Machine Learning for Targeted Assimilation of Satellite Data

Yu-Ju Lee[1,2]([✉]) [iD], David Hall[1], Jebb Stewart[3], and Mark Govett[4]

[1] Department of Computer Science, University of Colorado Boulder,
Boulder, CO 80309-0430, USA
`Yuju.Lee@Colorado.EDU`
[2] Cooperative Institute for Research in Environmental Sciences,
NOAA/OAR/ESRL/Global Systems Division, University of Colorado Boulder,
Boulder, CO, USA
[3] Cooperative Institute for Research in the Atmosphere,
NOAA/OAR/ESRL/Global Systems Division, Colorado State University,
Boulder, CO, USA
[4] NOAA/OAR/ESRL/Global Systems Division,
325 Broadway, Boulder, CO, USA

Abstract. Optimizing the utilization of huge data sets is a challenging problem for weather prediction. To a significant degree, prediction accuracy is determined by the data used in model initialization, assimilated from a variety of observational platforms. At present, the volume of weather data collected in a given day greatly exceeds the ability of assimilation systems to make use of it. Typically, data is ingested uniformly at the highest fixed resolution that enables the numerical weather prediction (NWP) model to deliver its prediction in a timely fashion. In order to make more efficient use of newly available high-resolution data sources, we seek to identify regions of interest (ROI) where increased data quality or volume is likely to significantly enhance weather prediction accuracy. In particular, we wish to improve the utilization of data from the recently launched Geostationary Operation Environmental Satellite (GOES)-16, which provides orders of magnitude more data than its predecessors. To achieve this, we demonstrate a method for locating tropical cyclones using only observations of precipitable water, which is evaluated using the Global Forecast System (GFS) weather prediction model. Most state of the art hurricane detection techniques rely on multiple feature sets, including wind speed, wind direction, temperature, and IR emissions, potentially from multiple data sources. In contrast, we demonstrate that this model is able to achieve comparable performance on historical tropical cyclone data sets, using only observations of precipitable water.

Keywords: Numeric weather prediction · Satellite
Machine learning · Data assimilation · Tropical cyclone
Precipitable water · Water vapor · Global Forecast System (GFS)

1   Introduction

Extreme weather has the potential to cause significant economic damage and loss of life. In order to minimize these losses, high precision weather predictions are needed. It is also important to make such predictions as far in advance as possible, to provide adequate advanced warning of impending storms. Improving the accuracy and time horizon of weather predictions are among the primary research objectives pursued by the National Oceanic and Atmospheric Administration (NOAA). To a significant degree, the accuracy of a numerical weather prediction is determined by the model initialization procedure, wherein data is assimilated from a variety of observational platforms. Over wide swaths of ocean or in remote areas of land, where in-situ observations are lacking, satellite data is used to augment and complete the construction of the initial conditions. As assimilation of satellite data is computationally expensive, the data resolution is typically reduced in order to accelerate its incorporation into the forecast. In the vicinity of an extreme weather event, such as a tropical cyclone, the situation can change rapidly. It is important to update the initial conditions more frequently using higher resolution data, in order to produce the most accurate forecasts. To this end, we are interested in automatically identifying specific regions of interests (ROI) where supplemental satellite observations could help increase the forecast's quality and overall impact.

At present, detection of extreme weather is primarily a manual process relying on difficult-to-quantify human expertise and experience, with no clear-cut definition for most weather phenomena. For example, it is well known that a tropical cyclone is characterized by a region of low surface-pressure surrounded by high speed winds and enhanced water vapor. However, there is no universally agreed upon combination of wind speed, pressure, and vapor that definitively identifies a tropical cyclone. If we attempt to hand-craft a definition identifying all tropical cyclones, we would have to deal with many edge cases that don't meet that definition. In addition to the challenge of constructing adequate definitions, there are also limits on the quality and quantity of observational data available. For example, the data needed for forecasting tropical cyclones is often provided by satellites in polar orbits. Those observations may be poorly timed, leading to images where the target cyclone is located in the periphery of the observed region, or absent from the image entirely (Fig. 1).

