Using Deep Learning to Estimate Tropical Cyclone Intensity from Satellite Passive Microwave Imagery

ANTHONY WIMMERS AND CHRISTOPHER VELDEN
Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin–Madison, Madison, Wisconsin

JOSHUA H. Cossuth
Remote Sensing Division, U.S. Naval Research Laboratory, Washington, D.C.

(Manuscript received 8 November 2018, in final form 22 March 2019)

ABSTRACT

A deep learning convolutional neural network model is used to explore the possibilities of estimating tropical cyclone (TC) intensity from satellite images in the 37- and 85–92-GHz bands. The model, called “DeepMicroNet,” has unique properties such as a probabilistic output, the ability to operate from partial scans, and resiliency to imprecise TC center fixes. The 85–92-GHz band is the more influential data source in the model, with 37 GHz adding a marginal benefit. Training the model on global best track intensities produces model estimates precise enough to replicate known best track intensity biases when compared to aircraft reconnaissance observations. Model root-mean-square error (RMSE) is 14.3 kt (1 kt = 0.5144 m s\(^{-1}\)) compared to two years of independent best track records, but this improves to an RMSE of 10.6 kt when compared to the higher-standard aircraft reconnaissance-aided best track dataset, and to 9.6 kt compared to the reconnaissance-aided best track when using the higher-resolution TRMM TMI and Aqua AMSR-E microwave observations only. A shortage of training and independent testing data for category 5 TCs leaves the results at this intensity range inconclusive. Based on this initial study, the application of deep learning to TC intensity analysis holds tremendous promise for further development with more advanced methodologies and expanded training datasets.

1. Introduction

Deep learning (DL) is a newly popular, powerful and often confounding computational tool for developing predictive models in the sciences. It builds on a long legacy of neural network modeling, with a key feature being the organization of neural connections into multiple layers of nonlinear operations, enabling models to apply high levels of abstraction in their tasks. New hardware innovations, particularly in accessing graphical processing units (GPUs), have enabled DL algorithms to become powerful enough to rival human performance at complicated tasks such as image classification (He et al. 2016). However, unlike most other computational methods that are applied to scientific problems, DL models can paradoxically achieve high predictive performance (Schmidhuber 2015; LeCun et al. 2015) but a DL model’s methods are difficult to practically untangle from its network architecture in a way that derives meaningful scientific information.

While the broader field of machine learning has had a long and fruitful application to meteorology (see Haupt et al. 2008; McGovern et al. 2017 and references therein), the state of the science of DL applied to meteorology is limited but rapidly growing. A large portion of work in this field to date applies to short-term forecasting for renewable energy (Diaz et al. 2015; Wan et al. 2016; Sogabe et al. 2016; Hu et al. 2016). Other major research includes feature identification for long-term climate analysis (Kurth et al. 2018) and augmenting the output of precipitation models with two-level DL techniques (Tao et al. 2016, 2018; Tao and Gao 2017). Recently, Pradhan et al. (2018) and Chen et al. (2018) introduced DL models applied to infrared and morphed rain rate images of
hurricanes, and these are discussed in detail later. The proliferation of recent conference abstracts on DL in meteorology also shows the impressive growth of this field in the past year (e.g., Prabhat et al. 2019; Lagerquist et al. 2019; Stewart et al. 2019; Gagne et al. 2019; Wu et al. 2019; Boukabara et al. 2019; Hall et al. 2019).

Tropical cyclone (TC) nowcasting is particularly well suited for DL applications. Operational TC intensity analysis remains largely dependent on quasi-subjective techniques such as the Dvorak technique (Dvorak 1984; Velden et al. 2006), requires intensive analyst training, and lacks a formal incorporation of imagery beyond visible and infrared window frequencies. Furthermore, a new generation of infrared imagers (Schmit et al. 2005, 2017) and emerging cubesat sensors (Blackwell et al. 2012; Cahoy et al. 2015; Reising et al. 2016) require new, robust techniques for TC nowcasting and forecasting that can assimilate a spectrally diverse variety of images all at once.

It is well known that the microwave imaging frequencies available on polar-orbiting meteorological satellites since the late 1980s offer unique perspectives for observing TCs. Uses of the imagery for TC applications have been well documented in earlier studies (e.g., Velden et al. 1989; Cecil and Zipser 1999; Hawkins et al. 2001; Kieper and Jiang 2012; Jiang 2012; Edson et al. 2008; Ritchie et al. 2012, 2014), and automating and enhancing the logic of the Dvorak technique with Geo IR (Olender and Velden 2007, 2018). The root-mean-square error (RMSE) of these methods ranges from 10 to 15 kt (1 kt $\approx 0.5144 \text{ m s}^{-1}$). Each has advantages and limitations discussed in section 6, so further improvements in satellite-based objective intensity estimates are desirable.

This paper introduces a DL model designed to estimate TC intensity using 37- and 89-GHz band imagery. In doing so, we explore the potential of this application of DL to 1) classify satellite microwave images to compete in performance with existing subjective and automated analysis methods, 2) provide reliable uncertainty information to accompany its output, and 3) improve the understanding of certain TC-related aspects of 37- and 89-GHz sensor capabilities in areas that have not been well addressed. Even though the technique requires an introduction to many new concepts in machine learning, the particular DL model applied here is a simple one. As such, this study focuses on the kind of model performance that is easily achievable and able to be improved with further development. Thus we present the DL model that follows as a demonstration of capabilities and not a final operational version.

