP algorithms when compared to each other.

Outputs from the model when trained on the domains seen in Figures 5a–5c were cropped to equal the domain in 5d before being compared to outputs from each of the ARTMIP algorithms. The domain in 5d was chosen because it is the largest domain that fits within the domains used in all eight ARTMIP algorithms. We found that there was not a discernible difference in AR IoUs between any of the three domains trained on CG-Climate and eight ARTMIP algorithms (Figure 6b). The AR IoU values between each of the outputs and all eight ARTMIP algorithms was slightly less than that of all eight ARTMIP algorithms when compared to each other.

Running CG-Climate on the training and inference domain in the small domain (Figure 5b) was 4.5 times as fast and required 4.5 times less computational memory than running it on the full domain (Figure 5a) when the same number of training images were used. The time and computational memory saved by using a smaller domain does not scale linearly to the difference in the number of grid points, as the ratio of the number of grid points in the small domain to the number of grid points in the full domain is roughly 1:12. Training the model with the small domain results in noticeably less consistency with ARTMIP algorithms when there are less than 150 images in the training set. However, once more images are added to the training set, CG-Climate appears to have similar consistency with ARTMIP algorithms across all three domains. All three domains had similar consistency to the ARTMIP algorithms on average given the amount of available training data. We therefore further evaluate performance based on the output of AR masks run on MERRA-2 data in the small domain (Figure 5b) and trained on human hand labels within the small domain at the same MERRA-2 spatial resolution as that of the inference set.

4. Results and Discussion

To estimate the comparative speed and computational saving of CG-Climate, we ran Guan and Waliser (2015) version 3 on MERRA-2 data to serve as a characteristic algorithm with similar computational efficiency to other traditional heuristic methods. CG-Climate completed processing of the Weather@Home data (4,000 winter simulations) in 10 hr and allocated 0.36 TB of computational memory. By scaling the result of running Guan and Waliser on a lesser amount of MERRA-2 data in the same spatial domain, we estimate Guan and Waliser would take 23 days to process the Weather@Home data set and allocate 92 TB of computational memory if the required input variables were available. This efficiency comes in addition to the computational resources saved from Weather@Home only producing wind velocity at one vertical level in its output.

Although CG-Climate had similar IoU values to those of the ARTMIP methods (Figures 6 and 7), the earlier noted areal extent bias calls for an additional metric to be used to understand the performance of CG-Climate. An alternative useful metric to look at is the consistency in which algorithms identify the same AR events. We classify multiple algorithms to identify the same event when any two individual AR masks have at least one overlapping grid point. Figure 8a shows the consistency of event overlap between each detection algorithm and the remaining ARTMIP algorithms. Each data point represents the precision and recall of a given algorithm at a

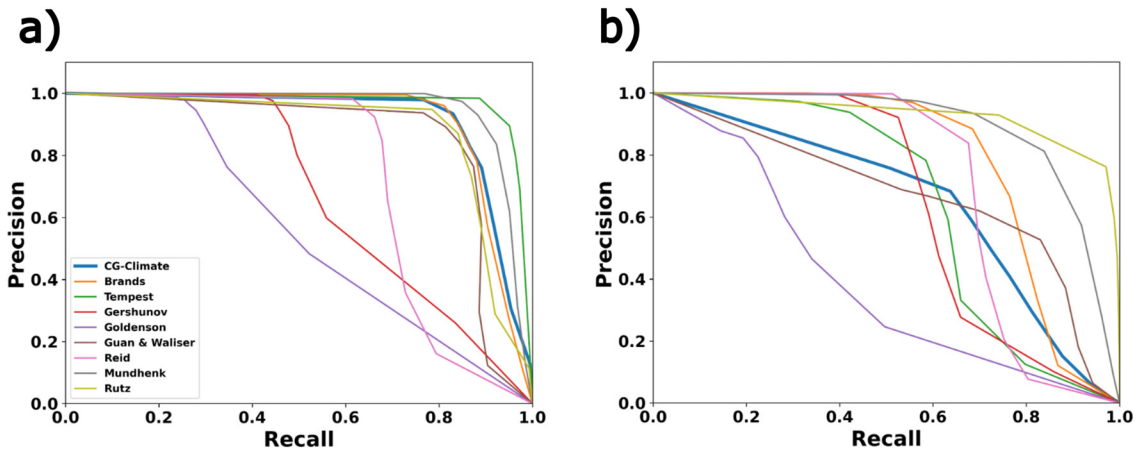

Figure 8. Precision and recall between each detection algorithm and all other ARTMIP algorithms (truth). The thresholds used are the number of algorithms that overlap for any given AR. (a) Precision and recall values between AR events. (b) Precision and recall values between AR grid points.