In this article, we propose a tropical cyclone detection algorithm that requires only observations of water vapor which is the primary data source to be provided by the new geostationary GOES-16 satellite. As it doesn't require measurements of wind speed or direction, we can avoid intermittent data from non-geostationary satellites. By using only high-frequency geostationary satellite data, we ensure continuous tracking. The proposed algorithm also employs a sliding window data augmentation strategy to overcome data sparsity, as discussed in Sect. 5.
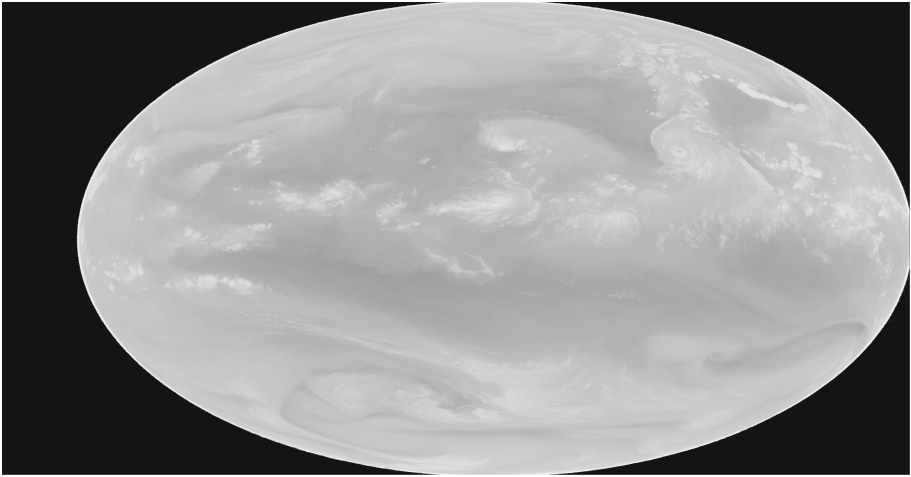
(a) Water Vapor image of GOES-15
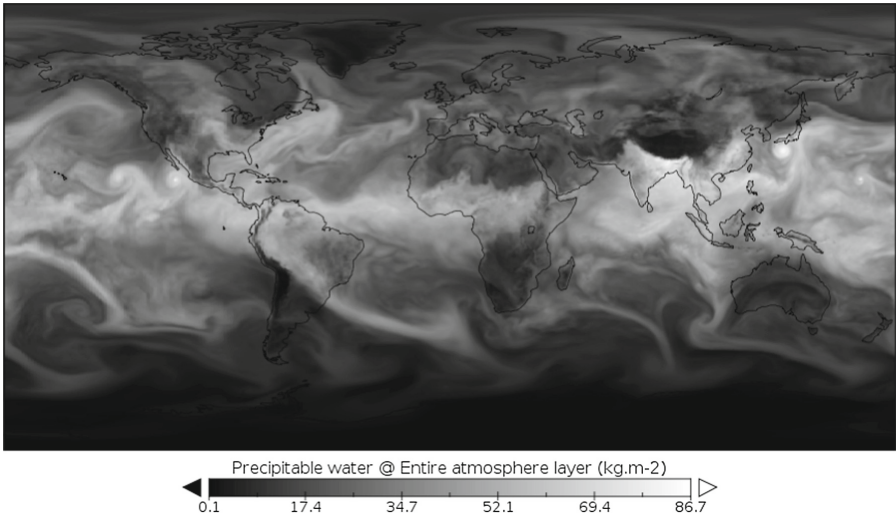
(b) Precipitable water of GFS data

Fig. 1. Precipitable water is a good approximate of satellite water vapor channel for preliminary study

2   Related Work

Many researchers have investigated extreme weather detection using both remote sensing data and Numerical Weather Prediction (NWP) models. A technique introduced by Dvorak in the 1970s is a widely accepted approach for classifying the intensity of tropical cyclones [3]. This technique uses visual identification of images of tropical cyclones in the visible and infrared bands for classification.

However, cyclone images vary a great deal, and expert opinion is required in order to properly apply this method. Although Dvorak's technique has been modified, with an increased reliance on NWP models, meteorologists still rely primarily on judgment and experience to identify and locate tropical cyclones using meteorological data. Since that time, significant research has been conducted in an effort to develop improved tropical cyclone indicators [2,9,10,12,14]. Research into the estimation of tropical cyclone intensity using satellite data has also been an active area of investigation. For example, Velden et al. [13] made use of sequences of GOES-8 infrared images to infer tropical cyclone wind speed and track. Jaiswal et al. [6] suggested matching infrared images with a database of known intensities as a means to estimate tropical cyclone strength. Disadvantages of these techniques include the need for manual adjustment of the matching-index threshold, and the requirement that the cyclone to be measured is well centered in the image.

