IMPROVING EXTREME WEATHER EVENTS DETECTION WITH LIGHT-WEIGHT NEURAL NETWORKS

Romain Lacombe
Stanford University
Plume Labs
rlacombe@stanford.edu

Hannah Grossman
Stanford University
hlg@stanford.edu

Lucas Hendren
Stanford University
hendren@stanford.edu

David Lüdeke
Stanford University
dludeke@stanford.edu

ABSTRACT

To advance automated detection of extreme weather events, which are increasing in frequency and intensity with climate change, we explore modifications to a novel light-weight Context Guided convolutional neural network architecture trained for semantic segmentation of tropical cyclones and atmospheric rivers in climate data. Our primary focus is on tropical cyclones, the most destructive weather events, for which current models show limited performance. We investigate feature engineering, data augmentation, learning rate modifications, alternative loss functions, and architectural changes. In contrast to previous approaches optimizing for intersection over union, we specifically seek to improve recall to penalize under-counting and prioritize identification of tropical cyclones. We report success through the use of weighted loss functions to counter class imbalance for these rare events. We conclude with directions for future research on extreme weather events detection, a crucial task for prediction, mitigation, and equitable adaptation to the impacts of climate change.

1 INTRODUCTION

Climate action failure and extreme weather are two of the most severe global risks today (IPCC, 2022; World Economic Forum, 2022). Tropical cyclones, the most destructive extreme weather events (NOAA, 2022), have a rising and disproportionate impact on low and medium income countries (LMICs), yet research into their effects focuses mostly on high-income countries (Parks & Guinto, 2022). Studies of extreme weather and climate change rely on heuristics or expert judgment to label data which leads to an inequitable global scientific focus, as well as discrepancies in predicted frequency, intensity, and attribution estimates. Improving automated detection of extreme weather events is thus paramount to fair attribution of climate loss and damages (Philip et al., 2020), and to develop the early warning and detection systems that will be critical for equitable adaption to climate change (IPCC, 2022; Nguyen et al., 2013).

Since 2020, deep learning has shown great promise for semantic segmentation of weather patterns in climate simulation data (Prabhat et al., 2021). However, initial approaches have relied on complex architectures and hard to train models with very large numbers of parameters. A key area of research is the application of lighter-weight neural networks to semantic segmentation of tropical cyclones (TC) and atmospheric rivers (AR) (Kapp-Schwoerer et al., 2020b).

Here we explore the application of the light-weight Context Guided convolutional neural network (CGNet) architecture to semantic segmentation of tropical cyclones in climate data. Input to our model is hand-labeled climate simulation data with channels that contain key atmospheric variables such as wind speed, moisture content, and atmospheric pressure for different time steps, latitudes, and longitudes. The output is a segmentation mask where each pixel takes a value corresponding to the background (BG), TC, or AR classes.

Specific challenges include the very small dataset size, inherent class imbalance of infrequent extreme events, unavoidable bias due to subjective human labeling, and limited capacity of the light-weight network. We report experiments with different hyper-parameters (loss function, learning rate), architecture (up-sampling), data augmentation, and feature engineering. We find that weighted loss functions aimed at compensating class imbalance provide the most significant improvement on recall of extreme weather events.

2 RELATED WORK

Initial inspiration for this work came from Prabhat et al. (2021) which trained a DeepLabV3+ convolutional neural net on the *ClimateNet* expert-label dataset. This $\sim$50 million parameters model achieved an intersection over union (IoU) score (1) of 0.24 for TCs, and was the first to demonstrate that deep learning models trained on hand-labeled climate data could effectively perform semantic segmentation of extreme weather patterns. However, the DeepLabV3+ architecture is complex, heavy, and thus costly in terms of memory, training time, and associated carbon footprint.

In *Spatio-temporal segmentation and tracking of weather patterns with light-weight Neural Networks*, Kapp-Schwoerer et al. (2020b) attempt to perform the same segmentation task on the *ClimateNet* dataset with the much lighter-weight ($\sim$500,000 parameters) Context Guided neural architecture. They improve on Prabhat et al. (2021) with a IoU score of 0.34 and a recall of 0.57 for TCs, our primary class of interest. This model and its associated metrics form our performance baseline.

For a detailed presentation of Context Guided convolutional neural networks, we refer the reader to the original paper that introduced the CGNet architecture, *A light-weight Context Guided Network for semantic segmentation* by Wu et al. (2021). To solve the class imbalance problem, we experimented with various loss functions reviewed in *Survey of loss functions for semantic segmentation* (Jadon, 2020). Lastly, we relied on *Deep Learning for the Earth Sciences* (Mudigonda et al., 2021) for general background on applying deep learning techniques to Earth Sciences.

