MistNet: Measuring historical bird migration in the US using archived weather radar data and convolutional neural networks

Tsung-Yu Lin[1], Kevin Winner[1], Garrett Bernstein[1], Abhay Mittal[1], Adriaan M. Dokter[2], Kyle G. Horton[2], Cecilia Nilsson[2], Benjamin M. Van Doren[3], Andrew Farnsworth[2], Frank A. La Sorte[2], Subhransu Maji[1], and Daniel Sheldon [a,1,4]

[1]College of Information and Computer Sciences, University of Massachusetts Amherst, 140 Governors Drive, Amherst, MA 01003, USA
[2]Cornell Lab of Ornithology, Cornell University, Ithaca, NY 14850, USA
[3]Edward Grey Institute, Department of Zoology, University of Oxford, Oxford, OX1 3PS, United Kingdom
[4]Department of Computer Science, Mount Holyoke College, 50 College Street, South Hadley, MA 01075, USA

Running title: MistNet: Measuring historical bird migration

Article Type: Research article

Words in the abstract: 291	**Number of tables:** 6 + 3 supplemental
Words in the main text: 6741	**Number of text boxes:** 0
Number of references: 77	**Supplementary material:** 3 appendices
Number of figures: 4 + 4 supplemental	

Author contributions: DS, SM, FLS, and AF conceived the study. TYL, KW, GB, AM, SM, and DS designed methodology. KGH, AD, and CN collected data. TYL, KW, GB, AD, KGH, CN, and BVD analyzed data. DS, SM, TYL, CN, and KGH wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data accessibility statement: The MistNet model, source code, and evaluation data will be added to the publicly available WSRLIB software package on github (`https://github.com/darkecology/wsrlib`) prior to final publication.

[a] **Corresponding Author**: Tel: +1 413-545-4843; Email: sheldon@cs.umass.edu

Abstract

1. Large networks of weather radars are comprehensive instruments for studying bird migration. For example, the US WSR-88D network covers the entire continental US and has archived data since the 1990s. The data can quantify both broad and fine-scale bird movements to address a range of migration ecology questions. However, the problem of automatically discriminating precipitation from biology has significantly limited the ability to conduct biological analyses with historical radar data.

2. We develop MistNet, a deep convolutional neural network to discriminate precipitation from biology in radar scans. Unlike prior machine learning approaches, MistNet makes fine-scaled predictions and can collect biological information from radar scans that also contain precipitation. MistNet is based on neural networks for images, and includes several architecture components tailored to the unique characteristics of radar data. To avoid a massive human labeling effort, we train MistNet using abundant noisy labels obtained from dual polarization radar data.

3. In historical and contemporary WSR-88D data, MistNet identifies at least 95.9% of all biomass with a false discovery rate of 1.3%. Dual polarization training data and our radar-specific architecture components are effective. By retaining biomass that co-occurs with precipitation in a single radar scan, MistNet retains 15% more biomass than traditional whole-scan approaches to screening. MistNet is fully automated and can be applied to data sets of millions of radar scans to produce fine-grained predictions that enable a range of applications, from continent-scale mapping to local analysis of airspace usage.

4. Radar ornithology is advancing rapidly and leading to significant discoveries about continent-scale patterns of bird movements. General-purpose and empirically validated methods to quantify biological signals in radar data are essential to the future development of this field. MistNet can enable large-scale, long-term, and

reproducible measurements of whole migration systems.

Keywords: weather radar, aeroecology, ornithology, machine learning, deep learning, convolutional neural networks

1 Introduction

Researchers discovered more than 70 years ago that radars, originally designed for military purposes, can also detect bird movements (Brooks, 1945; Lack & Varley, 1945). As radar technology developed, ornithologists used radars to document and measure previously difficult-to-observe aspects of bird movements, such as flight patterns and behaviors at high altitudes, at night, and over the sea (Harper, 1958; Casement, 1966; Eastwood, 1967). With the advent of large networks of weather radars, the possibility arose to use radar as a distributed instrument to quantify whole migration systems (Gauthreaux, 1970; Bruderer, 1997; Gauthreaux & Belser, 1998; Gauthreaux *et al.*, 2003; Dokter *et al.*, 2011; Bauer *et al.*, 2017; Nilsson *et al.*, 2018b).

The US WSR-88D[1] weather radar network (Crum & Alberty, 1993) stands out as one of the most comprehensive instruments for studying migration due its size, uniformity, and historical data archive. Installation began in the 1990s and the network currently includes 159 radars with nearly complete coverage of the continental US. Each radar scans its surroundings every 6 to 10 minutes. The radars and data collection are standardized, and essentially all of the data—over 200 million individual files—has been archived over more than 25 years (Ansari *et al.*, 2018). It is well known that these radars regularly detect birds and can be used to quantify their movements (Gauthreaux & Belser, 1998). The result is an unparalleled historical record of bird migration.