More recently, there has been an increased effort to apply machine learning techniques to automate the identification of severe weather phenomena. Liu et al. used data generated by the CAM-5 atmosphere model to automate extreme weather detection in climate simulations [8]. However, this technique also required the target object be well centered in the image, which is not well suited for use with GOES satellite images. Ho et al. identified tropical cyclones from QuickScat satellite data using support vector machines (SVMs) applied to wind speed and direction [5]. They also built a system to combine the data from QuickScat with TRMM precipitation measurements using a Kalman filter for cyclone tracking [4]. The technique of Panangadan et al. uses multiple satellite images with a graph-based algorithm to detect the eye of the cyclone and a Kalman filter or particle filter for cyclone tracking [11]. Zou et al. employed wind circulation as an additional feature for cyclone detection using QuickSCAT data [15]. Their technique used a wind speed and direction histogram to perform coarse identification, and then use the wind circulation path to refine the classification. However, common drawbacks in these techniques include their reliance on multiple data sources and their focus on wind-speed and direction for cyclone identification. In contrast, the technique described in this work achieves high accuracy identification of tropical cyclones using only water vapor images from a single geostationary satellite source.

3   Data

In this section, we describe the Global Forecast System (GFS) data and International Best Track Archive for Climate Stewardship (IBTrACS) data. From this data, we extract information covering tropical cyclones in the west pacific basin. The IBTrACS and GFS datasets are combined to form our labeled dataset used in training, validation and prediction.

3.1 Global Forecast System Data

The Global Forecast System (GFS) is a weather forecast model produced by the National Centers for Environmental Prediction (NCEP). The main goal of GFS is the production of operational forecasts for weather prediction over both short-range and long-range forecasts. GFS proves a large set of variables which have the potential to impact global weather. Examples variables include temperature, winds, precipitation, atmospheric ozone, and soil moisture. GFS data is provided in a gridded format on the regular latitude-longitude grid. GFS provides data with a horizontal resolution of 18 miles (28 km) for weather forecasts out to 16 days, and 44 mile resolution (70 km) for forecasts of up to two weeks. GFS data is produced 4 times per day at 00, 06, 12, 18 UTC time [1].

3.2 International Best Track Archive for Climate Stewardship (IBTrACS)

International Best Track Archive for Climate Stewardship (IBTrACS) is a project aimed at providing best-track data for tropical cyclones from all available Regional Specialized Meteorological Centers (RSMCs) and other agencies. It combines multiple datasets into a single product, describing the track of each tropical cyclone in latitude-longitude coordinates. This dataset makes it simple for researchers to locate a given tropical cyclone in both space and time [7] (Fig. 2).
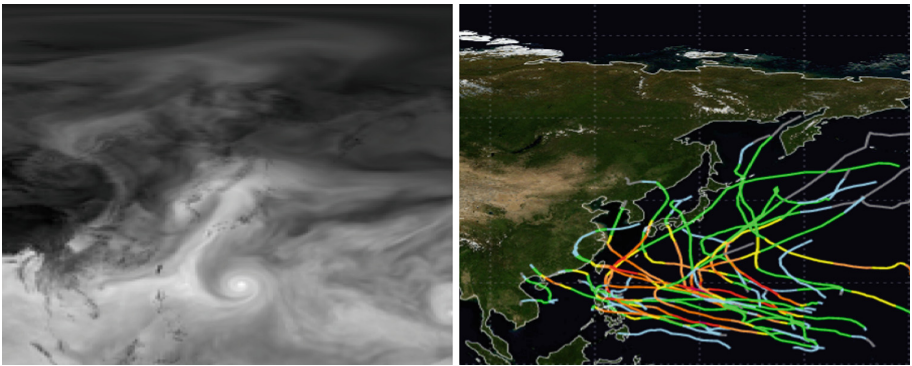

Fig. 2. Combining visualizations of the GFS precipitable water field (left) with IBTrACS cyclone track data (right) produces a labeled dataset of tropical cyclones in the west Pacific basin.

4 Issues and Challenges

Although the atmosphere is continuous, computational constraints force use to approximate it using discrete, gridded data. In order to identify severe weather

phenomenon, we naturally need to separate a large region data into small areas to analyze and prioritize data. Those areas we identify as having high potential impact become our regions of interest (ROI). We define these areas containing a cyclone center as positive samples. However, such identification can cause some challenges.

4.1   Area Size Determination Problem

There is no clear rule to decide the appropriate grid size which could clearly describe different severe weather phenomenon. If we use the smallest unit of data set, a single point of GFS data, it contains only one value and could not be used to describe or identify a weather phenomenon. However, if we use an overly large selection, it may contain too much information which is not related to a weather phenomenon. Therefore, to discuss tropical cyclones in this paper, we adapt a 32×32 image as our area size because this scale is large enough for us to capture each tropical cyclone in our GFS data.