2. Data

Generally, DL performs best with at least tens of thousands of training samples, and model performance scales logarithmically with the training sample size (Sun et al. 2017). Thus we have sought out the largest available dataset of TC observations in the 37- and 89-GHz bands. This is available in the Microwave
Imagery from NRL TC (MINT) collection, which covers global conical scanner observations from 1987 to 2012. As described in Cossuth et al. (2013), the dataset includes brightness temperatures from DMSP SSM/I and SSMIS, TRMM TMI, and Aqua AMSR-E (Table 1, with acronyms defined). The diverse frequency bands from these sensors centered on 85, 89, and 91 GHz are normalized in MINT to the sensitivity of AMSR-E (89 GHz) using the technique of Yang et al. (2014). By contrast, the small variation in spectral response between various sensors’ 37-GHz bands does not require renormalization.

The imagery in MINT includes observations of TCs ranging in intensity from 15 to 160 kt. Because of the intensity distributions of the typical TC life cycle, the images are most frequently of the tropical depression stage, then decline in number with increasing intensity. We filter the dataset for three factors to prevent misleading structural patterns: 1) TCs must be centered over water, 2) scan coverage must be >65% of the image box, and 3) the TC must not be labeled as “extratropical transition” in the records. The images are cropped to a width of 3.68 (~400 km) in both dimensions and sampled to 0.058 (~5 km) resolution using nearest-neighbor interpolation, making the working images $72 \times 72$ pixels in size.

A standard measure of TC intensity is the maximum sustained wind (MSW), defined as the TC’s maximum 1-min sustained wind at 10 m above the surface. Here we use the estimated MSW from the best track records of two sources: 1) the HURDAT2 database, which uses estimates from the National Hurricane Center for the North Atlantic or east Pacific basins, and the Central Pacific Hurricane Center in the central Pacific basin (Landsea et al. 2013); and 2) the Joint Typhoon Warning Center best tracks in the west Pacific, north Indian Ocean, and Southern Hemisphere (Chu et al. 2002). Each image is assigned the best track MSW matching the image time.

### Table 1. Satellite microwave instruments used in the MINT dataset.

<table>
<thead>
<tr>
<th>Satellite</th>
<th>Instrument</th>
<th>Frequency (GHz)</th>
<th>Orbit</th>
<th>Footprint (km x km)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMSP F16–F18</td>
<td>SSMIS</td>
<td>37.0, H 91.7, H</td>
<td>Polar, sunsynchronous</td>
<td>45 × 28, 14 × 13</td>
<td>Kunkee et al. (2008)</td>
</tr>
<tr>
<td>TRMM</td>
<td>TMI</td>
<td>37.0, H 85.5, H</td>
<td>Equatorial, between 38°S and 38°N</td>
<td>16 × 9, 7 × 5</td>
<td>Kummerow et al. (1998)</td>
</tr>
<tr>
<td>Aqua</td>
<td>AMSR-E</td>
<td>36.5, H 89.0, H</td>
<td>Polar, sunsynchronous</td>
<td>14 × 8, 6 × 4</td>
<td>NASA MSFC (2001)</td>
</tr>
</tbody>
</table>


![Fig. 1. Histogram of TC passive microwave image samples used in the model training.](image-url)
through linear interpolation. Note that using these best track MSW estimates as “truth” is not optimal because only a minority of these values has in situ confirmation; however, this was necessary for maintaining a large sample size of relatively homogeneous data. Analysis issues stemming from the limitations in this dataset are discussed in sections 5 and 6.

The standard practice for DL modeling is to split the dataset into independent training, validation, and “testing” components. Here, “validation” has a different meaning than usual for atmospheric science applications. It denotes an independent confirmation of model accuracy during training. By contrast, the “testing” component uses independent data and ultimately measures the accuracy of the finished model. In this study we use year 2010 data for the validation, years 2007 and 2012 data for testing, and all other years for training (Fig. 1, Table 2). These years were selected because 1) later years have more data and better best track estimates; and 2) 2007, 2010, and 2012 have an adequate amount of category 4–5 aircraft reconnaissance observations for statistical evaluation. In addition, we create a so-called “balanced” dataset during training and validation, which is composed of duplicated data at lesser-sampled intensities so that all values of MSW are equally represented. For example, referring to respective sample sizes in Fig. 1, the 30-kt bin remains as is, the 50-kt bin is duplicated 2.1 times and the 160-kt bin is duplicated 1233 times. Furthermore, all imagery including the duplicated imagery is varied through rotation and offsetting as described in section 4. This approach is necessary for the model to assign equal importance to all TC intensities during training, and is common practice in CNN training in order to prevent the model from inherently increasing its skill with the greater-sampled image types at the expense of the lesser-sampled image types. However, the testing dataset does not need to be balanced in this way, because the determination of model skill follows a different (and more conventional) approach.


The common Deep Learning network model applied to image data, and used here, is a convolutional neural network (CNN). A full tutorial on how CNNs operate

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>N (balanced)</th>
<th>Years</th>
<th>Basins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation</td>
<td>3016</td>
<td>13,041</td>
<td>2010</td>
<td>Global</td>
</tr>
<tr>
<td>Testing</td>
<td>6705</td>
<td>-</td>
<td>2007, 2012</td>
<td>Global</td>
</tr>
<tr>
<td>Testing (recon only)</td>
<td>404</td>
<td>-</td>
<td>2007, 2012</td>
<td>North Atlantic, east Pacific</td>
</tr>
</tbody>
</table>

