CLIMATENET: BRINGING THE POWER OF DEEP LEARNING TO WEATHER AND CLIMATE SCIENCES VIA OPEN DATASETS AND ARCHITECTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Pattern recognition tasks such as classification, object detection and segmentation have remained challenging problems in the weather and climate sciences. While there exist many empirical heuristics for detecting weather patterns and extreme events, the disparities between the output of these different methods even for a single event are large and often difficult to reconcile. Reliable, labeled training data is scarce in climate science. 'ClimateNet' is an effort to solve this problem by creating open, community-sourced, expert-labeled datasets that capture segmentation masks for different weather patterns in high-res climate datasets.

Here we present an update on the ClimateNet project and results from deep learning models trained on ClimateNet for segmenting Atmospheric Rivers (ARs) and Tropical Cyclones (TCs). We discuss the outcomes of various labeling campaigns with several groups of climate scientists, including challenges faced, lessons learned and opportunities for the future. Upon training with expert-labeled data obtained through the ClimateNet project, we find the segmentation masks predicted by the DL model to be qualitatively similar to the labels of expert climate scientists, and far superior to those obtained from training on empirical heuristics. The mean IOU from the segmentation model (0.52) performing similarly to human experts (0.51).

1 INTRODUCTION

Climate change is arguably one of the most pressing challenges facing humanity in the 21st century. cc For instance, the state of California receives over 50% of its rainfall through ARs, and water resource managers are interested in understanding if and how AR intensities and tracks will shift in the future; potentially resulting in devastating floods or droughts or both. In the state of Florida, homeowners are interested in understanding if hurricanes will become more intense in the future, and/or make landfall more often. This has direct impact on society and the environment, as well as the economy via home prices and the insurance industry. Hurricanes have caused the US economy over $200B worth of damage in 2017; and a range of stakeholders are interested in a more careful characterization of the change in number, frequency and intensity of such devastating weather phenomena in the coming decades.

In order to address these important questions, climate scientists routinely configure and run high-resolution, high-fidelity simulations under a variety of climate change scenarios. Each simulation produces tens of TBs of output; which requires fast, precise and reliable automated analysis. Thus far, climate scientists have relied upon multi-variate threshold conditions for prescribing extreme weather patterns Prabhat et al. (2015). However, even within a single class of extreme weather events, there often is no consensus on the precise and accurate definition of that event. For example, the ARTMIP project has shown that across the dozen different AR detection methods and algorithms, results can differ by an order of magnitude Shields et al. (2018). Such methodological disparity is unacceptable for societal and environmental planning.

Since the beginning of this decade, DL has been applied successfully to solve challenging pattern recognition problems in computer vision, speech recognition, robotics and control systems LeCun et al. (2015); Levine et al. (2016). A key requirement for the success of supervised DL is the availability of plentiful high-quality labelled data. Further, the computer vision community has

established that DL is effective at learning relevant features for solving pattern recognition tasks without requiring application-specific tuning. Recent work has demonstrated that DL can be used for solving pattern classification, localization and segmentation problems for climate datasets Liu et al. (2016); Hong et al. (2017); Racah et al. (2017); Kurth et al. (2018). Kurth et al. (2018) demonstrated that model predictions of ARs and TCs segmentation masks sometimes exceeded the quality and realism of heuristic-based training data, a rather promising outcome for a first attempt at DL-based segmentation of climate data. The success of the above applications, however, was limited by the lack of plentiful high-quality reliable labeled data.

Given (i) the shortcomings of existing heuristics of detecting weather and climate patterns; (ii) the power of DL in recognizing complex patterns *without* requiring engineered features; and (iii) the scarcity of reliable labeled data; we have developed ClimateNet – a community-sourced labeling strategy to prepare a vast and reliable database of weather and climate pattern labels to push the frontier of DL methods for variety of important and urgent pattern recognition tasks in the weather and climate sciences.