4.2   Occasional Weather Events Problem

Weather phenomenon like tropical cyclone does not happen occur in spatial and time domain. Therefore, events are not distributed evenly throughout the entire region and time. In Fig. 3, we could find that we only have one positive sample in total 36 small areas of west pacific basin at the same time. Also, in entire IBTrACS data set, it is quite common no positive samples on some days because there are no tropical cyclones for that time period. Positive samples are extremely rare compared to negative samples. As a result, we don't have sufficient number of positive samples which are needed for analysis and causes data sparsity problem.

4.3   Weather Events Location Problem

In data preprocessing, we divide a large region into many small areas equally. Throughout tropical cyclone life cycle, location and size vary through time, therefore, center of a tropical cyclones is not permanently located in the center of area without external adjustment. In Fig. 4, we show that there are many situations, a tropical cyclone located near the edge and area doesn't cover entire tropical cyclone.

4.4   No Labeled Data for Negative Samples Problem

Best tracking data set are designed to provide information of tropical cyclones such as location, wind speed, wind direction, timestamps. However, scientists are only interested in regions with tropical cyclones, regions without tropical cyclones are not recorded in any best tracking data set. For example, we can easily find amount of tropical cyclones images from the Internet, but there are
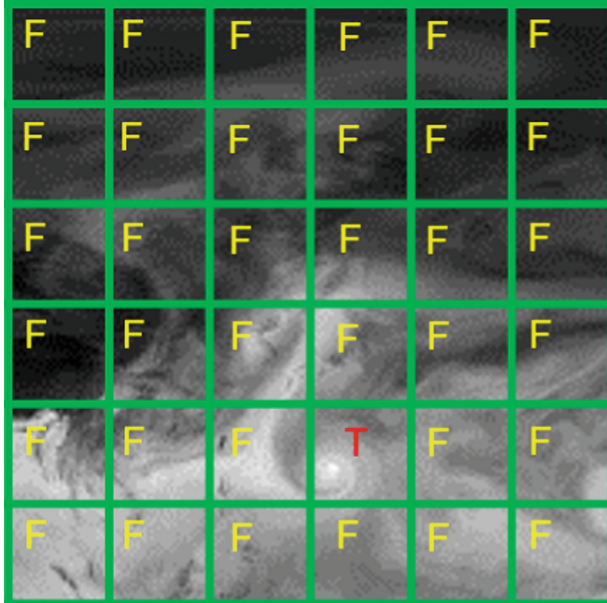
Fig. 3. The west pacific basin is divided into areas and mark a area with a tropical cyclone as true, others as false through cyclone tracking data set

no images described as no tropical cyclones from the Internet. Therefore, there is no clear definition of samples without tropical cyclones from scientists. It leads no definition of negative samples for our training set. In this paper, we randomly select areas which doesn't include cyclones at best tracking data as negative samples. In order to improve the confidence level of prediction, we increase the scale of time and space to make our random selection samples are well distributed.

5 System

Our system is designed to output probability prediction as reference for data assimilation. The input data of system only comes from precipitable water. In order to identify cyclone from a region, our system needs to apply sliding window technique to solve problems we list in Sect. 4. We use center radiation feature extraction to increase accuracy rate. After that, an ensemble algorithm helps the system to reduce the penalty for area prediction error. Figure 5 shows system flow and processing at each step.

5.1   Data Augment with Sliding Window

We want to identify tropical cyclones within an area regardless the cyclone size is large or small, cyclone image is part or all, and cyclone position is in the middle or edge of the image. In order to train our model to identify all these cases, we need to provide training set to cover as more variability as possible. When we form our labeled training set, we naturally don't have too many positive samples with high variability and generate lots or samples which cyclones locate on edge of images if we separate entire region equally. Our method is to use sliding window technique to solve these two issues. We treat the target as a small sight and then sweep the entire picture in the order of upper left, upper right, lower left and lower right. Each area we slide would be treat as a positive or negative sample. This process would produce $W \times H$ samples