Weather radar data can answer a wide range of important migration ecology questions. Previous studies have used weather radar data to understand patterns and determinants of nocturnal migration (Gauthreaux *et al.*, 2003; Kemp *et al.*, 2013; La Sorte *et al.*, 2015a; Farnsworth *et al.*, 2016), identify critical stopover habitat (Buler & Diehl, 2009; Buler & Dawson, 2014), locate on-the-ground roosting sites of birds (Winkler, 2006; Buler *et al.*, 2012; Laughlin *et al.*, 2013, 2016; Bridge *et al.*, 2016), understand flyways (Horton *et al.*, 2018; Nilsson *et al.*, 2018b) and flight behavior (Dokter *et al.*, 2013; Horton *et al.*, 2016; La Sorte *et al.*, 2015b), quantify demography (Dokter *et al.*, 2018b), doc-

[1] **W**eather **S**urveillance **R**adar, 19**88**, **D**oppler

ument the effects of artificial light (Van Doren *et al.*, 2017; McLaren *et al.*, 2018) and disturbance (Shamoun-Baranes *et al.*, 2011) on migration, explore the projected implications of climate change (La Sorte *et al.*, 2019), and forecast migration at continent scales (Van Doren & Horton, 2018). Researchers worldwide recognize the potential of radar data to provide new and urgently needed information about the migration ecology of birds, bats, and insects, including detailed information about: routes, phenology, and mechanisms of migration; ecosystem services; the impacts of human activities and climate change on migration systems; conservation prioritization; aviation safety; and agricultural pests (Kelly & Horton, 2016; Bauer *et al.*, 2017, 2018).

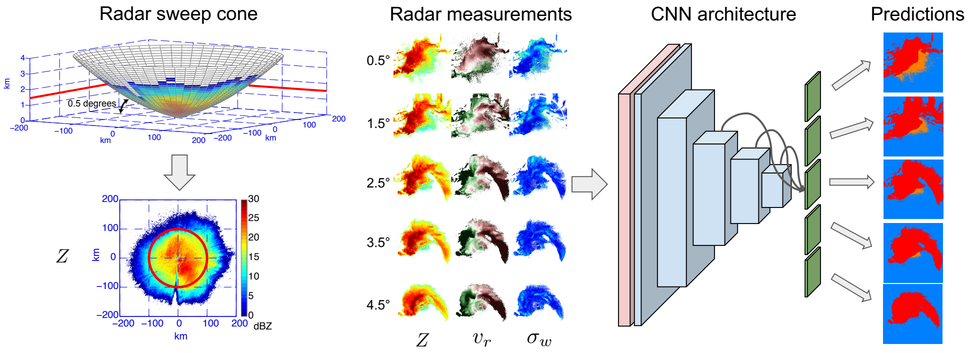
Significant methodological challenges have slowed the full and widespread use of weather radar data as a biological instrument. Early WSR-88D studies demonstrated how to detect and quantify bird movements but required substantial manual effort, primarily to screen radar images for precipitation and other unwanted targets prior to analysis (Gauthreaux & Belser, 1998; Gauthreaux *et al.*, 2003). Human interpretation of images has persisted into most modern analyses (Buler & Diehl, 2009; Buler *et al.*, 2012; Buler & Dawson, 2014; Farnsworth *et al.*, 2016; Van Doren *et al.*, 2017; Horton *et al.*, 2018; McLaren *et al.*, 2018) and is a substantial barrier to very large-scale research with WSR-88D data, for example, the complete analysis of 200 million historical data files.

Recent advances have led to the first fully automated methods to extract biological information from weather radar data. In 2012–2013, the WSR-88D network was upgraded to dual polarization technology, which makes it significantly easier to separate biology from precipitation in modern data (Stepanian *et al.*, 2016), but leaves open the problem of extracting biological information from historical data. Dokter *et al.* (2011) developed an algorithm to separate precipitation from biology in European C-band radars; this was later extended to US dual polarization and S-band data (Dokter *et al.*, 2018b,a), but currently cannot fully separate precipitation from biology in historical US data. Roy-Chowdhury *et al.* (2016), Van Doren & Horton (2018) and Horton *et al.* (2019) trained machine learning models to automatically identify radar scans that are contaminated with precipitation. These methods are useful in that they allow automated analysis prior to the dual polarization upgrade in 2012–2013, but discard all biology in scans where precipitation occurs; our results will show this to be about 19% of the total biomass. Whole-scan classifiers are also inflexible: they are tailored to a specific spatial extent (e.g., rain within 37.5 km of the radar), and would require substantial additional labeling effort and model training to adapt to a slightly different analysis (e.g., rain within 150 km of the radar).

In this paper we develop MistNet, a deep convolutional neural network (CNN) to separate precipitation from biology at a fine spatial resolution in historical WSR-88D data. MistNet has a false discovery rate of at most 1.3% and retains 15% more of the total biomass than whole-scan classification. Radar images contain clear visual patterns that allow humans to discriminate precipitation from biology. Deep learning has revolutionized the ability of computers to mimic humans in solving similar recognition tasks for images, video and audio (Krizhevsky *et al.*, 2012; Graves *et al.*, 2013; Simonyan & Zisserman, 2014). MistNet is based on models for images, but includes several innovations that are specifically tailored to weather radar data. To avoid the cost of collecting a massive human-labeled data set, we use "weak" labels from dual polarization data. We develop a novel "adapter" architecture to handle the large number of input channels in radar data compared to RGB images, and the need to predict at different elevations. We conduct a large-scale empirical validation of MistNet and competing approaches on two evaluation data sets. MistNet makes fine-grained predictions and can be used within radar ornithology workflows to address a range of biological questions at different scales. We present several case studies to illustrate the flexibility of the approach.