3 DATASET & FEATURES

We trained our neural net on *ClimateNet*, an open, community-sourced, human expert-labeled dataset of outputs from Community Atmospheric Model (CAM5.1) climate simulation runs for 459 time steps from 1996 to 2013. Each sample is a netCDF file containing a 1152×768 array for one simulation time step, with each pixel mapping to: one (latitude, longitude) point with 34.8 km/pixel horizontal and 26.1 km/pixel vertical resolution near the Equator; 16 channels for key atmospheric variables, described in table 2 and visualized in figure 4; and one ground truth class label. The dataset is split into a training set of 398 (map, labels) pairs from 1996 to 2010, and a test set of 61 (map, labels) pairs spanning 2011 to 2013. For learning rate scheduling, we created a validation set of 56 (map, labels) pairs spanning 2008 to 2010, which we set aside from the training set to keep the test set consistent with our baseline.

The implementation by Kapp-Schwoerer et al. (2020a) is trained on the following four channels: TMQ, total vertically integrated precipitable water; U850, zonal (east-west) winds at the 850 mbar pressure surface; V850, meridional (north-south) wind at the 850 mbar pressure surface; and PSL, atmospheric pressure at sea level. From the existing 16 channels, we engineered new features, *wind velocity* and *wind vorticity*, to help the model identify TCs since they are characterized by high wind speeds and rotation. Wind velocity is the L_2 norm of zonal and meridional components of the wind vector field (equation 11). Wind vorticity is the curl of the wind vector field around the earth radius axis (equation 10), a measure of the local rotation (Simpson, 2010). We pre-computed these engineered features at the 850 mbar pressure level and at the lowest altitude level.

The output of the model is a (1152×768) tensor of softmax probabilities for background, TC, or AR classes. Importantly, labels for the supervised learning of this task are segmentation maps that were hand-drawn by climate scientists as part of a community labeling exercise described in Prabhat et al. (2021). Figure 2 illustrates how labels were generated as a consensus between experts.

In an effort to reduce over-fitting to the relatively small training set, we explored data augmentation techniques. While transforming the image based on randomized longitude increments seemed promising, we observed that random translations along the longitudes dimension immediately de-

creased performance. We hypothesize that this may be due to the importance of geography (relative positioning of continents and oceans) for atmospheric circulation and weather patterns. As a consequence, rather than providing additional data for training, data augmentation may act as a detriment to learning by precluding the learning of accurate geographical representations.

4 METHODS

4.1 BASELINE IMPLEMENTATION AND PERFORMANCE

We established our baseline by training the Kapp-Schwoerer et al. (2020a) implementation of the CGNet architecture for 15 epochs over the *ClimateNet* training set, with a Jaccard loss (equation 4) based on the IoU for the 3 classes (background, AR, and TC).

We report recall as a key performance metric to minimize false negatives, which is especially important for identification of infrequent events. The baseline performance for TCs reaches an IoU score of 0.3396 and a recall of 0.5708 on the test set (see table 1). A higher performance on the train set (IoU score of 0.38 for TCs) indicates the model may also display some variance and over-fitting.

A fundamental challenge for climate event identification is the inherent imbalance of the data, since, by definition, the extreme events we aim to detect are very rare. We conclude from this analysis that the baseline implementation exhibits high bias, some variance, and relatively low recall.

4.2 CGNET ARCHITECTURE

The light-weight CGNet architecture introduced by Wu et al. (2021) follows the principle of "deep and thin" and is designed specifically for high-accuracy semantic segmentation while maintaining a small memory footprint. This is advantageous for reducing training time and model complexity.

Context Guided block. The basic unit of CGNet is a Context Guided (CG) block, presented in figure 1, which concatenates the output of normal and dilated convolutional layers to integrate local and surrounding context respectively. It uses 1x1 convolutions and average pooling to further refine the representation using a global context. The CG block reduces parameter count and memory footprint by employing channel-wise convolutions to lower computational cost across channels.