$$W = H = (R - A)/S \tag{1}$$

where, R = length of region
A = length of area
S = Interval points between area

Figure 6 shows the results after using sliding window. With sliding window process, image with a cyclone in west pacific basin produces continuous positive samples with the same cyclone in different location of areas. Continuously
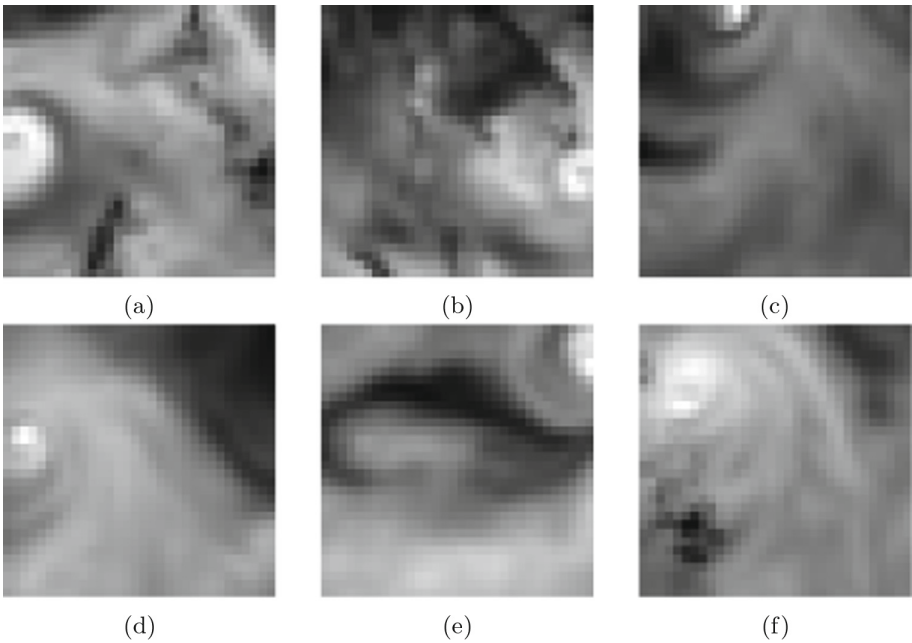

(a)	(b)	(c)
(d)	(e)	(f)

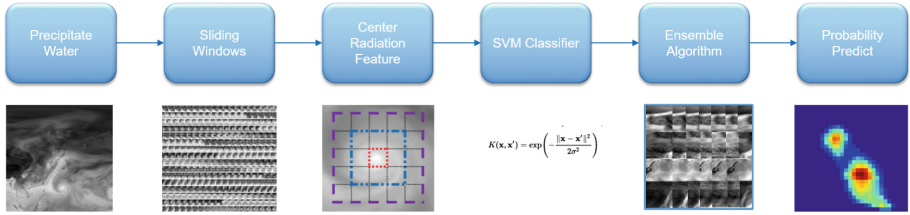**Fig. 4.** Center of tropical cyclone located near or on area edge.

Fig. 5. System overview

slide entire region not only solving the boundary issues because at least one area including entire tropical cyclone located in center, but also generating many positive samples with variability for training. Data augment methods like rotation, flips and scaling are useful to deal few samples issues. However, for our GFS data set, it is more suitable to use sliding window technique for data augment. The reason is that GFS data set cover entire tropical cyclone life cycle and has time relationship between each data. As time changes, the location, shape and size of a tropical cyclone will change. Another advantage of sliding window technique is tropical cyclones would rotate and varies with time, it would generate different variance samples naturally. If we rotate and flips manually, it may generate non-existent tropical cyclone phenomenon.
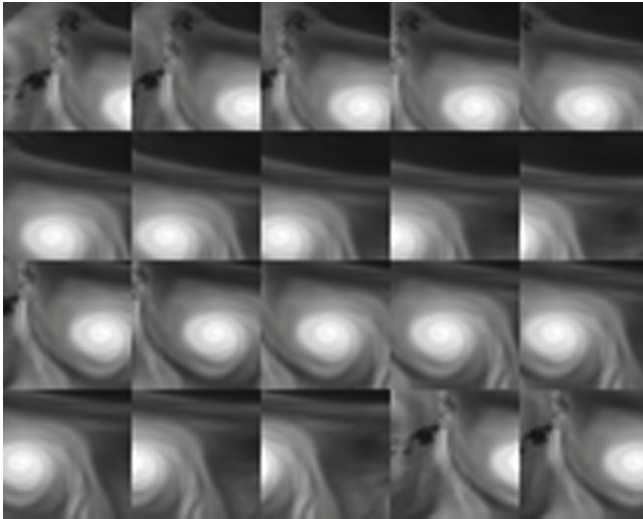

Fig. 6. Output images of tropical cyclone at west pacific basin through sliding windows.

5.2 Feature Extraction Through Center Radiation

Although separating a region into small areas could reduce difficulty to identify a tropical cyclone, it still is not enough to provide high accuracy to determine