different threshold. Greater areas under precision-recall curves indicate more consistency with the truth (Boyd et al., 2013). By this metric, there was a high amount of variability between all algorithms. In less than 10% of AR instances, all eight ARTMIP algorithms agreed that there was an AR. When CG-Climate found an AR, at least half of the ARTMIP algorithms also found an AR in that same location 82% of the time. When at least half of the ARTMIP algorithms found an AR, CG-Climate found an AR 82.5% of the time. This indicates that CG-Climate does not have a large bias in frequency of AR events and generally detects events in the same location as ARTMIP algorithms.

Figure 8b shows precision and recall curves for AR grid points rather than AR events. In this plot, precision is defined as the percentage in which AR grid points detected by a specified algorithm are overlapping within a given threshold of ARTMIP algorithms. Recall is the percentage in which AR grid points detected within a given threshold of ARTMIP algorithms are overlapping with a specified detection algorithm. By this metric, CG-Climate is far less consistent with other algorithms than the previous metric while still being within the range of other algorithms. Overall, precision and recall values are far higher when used as a measure of consistency between events instead of consistency between grid points, indicating that spatial area bias generally accounts for more inconsistency between overall AR IoU values than event frequencies and event locations. To further understand the primary driver of IoU inconsistencies for each algorithm, we refer to Figure 9.

When applied to weather event identification, the components that are used to calculate IoU can be broken down into three categories. The first being the area in which grid points intersect with each other, the second being the area in which grid points do not intersect with each other, but are associated with an event that intersects, and the third being the area in which grid points do not intersect with each other and are not associated with an event that intersects (Figure 1). Figure 9 shows the percentage of total non-intersecting grid points that are associated with a matching event for each algorithm. Out of all the AR grid points used to calculate the AR IoU between CG-Climate and the ARTMIP algorithms, 67.6% of them were associated with matching events.

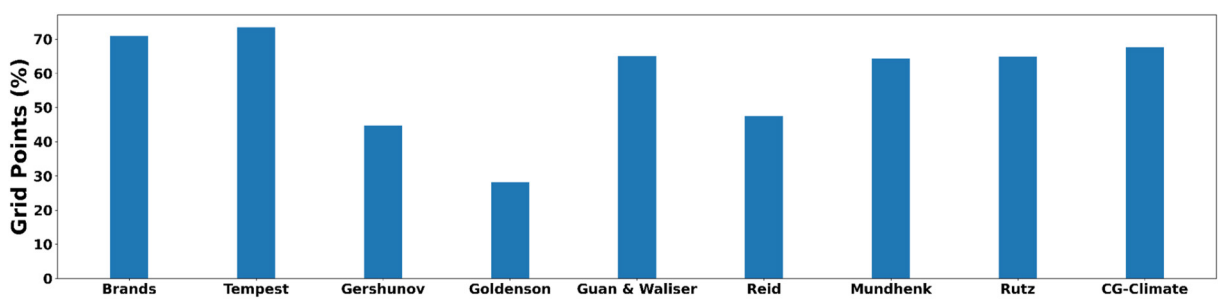

Figure 9. The percentage of non-intersecting grid points that are associated with matching events in IoU calculations for each algorithm.