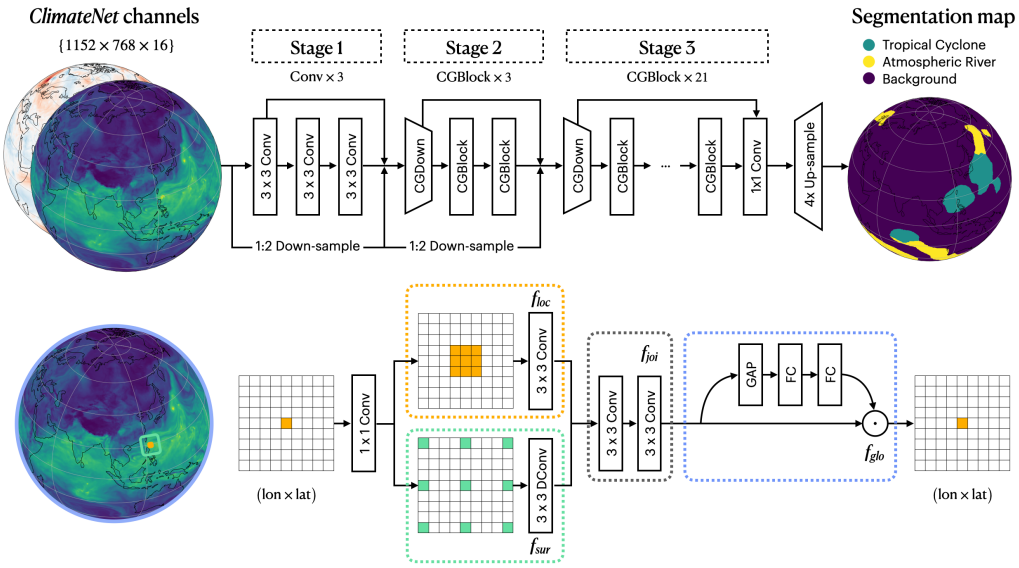

Figure 1: Above: Context Guided convolutional neural network (CGNet). Below: Context Guided block (CG) consisting of local feature extractor f_{loc}, surrounding context extractor f_{sur}, joint feature extractor f_{join}, and global context extractor f_{glo} where $\odot$ represents element-wise multiplication.

Architectural experimentations. In order to improve performance, we experimented with additional CNN + BatchNorm + ReLU layers to the model to produce a deeper network with the goal of learning more complex features. We also experimented with doubling the final up-sampling layer to increase resolution of the output predictions. Both of these attempts were unsuccessful at significantly improving performance.

Learning rate scheduler. Experimenting with learning rates greater and lower than the original (0.001) negatively affected IoU and Dice scores. To limit variance, we implemented learning rate (LR) scheduling and early termination for the Adam optimizer. This proved successful in reducing the over-fitting observed in the baseline.

4.3 Addressing Imbalanced Classes

The foremost challenge presented by this task is the extreme data imbalance inherent to rare weather events. Prabhat et al. (2021) report 94% of pixels in the *ClimateNet* data belonging to the background class. We find that TCs represent only 0.462% of pixels of the entire dataset (and ARs only 5.674%). This means that a naive model assigning *every pixel* to the background class would reach 94% accuracy despite failing at its task.

To address this class imbalance, we experimented with modifying the loss landscape to better account for under-represented classes and improve performance on rare events such as TC and AR pixels. To that end, we leaned on the literature review by Jadon (2020) to select and implement additional performance metrics and loss functions for training and evaluation.

4.3.1 Performance metrics

To fulfill our problem statement of improved detection of rare weather events in climate data, we explored performance metrics that better represent the model's capacity to learn that task. Specifically, we value detecting extreme events more than identifying their exact boundaries hand-labeled by experts, and aim to penalize missing relevant events more than over-predicting their geographical extent. Specifically, we implemented the following performance metrics:

- **Intersection over union:** our baseline model was trained to optimize for the IoU metric (equation 1), as usual for many computer vision problems.
- **Sørensen–Dice similarity** or Dice coefficient (equation 2) is a measure of the similarity between class predictions and ground truth that is widely used for image comparison.
- **Recall** or **Sensitivity:** we devised our training strategy to optimize for recall (equation 3) as a proxy for the ability to detect most true positives of the TC class.

4.3.2 Weighted loss functions

To optimize for these metrics, we explored and implemented a broad set of loss functions designed to assign higher weights to rare classes, building on a review by Jadon (2020):

- **Jaccard loss:** used by our baseline mode. Computes a derivable prediction of segmentation map IoU from the softmax probabilities output of the classifier (equation 4).
- **Dice loss:** derivable Dice coefficient from the softmax probabilities (equation 5).
- **Cross-entropy loss:** canonically used in multi-class classification problems, it helps balance under-represented classes. We used the pyTorch implementation of the cross entropy loss (equation 6) and weighted cross entropy loss (equation 7).
- **Focal Tversky loss:** a tunable loss function which gives higher weight to false positives, false negatives, and hard examples, by introducing hyper-parameters β and γ (equation 8).
- **Weighted Jaccard loss:** to normalize the relative weights of each class in the IoU estimate, we experimented with a custom loss function inspired by the Jaccard loss (equation 9).

5 RESULTS & DISCUSSION

We report summary results for the baseline and six experiments in table 1, and corresponding precision-recall and ROC curves in figure 3. Table 4 reports detailed performance metrics for our experiments (except data augmentation due to performance drop), and figure 5 compares ground truth labels and baseline results with our predicted segmentation maps on a test set sample.

Table 1: Summary results for baseline model and six experiments.