whether there is a tropical cyclone in the area. In order to solve this problem, we provide a center radiation algorithm for feature engineering. Figure 7 shows the graphical representation of this algorithm. This algorithm has a two steps mechanism to generate features. First, Use Algorithm 1 to locate maximum value of water vapor within an area and align it as a center. The central point would be a first feature. Second, after determining the center, we take average value of surrounding 8 points as a first layer feature. Then we take average value of surrounding 16 points as a second layer feature. We increase features until reach 8 layers from center. We think this feature engineering could represent the characteristic of tropical cyclones in precipitable water data set. Although a tropical cyclone changing its size and rotating over time, we assume that its shape is still roughly symmetrical. This means that its center will have the maximum value of precipitable water and decrease as the radius widens. Although a tropical cyclone can have more than one maximum value of precipitable water, these values are not randomly distributed but are gathered together. We can arbitrarily choose one of the maximum values without losing representativeness in Algorithm 1.
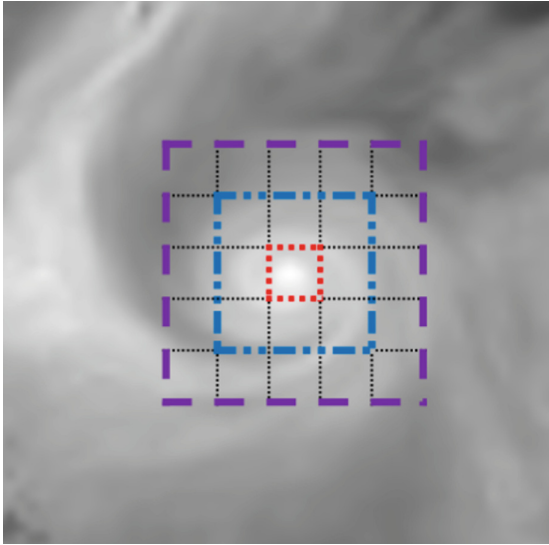

Fig. 7. Think of the center point as a feature and the surrounding point layer as a feature.

5.3   Ensemble Algorithm

With sliding window technique, we separate a region into many small areas, and we need to ensemble classification results of those areas into one probability predict result of entire region. We want to design an algorithm to reduce the cost of predicting errors and increase the weight of correct predictions. The idea is the

Data: Precipitable water value of an area of west pacific basin region

for *Each point in areas* **do**

    From Top left to Bottom right, find the maximum points;

    if *There is only one maximum value* **then**

        Select that point as the center;

    else

        Find the first point with maximum value

        Select that point as the center;

    end

end

<p align="center">Algorithm 1. Determine center point in an area.</p>

more prediction results for one point, the less predictive error will pay. Therefore, we leverage the sliding window result to give a point multiple predictions and use an algorithm to ensemble those predictions. Algorithm 2 shows how we ensemble points in the same area. At present we take $\Delta = 1$ in our algorithm.

Data: Precipitable water value of west pacific basin region

Calculating how many time each point be slided by a sliding window; Total sliding times = the number of times passed by sliding window of a point Value = the value of a point **for** *Each area within the region* **do**

    Identify each area through classifier;

    if *The area is identified to have a tropical cyclone* **then**

        All points within area Value + Δ;

    else

        All points within area Value - Δ;

    end

end

if *The Value of any point ¡ 0* **then**

    set Value of that point to 0

end

Calculation of probability for each points;

Prob of a point = Value / Total sliding times ;

<p align="center">Algorithm 2. Ensemble classified areas into Probability Prediction.</p>

6 Experiment

6.1 Experiment Setup

We combined GFS precipitable water data and IBTrACS best tricking data and generated five years labeled data from 2010 to 2014. With sliding window for data preprocessing, we solved data sparsity problem and could have enough training samples. Also, we randomly selected negative samples to balance positive and negative samples. Because number of tropical cyclones per year are not equal, number of samples from 2010 to 2015 are also not equal, too. In order to prove

our center radiation feature engineering and models work well for tropical cyclone characteristics and can use without worrying about concept drift over time, we treated 2015 year samples, processing with the same sliding window and feature engineering as another test set (Table 1).

Table 1. Data source, training and test data set information.

Data	Event	Resolution Feature
GFS data Tropical cyclones 32x32		Precipitable Water

Data Set	Year	Positive Samples	Negative Samples
Train	2010-2014	105000	105000
Test1	2010-2014	35000	35000
Test2	2015	10000	10000

6.2 Experiment Result and Discussion

In Sect. 5.2, we explained how to extract features through origin data, starting from the center point and layer by layer to extract our features. However, it is hard to estimate how many features we needed. We didn't know if more features can help us better to identify tropical cyclones. An experiment is designed to verify whether this kind of feature extraction works. This experiment began with only one feature, the maximum value we found from Algorithm 1. Then, we increased more features in each step. In order to reduce the influence of very similar points, we obtained the feature by adding two points to the radius at a time. For example, the first added feature is points with radius = 2 from their center point, the second added feature is points with radius = 4 from their center point. We use the same approach for both training and test data.