2 CLIMATENET

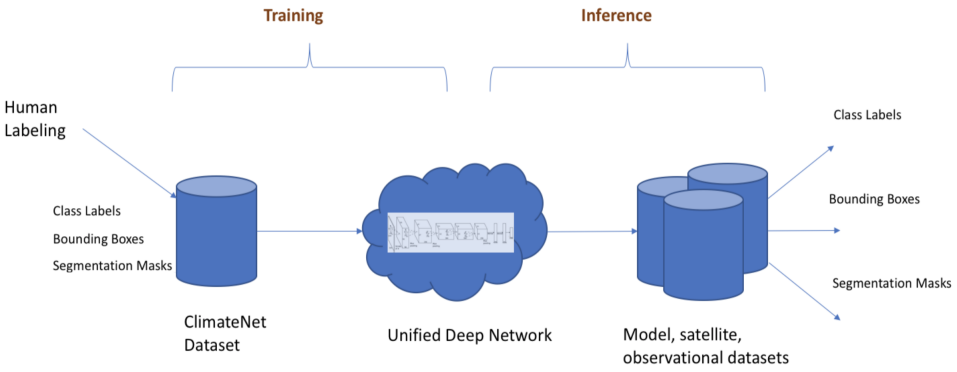


Figure 1: ClimateNet schematic: Training and Inference phases

Figure 1 presents the workflow for augmenting existing weather and climate datasets with ground truth label information. The unified DL workflow, as relevant to weather and climate sciences, can be split into two pieces: the Training phase and the Inference phase. First, the ClimateContours tool is used by human experts to produce a 'hand'-labeled database of weather patterns. Second, a unified DL model is trained to learn a common architecture representing various weather patterns. Finally, the trained model is applied to disparate datasets to obtain predictions.

Recent work in the DL community has shown that training on a multitude of tasks for the same underlying datasets can help with generalization of representations Luong et al. (2015). ClimateNet

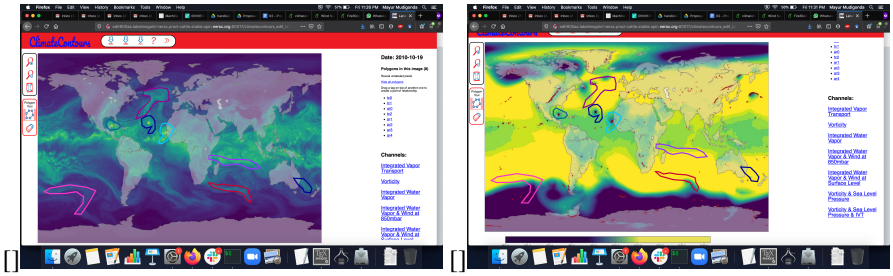


Figure 2: (Left) Showing integrated water vapor (Right) Showing vorticity and sea level pressure. These figures depict two of seven fields available to experts while labeling events. The seven fields (variables) were chosen based on expert advice on the most useful physical fields that characterize ARs and TCs.

attempts to provide a similar framework by addressing a multitude of pattern recognition problems in weather and climate data simultaneously.

The goals of ClimateNet are twofold: (i) to create a vast and reliable database of high-quality expert-labeled datasets across a variety of weather and climate phenomena and dataset types, (ii) create and release trained deep neural network architectures, which can be further adapted and customized by the climate science community.

2.1 DATA INGEST

ClimateNet currently takes as input raw climate simulation data from the high-resolution 25-km CAM5.1 climate model Wehner et al. (2014). This dataset contains tens of physical quantities of interest to weather and climate scientists; for example, wind velocities, temperatures, pressures, and humidities at different vertical levels and across the globe. All of these variables (channels) contain information relevant to the dynamics of weather and climate phenomena, but not all variables are needed to detect a weather event. Based on the experience and wealth of knowledge accumulated by meteorologists and climate scientists, and for relative ease of use, we provide a subset of seven variables to the user to aid them in creating labels for TCs and ARs through an online interactive web-based tool called *ClimateContours*.

2.2 CLIMATECONTOURS

ClimateContours adapts the *LabelMe* Russell et al. (2008) tool developed at MIT for crowdsourcing the 'hand'-labeling task that is all-important for training DL models. Figure 2 shows a screenshot of the online web-based labeling tool. The tool provides multiple variables that an expert can use such as integrated water vapor Figure 2 (left), vorticity and sea level pressure as seen in Figure 2 (right), integrated vapor transport, and a few more.

The task presented to the user is to draw bounding polygons around all TCs and ARs that exist in any given timestep. The user is also given an option to rate their confidence level (low, medium, high) for each event, to obtain information about the reliability of any label. All of this information is stored in a *xml* file that is used to create the annotated dataset through post-processing.