Models & Metrics		1: Baseline model	2: Learning rate decay	3: Feature engineering	4. Cross entropy	5. Weighted cross entropy	6. Focal Tversky	7. Weighted Jaccard
TC	**IoU**	0.3396	0.3492	0.3161	0.2228	0.2025	0.3160	0.2245
	Precision	0.4560	0.5346	0.4933	0.7134	0.2145	0.3701	0.2384
	Recall	0.5708	0.5016	0.4681	0.2447	0.7836	0.6836	**0.7944**
	Specificity	0.9962	0.9976	0.9973	0.9995	0.9841	0.9936	0.9860
AR	**IoU**	0.3983	0.4128	0.4147	0.3575	0.2932	0.3839	0.3411
	Precision	0.5429	0.5344	0.5425	0.6896	0.3069	0.4479	0.3714
	Recall	0.5993	0.6448	0.6377	0.4261	0.8680	0.7287	0.8068
	Specificity	0.9701	0.9667	0.9681	0.9886	0.8839	0.9468	0.9191

Tropical cyclones recall. While we measured IoU, Dice, precision, recall/sensitivity, and specificity scores for TC and AR events, our key results focus on: (i) recall performance to prioritize detection of positives given the severity of a positive event; and (ii) TCs specifically, the most destructive extreme weather events, for which previous models showed limited performance.

Key results. After comparing our models on the precision-recall and specificity-sensitivity curves, we found that our weighted Cross Entropy and weighted Jaccard loss models with engineered features and a learning rate scheduler achieve better recall than the baseline (0.7836 and 0.7944 compared to 0.5708, a performance gain of +37.3% and +39.2%, respectively). Our experiments with the baseline model with LR scheduler, with baseline loss on engineered data with LR scheduler, and with cross entropy loss on engineered data with LR scheduler performed worse or no better than the baseline (0.2447, 0.4681, and 0.5016, respectively).

Carbon footprint. Given the climate focus of this model and our goal of keeping it light-weight, we tracked and evaluated our carbon footprint during our experiments. Based on emissions factors from Lacoste et al. (2019), and approximately 40 hours of usage of an NVIDIA A100 GPU VM with 40GB of RAM, we estimate our model training emissions at around 6.24 kg CO_{2e}.

6 CONCLUSION

In conclusion, semantic segmentation of extreme weather events in climate data is a challenging problem. The small and imbalanced dataset makes improving on task performance difficult, and CGNet is an intentionally light-weight model with limited capacity. IoU alone is a poor performance metric for identification of rare extreme weather events and should be paired with recall to reflect the priority given to true positive predictions on under-represented classes.

We found success with weighted loss functions, and showed a significant (+39.2%) improvement in recall for our class of interest. We demonstrated that careful matching of loss functions and optimization algorithms with the task at hand can yield important performance gains, even for light-weight architectures with a much lower resource footprint than current trends in machine learning.

Because advances in light-weight segmentation are so new (the seminal CGNet paper was published in 2021), we have found no other applications of these novel architectures to climate data so far beyond the reported baseline. We hope our results will contribute to improving automated extreme weather events detection, which is of crucial importance to prediction, mitigation, and equitable adaptation to the increasing destructiveness of anthropogenic climate change.

ACKNOWLEDGMENTS

We would like to thank Lukas Kapp-Schwoerer, Andre Graubner, and their co-authors in Kapp-Schwoerer et al. (2020b) for their implementation of CGNet on *ClimateNet* data, the authors of Wu et al. (2021) for the original light-weight Context Guided network architecture, and the authors of Prabhat et al. (2021) and the climate sciences expert-labeling community for creating and annotating the *ClimateNet* dataset, which made this study possible. We are also grateful to Andrew Ng, Kian Katanforoosh, and Sarthak Consul at Stanford University for their guidance and support.

DATA AVAILABILITY

The original *ClimateNet* dataset is available at `https://portal.nersc.gov/project/ClimateNet/`. The dataset with engineered features is available at `https://huggingface.co/datasets/rlacombe/ClimateNet/`.

We provide an online repository at `https://github.com/hannahg141/ClimateNet` with: (i) our modified implementation of the CGNet model building on Kapp-Schwoerer et al. (2020a); (ii) notebooks for download, exploration, and visualization of the *ClimateNet* data set, generation of engineered features, and flexible model training on a Google Colab instance; and (iii) a baseline and six experimental models along with their training and evaluation metrics history.

FUTURE WORK

A critical issue with model training on *ClimateNet* is the small and imbalanced nature of the dataset. Also, as is apparent in figure 2, individual labels appear to have some degree of subjectivity, and we suspect human-expert consensus labeling leads to unavoidable bias and high Bayes error. Training on historical observational data, expanding expert-labeling efforts, or learning event identification with more objective ground truth labels (e.g. building on previous work on TC centers identification (Nguyen et al., 2014)) has the potential to improve performance on this task.

A promising direction for that purpose is the *International Best Track Archive for Climate Stewardship* (IBTrACS) dataset, a historical database of TC positions, wind speeds, and geographical extents maintained by NOAA (Knapp et al., 2018). In conjunction with weather re-analysis data services such as ERA5 (Copernicus Climate Change Service, 2017), this set of labels could enable training on a large corpus of observational data. Crucially, the *IBTrACS* data set is global and covers oceanic basins where tropical cyclones with the most destructive impact on LMICs are forming.