2 Materials and Methods

Our goal was to develop a system to discriminate biology from weather in radar data (Figure 1). Convolutional neural networks (CNNs) have achieved outstanding performance on recognition tasks in related domains such as image classification (Krizhevsky *et al.*,



(a) Radar geometry

(b) Overview of the processing pipeline

Figure 1: **(a) Radar geometry.** A sweep, shown here at an elevation of 0.5 degrees, traces out an approximately conical surface and is usually rendered as top-down image or "plan-position indicator" (i.e., PPI). **(b) Overview of processing pipeline.** Radar measurements are collected on a three-dimensional polar grid (3 products x 5 elevations) and rendered as a 15-channel "image" in Cartesian coordinates. An adapter network maps 15 channels to 3 channels to match a conventional RGB image. The CNN processes the image and outputs five segmentation masks, one for each elevation. Each segmentation mask delimits areas containing biology and weather (red: rain, orange: biology, blue: background). The inputs, intermediate activations, and outputs of the CNN are three-dimensional arrays arranged in layers and depicted as boxes (pink: input, light blue: intermediate, green: output; see also Section 2.1). Activations at each layer are a function of those at the preceding layer. The activations in output branches (green boxes) are functions of several earlier layers, shown for one branch with black curved arrows.

2012), face recognition (Taigman *et al.*, 2014), speech recognition (Graves *et al.*, 2013; Hinton *et al.*, 2012), and video understanding (Simonyan & Zisserman, 2014; Tran *et al.*, 2015)), and are therefore an excellent candidate for this task. However, we faced several radar-specific challenges.

First, most existing CNNs are designed for three-channel color images (RGB) with pixels arranged in a Cartesian grid. Unlike images, weather radar data is collected on a three-dimensional polar grid with many channels, so there were a number of unresolved questions about how to represent radar data and design CNNs for this task.

A second challenge was the availability of training data. CNNs require a large number of labeled examples to train due to the vast number of parameters. For image classification, data sets of more than a million labeled images are routine (Krizhevsky *et al.*, 2012). We wish to make pixel-level predictions to segment areas of precipitation and biology. Curating enough high-quality pixel-level annotations to train a segmentation network would be costly and impractical. We therefore investigated two alternatives. First, we collected noisy training labels automatically using dual polarization radar products. Second, we adopted a common transfer learning technique: instead of training a CNN "from scratch" with randomly initialized parameters, we started with CNNs trained for image recognition tasks and then updated the parameters using training labels for the radar task. This allows a model to learn faster with fewer labels (Razavian *et al.*, 2014).

Note that the issues of architecture and training are intertwined. To take advantage of high-quality pre-trained CNNs from image recognition tasks, our architecture must render radar data as three-channel images.

2.1 Radar Data and CNN Preliminaries

Radar Data The US National Weather Service operates the WSR-88D (Weather Surveillance Radar-1988 Doppler; also called NEXRAD) network of radars (Crum & Alberty, 1993; Doviak & Zrnić, 1993). The network currently includes 143 radars in the contiguous US and an additional 16 radars in Alaska, Hawaii, and other US territories and military installations.

Scanning strategy and geometry. Radars in the WSR-88D network conduct volume scans to sample the surrounding airspace. Each volume scan (hereafter: scan) takes from four to ten minutes. During one scan, the radar conducts a sequence of 360-degree sweeps where it rotates its antenna around a vertical axis with fixed elevation angle to sample a cone-shaped slice of the airspace (Figure 1a top); conventional radar images are top-down views of these sweeps (Figure 1a bottom). A typical scanning strategy during clear-air conditions includes five sweeps at elevation angles from 0.5 to 4.5 degrees. From each sweep come a set of gridded data products summarizing the radar signal returns within discrete sample volumes, which are the portions of the atmosphere sensed at a particular antenna position and range from the radar (Doviak & Zrnić, 1993).

Data products and dual polarization. WSR-88D radars collect six data products.

Of these, three "legacy" products have been collected since the installation of the system in the early 1990s, and three dual polarization or "dual-pol" products became available when the system was upgraded during the period from 2011 to 2013 (Stepanian, 2015; Stepanian *et al.*, 2016). The legacy data products are *reflectivity factor* (Z), *radial velocity* (v_r), and *spectrum width* (σ_w). Reflectivity factor is related to the density of objects in the atmosphere and their radar cross sections (which are related to their sizes); radial velocity and spectrum width are the reflectivity-weighted mean and standard deviation, respectively, of the radial velocity, computed from the Doppler spectrum of the returned radio waves, and provide information about the velocity of scatterers. Dual-pol radars emit and detect radio waves both in vertical and horizontal polarizations (Stepanian *et al.*, 2016). The relationship between backscatter in the two polarizations provides information about the object shape (height-to-width ratio) and uniformity within a pulse volume, which help discriminate different types of objects (rain, birds, etc.) (Zrnić & Ryzhkov, 1998; Stepanian *et al.*, 2016; Dokter *et al.*, 2018a). The dual polarization products are *differential reflectivity* (Z_{DR}), *differential phase* (ψ_{DP}), and *correlation coefficient* (ρ_{HV}).