2.3 LABELING CAMPAIGNS

In order to capture the expertise of climate scientists in characterizing ARs and TCs, we conducted multiple labeling campaigns across several institutions and events (LBNL, UC Berkeley, NCAR, Scripps/UCSD, the ARTMIP workshop and Climate Informatics conference) to produce the ClimateNet dataset. The project currently boasts of over 1000 carefully curated data labeled by climate experts using the ClimateContours tool.

Further, from the collected samples two separate "Quality Assurance (QA)/Quality Control (QC)" processes were run on about 600 samples to correct for any inadvertent mistakes such as misclassification or missed events. The details of the QA/QC process will be detailed in a subsequent paper, but it is worth noting that a rigorous principled approach was followed for this process.

2.4 MODEL TRAINING AND INFERENCE

We developed a reference implementation of the segmentation architecture developed in Kurth et al. (2018) in TensorFlow. We anticipate users leveraging the trained network for 'out-of-the-box' segmentation applications on their datasets, as well as transfer learning applications for datasets with other events (weather fronts, extra-tropical cyclones, etc) and modalities (i.e. observational products).

2.4.1 MODEL

The deep neural network architecture used in this work is the DeepLabv3+ architecture (see Figure 3) developed by Chen et al. (2018). This architecture has attained state-of-the-art results across various semantic segmentation benchmarks such as PASCAL VOC 2012 and Cityscapes. DeepLabv3+ consists of an encoder which captures rich semantic information across multiple scales, and a decoder module that upsamples the learned representation and refines the object boundaries. There are a

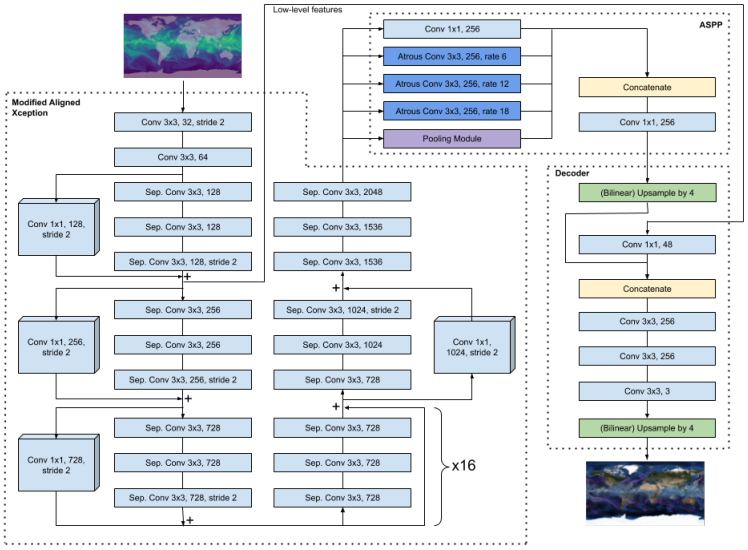


Figure 3: DeepLabv3+: All convolutional layers are followed by a batch normalization and a ReLU activation layer, which are omitted from this schematic for the sake of brevity. "Sep. Conv" denotes depthwise separable convolution. The pooling module consists of a two-dimensional pooling layer, followed by a convolutional layer, a batch normalization layer and a ReLU activation layer. For more details, we refer the reader to Chen et al. (2018).

number of possible choices for the network backbone. Following the results presented in Chen et al. (2018), we employ the Xception model Chollet (2016), as modified by Chen et al. (2018).

2.4.2 TRAINING

We examine the performance of DeepLabv3+ on the task of semantic segmentation. Specifically, we train a model to produce segmentation masks for a grid of size 1152 x 768, where the 3 possible classes for each grid cell are atmospheric rivers (ARs), tropical cyclones (TCs), and background.

We perform the training in two different settings. In the first setting, we train the model on a large set of heuristically annotated data (Prabhat et al., 2015). In the second setting, we use the expert-labelled ClimateNet dataset introduced in this paper, which consists of far fewer samples.

The heuristically annotated dataset consists of 128k samples, where each sample is produced by CAM (cite) and annotated by TECA (Prabhat et al., 2015). For training, we randomly split this dataset into a training set that consists of 51.2k samples (40%), and validation and test sets which are subsets of the remaining 60%. In both cases the data contains 16 channels describing different physical quantities, of which we select four which are relevant to the events we are trying to detect: TMQ (Total vertically integrated precipitable water), U850 (Zonal wind at 850 mbar pressure surface), V850 (Meridional wind at 850 mbar pressure surface), PRECT (total convective and large-scale precipitation rate).