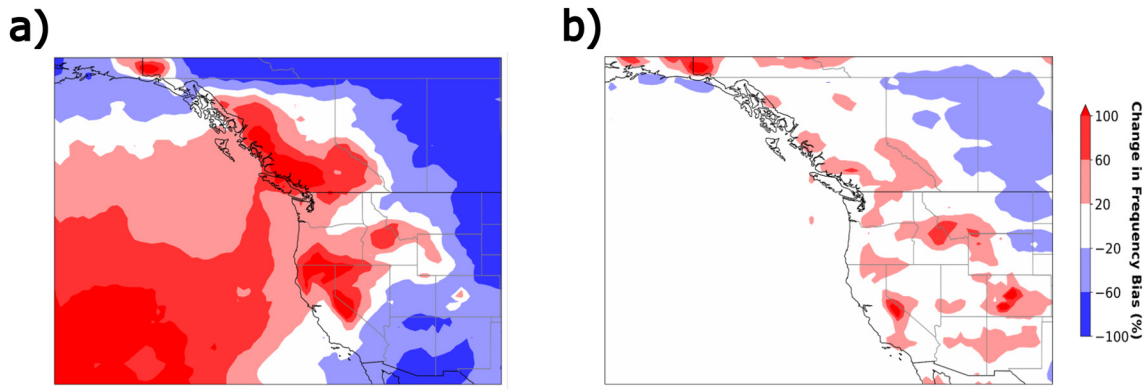

Figure 10. Spatial plots of the change in mean ARTMIP detected AR gridpoint frequency to mean CG-Climate AR gridpoint frequency. (a) CG-Climate trained on human hand labels (b) CG-Climate trained on ARTMIP labels. All training sets contained 400 feature label pair images.

The spatial extent of ARs is a relevant characteristic of ARs that may change in the future (Inda-Díaz et al., 2021; O'Brien et al., 2022). The mean area of ARTMIP ARs in this study is 1.6 million km^2 and individual algorithm area means range from 640,000 km^2 to 2.5 million km^2. The mean area of ARs detected by CG-Climate is 2.1 million km^2. While the average area of ARs detected by CG-Climate is certainly above the ARTMIP mean, it is still within the range of variance. To further investigate the spatial area bias of CG-Climate, we train it again on ARTMIP labels from each algorithm rather than expert labels. For this experiment, the training data consisted of MERRA-2 data from January to February of the years 1995–2004 and the inference data once again consisted of January to February of the years 2006–2015. Each output was only compared to the ARTMIP algorithm that was used to detect the labels that it was trained on. When CG-Climate was trained on ARTMIP labels rather than human hand labels, there was a 63.5% reduction in non-intersecting grid points that were associated with matching events and a 43.8% reduction in the number of events that do not match at all. This result shows that CG-Climate's spatial area bias can be largely attributed to the spatial area bias of the training data labels rather than a deficiency in the neural network architecture.

The use of human expert hand labels is a useful method of finding an alternative perspective to track ARs and may indicate that there is relevant missing information in traditional tracking algorithms that is critical to the identity of ARs, but its differences must be acknowledged when used. Figures 10a and 10b shows the percentage change between the average frequency in which each ARTMIP algorithm detects an event and the average frequency in which CG-Climate detects an event at each grid point in the spatial domain. There is a sharp decrease in AR grid point frequency biases when CG-Climate is trained on ARTMIP labels in comparison to human hand labels. While training CG-Climate on hand labels is still an effective neutral solution toward choosing a neutral training set that is not particularly biased toward any particular ARTMIP algorithm, it must be noted that using hand labels also causes it to detect AR grid points more frequently than ARTMIP algorithms, largely resulting from the tendency of hand labels to be larger than heuristic-based calculations.

The origin of the spatial area bias in human hand labels remains a question. One possible explanation is that it is rooted in the ambiguity of the task of identifying ARs. The overwhelming majority of semantic segmentation image identification tasks focus on solid objects with rigid edges. ARs are objects that exist in a fluid medium and do not have clean and well-defined edges. In the application of AR identification, they exist in a space with the background class accounting for most pixels and the event class accounting for a relatively small minority. This could create a dynamic in which labelers mentally classify the background area to have less importance than the event area and prioritize the inclusion of event pixels in their labels more than the exclusion of background pixels. Additionally, contour labels were made on a global domain, which could create the challenge of making precise labels. Despite deviations from ARTMIP methods, the inclusion of additional space close to ARs is likely a result of a unique perspective added to AR detection and may help benefit future studies that aim to quantify interactions between ARs and large-scale synoptic systems. The immediate space around ARs could be highly relevant to the flow, the development, the lifetime, and the structure of them.