Resolution and Rendering. A WSR-88D data product is stored as a collection of sweeps. Each sweep is a two-dimensional data array corresponding to a polar grid indexed by range r and azimuth ϕ. Data prior to 2008 had "legacy" resolution of $1000\,\mathrm{m} \times 1°$; during 2008 the radars were upgraded to "super-resolution" of $250\,\mathrm{m} \times 0.5°$. To standardize data for analysis, we aligned all products to a fixed three-dimensional grid using nearest neighbor interpolation (Parker *et al.*, 1983). We used a super-resolution grid ($250\,\mathrm{m} \times 0.5°$) with third dimension corresponding to the five elevation angles 0.5°, 1.5°, 2.5°, 3.5° and 4.5°. Higher sweeps, which are only available in certain operating modes, were discarded. We then resampled each sweep onto a Cartesian grid with resolution of 500 m and radius of 150 km using nearest neighbor interpolation, resulting in a 600×600 grid centered at the radar station, where the x- and y- dimensions correspond to distance along the earth's surface in the east-west and north-south directions. The result is a set of aligned 600×600 arrays (Figure 1), one for each product and elevation. We used the same units as the original data files: in particular, reflectivity factor used a decibel scale (dBZ). “NODATA” values were replaced by numeric defaults.

CNNs A deep neural network transforms an input array (e.g., an image) into one or more output values through a sequence of linear and nonlinear transformations. The computation is arranged into L layers, where $\mathbf{z}^{(0)}$ is the input array, and, for each layer ℓ, the network computes an array of values $\mathbf{z}^{(\ell)}$—termed the “activations” at layer ℓ—as a function of the activations of previous layers. The output array is $\mathbf{z}^{(L)}$. In networks that we will consider, each $\mathbf{z}^{(\ell)}$ is a three-dimensional array of dimension $c_\ell \times m_\ell \times n_\ell$ conceptualized as an image with c_ℓ channels and size or “spatial dimension” $m_\ell \times n_\ell$. These are illustrated as colored boxes in Figure 1 (pink for the input image, and light blue for intermediate activations). The typical operations used to compute the activations at a single layer in a CNN from its predecessors involve convolutions, downsampling, elementwise nonlinear transformation, pooling, and fully connected layers. A convolution is a linear operation that slides a filter across each position of the input image and produces one output value for each image location. This is done simultaneously with many filters to produce multi-channel output. With appropriate filters, convolutions can implement a wide range of basic image operations including smoothing, Fourier analysis and other changes of basis, edge extraction, texture extraction, and template matching (Forsyth & Ponce, 2003). It can be combined with downsampling to produce an output image of smaller size. A linear operation such as convolution is typically followed by an elementwise nonlinear transformation or “nonlinearity” such as the ReLU nonlinearity, which transforms each value as $z' = \max(0, z)$. Pooling reduces the spatial dimension by aggregating over small blocks of the image. A fully connected layer is one where each value is a linear function of *all* values in the previous layer followed by a nonlinearity. Each convolutional layer and fully connected layer has weights controlling the linear transformations; these are the parameters to be learned. See the book of Goodfellow *et al.* (2016) for more background on CNNs.

In MistNet, the input $\mathbf{x} = \mathbf{z}^{(0)}$ has dimension $15 \times 600 \times 600$ and the output $\mathbf{z}^{(L)}$

has dimension $3 \times 5 \times 600 \times 600$, which corresponds to the class probability for each of 3 classes (precipitation, biology, background) at each position in five 600×600 images, one for each elevation angle. At prediction time, the class with the highest probability is predicted. Let $f(\mathbf{x}; \boldsymbol{\theta})$ be the function describing the entire mapping from the CNN's input to its output, so that $\mathbf{z}^{(L)} = f(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ contains the parameters of all layers. During learning, the loss function $L(f(\mathbf{x}; \boldsymbol{\theta}), \mathbf{y})$ is computed to calculate the disagreement between the network output and the true class labels $\mathbf{y}$ (which have dimension $5 \times 600 \times 600$). We train all models using the cross-entropy loss function and stochastic gradient descent (SGD, Goodfellow *et al.*, 2016), which adjusts $\boldsymbol{\theta}$ in the direction $-\sum_{i \in B} \frac{\partial}{\partial \boldsymbol{\theta}} L(f(\mathbf{x}_i; \boldsymbol{\theta}), \mathbf{y}_i)$. Here, i indexes training examples and B is set of indices corresponding to the batch of examples used for one update. Gradients are computed by backpropagation (Rumelhart *et al.*, 1986).

2.2 Training and Evaluation

Weak Training Labels Because we do not have a large data set of radar images with pixel-level labels, we conducted transfer learning from image classification models trained on the ImageNet dataset (Deng *et al.*, 2009). We initialized MistNet's model parameters using those models, and then adapted the parameters by training with weak annotations obtained from dual-pol data.