Fig. 2. Simplified diagram of the standard convolutional neural network architecture, with terms explained in section 3.
is beyond the scope of this paper, but for background we recommend O’Shea and Nash (2015) and Deshpande (2016) to reach a more complete understanding of the terminology and methods used in this section. Dieleman et al. (2015) also provide an excellent demonstration of CNNs to the problem of galaxy classification along with a thorough introduction of the concepts. In broad terms, a CNN operates by applying two- or three-dimensional “filters” that scan an input image for elemental features of relevance to the task, and then producing an intermediate product with multiple channels (“feature maps”) of information in a third dimension (Fig. 2), while maintaining the spatial structure of the image information. This operation repeats many times with each operation producing a new intermediate layer. The progressively larger-scale image representation of successive layers allows them to work at a higher level of abstraction than the layer before. Layer-wise operations include convolution (to resolve relevant patterns and shapes), pooling (to reduce the size of the information), normalization (to prevent runaway weighting) and activation (to introduce nonlinearity with simple signal gates). In the final stages of a CNN model, this three-dimensional intermediate product is then remapped as a flat, one-dimensional array of nodes. This array connects to a set of “fully connected” layers and then to the output layer. Together, all these layers make up an architecture of interconnected nodes, or “neurons,” found in the convolutional filters and layered connections. Each connection between neurons is calibrated with weights and offsets during the training of the model in order to produce an optimized output. The primary function of the convolution and pooling segment, shown in Fig. 2, is for feature identification, whereas the primary function of the fully connected segment is classification. Although normalization and activation are not explicitly shown in this diagram, they are applied throughout the model.

The particular architecture of our CNN model, called “DeepMicroNet” (Table 3), is loosely based on the AlexNet design for image classification (Girshick et al. 2014). The “stride” column accounts for how the intermediate

### Table 3. DeepMicroNet model architecture.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Type</th>
<th>Filter size</th>
<th>Stride</th>
<th>Activation</th>
<th>Feature maps</th>
<th>Feature size</th>
</tr>
</thead>
<tbody>
<tr>
<td>In</td>
<td>Input</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1 or 2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>64 × 64</td>
</tr>
<tr>
<td>C1</td>
<td>Convolution</td>
<td>5 × 5</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>16</td>
<td>64 × 64</td>
</tr>
<tr>
<td>P1</td>
<td>Max pooling&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2 × 2</td>
<td>2</td>
<td>—</td>
<td>16</td>
<td>32 × 32</td>
</tr>
<tr>
<td>BN1</td>
<td>Batch normalization</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>16</td>
<td>32 × 32</td>
</tr>
<tr>
<td>C2</td>
<td>Convolution</td>
<td>5 × 5</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>32</td>
<td>32 × 32</td>
</tr>
<tr>
<td>P2</td>
<td>Max pooling</td>
<td>2 × 2</td>
<td>2</td>
<td>—</td>
<td>32</td>
<td>16 × 16</td>
</tr>
<tr>
<td>BN2</td>
<td>Batch normalization</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>32</td>
<td>16 × 16</td>
</tr>
<tr>
<td>C3</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>64</td>
<td>16 × 16</td>
</tr>
<tr>
<td>C4</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>64</td>
<td>16 × 16</td>
</tr>
<tr>
<td>C5</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>64</td>
<td>16 × 16</td>
</tr>
<tr>
<td>P5</td>
<td>Max pooling</td>
<td>2 × 2</td>
<td>2</td>
<td>—</td>
<td>64</td>
<td>8 × 8</td>
</tr>
<tr>
<td>C6</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>128</td>
<td>8 × 8</td>
</tr>
<tr>
<td>C7</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>128</td>
<td>8 × 8</td>
</tr>
<tr>
<td>C8</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>128</td>
<td>8 × 8</td>
</tr>
<tr>
<td>P8</td>
<td>Max pooling</td>
<td>2 × 2</td>
<td>2</td>
<td>—</td>
<td>128</td>
<td>4 × 4</td>
</tr>
<tr>
<td>C9</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>256</td>
<td>4 × 4</td>
</tr>
<tr>
<td>C10</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>256</td>
<td>4 × 4</td>
</tr>
<tr>
<td>C11</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>256</td>
<td>4 × 4</td>
</tr>
<tr>
<td>P11</td>
<td>Max pooling</td>
<td>2 × 2</td>
<td>2</td>
<td>—</td>
<td>512</td>
<td>2 × 2</td>
</tr>
<tr>
<td>C12</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>512</td>
<td>2 × 2</td>
</tr>
<tr>
<td>C13</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>512</td>
<td>2 × 2</td>
</tr>
<tr>
<td>C14</td>
<td>Convolution</td>
<td>3 × 3</td>
<td>1</td>
<td>Leaky ReLU</td>
<td>512</td>
<td>2 × 2</td>
</tr>
<tr>
<td>C15</td>
<td>Convolution</td>
<td>2 × 2</td>
<td>2</td>
<td>Leaky ReLU</td>
<td>1024</td>
<td>1 × 1</td>
</tr>
<tr>
<td>FC1</td>
<td>Fully connected</td>
<td>—</td>
<td>—</td>
<td>Leaky ReLU</td>
<td>—</td>
<td>1000</td>
</tr>
<tr>
<td>FC2</td>
<td>Fully connected</td>
<td>—</td>
<td>—</td>
<td>Leaky ReLU</td>
<td>—</td>
<td>200</td>
</tr>
<tr>
<td>Out</td>
<td>Fully connected</td>
<td>—</td>
<td>—</td>
<td>Softmax</td>
<td>—</td>
<td>29</td>
</tr>
</tbody>
</table>

<sup>a</sup> Terms—Filter Size: Size of the convolutional filter, Stride: Spacing of the convolutional/pooling filter application, Activation: Nonlinear “signal gate” processing method following the layer process, Feature Maps: Layer depth of the product of convolution/pooling/etc., Feature Size: Row × column size of the product of convolution/pooling/fully connected layer, Leaky ReLU: “Leaky Rectified Linear Unit” (a type of activation function that prevents neuron saturation in deep networks), Max Pooling: Calculation a local maximum of data within feature maps (here in a 2 × 2 window, stepped by a stride of 2) in order to efficiently condense the information to fewer variables.