The ClimateNet dataset contains 219 unique samples and 459 labelings, i.e. some of the 219 unique samples were labeled by multiple experts. The median number of different labelings per sample is 2. For training, we split this dataset into a training set which contains 422 (91%) labelings, and validation and test sets which consist of 18 and 19 labelings, respectively. Note that all labelings in the training set have timestamps that are earlier than the timestamp of any labelings in the validation and test sets, thus avoiding performance artefacts from persistence and auto-correlation of weather phenomena.

Since both datasets are heavily unbalanced, (e.g., 93.864% of the labels in ClimateNet are background) we train by optimizing the weighted cross-entropy loss using the Adam optimizer (Kingma & Ba, 2014), where the class weight is a function of the inverse class frequency. In particular, the weight of a given class is the inverse of its class frequency, squared.

Finally, we use a learning rate scheduler that multiplicatively reduces the learning rate each time the performance on the validation set does not improve for 3 epochs in a row.

2.4.3 RESULTS

We measure the performance of our model using the mean Intersection-over-Union (mIoU) metric. While the mIoU is a useful metric for semantic segmentation tasks in general, the high level of variance in different experts' opinions on the same sample requires us to also judge the plausibility of the resulting masks directly.

For the heuristic setting, we achieve an IOU of 99.6 on background and 76.6 on the AR's. However, due to an IOU of only 44.4 on TC's the mean IOU is only 73.5. These results are comparable to those reported in Kurth et al. (2018).

Heuristic Setting			
mean IoU	Background IoU	TC IoU	AR IoU
0.7354	0.9958	0.4438	0.7667

Table 1: In the heuristic setting our model achieves IoU scores similar to those reported in Kurth et al. (2018). Here the IoU score is bottlenecked by the TC IoU.

In the hand-labelled setting, we observe significantly worse IOU's. However, these must be taken with a grain of salt: since the same image is often labelled quite differently by different experts, the IOU provides an excessively pessimistic estimate of performance. It is therefore useful and necessary to also gauge the quality of the predictions by manually comparing them to the ground truths. We see that the IOU that our model achieves is similar to that of a human expert, and we again observe that the model performs far better on AR's than on TC's in terms of IOU.

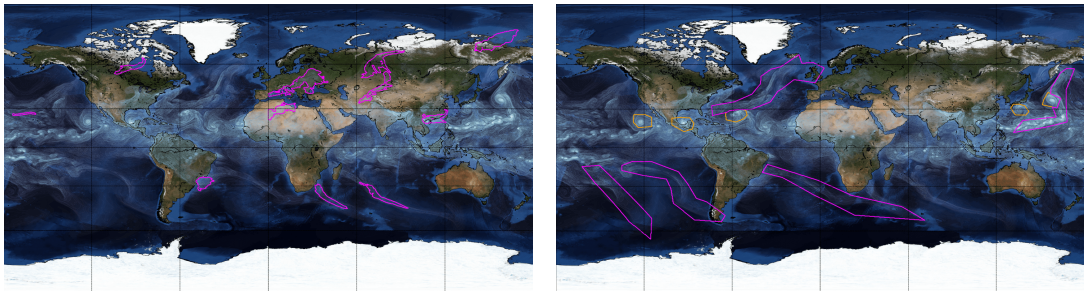
ClimateNet				
	mean IoU	Background IoU	TC IoU	AR IoU
DeepLabv3+	0.5247	0.9389	0.2441	0.3910
Mean Expert Performance	0.5120	0.9382	0.2567	0.3412

Table 2: After training on ClimateNet, our model performs similar to human experts.

Figure 4 captures the significance of the ClimateNet project and the premise for developing a carefully curated expert-labelled dataset for training DL models for use in the weather and climate sciences. It is evident from this figure that the quality, accuracy and precision of the segmentation masks predicted by the DL model trained on expert-labeled data is far superior to the predictions of a model trained on heuristics. This figure also shows conclusively that the current state-of-the-art DL models are able to emulate the quality of the training data, strengthening the case for producing high-quality curated expert-labeled datasets for a variety of other weather and climate phenomena at different spatial and temporal scales.