There are several simple rules to discriminate precipitation from biology with reasonable accuracy using dual-pol products. Biological scatterers tend to have a much lower correlation coefficient than hydrometeors because their orientation, position, and shape are much more variable in time (Stepanian *et al.*, 2016; Kilambi *et al.*, 2018). It has become common practice among radar biology practitioners to use a threshold of $\rho_{\mathrm{HV}} \leq 0.95$ to identify biological scatterers (Dokter *et al.*, 2018a). Although weather events such as mixed precipitation can also produce ρ_{HV} values this low (Lim *et al.*, 2005), this rule is believed to have reasonable accuracy in general, and has been validated through comparisons with a colocated bird radar (Dokter *et al.*, 2011; Nilsson *et al.*, 2018a). Little is known about the best threshold value or pixel-level accuracy of this method.

Recently, Kilambi *et al.* (2018) proposed a refined thresholding rule using the *depolarization ratio* (in decibel units)

$$\mathrm{DR} = 10\log_{10}\left(\frac{Z_{\mathrm{DR}} + 1 - 2Z_{\mathrm{DR}}^{1/2}\rho_{HV}}{Z_{\mathrm{DR}} + 1 + 2Z_{\mathrm{DR}}^{1/2}\rho_{\mathrm{HV}}}\right), \tag{1}$$

which is a proxy for the *circular depolarization ratio* (Ryzhkov *et al.*, 2017), a quantity that is useful for discriminating meteorological targets but is not measured directly by WSR-88D radars. Kilambi *et al.* (2018) showed that meteorological targets have smaller DR values and suggested a classification threshold of DR$= -12\,\mathrm{dB}$ based on a quantitative evaluation of 32 volume scans.

We evaluated the performance of a range of threshold rules for both ρ_{HV} and DR on a large set of manually labeled evaluation scans (described below). Both ρ_{HV}-thresholding and DR-thresholding performed well, with DR-thresholding being slightly more accurate (Section 3). MistNet was developed prior to the publication of Kilambi *et al.* (2018) and uses ρ_{HV}-thresholding for training; replacing this with DR-thresholding is a possible avenue for future improvement.

Finally, we used the following rules to generate training labels for MistNet. For each pixel, if the reflectivity factor Z is reported as "no data" (below signal-to-noise threshold) then we set the label to "background". Otherwise, if $\rho_{\mathrm{HV}} > 0.95$ we set the label to "precipitation". All remaining labels are "biology".[2] We included the background class during training to avoid semantic confusion resulting from forcing the model to predict background pixels as either weather or biology. At prediction time, it is known whether or not a pixel belongs to the background class, and predictions are only made on non-background pixels.

[2] Note that the biology class includes all non-hydrometeor scatterers, including insects, dust, debris, etc.; our goal is to eliminate precipitation, not to make fine-grained distinctions among non-hydrometeor scatterers. This class will also include pixels containing hydrometeor scatterers that happen to have low ρ_{HV} values, such as mixed precipitation.

Training Set We downloaded radar scans for training from Amazon Web Services (Ansari *et al.*, 2018)[3]. The scans were selected from all 161 radar stations[4] in spring (April and May) and fall (September and October) from 2014 through 2016. For each station, we sampled scans at 30-minute intervals within a 3-hour period starting at local sunset, resulting in a training set of 239 128 scans.

Evaluation data We collected two separate evaluation data sets of human-labeled ground truth data: a geographically representative *contemporary* set, and a historically representative *historical* set. Data was labeled using a slight modification of a web-based tool designed for interactive image segmentation (Tangseng *et al.*, 2017).[5] We used the tool to delineate areas of precipitation out to a ground range of 150 km in selected sweeps.

The contemporary set includes data from 16 geographically representative stations for two one-month periods during spring (15 April to 15 May) and fall (15 September to 15 October) of 2017. Thirteen stations were selected using a stratified random design and three additional stations were selected manually; details and a list of stations are given in Appendix A. On each day and for each station, the scan closest to three hours after local sunset was selected to approximate the time of peak nocturnal migration (e.g., Farnsworth *et al.*, 2016; Horton *et al.*, 2015). This resulted in a total of 971 scans. For each scan, we labeled the lowest elevation sweep and one randomly selected higher sweep from the lowest five elevation angles. We later discarded some labeled sweeps due to changes in our rendering process. In each volume scan, only the sweeps closest to one of the desired elevations (0.5°, 1.5°, 2.5°, 3.5° and 4.5°) were retained; some labeled sweeps between these elevation angles were discarded. The final number of labeled sweeps at each elevation was 971 (0.5°), 254 (1.5°), 235 (2.5°), 167 (3.5°), and 173 (4.5°).

The historical set includes data from stations KMOB (Mobile, AL) and KBGM (Binghamton, NY) from 1995 to 2017. These scans are drawn from an existing data set of manually screened scans (Van Doren & Horton, 2018). Scans were selected from a

[3] `https://s3.amazonaws.com/noaa-nexrad-level2/index.html`

[4] This is the total number of stations reporting data during that time, including those outside the contiguous US.