<sup>b</sup> 1 if one channel only, 2 if 37 and 89 GHz.

<sup>c</sup> Steps that reduce the data size within the CNN are indented.
layers are subsampled. The “Leaky ReLU” activation is a fairly rapid and streamlined activation function (signal gate) among several available options. The final signals from the model (“scores”) are then processed with the softmax function $S$, which normalizes the output into a probability density function (PDF) in the output dimension:

$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}},$$

(1)

where $e^{y_i}$ is the exponent of the raw score coming from the CNN model for the classification $i$ (such as 70, 75 kt, etc.), and the denominator is the sum of the terms for all classifications.

This model design generally follows the best practices of simple CNN construction. Specifically, convolutional filter sizes are kept small in order to follow the “narrow and deep” approach rather than “wide and shallow” because larger filters tend to increase computational time. The addition of a pooling step is typical at every other layer in the beginning in order to reduce dimensionality of the data early in processing, and then less frequently afterward. The many parameters (overall number of layers, feature map number, fully connected neuron number, filter size and stride) were optimized by trial and error, but are near the typical values for similar CNN models.

Most CNN architecture is designed to make classifications that have little to no interrelationship, such as “car,” “pedestrian,” and “street sign,” and a CNN model is optimized to simply maximize the probability of correct classification. However, it would be unsuitable to adopt this strict convention to the TC intensity problem because then the model would not understand that “70 kt” is a close approximation of “75 kt,” and so on.

An alternative approach is to produce a scalar value as output (such as MSW) and optimize the model by minimizing the least squares error with the training data. However, it would be less helpful to end users to supply just a MSW value as output with no context. Instead, we take the following unique approach that retains the probabilistic output of the classification approach along with a built-in consideration for the relationship between similar values of MSW. Here, the “truth” value of MSW is treated as a weighted distribution of MSW values in rough accordance with the best track uncertainty reported by Torn and Snyder (2012). For instance, one case of a best track value of “70 kt” is treated in this model as a range of values [60, 65, 70, 75, 80 kt] weighted by importance with the input filter $W = [0.10, 0.23, 0.34, 0.23, 0.10]$. The filter is applied universally, except that at the edges of the MSW range the filter is truncated and renormalized accordingly. This approach has the advantage of smoother and more robust guidance toward the lowest error during model training without the complicating influence of more distant values of MSW, which can happen in a least-squared error approach.

### Table 4. Training hyperparameter values in DeepMicroNet.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch size</td>
<td>1000</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.006</td>
</tr>
<tr>
<td>Learning rate decay</td>
<td>0.93 over 500 batch steps</td>
</tr>
<tr>
<td>Momentum</td>
<td>0.9</td>
</tr>
<tr>
<td>Dropout rate on fully connected layers</td>
<td>50%</td>
</tr>
</tbody>
</table>

**FIG. 3.** Learning curve for the two-channel DeepMicroNet model. Training progresses at 1000 images per step, hence the number of training steps per epoch is $240 \times 746/1000 = 241$. The model state at the point “Best loss” is the state used in this study.

**FIG. 4.** Samples of (top) 37-GHz-only model results and (bottom) 89-GHz-only model results imagery for Hurricane Felix (category 4). Image titles are [basin][TC No.][yr][month][day][h][min][s]_[satellite sensor]. Grayscale value varies from 160 (bright) to 280 K (dark). The green line on the histograms corresponds to the recon-aided best track intensity (see section 5b) interpolated to image time.
We have set up 33 possible classifications of best track MSW values from 10 to 170 kt, stepped by 5 kt. Note that during training we use duplicates of the 20-kt data to train the 10- and 15-kt classifications, and likewise the 165- and 170-kt classifications are trained with copies of the 160-kt data. We developed this edge-buffer approach in order to allow PDFs to spread naturally around the actual limits of 20 and 160 kt for more accurate training. This would admittedly fail to train the exceedingly rare cases of >160-kt intensity, but this issue is not important for the proof-of-concept purpose of this study.

The DeepMicroNet model is built in Python with the Tensorflow library (Abadi et al. 2016), which was selected for its speed and versatility. The model training runs on a server with 3 Intel Haswell CPUs, and always in GPU-enabled mode employing one GeForce GTX 1080 8 GB video card.

4. Model training

Model training in a DL architecture begins with a random initialization of weights between model nodes,
and then iteratively improves performance by shaping the network toward an optimal fit to the training dataset through stochastic gradient descent. This is very similar to the approach of adjusting the state of an atmospheric model with observational data and converging on a new solution with gradient descent along an adjoint model (Thacker 1988; Errico 1997). Here, the goal of the model is to reach optimal performance by minimizing the loss function, defined as the “cross entropy” of the output PDFs:

\[
\text{loss} = -\frac{1}{N} \sum_{i=0}^{N} W(y_i) \log S'(y_i),
\]

where \(N\) is the number of classifications, \(W\) is the input filter applied to the “truth” classification \(y_i\) and \(S'(y_i)\) is the softmax function applied to the model output. The term “cross entropy” is appropriate here because it parallels the mathematics of entropy along various states of energy in statistical mechanics.