By running CG-Climate on multiple datasets, we can better understand its flexibility. CG-Climate was run on over 1,200 January and February months of data from the Weather@Home project historical scenario that uses

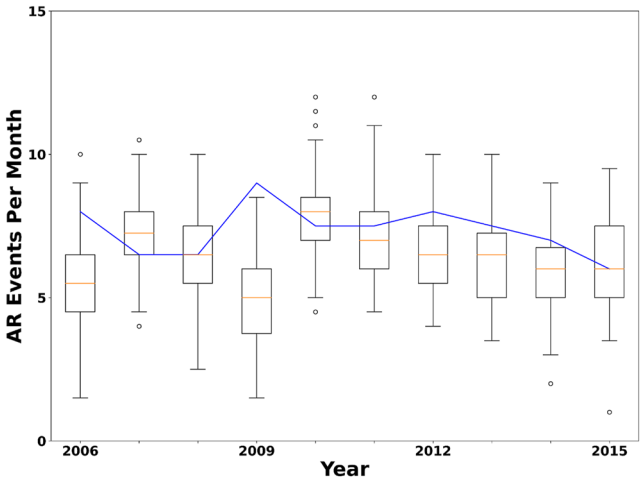

Figure 11. Amount of ARs detected per month in MERRA-2 data (line) and Weather@Home data (box whisker plots) using CG-Climate.

radiative forcing from 2006 to 2015. Figure 11 shows the ensemble variability of the amount of ARs detected per month in the Weather@Home data (box and whisker) and the amount of ARs detected per month in MERRA-2 data (blue line) during the same time period at a 6 hr temporal resolution. Overall, the number of ARs found per month was similar in both datasets, indicating that CG-Climate can be compatible with various simulations and spatiotemporal resolutions. An average of 6.35 ARs were detected per month in the Weather@Home data, which is slightly less than the average of 7.35 ARs per month found in the MERRA-2 data. The slight decrease in ARs per month is consistent with the slight negative winter precipitation and extreme precipitation bias that HadAM4 has in the North Pacific (Bevacqua et al., 2021; Watson et al., 2020) and the relatively low sample size of reanalysis data enabling natural variability to have a substantial impact on the average.

5. Conclusion

CG-Climate is able to identify atmospheric rivers at fast speeds and requires a relatively low amount of computational resources. It can do this without some key input variables that are required in other popular detection methods. Given the amount of available AR expert hand-labeled contour data available for public use, there is not a noticeable decline in performance when CG-Climate is trained and tested on regional domains. This may become increasingly relevant to future studies, as regional climate change projections are particularly complicated to make and are currently not well understood (Collins et al., 2018). The model has a high level of consistency with ARTMIP algorithms for varying spatial domains. Although some heuristics are found in multiple ARTMIP algorithms, CG-Climate still has IoU values that are similar to various detection algorithms and frequently detects ARs in the same location and time as ARTMIP algorithms. The main source of inconsistency between CG-Climate and ARTMIP algorithms is its spatial area bias. However, we found that the spatial area bias can largely be attributed to the choice in training data rather than the neural network architecture itself. CG-Climate is a useful tool that can be used to identify ARs in large climate datasets and will likely facilitate future studies of shifts in western precipitation resulting from a changing climate as datasets continue to get larger and the computational and time cost of tracking processes within them continues to grow.

Data Availability Statement

The source code and hand-labeled training data used to run CG-Climate is available at NERSC Science Gateways and can be found at https://portal.nersc.gov/project/ClimateNet/ (Kapp-Schwoerer & Graubner, 2020). Source data for the full MERRA-2 Tier 1 catalogues are available from the Climate Data Gateway (CDG), https://doi.org/10.5065/D62R3QFS (NCAR/UCAR Climate Data Gateway, 2018). Participation in ARTMIP is open to any person or group with an AR detection scheme and/or interest in analyzing data produced by ARTMIP (Christine Shields, 2018). The MERRA-2 data used in this study can be found at https://rda.ucar.edu/datasets/ds313-3/ (Atmospheric Chemistry Observations & Modeling, National Center for Atmospheric Research, University Corporation for Atmospheric Research & Climate and Global Dynamics Division, National Center for Atmospheric Research, University Corporation for Atmospheric Research, 2018). The Weather@Home data used in this study are available through Bevacqua, Watson, et al. (2020) at http://doi.org/10.5281/zenodo.4311221 (Bevacqua et al., 2020).