[5] `https://github.com/kyamagu/js-segment-annotator`

2.5-hour period centered on three hours after local sunset on March 15th, April 15th, May 15th, September 1st, October 1st, and November 1st for each station, resulting in 4891 scans (spring, 2549; fall, 2342), and then manually classified as either "clear" or "weather" based on the presence or absence of precipitation within 37.5 km of the radar station. We used all clear scans for computing pixel-level performance, by assuming all non-background pixels with 37.5 km of the radar belonged to the "biology" class. For weather scans, we randomly selected 50 scans per station for spring and fall to manually segment using our web-based tool so we could compute pixel-level performance metrics (200 scans × 5 elevations = 1000 fully segmented images).

All evaluations were performed for the region within 37.5 km ground range from the radar. To measure pixel-level performance we first computed a confusion matrix in which each pixel was weighted by reflectivity factor on a linear scale ($\mathrm{mm}^6\,\mathrm{mm}^{-3}$) after first capping values at 35 dBZ to limit the effect of extremely high values that are typically discarded in biological analyses. For the historical data set, clear and weather scans were subsampled from the original sample at different rates (see above); when computing the confusion matrix, we weighted the contribution of each type of scan to be proportional to its representation in the original sample (29.7 % weather for KBGM and 37.2 % weather for KMOB). Entries in the confusion matrix correspond to the fraction of total reflectivity—which represents the total biomass—classified a certain way. From this matrix we computed the standard metrics of *precision* (fraction of predicted biology that is actually biology), *recall* (fraction of true biology that is predicted to be biology), and *F-score* (harmonic mean of precision and recall). Precision is equal to one minus the *false discovery rate*. Recall is the same as *sensitivity*.

2.3 CNN Experiments

We performed several iterations of preliminary experiments, which suggested that the following elements would be important to MistNet: (1) using deep convolutional networks, (2) leveraging models pre-trained on ImageNet, (3) using data from all 15 modalities (5 elevations × 3 legacy products) to make predictions at each elevation, and (4) large training data sets assembled using dual-pol labels.

This led to the following design for MistNet. It is based on the FCN8 (fully convolutional network with predictions at a spatial granularity of 8 pixels) architecture from (Long *et al.*, 2015) with an ImageNet pre-trained VGG-16 "backbone" (Simonyan & Zisserman, 2015). See also Figure 1. We added a linear adapter network to map the input data from 15 to 3 channels at each spatial location for compatibility with the input dimensions of the VGG-16 network, and trained the parameters of the linear adapter. Unlike the standard FCN8 network, which predicts one value per spatial location, MistNet makes five predictions, one per elevation. This is accomplished by creating five separate branches that take as input the activations of several preceding convolutional layers and output class probabilities (the curved arrows in Figure 1 represent one of these branches).

MistNet was trained in MATLAB using MatConvNet (Vedaldi & Lenc, 2015). All parameters except those for the adapter and prediction branches were initialized from a VGG-16 architecture (Simonyan & Zisserman, 2015) pre-trained on ImageNet. The parameters of the adapter and prediction branches were initialized randomly and the entire model was trained end-to-end with stochastic gradient descent using the full data set of 239 128 scans. We augmented the training data by including a second version of each scan that was downsampled to legacy resolution prior to rendering in Cartesian coordinates, as a means to improve the ability of MistNet to generalize to older data, since all training data comes from 2014 and later. Additional details of the MistNet architecture and training procedure are provided in Appendix B.

We conducted a range of experiments to examine the benefits of different MistNet design choices:

- *Deep vs. shallow architectures.* We compared MistNet, a deep model, to two "shallow" baselines, which are both two-layer convolutional networks. The first has filter size 1 × 1, which means that predictions for each spatial location depend only on the radar measurements at that location. The second has filter size 5 × 5, which makes predictions at each location using all data from a 5 × 5 window centered at the location.

- *Predicting at 5 elevations using 15 channels.* We compared MistNet to two baseline approaches that use a standard FCN8 architecture for RGB images to make predictions at five elevations. Both baselines make predictions separately at each elevation using three selected input channels. The first baseline uses the Z, v_r, and σ_w products for the target elevation as the three input channels. This method lacks access to information from other elevations, which can be highly discriminative, since rain typically spans multiple sweeps while biology is concentrated at the lowest sweeps. The second baseline uses reflectivity from the target elevation and the two closest elevations as its three input channels; it gains access to information from adjacent sweeps but loses access to information from the v_r and σ_w products.

- *Size of training set.* We compared models trained on data sets of consisting of 100, 1000, 10 000 and 100 000 scans.

- *Post-processing predictions.* The standard prediction rule is to classify a pixel as precipitation if the predicted class probability for precipitation exceeds 0.5. In preliminary experiments we observed that MistNet underpredicted precipitation at the boundaries of rain storms and sometimes missed rain mixed with biology at low elevations. We developed the following postprocessing rules to improve these cases: we predict a pixel as rain if the class probability for rain exceeds 0.45 *or* if the average class probability for rain across the five elevations at that spatial location exceeds 0.45. We further compute a "fringe" of 8 pixels surrounding any rain pixel and classify the fringe as rain, with the goal of conservatively removing as much rain as possible due to its possible adverse impacts of biological analysis.

- *Pre-training.* We compared models trained with parameters initialized from ImageNet models to ones trained from randomly initialized parameters.

- *Low-resolution rendering.* We trained models with and without augmentation by low-resolution rendering.