To reduce overfitting during model training, we also employed three dataset augmentation techniques to the training data. First, the positions of the training images are randomly shifted -4 to 4 pixels in each direction (±20 km, to match the uncertainty in best track positions) and cropped to 64 pixels in size from the original 72-pixel template. Second, the images are randomly flipped with 50% probability along the horizontal axis so that Northern and Southern Hemisphere configurations are represented equally, and so that the model is rotationally invariant. Third, the images are rotated randomly up to 40° in either direction. The training process works through many iterations of the dataset so that many combinations of offset, north/south orientation and rotation are represented for each image.
The training process for the two-channel model (discussed in section 5b) converges to an optimum solution after 34 iterations, or “epochs,” of the training data (Fig. 3). This lasted about 3 h on our system. Note that the optimum solution is the state in which the validation dataset, not the training dataset, has the lowest computed loss, which is standard practice for model development in machine learning. Further model training generally results in overfitting to the training dataset.

Finally, the training hyperparameters, which control the learning process of the training stage (and are rather esoteric to nonmachine learning researchers), were finalized by recursively searching for the optimum hyperparameter value that efficiently minimized the loss, one at a time. These values are shared in Table 4 simply to help with the reproducibility of the model.

5. Independent case testing of the model

This section primarily describes the results of independent model testing and highlights the most important features. For better clarity, the further discussion of their meaning follows in the next section.

a. Comparative testing of 37- and 89-GHz channels

To understand the relative value of each channel to estimating TC intensity, the DeepMicroNet model was run in three versions: 1) using only the 37-GHz channel, 2) using only the 89-GHz channel, and 3) using both channels. Because model performance varies slightly in each training session, we trained the model three times for each scenario and selected the best performer (with respect to the validation dataset) to represent the group. An example of 37-GHz-only and 89-GHz-only performance (Fig. 4) shows how the model produces estimates from image input. The probability density function (PDF) is much wider for the 37-GHz-only scenario than for the 89-GHz-only scenario, due to lower precision in the model’s ability to train on the data, and resulting in higher uncertainty in the model’s intensity estimate.

Full results clearly show that the 89-GHz-only model has lower error and lower uncertainty than the 37-GHz-only
model overall (Fig. 5). The 2-channel model, on the other hand, slightly outperforms the 89-GHz-only model, but shows no substantial difference in precision (width of PDFs). Evidently the 89-GHz band supplies the more relevant information to TC intensity, and the model uses the 37-GHz band in a similar but less impactful way. Only a marginal amount of information from the 37-GHz band contributes uniquely to the 2-channel model, with the largest improvement coming in the category 5 intensity range. However, it is difficult to generalize this difference because of the small sample size for category 5. Overall, the improvement is enough to justify limiting the remaining model evaluation to only the two-channel version of DeepMicroNet going forward.

b. Model performance

The following describes a two-channel model selected as the best of 12 runs based on performance using the pretesting (validation) data. (The model was found to perform differently on each run due to random initialization and stochastic iterations toward a local optimum.) Model performance is evaluated against both the best track record and a more limited dataset of best track intensity estimates within 3 h from an aircraft reconnaissance observation (hereafter the “recon-aided best track”). The first part evaluates the ability of the model to replicate the best track intensity estimates, whereas the second part evaluates the ability of the model to estimate more accurate recon-aided best track intensity. These two comparisons are different insofar as the best track has errors and biases of its own, largely due to the predominant influence of the Dvorak technique on the best track record.

A small, representative sample from the 6705 test cases demonstrates the variety of model output (Figs. 6–12). (This sample was selected from the recon-aided dataset of the east Pacific and North Atlantic basins in order to insure an accurate “truth” MSW value, but we have found that the following analysis applies to all basins.) The figures are ordered by Saffir–Simpson hurricane wind scale category, and each figure shows one underestimation of MSW (top row), two high-accuracy
The shape of the PDFs is approximately a normal curve, often with a slight skew toward the center of the x-axis. Inspection of model output indicates that the model predominantly responds to the characteristics of organization in the images, including eyewall definition, banding and the extent of cyclonic signatures, which increases consistently from Fig. 6 to Fig. 12. Similarly, the main reason for a departure in model accuracy (top and bottom rows) is that the image is either less organized or more organized than its corresponding MSW typically indicates. Consistent with the training dataset characteristics, many TC eyes are not perfectly aligned in the image, which demonstrates that the model is robust to off-center eyes. Note also that partial satellite scans (Fig. 7, third row) do not prevent normal model performance.

Viewed as a scatterplot, the model shows fairly consistent accuracy across nearly all intensities (Fig. 13). In both the best track and the recon-aided comparisons, the most prominent departure from the 1:1 line occurs above approximately 135 kt in the best track MSW, where the model no longer shows a convincing trend of increasing MSW estimates. The other patterns of bias and random error are detailed more clearly in the statistical plots of Fig. 14 (best track comparison) and Fig. 15 (recon-aided comparison).

Starting with the comparison to best track intensity for all microwave images, note that tropical depressions and tropical storms are well sampled at every 5-kt step, whereas higher intensities are less common and are better analyzed in larger groups (Fig. 14a). For this reason we include 15-kt bins of the results in the subplots that follow (similar to a running average). The model bias (Fig. 14b) shows a predictable consequence of “regression toward the mean” at the extremes, where the model is biased high at the low extremes of MSW, and is biased low at the high extremes of MSW. This shows that the model training technique intended to avoid such bias (section 3) was only partly successful at the extremes. Otherwise, the main feature of model bias is the slightly

---

1 The one exception to this arrangement is with category 5 examples (Fig. 12), where no positive bias examples existed.
(but significantly) positive values in the middle range (30–110 kt), which we will consider along with the bias against recon-aided best track later on. Predictably, the modes of the PDFs have less bias than the means because the means are more sensitive to skew, but it is worthwhile to analyze the statistics of the mean and mode together because the mean has less error away from the extremes of MSW (Fig. 14c). However, error from negative bias clearly dominates the model RMSE in the category-5 range of intensities for both the mode and mean. Likewise, positive bias is also significant for the RMSE, but less extreme, at the tropical depression (TD)–tropical storm (TS) range.