Acknowledgments
This work is supported by the California Department of Water Resources Ph3 Atmospheric River Research Program (Award 4600014294) and the Forecast Informed Reservoir Operations Award (USACE W912HZ1920023). This work is also supported by the Natural Environmental Research Council Independent Research Fellowship (Grant NE/S014713/1). The authors thank the reviewers of this work for their dedication towards supporting the research community.

References

Atmospheric Chemistry Observations & Modeling, National Center for Atmospheric Research, University Corporation for Atmospheric Research, & Climate and Global Dynamics Division, National Center for Atmospheric Research, & University Corporation for Atmospheric Research. (2018). *Merra2 global atmosphere forcing data*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/XVAQ-2X07

Bevacqua, E., Shepherd, T. G., Watson, P. A. G., Sparrow, S., Wallom, D., & Mitchell, D. (2021). Larger spatial footprint of wintertime total precipitation extremes in a warmer climate. *Geophysical Research Letters*, *48*(8), e2020GL091990. https://doi.org/10.1029/2020GL091990

Bevacqua, E., Watson, P., Sparrow, S., & Wallom, D. (2020). Multi-thousand-year simulations of December-February precipitation and zonal upper-level wind. *Zenodo*. https://doi.org/10.5281/ZENODO.4311221

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals, (Eds.). In D. Hutchison, et al., *Advanced information systems engineering* (Vol. 7908, pp. 451–466). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40994-3_29

Brands, S., Gutierrez, J. M., & San-Mart in, D. (2017). Twentieth-century atmospheric river activity along the west coasts of Europe and North America: Algorithm formulation, reanalysis uncertainty and links to atmospheric circulation patterns. *Climate Dynamics*, *48*, 2771–2795. https://doi.org/10.1007/s00382-016-3095-6

Chikamoto, Y., Kimoto, M., Ishii, M., Mochizuki, T., Sakamoto, T. T., Tatebe, H., et al. (2013). An overview of decadal climate predictability in a multi-model ensemble by climate model MIROC. *Climate Dynamics*, *40*(5–6), 1201–1222. https://doi.org/10.1007/s00382-012-1351-y

Christine Shields, S. (2018). 3-hourly MERRA2 IVT, uIVT, vIVT, IWV data computed for ARTMIP. [Dataset]. NCAR/UCAR Climate Data Gateway. https://doi.org/10.5065/D62R3QFS

Collins, M., Minobe, S., Barreiro, M., Bordoni, S., Kaspi, Y., Kuwano-Yoshida, A., et al. (2018). Challenges and opportunities for improved understanding of regional climate dynamics. *Nature Climate Change*, *8*(2), 101–108. https://doi.org/10.1038/s41558-017-0059-8

Delworth, T. L., Rosati, A., Anderson, W., Adcroft, A. J., Balaji, V., Benson, R., et al. (2012). Simulated climate and climate change in the GFDL CM2.5 high-resolution coupled climate model. *Journal of Climate*, *25*(8), 2755–2781. https://doi.org/10.1175/JCLI-D-11-00316.1

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., et al. (2020). Publisher correction: Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, *10*(8), 791. https://doi.org/10.1038/s41558-020-0854-5

Espinoza, V., Waliser, D. E., Guan, B., Lavers, D. A., & Ralph, F. M. (2018). Global analysis of climate change projection effects on atmospheric rivers. *Geophysical Research Letters*, *45*(9), 4299–4308. https://doi.org/10.1029/2017GL076968

Gershunov, A., Shulgina, T., Clemesha, R. E. S., Guirguis, K., Pierce, D. W., Dettinger, M. D., et al. (2019). Precipitation regime change in Western North America: The role of Atmospheric Rivers. *Scientific Reports*, *9*(1), 9944. https://doi.org/10.1038/s41598-019-46169-w

Gershunov, A., Shulgina, T., Ralph, F. M., Lavers, D. A., & Rutz, J. J. (2017). Assessing the climate-scale variability of atmospheric rivers affecting western North America. *Geophysical Research Letters*, *44*(15), 7900–7908. https://doi.org/10.1002/2017GL074175