2.4 Comparison with Whole-Scan Classification

MistNet segments radar scans, which allows for pixel-level screening of weather. Most previous biological analyses of historical weather radar data use *scan-level* screening. A scan is accepted and used in the analysis if it is free from precipitation and clutter, otherwise it is rejected. The screening step is conducted either by a human (Buler & Diehl, 2009; Buler *et al.*, 2012; Buler & Dawson, 2014; Farnsworth *et al.*, 2016; Van Doren *et al.*, 2017; Horton *et al.*, 2018; McLaren *et al.*, 2018) or using a machine learning classifier (Roy-Chowdhury *et al.*, 2016; Van Doren & Horton, 2018; Horton *et al.*, 2019). However, even a perfect whole-scan classifier will miss biology that co-occurs with precipitation. We compared MistNet to whole-scan classification by computing the implied pixel-level performance of whole-scan classification: all pixels in a weather scan were considered as precipitation and removed from analysis, and all pixels in a clear scan were retained. Instead of comparing to an automated classifier, we compared the performance of MistNet to an "oracle" whole-scan classifier that uses the human whole-scan labels (weather or clear). This is considered an upper bound on the performance of an automated whole-scan classifier.

We also compared MistNet to (oracle) whole-scan classification on an end-to-end performance measure. We computed vertical profiles of reflectivity (VPRs) using three different methods to exclude precipitation: (1) MistNet, (2) the implied pixel-level segmentation of the oracle whole-scan classifier, and (3) the ground-truth segmentation. Each VPR consists of average reflectivity measurements (in units of η, $\mathrm{cm}^2\,\mathrm{km}^{-3}$, cf. Chilson *et al.*, 2012) of sample volumes in each 100 m height bin up to 3000 m and within 37.5 km of the radar; we used WSRLIB (Sheldon, 2017) and followed (Farnsworth *et al.*, 2016; Horton *et al.*, 2018; Van Doren & Horton, 2018) to compute VPRs. For each segmentation method, pixels labeled as precipitation were set to zero reflectivity. We then measured the error of the VPRs with automatic segmentation (MistNet and whole-scan classification) compared to the VPR with ground-truth segmentation. Performance was measured as root mean-squared error (RMSE, $\mathrm{cm}^2\,\mathrm{km}^{-3}$) over the height bins.

2.5 Biology Case Studies

There are a wide range of science and conservation uses for continent-wide historical measurements of bird migration. We used MistNet to prepare a 19.5-year data set of spring and fall migration intensity. From each of the 143 radar stations in the contiguous US, we processed nighttime scans at 30-minute increments starting at local sunset for spring (1 March–15 June) and fall (1 August–15 November) from 1999 through the middle of 2018—a total of approximately 10 million scans. We used MistNet to segment each scan and then computed vertical profiles of reflectivity for further analysis. We conducted several case studies to demonstrate biological uses of this data, including spatial mapping of migration traffic as well as visualization of seasonal phenology across many years, the within-season temporal patterns of migration, and within-night patterns of airspace usage by migrants.

3 Results

Method	Precision	Recall	F-score
$\rho_{\mathrm{HV}} > 0.95$	90.1	93.4	91.7
$\mathrm{DR} < -15$	89.0	96.6	93.1
MistNet	99.1	96.7	97.9

Table 1: Performance of dual-pol thresholding on contemporary evaluation set. MistNet performance is shown for comparison. The threshold values of 0.95 for ρ_{HV} and -15 for DR led to the best performance among a range of alternatives.

Dual-pol thresholding and training labels Table 1 shows the classification performance of dual-pol based rules. Thresholding based on ρ_{HV} and DR both achieve F-score greater than 90% on the contemporary evaluation set. The best thresholds were 0.95 for ρ_{HV} and $-15\,\mathrm{dB}$ for DR (Figure C.3). DR-based thresholding achieves higher F-score and is an attractive alternative to ρ_{HV}-based thresholding. For comparison, MistNet achieves F-score of 97.87% on the same evaluation set using only legacy data products for prediction. This shows that simple dual-pol based thresholding provides effective training signal: the trained model only has access to legacy data and *exceeds* the performance of the dual-pol based training labels. However, note that more accurate predictions could be obtained using dual-pol data if this were the final goal, for example, using despeckling (Kilambi *et al.*, 2018) or spatial postprocessing (Dokter *et al.*, 2018a). Our goal is only to obtain a cheap and "good enough" training signal.