Table 2 shows the predict results of 2010–2014 test set. In general, the more feature is trained, the more accurate the prediction is for that model. More interesting, from Table 2 we can find the best number of feature when adding more layers will not cause better results. Because number of features may have a close relationship with the size of tropical cyclones, and size of each tropical cyclone varies, too many features will reduce the tolerance of the model for different tropical cyclones. In addition, the balanced training data has a large impact. In our experience, if we don't make the training samples equal, it would be very biased to negative samples as natural of meteorology data.

Meteorological data has a strong relationship with time, therefore, we designed another experiment to discuss our models with concept drift over time. In this experiment we used 2010–2014 training model to predict 2015 tropical cyclone data. As Table 3 shows, it is still consistent with our previous experimental results. The accuracy is proportional to the number of features.

Table 2. Results of prediction of 5 different training models of 2010–2014 test set.

Configuration	True/false	Predict false	Predict true	Accuracy
Center point	False	31223	3777	0.892
	True	934	9690	0.948
Center point + 1 radiation feature	False	31454	3546	0.899
	True	1732	33268	0.951
Center point + 2 radiation features	False	31572	3428	0.952
	True	1682	33318	0.952
Center point + 3 radiation features	False	31634	3366	0.902
	True	1679	33330	0.952
Center point + 4 radiation features	False	31693	3307	0.923
	True	1639	33361	0.953

Table 3. Results of prediction of 5 different training models of 2015 test set.

| Configuration	True/false	Predict false	Predict true	Accuracy
Center point	False	9127	873	0.913
	True	328	9627	0.967
Center point + 1 radiation feature	False	9107	893	0.910
	True	229	9771	0.977
Center point + 2 radiation features	False	9117	883	0.912
	True	190	9810	0.981
Center point + 3 radiation features	False	9120	880	0.912
	True	193	9807	0.981
Center point + 4 radiation features	False	9134	886	0.912
	True	193	9807	0.981

We designed a system to find region of interest of serve weather phenomenon and gave these regions high probability. In our system, we used Algorithm 2 to ensemble areas prediction into region probability prediction. Figure 8 are the outcome of our system. Left side of Fig. 8 are origin precipitable water data from GFS data and right side are output probability predict results which data assimilation process would refer to. Figure 8a and c are one day at 2015 and 2016 without any tropical cyclone. On the other hand, Fig. 8e has two tropical cyclones (one is located at edge) and Fig. 8g has one tropical cyclone. Ideally, Fig. 8b and d should be all 0; however, these two figures still are affected by false predictions and have some high probability points which should not exist. In Fig. 8f and h, we covered most tropical cyclones center and regions around tropical cyclones although Fig. 8h is affected by two major false prediction in corners. Fortunately, on the premise that computing power can be handled, a small amount of false prediction will only consume more computing power
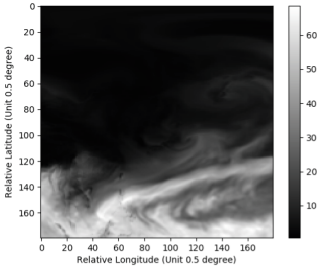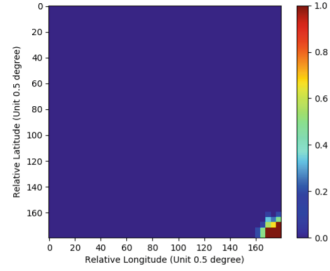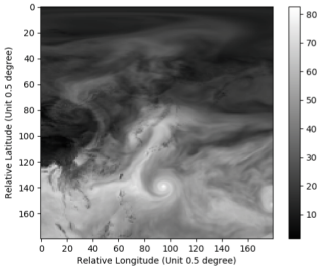
(a) Precipitable water data without tropical cyclone.
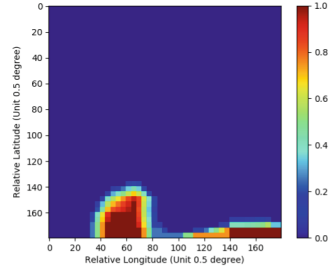
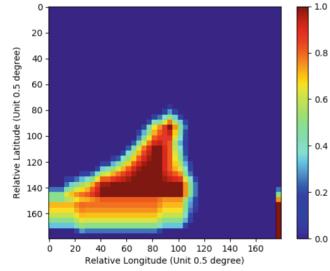(b) Probability predict result without tropical cyclone.
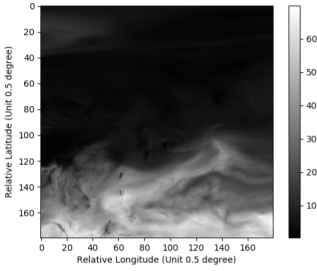
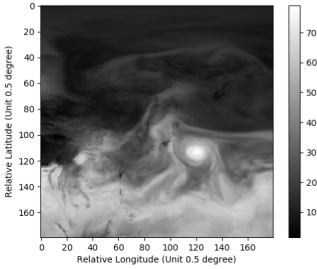(c) Precipitable water data without tropical cyclone.