This avenue for future work could generalize our models from simulations to observational data, a key step towards early warning and detection systems for equitable adaptation to climate change.

REFERENCES

Copernicus Climate Change Service. ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS). 2017. URL `https://cds.climate.copernicus.eu/cdsapp#!/home`.

IPCC. Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. 2022. doi: 10.1017/9781009325844. URL `https://www.ipcc.ch/report/ar6/wg2/`.

Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, october 2020. doi: 10.1109/cibcb48159.2020.9277638. URL `https://doi.org/10.1109%2Fcibcb48159.2020.9277638`.

Lukas Kapp-Schwoerer, Andre Graubner, Sol Kim, and Karthik Kashinath. ClimateNet, a Python library for deep learning-based climate science. 2020a. URL `https://github.com/andregraubner/ClimateNet`.

Lukas Kapp-Schwoerer, Andre Graubner, Sol Kim, and Karthik Kashinath. Spatio-temporal segmentation and tracking of weather patterns with light-weight neural networks. *AI for Earth Sci-*

ences Workshop at NeurIPS 2020., 2020b. URL https://ai4earthscience.github.io/neurips-2020-workshop/papers/ai4earth_neurips_2020_55.pdf.

K. R. Knapp, H. J. Diamond, J. P. Kossin, M. C. Kruk, and C. J. Schreck. International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4, NOAA National Centers for Environmental Information, 2018. URL https://www.ncei.noaa.gov/products/international-best-track-archive.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning, 2019. URL https://arxiv.org/abs/1910.09700.

Mayur Mudigonda, Prabhat Ram, Karthik Kashinath, Evan Racah, Ankur Mahesh, Yunjie Liu, Christopher Beckham, Jim Biard, Thorsten Kurth, Sookyung Kim, Samira Kahou, Tegan Maharaj, Burlen Loring, Christopher Pal, Travis O'Brien, Kenneth E. Kunkel, Michael F. Wehner, and William D. Collins. *Deep Learning for the Earth Sciences*. John Wiley & Sons, Ltd, 2021. doi: https://doi.org/10.1002/9781119646181. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119646181.

Leon T Nguyen, John Molinari, and Diana Thomas. Evaluation of tropical cyclone center identification methods in numerical models. *Monthly Weather Review*, 142(11):4326–4339, 2014.

Thanh Cong Nguyen, Jackie Robinson, Shinji Kaneko, and Satoru Komatsu. Estimating the value of economic benefits associated with adaptation to climate change in a developing country: A case study of improvements in tropical cyclone warning services. *Ecological Economics*, 86:117–128, 2013. ISSN 0921-8009. doi: https://doi.org/10.1016/j.ecolecon.2012.11.009. URL https://www.sciencedirect.com/science/article/pii/S0921800912004508. Sustainable Urbanisation: A resilient future.

NOAA. National oceanic and atmospheric administration. fast facts: Hurricane costs. 2022. URL https://coast.noaa.gov/states/fast-facts/hurricane-costs.html.

Robbie M. Parks and Renzo R. Guinto. Invited perspective: Uncovering the hidden burden of tropical cyclones on public health locally and worldwide. *Environmental Health Perspectives*, 130(11):111306, 2022. doi: 10.1289/EHP12241. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/EHP12241.

S. Philip, S. Kew, G. J. van Oldenborgh, F. Otto, R. Vautard, K. van der Wiel, A. King, F. Lott, J. Arrighi, R. Singh, and M. van Aalst. A protocol for probabilistic extreme event attribution analyses. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(2):177–203, 2020. doi: 10.5194/ascmo-6-177-2020. URL https://ascmo.copernicus.org/articles/6/177/2020/.

Prabhat, K. Kashinath, M. Mudigonda, S. Kim, L. Kapp-Schwoerer, A. Graubner, E. Karaismailoglu, L. von Kleist, T. Kurth, A. Greiner, A. Mahesh, K. Yang, C. Lewis, J. Chen, A. Lou, S. Chandran, B. Toms, W. Chapman, K. Dagon, C. A. Shields, T. O'Brien, M. Wehner, and W. Collins. 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, 2021. doi: 10.5194/gmd-14-107-2021. URL https://gmd.copernicus.org/articles/14/107/2021/.

Isla Simpson. Circulation and Vorticity (class lecture, Advanced Atmospheric Dynamics, University of Toronto), 2010. URL https://www2.cgd.ucar.edu/staff/islas/teaching/3_Circulation_Vorticity_PV.pdf.

World Economic Forum. *Global Risks Report*. 2022. URL https://www.weforum.org/reports/global-risks-report-2022/.

Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. CGNet: A light-weight context guided network for semantic segmentation. *IEEE Transactions on Image Processing*, 30:1169–1179, 2021. doi: 10.1109/TIP.2020.3042065. URL https://ieeexplore.ieee.org/document/9292449.

APPENDIX

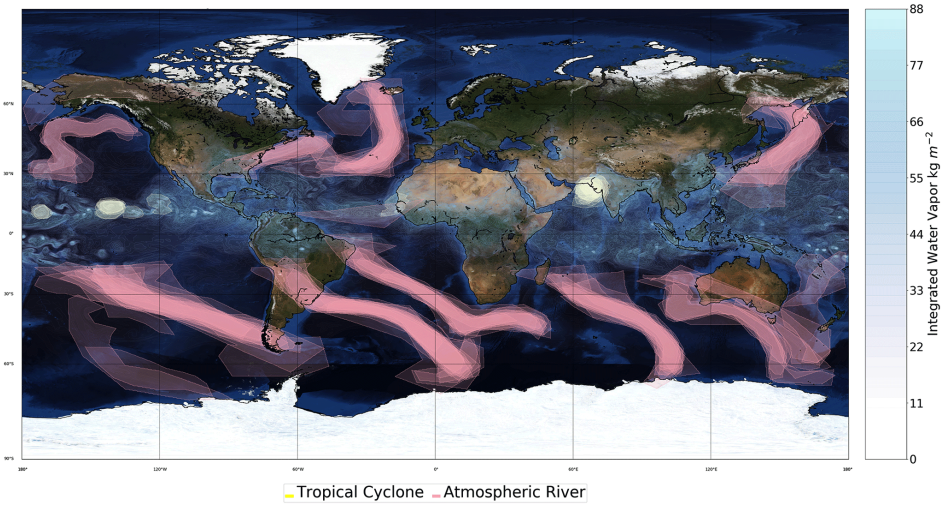

Figure 2: Example image from Prabhat et al. (2021) showing 15 different expert labelings (TC labels in white/yellow masks seen near the equator; AR labels in pink masks). The background "blue marble" map included via Matplotlib's Basemap library is (c) NASA.

Table 2: *ClimateNet* dataset channels (Prabhat et al., 2021).

CHANNEL	DESCRIPTION	UNITS
TMQ	Total (vertically integrated) precipitable water	kg/m^2
U850	Zonal wind at 850 mbar pressure surface	m/s
V850	Meridional wind at 850 mbar pressure surface	m/s
UBOT	Lowest level zonal wind	m/s
VBOT	Lowest model level meridional wind	m/s
QREFHT	Reference height humidity	kg/kg
PS	Surface pressure	Pa
PSL	Sea level pressure	Pa
T200	Temperature at 200 mbar pressure surface	K
T500	Temperature at 500 mbar pressure surface	K
PRECT	Total (convective and large-scale) precipitation rate	m/s
TS	Surface temperature (radiative)	K
TREFHT	Reference height temperature	K
Z1000	Geopotential Z at 1000 mbar pressure surface	m
Z200	Geopotential Z at 200 mbar pressure surface	m
ZBOT	Lowest model level height	m
LABELS	0: Background, 1: Tropical Cyclone, 2: Atmospheric River	-
Engineered features		
WS850	Wind speed (equation 10) at 850 mbar pressure surface	m/s
VRT850	Relative wind vorticity (eq. 11) at 850 mbar pressure surface	m/s
WSBOT	Lowest level wind speed	m/s
VRTBOT	Lowest level relative wind vorticity	m/s

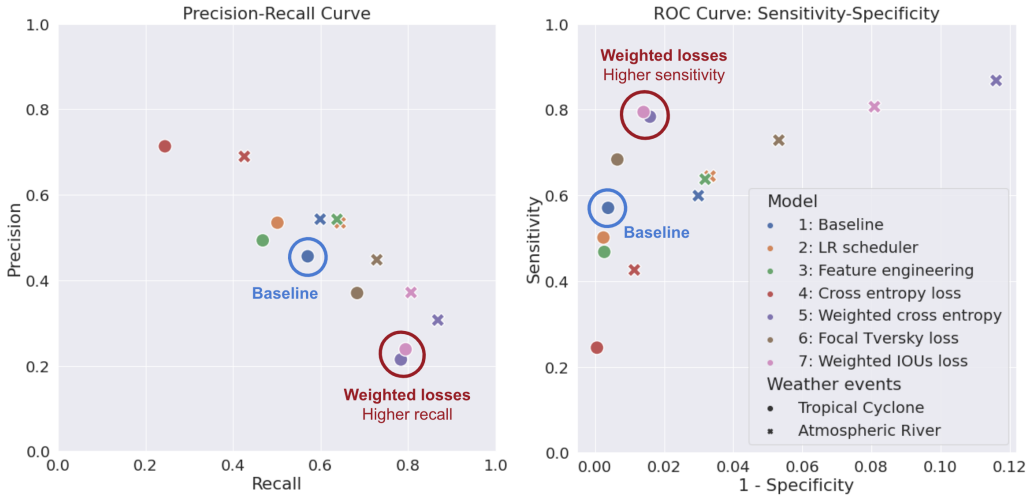

Figure 3: (i) Precision-Recall Curve (left); (ii) ROC Curve (Sensitivity vs 1-Specificity, right). (●): tropical cyclone. (×): atmospheric river.