All the statistics up to this point evaluate the accuracy of a single center value in the PDF distribution, but what about the distribution itself? Here the reliability of the PDF is measured by the percentage of samples whose best track MSW fall inside the innermost 50% of the PDFs (Fig. 14d). The percentage is above 50% if the PDFs are too wide and below 50% if the PDFs are too narrow. For most TC intensities the value is close to 50%, indicating that the PDF distributions accurately reflect the true uncertainty of the model MSW estimate. Only at category 5 intensities are the values too low, confirming that here the distributions do not extend far enough.

Recon-aided best track intensities are fewer in number by more than an order of magnitude (Fig. 15a), but still numerous enough that it is at least possible to discuss the trends according to the Saffir-Simpson hurricane wind scale category. In the TS–category 1 range, the model is biased slightly low, and then biased high at categories 2–4 (Fig. 15b). The strong negative bias at the category 5 range is similar to the best track analysis above, but here it is even more extreme. RMS error (Fig. 15c) generally varies more with TC intensity, with one relative maximum around categories 1–2, and another maximum at category 5. Note also that error is consistently lower for the model-derived mean than for the mode, likely because of its lower magnitude of bias (except for category 5). Finally, the evaluation of the model PDF (Fig. 15d) has well over 50% of the

Fig. 10. As in Fig. 6, but for category 3 strength TCs. Note that the most significant low bias case in this range (top row) has a mean error of only –5.4 kt.
observations falling within the inner 50% of the model PDF for most intensities. This reveals that over most intensities the model PDF is too wide, or in other words, the model intensity estimate is actually more precise than its PDF indicates.

Error statistics from Figs. 14 and 15 using the mean of PDF values are summarized in Table 5.

c. Sensitivity of the model to common TC variables

A further examination of the model sensitivity to latitude, translation speed, and image resolution applied to recon-aided observations is provided in Table 6. Pairs of statistics with significant differences in bias or RMSE are identified in bold. However, the groupings according to latitude and translation speed have substantial differences in average MSW, which indicates a serious potential for cross correlation between model error and TC intensity. For example, the RMSE is significantly lower in each of the groupings for latitude and translation speed that have lower MSW (and lower MSW alone corresponds to lower RMSE as in Fig. 15c). A larger sample of observations would be necessary to control for this influence, and so these results for latitude and translation speed must be taken with caution.

The most counterintuitive result in this table is the similarity in bias with translation speed from 0–10 to 10–20 kt (−0.6 and −0.7 kt, respectively). This result strongly suggests that the model has a mechanism for incorporating the effects of translation speed within 0–20 kt. Whether this result persists in the (rather extreme) 20–30-kt range is not certain because of the small sample size (N = 15).

Finally, the groupings by image resolution are the most likely to be sampled fairly, as their average MSW values differ by less than 1 kt. The RMSE is larger when applied to lower-resolution images than when applied to higher-resolution images to a confidence of 90%. Moreover, the reduction in RMSE from 11.0 kt (lower resolution) to 9.6 kt (higher resolution) is a substantial improvement that approaches the accuracy of current state-of-the-art methods (section 6b). Given the significance of this result, it is reasonable to ask whether
the 37-GHz channel is less impactful in DeepMicroNet simply because of its lower resolution in most of the training dataset, or because of something more fundamental to what it depicts? The lower precision of the 37-GHz-based model in estimating MSW (wider PDFs, Fig. 4) at both high resolution and low resolution strongly suggests that the reason is fundamental to the 37-GHz channel. The reason for this is likely that the 37-GHz signal captures low-level features (warm rain and shallow convection), whereas the 89-GHz signal captures more of the deep convection associated with the eyewall and inner rainbands that relate more to MSW.

d. Results of models trained to forecast TC intensity

After optimizing the CNN model to estimate MSW at the time of the image, it is a simple task to retrain the model with best track MSW forward in time, so that the resulting model estimates the future MSW. When comparing to the future MSW from all best track cases, the same sample was used as previously, omitting only those with forecasts that were later than the latest best track point. When comparing to recon-aided best track MSW, the same sample was used as before, and was not adjusted for the differing availability of reconnaissance observations at the forecast time. This was done in order to keep the sample observations consistent in each version of the model, in spite of the trade-off of further uncertainty in the “truth” MSW value. As with the experiment in section 5a, we used three model runs for each forecast time and selected the best one to represent the results for that time. This insures a more authentic trend with time because it is less influenced by random error.

Results show a steady increase in model error at forecasting best track MSW beyond 6 h, and an abrupt increase in error at forecasting recon-aided MSW beyond 6 h (Fig. 16). However, in both cases the models had slightly more skill at +6 h than at the exact time of the satellite observation, which was also found with the statistical prediction scheme of Jiang et al. (2019) based on TRMM TMI imagery. This improvement in skill...
FIG. 13. (a) Heat map scatterplot of independent model-estimated MSW for 2007 and 2012 vs global best track; (b) as in (a), but limited to recon-observed times/locations.
at +6 h makes sense in light of the time required for the latent heat release from deep convection in a TC (evident in the 89-GHz imagery) to affect the gradient balance and surface pressure/wind.

The stark difference in error between best track and recon-aided cases beyond 6 h is not easily accounted for, but factors that could explain the difference include: a lower temporal variability in the full best track MSW dataset due to lack of reconnaissance, the higher occurrence of landfall shortly after reconnaissance (leading to higher MSW variability), and the higher average variability in MSW in the recon-aided dataset due to its higher average MSW. In other words, the best track MSW is a more homogeneous dataset for testing accuracy with forecast time, indicating that this trend is actually better represented with the best track than with the recon-aided dataset.