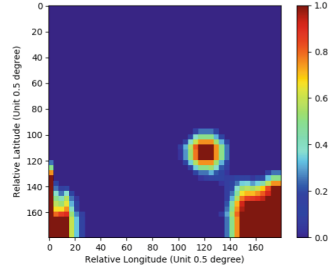(d) Probability predict result without tropical cyclone.

(e) Precipitable water data with two tropical cyclones.

(f) Probability predict result with two tropical cyclones.

(g) Precipitable water data with a tropical cyclone.

(h) Probability predict result with a tropical cyclone.

Fig. 8. Probability prediction

without affecting the predict result of NWP. Our system is designed to feed high resolution satellite data into region of interesting, but still keep update original resolution data for others.

7   Conclusion

In this paper, we apply machine learning in GFS precipitable water data to identify tropical cyclones, facilitate sliding window technique to solve data sparsity and overcome the traditional limitation that tropical cyclones were needed to be put in the center of figures for identification. The center-radiation methods for feature engineering of precipitable water data was proved to achieve fairly high classification accuracy rate for 2010–2015 GFS data set around 94%. Our system produces the probability predict result as a reference of data assimilation process. The probability predict results indicate the region of interest where we may need to put high resolution satellite data and increase initial value update frequency to help NWP with better weather prediction. This successful result could be a good pilot study for further GOES-15 or GOES-16 satellite image research.

References

1. Global forecast system (GFS). https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs
2. Cecil, D.J., Zipser, E.J.: Relationships between tropical cyclone intensity and satellite-based indicators of inner core convection: 85-GHZ ice-scattering signature and lightning. Mon. Weather Rev. **127**(1), 103–123 (1999)
3. Dvorak, V.F.: Tropical cyclone intensity analysis and forecasting from satellite imagery. Mon. Weather Rev. **103**(5), 420–430 (1975)
4. Ho, S.-S., Talukder, A.: Automated cyclone discovery and tracking using knowledge sharing in multiple heterogeneous satellite data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 928–936. ACM (2008)
5. Ho, S.-S., Talukder, A.: Automated cyclone identification from remote QuikSCAT satellite data. In: 2008 IEEE Aerospace Conference, pp. 1–9. IEEE (2008)
6. Jaiswal, N., Kishtawal, C.M., Pal, P.K.: Cyclone intensity estimation using similarity of satellite IR images based on histogram matching approach. Atmos. Res. **118**, 215–221 (2012)
7. Knapp, K.R., Kruk, M.C., Levinson, D.H., Diamond, H.J., Neumann, C.J.: The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data. Bull. Am. Meteorol. Soc. **91**(3), 363–376 (2010)

8. Liu, Y., et al.: Application of deep convolutional neural networks for detecting extreme weather in climate datasets. arXiv preprint arXiv:1605.01156 (2016)
9. McBride, J.L., Zehr, R.: Observational analysis of tropical cyclone formation. Part II: comparison of non-developing versus developing systems. J. Atmos. Sci. **38**(6), 1132–1151 (1981)
10. Mueller, K.J., DeMaria, M., Knaff, J., Kossin, J.P., Vonder Haar, T.H.: Objective estimation of tropical cyclone wind structure from infrared satellite data. Weather Forecast. **21**(6), 990–1005 (2006)
11. Panangadan, A., Ho, S.-S., Talukder, A.: Cyclone tracking using multiple satellite image sources. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 428–431. ACM (2009)
12. Ryan, B.F., Watterson, I.G., Evans, J.L.: Tropical cyclone frequencies inferred from gray's yearly genesis parameter: validation of GCM tropical climates. Geophys. Res. Lett. **19**(18), 1831–1834 (1992)
13. Velden, C.S., Olander, T.L., Wanzong, S.: The impact of multispectral goes-8 wind information on Atlantic tropical cyclone track forecasts in 1995. Part I: dataset methodology, description, and case analysis. Mon. Weather Rev. **126**(5), 1202–1218 (1998)
14. Zhuge, X.-Y., Guan, J., Yu, F., Wang, Y.: A new satellite-based indicator for estimation of the western North Pacific tropical cyclone current intensity. IEEE Trans. Geosci. Remote Sens. **53**(10), 5661–5676 (2015)
15. Zou, J., Lin, M., Xie, X., Lang, S., Cui, S.: Automated typhoon identification from QuikSCAT wind data. In: 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4158–4161. IEEE (2010)