Table 3: Description of experiments: baseline and six select models we trained.

EXPERIMENT	LOSS FUNCTION	FEATURES	LR DECAY
1. Baseline	Jaccard loss	Baseline dataset	No
2. Learning rate decay	Jaccard loss	Baseline dataset	Yes
3. Feature engineering	Jaccard loss	Engineered	Yes
4. Cross entropy	Cross-entropy loss	Engineered	Yes
5. Weighted cross entropy	Weighted cross-entropy loss	Engineered	Yes
6. Focal Tversky	Focal Tversky loss	Engineered	Yes
7. Weighted Jaccard	Weighted Jaccard loss	Engineered	Yes

Table 4: Detailed results for baseline model and six experiments.

Models & Metrics		1: Baseline model	2: Learning rate decay	3: Feature engineering	4. Cross entropy	5. Weighted cross entropy	6. Focal Tversky	7. Weighted Jaccard
TC	**IoU**	0.33955127	0.34916790	0.31608774	0.22278776	0.20251324	0.31599551	0.22453519
	Dice	0.50696270	0.51760482	0.48034448	0.36439318	0.33681665	0.48023798	0.36672721
	Precision	0.45598923	0.53463995	0.49327914	0.71342764	0.21450748	0.37011321	0.23838463
	Recall	0.57076677	0.50162175	0.46807083	0.24468465	0.78363352	0.68365510	0.79444291
	Specificity	0.99623709	0.99758723	0.99734295	0.99945687	0.98414286	0.99357050	0.98597405
AR	**IoU**	0.39832633	0.41285328	0.41468760	0.35750965	0.29317686	0.38387118	0.34108058
	Dice	0.56971870	0.58442485	0.58626032	0.52671397	0.45342113	0.55477878	0.50866530
	Precision	0.54289729	0.53435760	0.54252858	0.68965182	0.30685734	0.44789329	0.37141691
	Recall	0.59932803	0.64484427	0.63766037	0.42605401	0.86800497	0.72866865	0.80679887
	Specificity	0.97010985	0.96671544	0.96815083	0.98864332	0.88386163	0.94679579	0.91912138

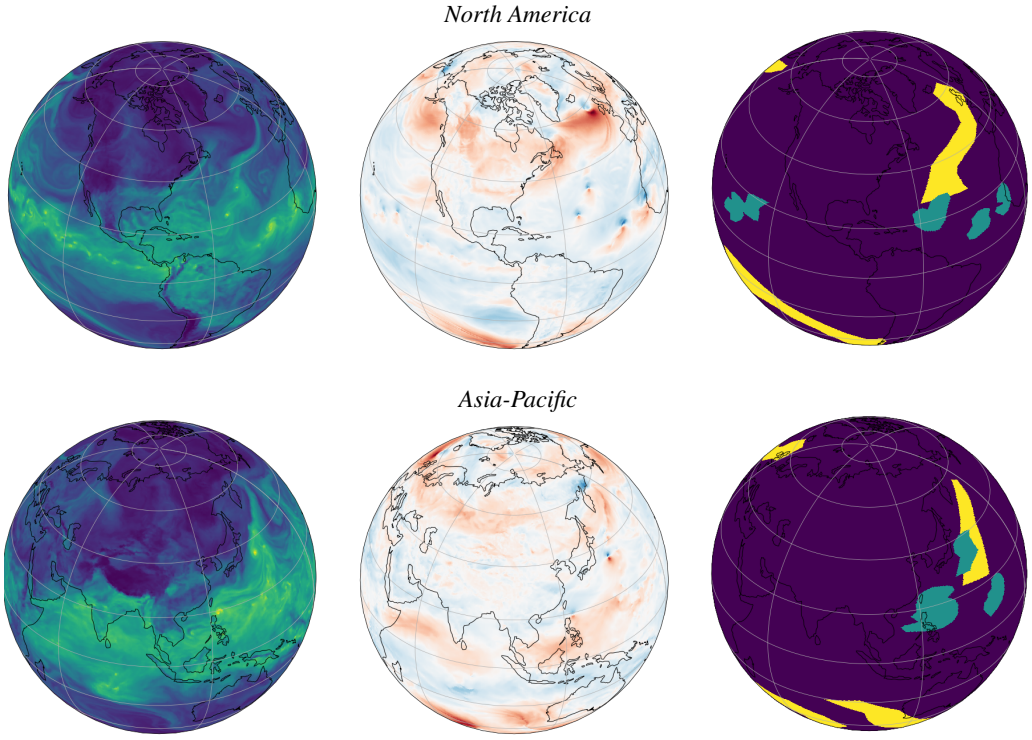

Figure 4: **Sample *ClimateNet* channels and associated ground truth labels** (TMQ & U850). AR: yellow; TC: green; BG: purple. Viewed from 35°N 80°W (above) and 35°N 100°E (below).