6. Discussion

a. Analysis of error

Perhaps counterintuitively, the DeepMicroNet model performs better against TC intensity data informed by reconnaissance (RMSE = 10.6 kt) than against the kind of best track data on which the model was directly trained (RMSE = 14.3 kt). This is in spite of the fact that the recon-aided best track has a higher proportion of category 1, 2, and 5 TCs, which have higher model error overall, and a lower proportion of TDs, which have lower error. This result is very encouraging in light of our understanding of the model training process. On one hand, the model is expected to incorporate the biases inherent in the best track training dataset. Indeed, Knaff et al. (2010) observe that the Dvorak technique, which has a heavy influence on the best track, exhibits a bias reaching as low as $-2$ to $-6$ kt where MSW = 30–80 kt. Sheets and McAdie (1988) and Koba et al. (1990, 1991) also note this tendency. Knaff et al. (2010) also note a high bias in the Dvorak technique of as much as 4 kt around MSW = 100 kt, and another low bias of $-8$ kt at MSW = 140 kt. All three of these trends are captured very closely by our own comparison to recon-aided best track intensity (Fig. 15b). The low-intensity bias in particular (around MSW = 30–80 kt) is often apparent in the microwave imagery as a system with little convective organization but nonetheless high surface winds, such as in Fig. 8 (top row). These biases could be mitigated in the model by training on recon-aided best track only, provided the sample size is large enough for a Deep Learning environment. Overcoming such a problem with sample size would be challenging, given that the current training dataset has sample size.
limitations of its own, and so it would require an entirely different model development approach.

On the other hand, the generally superior performance of the model against recon-aided MSW indicates that the model has overcome the effects of random error in the training dataset to a significant degree, leading to greater skill in the recon-aided comparison. This is especially true for category 2–4 intensities, where the bias is higher in Fig. 15 than in Fig. 14, yet the RMSE is lower. The relatively higher error for category 1 is understandable, because this intensity range has the highest variety of convective structures due to TCs organizing, breaking up or transitioning to extratropical systems, and so the model would be trained with less consistency in this range. Overall, the DL training process evidently reached a “middle ground” through some of the arbitrarily high and low estimates of the best track record as it worked to minimize the loss function [Eq. (1)]. Achieving further accuracy would probably require the use of images from a TC’s recent past to provide more context on its evolution. This general practice is already inherent in the IR-based Dvorak and ADT methods to estimate TC intensity from satellite data (Dvorak 1984; Olander and Velden 2007).

The bias and error for category 5 intensities involve more unique issues. Here the error with respect to the best track is much lower than the error with respect to the recon-aided best track, but as discussed, the recon-aided best track is generally more credible. However, the recon-aided best track for the category 5 range accounts for only two long-lived TCs—Dean (2007) and Felix (2007)—so the effects of undersampling are likely contributing. In the end it remains inconclusive whether the model error in the category 5 range is truly as high as Fig. 15 indicates. Furthermore, the model is trained on far fewer category 5 images (~500) than are normally used to define a classification type in a CNN model (>10000), so estimating MSW with better precision in the category 5 range using microwave imagery may still be possible, but is just not achieved with this model.

One would expect some model bias, particularly the average bias from Fig. 14b, to originate from interannual variability, bias from analysts over time, or differences in TC frequency by basin each year. The first two points are beyond the scope of this study, and to pursue these questions would require interfering with the independence of the training dataset. On the subject of differences due to basin frequency, we note we have found no significant improvement in performance when adding ancillary information of designated TC basin to the model. While such information would likely be necessary when training a model with

![Fig. 15](image-url)
Geo IR imagery because of varying interbasin tropopause temperatures and cloud cover size, this “negative” finding in our study suggests that microwave signatures are more homogeneous across TC basins. Specifically, the convective structures apparent in the microwave imagery may indicate a certain value of MSW regardless of the conditions in the larger environment.

b. Comparison to other satellite estimation methods

The DeepMicroNet model is competitive with most other automated satellite-based intensity estimation methods, some of which are listed in Table 7. Our initial results are on par with the Dvorak technique (Velden et al. 2006; Knaff et al. 2010), which has been the longstanding benchmark for operational satellite-based intensity. However, DeepMicroNet does not do as well as most other algorithms in the category 5 range, for reasons previously discussed. The current leading method for objective, computer-based methods is SATCON, which is a weighted consensus algorithm that employs estimates from several different satellite sources (Herndon and Velden 2018) and is now being used operationally at tropical cyclone analysis centers worldwide. The 37- and 89-GHz imager channels do not currently contribute to SATCON directly, so these results show that DeepMicroNet could provide a substantial contribution as an additional SATCON member in the future. Also, any lack of skill in DeepMicroNet above category-4 intensity could be balanced out by other information.

Several features of DeepMicroNet come more easily with DL than with existing approaches, such as probabilistic estimation and the ability to operate from partial scans. Olander and Velden (2007) state that one of the largest sources of error in the ADT is uncertain TC center-fixing, and it is a positive sign that this problem is ameliorated in DeepMicroNet.

The 7-bin CNN model of Pradhan et al. (2018), trained on Geo-IR imagery, would be a natural source for comparison to DeepMicroNet, but it is difficult to relate the results because their output is limited to TC category only (TD, TS, category 1, etc.) rather than MSW. It estimates TC category with an accuracy of 81% and a “Top-2” accuracy of 95%. Since TC categories are...
spaced by 13–30 kt, a model such as DeepMicroNet with an RMS error of 10.6 kt would score at a similar level of skill, though with higher precision. Further work would be needed to show whether this difference in precision also affects the inherent accuracy of the methods.