Goldenson, N., Leung, L. R., Bitz, C. M., & Blanchard-Wrigglesworth, E. (2018). Influence of atmospheric rivers on mountain snowpack in the western United States. *Journal of Climate*, *31*(24), 9921–9940. https://doi.org/10.1175/JCLI-D-18-0268.1

Guan, B., & Waliser, D. E. (2015). Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies: Detection of atmospheric rivers. *Journal of Geophysical Research: Atmospheres*, *120*(24), 12514–12535. https://doi.org/10.1002/2015JD024257

Guillod, B. P., Jones, R. G., Bowery, A., Haustein, K., Massey, N. R., Mitchell, D. M., et al. (2017). weather@home 2: Validation of an improved global–regional climate modelling system. *Geoscientific Model Development*, *10*(5), 1849–1872. https://doi.org/10.5194/gmd-10-1849-2017

Inda-Díaz, H. A., O'Brien, T. A., Zhou, Y., & Collins, W. D. (2021). Constraining and characterizing the size of atmospheric rivers: A perspective independent from the detection algorithm. *Journal of Geophysical Research: Atmospheres*, *126*(16). https://doi.org/10.1029/2020JD033746

Johnson, F., & Sharma, A. (2009). Measurement of GCM skill in predicting variables relevant for hydroclimatological assessments. *Journal of Climate*, *22*(16), 4373–4382. https://doi.org/10.1175/2009JCLI2681.1

Kapp-Schwoerer, L., & Graubner, A. (2020). Spatio-temporal segmentation and tracking of weather patterns with light-weight Neural Networks (Vol. *5*).

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (2015). The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Lavers, D. A., Waliser, D. E., Ralph, F. M., & Dettinger, M. D. (2016). Predictability of horizontal water vapor transport relative to precipitation: Enhancing situational awareness for forecasting western U.S. Extreme precipitation and flooding: Predictability of western U.S. Extremes. *Geophysical Research Letters*, *43*(5), 2275–2282. https://doi.org/10.1002/2016GL067765

Massey, N., Jones, R., Otto, F. E. L., Aina, T., Wilson, S., Murphy, J. M., et al. (2015). weather@home—Development and validation of a very large ensemble modelling system for probabilistic event attribution. *Quarterly Journal of the Royal Meteorological Society*, *141*(690), 1528–1545. https://doi.org/10.1002/qj.2455

Mitchell, D., AchutaRao, K., Allen, M., Bethke, I., Beyerle, U., Ciavarella, A., et al. (2017). Half a degree additional warming, prognosis and projected impacts (HAPPI): Background and experimental design. *Geoscientific Model Development*, *10*(2), 571–583. https://doi.org/10.5194/gmd-10-571-2017

Mundhenk, B. D., Barnes, E. A., & Maloney, E. D. (2016). All-season climatology and variability of atmospheric river frequencies over the North Pacific. *Journal of Climate*, *29*(13), 4885–4903. https://doi.org/10.1175/JCLI-D-15-0655.1

Neiman, P. J., Ralph, F. M., Wick, G. A., Lundquist, J. D., & Dettinger, M. D. (2008). Meteorological characteristics and overland precipitation impacts of atmospheric rivers affecting the West Coast of North America based on eight years of SSM/I satellite observations. *Journal of Hydrometeorology*, *9*(1), 22–47. https://doi.org/10.1175/2007JHM855.1

O'Brien, T. A., Wehner, M. F., Payne, A. E., Shields, C. A., Rutz, J. J., Leung, L., et al. (2022). Increases in future AR count and size: Overview of the ARTMIP tier 2 CMIP5/6 experiment. *Journal of Geophysical Research: Atmospheres*, *127*(6), e2021JD036013. https://doi.org/10.1029/2021JD036013

Payne, A. E., Demory, M.-E., Leung, L. R., Ramos, A. M., Shields, C. A., Rutz, J. J., et al. (2020). Responses and impacts of atmospheric rivers to climate change. *Nature Reviews Earth & Environment*, *1*(3), 143–157. https://doi.org/10.1038/s43017-020-0030-5