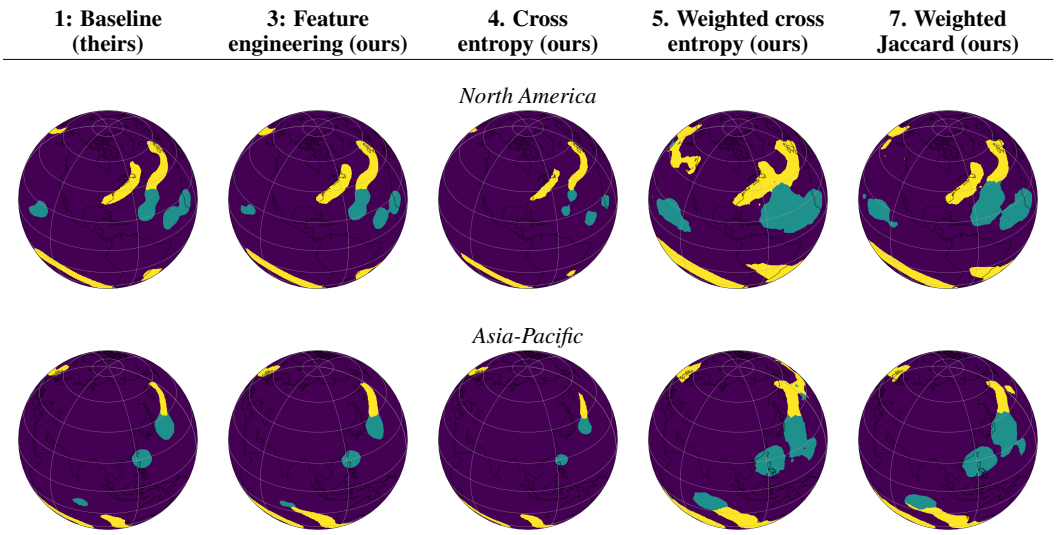

Figure 5: **Predicted segmentation maps** produced by the baseline and four of the models we trained (test set sample). Outputs aim to predict ground truth labels in figure 4.

EQUATIONS

METRICS

Performance metrics for a single sample.
Formalism: TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives.

$$\text{INTERSECTION OVER UNION} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{SØRENSEN–DICE SIMILARITY} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{RECALL/SENSITIVITY} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

LOSS FUNCTIONS

Loss functions for a single sample. Formalism: $y = y_{ij}$ is a one-hot encoded ground truth tensor for the three classes at longitude and latitude (i, j), and $\hat{y} = \hat{y}_{ij}$ is the 3-classes probabilities tensor computed as the softmax of the logits predicted by the network. Parameters are w_C, the tensor of weights used to balance under-represented classes, and β and γ, scalars which allow for the tuning of relative weights of false positives and false negatives and of hard examples in the focal Tversky loss. All operations here are element-wise.

$$\text{JACCARD LOSS}(y, \hat{y}) = 1 - \frac{\hat{y}y}{(\hat{y} + y) - \hat{y}y} \tag{4}$$

$$\text{DICE LOSS}(y, \hat{y}) = 1 - \frac{2y\hat{y} + 1}{y + \hat{y} + 1} \tag{5}$$

$$\text{CROSS ENTROPY LOSS}(y, \hat{y}) = -y \log(\hat{y}) \tag{6}$$

$$\text{WEIGHTED CROSS ENTROPY LOSS}(y, \hat{y}) = -w_C y \log(\hat{y}) \tag{7}$$

$$\text{FOCAL TVERSKY LOSS}(y, \hat{y}) = \left(1 - \frac{y\hat{y}}{\beta(1-y)\hat{y} + (1-\beta)y(1-\hat{y})}\right)^{\gamma} \tag{8}$$

$$\text{WEIGHTED JACCARD LOSS}(y, \hat{y}) = 1 - w_C \frac{\hat{y}y}{(\hat{y} + y) - \hat{y}y} \tag{9}$$

WIND VELOCITY

Wind speed is the L_2 norm of the zonal and meridional components of the wind vector field:

$$w_s = \sqrt{u^2 + v^2} \tag{10}$$

RELATIVE WIND VORTICITY

Wind vorticity is the rotation of the wind vector field, where λ = longitude and ϕ = latitude:

$$\zeta = \frac{\partial u}{\partial \lambda} - \frac{1}{\cos\phi} \frac{\partial v \cos\phi}{\partial \phi} \tag{11}$$