The recent DL regression model of Chen et al. (2018) is also noteworthy here. They report an RMSE of 9.45 kt compared to best track values for the west Pacific, east Pacific, and North Atlantic basins. However, the model requires the Climate Prediction Center Morphing (CMORPH, Joyce et al. 2004) rain rate imagery as input, which is interpolated from microwave observations from the past and future (CMORPH has a 24-h latency). Because of this latency and use of future data it is unfortunately not a near-real-time estimation like those in Table 7. Regardless, it shows great promise in the combination of passive microwave imagery with other modes of satellite observation.

7. Conclusions

By following a fairly standard recipe for CNN modeling, this project has produced operational-quality TC intensity estimates using satellite bands that have traditionally not played a major role in quantitative techniques. The 89-GHz band in particular is shown to be a first-order tool for intensity estimation—both in real time and in forecasting to at least 6 h. The model’s ability to find trends finer than the random error of the best track and replicate the known best track biases shows that further improvements are still possible. Additionally, the low errors of this neural network technique present a target for future analytical techniques to further explain the precise mechanisms behind the 37- and 89-GHz bands’ relationship to intensity.

The technique has major implications for estimating other aspects of TCs as well (e.g., radius of maximum winds, rapid intensification, eyewall replacement cycles) because the model design, and capabilities, would be nearly the same for each of these applications. For some tasks, only the training data and output classifications would need to change. Furthermore, there is no reason to believe that this level of performance is unique to 37- and 89-GHz band imagery, as others have already demonstrated a CNN application with Geo IR (Pradhan et al. 2018; Chen et al. 2018). Even higher model accuracy and increased capabilities are all but certain when more data sources, from satellite imagery to ancillary

### Table 7. Comparable statistics of other satellite-based TC intensity estimation methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Data source</th>
<th>Bias (kt)</th>
<th>RMSE (kt) (all TCs)</th>
<th>RMSE (kt) (categories 3–5)</th>
<th>References a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dvorak technique</td>
<td>Geo IR, Vis</td>
<td>2 to −8</td>
<td>10.5</td>
<td>12</td>
<td>Knaff et al. (2010)</td>
</tr>
<tr>
<td>Bias-corrected Dvorak technique</td>
<td>Geo IR, Vis</td>
<td>−</td>
<td>9.75</td>
<td>−</td>
<td>Knaff et al. (2010)</td>
</tr>
<tr>
<td>Deviation angle variance</td>
<td>Geo IR</td>
<td>−</td>
<td>14–15</td>
<td>&gt;15</td>
<td>Piñeros et al. (2008), Ritchie et al. (2012, 2014)</td>
</tr>
<tr>
<td>Advanced Dvorak technique (ADT)</td>
<td>Geo IR</td>
<td>−0</td>
<td>11</td>
<td>13.3</td>
<td>Olander and Velden (2018) b</td>
</tr>
<tr>
<td>Warm core anomaly (CIRA)</td>
<td>AMSU, ATMS</td>
<td>7 to −20</td>
<td>14</td>
<td>23</td>
<td>Demuth et al. (2004, 2006)</td>
</tr>
<tr>
<td>Warm core anomaly (CIMSS)</td>
<td>AMSU, SSMIS, ATMS</td>
<td>−2</td>
<td>12</td>
<td>13–14</td>
<td>Herndon and Velden (2018) b</td>
</tr>
<tr>
<td>Consensus approach (SATCON)</td>
<td>ADT, AMSU, SSMIS, ATMS</td>
<td>−0</td>
<td>8.9</td>
<td>9.2</td>
<td>Herndon and Velden (2018) b</td>
</tr>
<tr>
<td>Passive microwave</td>
<td>TRMM TMI</td>
<td>−</td>
<td>12.3</td>
<td>−</td>
<td>Jiang et al. (2019)</td>
</tr>
<tr>
<td>Feature-based decision tree</td>
<td>SSMI, SSMIS, TMI, AMSR-E</td>
<td>−</td>
<td>11.9</td>
<td>−</td>
<td>Bankert and Cossuth (2016) b</td>
</tr>
</tbody>
</table>

a Where not explicitly stated in the reference paper, we approximate the values from tables and figures.

b Nonrefereed source.
data, are brought into the system. More advanced methods of DL beyond CNN models are also likely to increase performance, and perhaps provide the next generation of SATCON-like multisensor integration.

Finally, it is worth returning to a common criticism of DL, which is that the methods of probing a model’s neural network still prevent one from directly learning enough about the physical meanings behind its behavior. While this makes it difficult to use innovations in DL to advance the science in more explicit areas such as parameterization of numerical weather prediction methods, recent research is beginning to make DL models less opaque and make more intuitive use of their inner workings (Olah et al. 2018), and DL continues to develop rapidly in many directions. Thus, future progress with DL applications in meteorology will likely advance in tandem with DL’s growing potential for scientific yield.

Acknowledgments. We thank Naval Research Lab Contract N00173-17-1-G007 for assisting in support of this project. We also thank John Knaff, Haiyan Jiang, and one anonymous reviewer for their suggestions that greatly improved the manuscript.

REFERENCES


Gagne, D.-J., H. Chrisensen, A. Subramanian, and A. H. Monahan, 2019: Evaluating generative adversarial network stochastic parameterizations of the Lorenz ’96 model at weather and climate time scales. 18th Conf. on Artificial Computation and Mecha


——, and ——, 2018: The UW-CIMSS advanced Dvorak technique (ADT)—Current status and future upgrades. 33rd Conf. on Hurricanes and Tropical Meteorology, Ponte Verdi,