Polade, S. D., Gershunov, A., Cayan, D. R., Dettinger, M. D., & Pierce, D. W. (2017). Precipitation in a warming world: Assessing projected hydro-climate changes in California and other Mediterranean climate regions. *Scientific Reports*, *7*(1), 10783. https://doi.org/10.1038/s41598-017-11285-y

Prabhat, P., Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., et al. (2021). ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development*, *14*(1), 107–124. https://doi.org/10.5194/gmd-14-107-2021

Ralph, F. M., Rutz, J. J., Cordeira, J. M., Dettinger, M., Anderson, M., Reynolds, D., et al. (2019). A scale to characterize the strength and impacts of atmospheric rivers. *Bulletin of the American Meteorological Society*, *100*(2), 269–289. https://doi.org/10.1175/BAMS-D-18-0023.1

Reid, K. J., King, A. D., Lane, T. P., & Short, E. (2020). The sensitivity of atmospheric river identification to integrated water vapor transport threshold, resolution, and regridding method. *Journal of Geophysical Research: Atmospheres*, *125*(20), e2020JD032897. https://doi.org/10.1029/2020JD032897

Rutz, J. J., Steenburgh, W. J., & Ralph, F. M. (2014). Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Monthly Weather Review*, *142*(2), 905–921. https://doi.org/10.1175/MWR-D-13-00168.1

Shields, C. A., & Kiehl, J. T. (2016). Atmospheric river landfall-latitude changes in future climate simulations. *Geophysical Research Letters*, *43*(16), 8775–8782. https://doi.org/10.1002/2016GL070470

Shields, C. A., Rutz, J. J., Leung, L.-Y., Ralph, F. M., Wehner, M., Kawzenuk, B., et al. (2018). Atmospheric River tracking method Intercomparison project (ARTMIP): Project goals and experimental design. *Geoscientific Model Development*, *11*(6), 2455–2474. https://doi.org/10.5194/gmd-11-2455-2018

Stansfield, A. M., Reed, K. A., & Zarzycki, C. M. (2020). Changes in precipitation from North Atlantic tropical cyclones under RCP Scenarios in the variable-resolution community atmosphere model. *Geophysical Research Letters*, *47*(12), e2019GL086930. https://doi.org/10.5194/gmd-10-1069-2017

Ullrich, P. A., & Zarzycki, C. M. (2017). TempestExtremes: A framework for scale-insensitive pointwise feature tracking on unstructured grids. *Geoscientific Model Development*, *10*(3), 1069–1090. https://doi.org/10.5194/gmd-10-1069-2017

Warner, M. D., Mass, C. F., & Salathé, E. P. (2015). Changes in winter atmospheric rivers along the North American west coast in CMIP5 climate models. *Journal of Hydrometeorology*, *16*(1), 118–128. https://doi.org/10.1175/JHM-D-14-0080.1

Watson, P., Sparrow, S., Ingram, W., Wilson, S., Marie, D., Zappa, G., et al. (2020). Multi-thousand member ensemble atmospheric simulations with global 60 km resolution using climateprediction.net (other). Oral. https://doi.org/10.5194/egusphere-egu2020-10895

Williams, A. P., Cook, B. I., & Smerdon, J. E. (2022). Rapid intensification of the emerging southwestern North American megadrought in 2020–2021. *Nature Climate Change*, *12*(3), 232–234. https://doi.org/10.1038/s41558-022-01290-z

Wu, T., Tang, S., Zhang, R., & Zhang, Y. (2019). CGNet: A light-weight context guided network for semantic segmentation. arXiv:1811.08201 [cs] Retrieved from http://arxiv.org/abs/1811.08201

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. (arXiv:1505.00853 [cs, stat]) Retrieved from http://arxiv.org/abs/1505.00853

Zhang, C., Tung, W., & Cleveland, W. S. (2021). In search of the optimal atmospheric river index for US precipitation: A multifactorial analysis. *Journal of Geophysical Research: Atmospheres*, *126*(10), e2020JD033667. https://doi.org/10.1029/2020JD033667

Zhao, M. (2020). Simulations of atmospheric rivers, their variability, and response to global warming using GFDL's new high-resolution general circulation model. *Journal of Climate*, *33*(23), 10287–10303. https://doi.org/10.1175/JCLI-D-20-0241.1