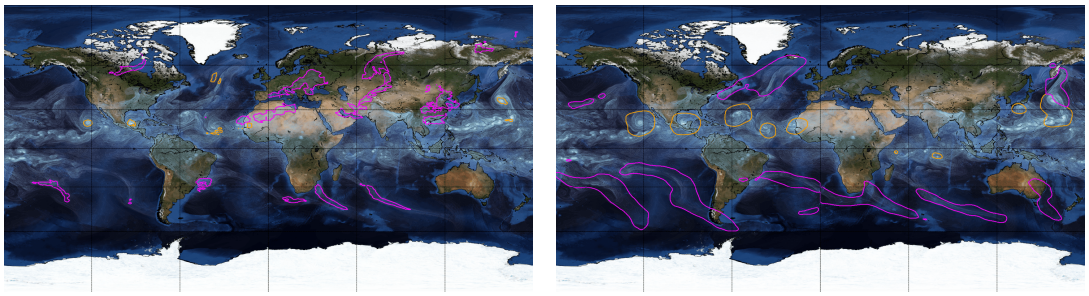
3 CONCLUSION

The goal of the ClimateNet project is to contribute an end-to-end learning system, including a curated and annotated database, and a DL model that can segment multiple event classes. In this paper we have shown initial successes with ARs and TCs on high-res global climate model output. We are developing capabilities for enabling users to load their own datasets into the ClimateContours tool and label a much wider range of phenomena (regional or global), which will help expand and diversify the ClimateNet database. We have begun combining forces with EnviroNet, a similar project for a much broader set of applications and problems in the environmental sciences. We believe that easy access to curated datasets and a trained DL architecture will be critical in advancing DL applications in weather and climate science; and in lowering the barrier of entry for weather and climate scientists who are interested in incorporating DL into their existing workflows.



(a) Heuristic Labels

(b) Expert Labels



(c) Predictions trained on Heuristic Labels

(d) Predictions trained on Expert Labels

Figure 4: Comparison of heuristic labels, expert labels, and predictions

REFERENCES

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv e-prints*, art. arXiv:1802.02611, Feb 2018.

Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv e-prints*, art. arXiv:1610.02357, Oct 2016.

Hong, S., Kim, S., Joh, M., and Song, S.-k. Globenet: Convolutional neural networks for typhoon eye tracking from remote sensing imagery. *arXiv preprint arXiv:1708.03417*, 2017.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv:1412.6980, Dec 2014.

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., Mahesh, A., Matheson, M., Deslippe, J., Fatica, M., et al. Exascale deep learning for climate analytics. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, pp. 51. IEEE Press, 2018.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Liu, Y., Racah, E., Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M., Collins, W., et al. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv:1605.01156*, 2016.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*, 2015.

Prabhat, M., Byna, S., Vishwanath, V., Dart, E., Wehner, M., and Collins, W. Teca: Petascale pattern recognition for climate science. 9257, 09 2015.

Racah, E., Beckham, C., Maharaj, T., Ebrahimi Kahou, S., Prabhat, M., and Pal, C. Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3402–3413. 2017.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

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., Ullrich, P., Gershunov, A., Goldenson, N., Guan, B., Qian, Y., Ramos, A. M., Sarangi, C., Sellars, S., Gorodetskaya, I., Kashinath, K., Kurlin, V., Mahoney, K., Muszynski, G., Pierce, R., Subramanian, A. C., Tome, R., Waliser, D., Walton, D., Wick, G., Wilson, A., Lavers, D., Prabhat, Collow, A., Krishnan, H., Magnusdottir, G., and Nguyen, P. Atmospheric river tracking method intercomparison project (ARTMIP): project goals and experimental design. *Geoscientific Model Development*, 11(6):2455–2474, 2018. doi: 10.5194/gmd-11-2455-2018. URL https://www.geosci-model-dev.net/11/2455/2018/.

Wehner, M. F., Reed, K., Li, F., Prabhat, Bacmeister, J., Chen, C.-T., Paciorek, C., Gleckler, P., Sperber, K., Collins, W. D., Gettelman, A., and Jablonowski, C. The effect of horizontal resolution on simulation quality in the community atmospheric model, cam5.1. *Journal of Modeling the Earth System*, 06:980–997, 2014. doi: 10.1002/2013MS000276.