	Predicted	
	Biology	Precipitation
Biology	17.9	0.8
Precipitation	0.2	81.2

Precision	Recall	F-score
98.7	95.9	97.3

(a) Historical evaluation set

	Predicted	
	Biology	Precipitation
Biology	50.6	1.8
Precipitation	0.3	47.3

Precision	Recall	F-score
99.1	96.7	97.9

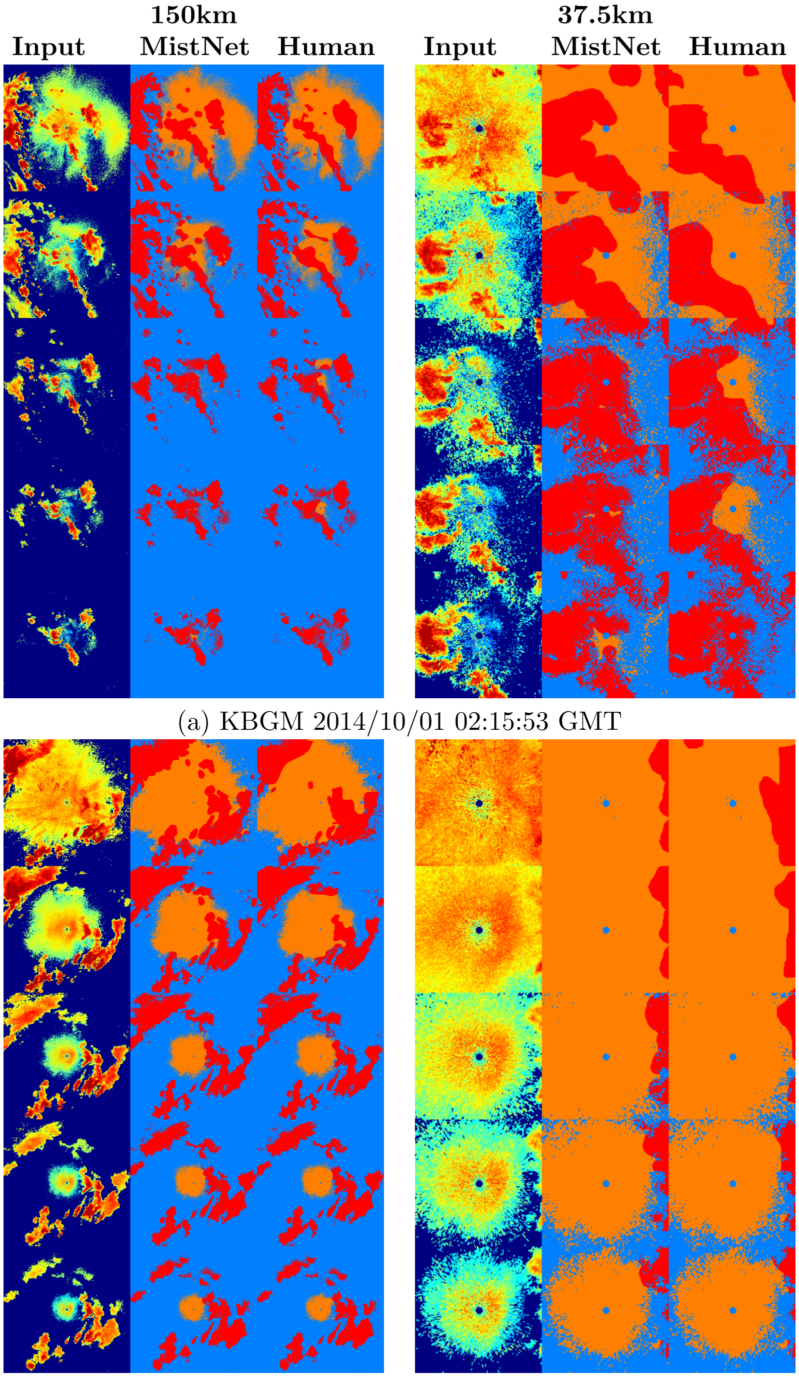
(b) Contemporary evaluation set

Table 2: Confusion matrices and overall performance measurements for MistNet on historical and contemporary evaluation sets.

Overall performance Figure 2 and C.4 shows several examples of predictions made by MistNet compared to the ground-truth human annotations. Table 2 gives the confusion matrices and overall performance measurements of MistNet on the historical and contemporary evaluation sets. The overall prevalance of precipitation is higher in historical data (81.4 % vs. 47.6 %).

CNN experiments The results of experiments to assess MistNet design choices are shown in Tables 3, 4, 5, C.2, C.3 and Figure 3. Unless stated otherwise, the results in this section use 10 000 scans, use ImageNet pre-training, do not post-process predictions, and do not include low-resolution augmentation of the training data.

- *Deep vs. shallow architectures.* Table 3 shows the performance of MistNet compared with the two shallow models. In the 2-layer networks, 5×5 convolutions performs better than 1×1 convolutions, which shows that the spatial context is helpful for prediction. MistNet's F-score is 5 to 10 percentage points better than both shallow networks, showing that the deep architecture, which considers more spatial context and at different scales, is beneficial. The difference is more pronounced on the weather subset of historical scans.



(a) KBGM 2014/10/01 02:15:53 GMT

(b) KMOB 2007/09/01 03:10:00 GMT

Figure 2: **MistNet Segmentation Results.** Segmentation results (red: rain, orange: biology, blue: background) predicted by MistNet are shown along with the human annotations in the ranges of 150km and 37.5km. Each example is shown as a stack of five rows from top to bottom corresponding to the elevation angles from 0.5 to 4.5 degrees.

Data set	Method	Precision	Recall	F-score
Historical (all)	2-layer (1 × 1)	91.0	78.3	84.2
	2-layer (5 × 5)	88.9	94.4	91.5
	MistNet	93.5	99.0	96.2
Historical (weather)	2-layer (1 × 1)	66.5	81.4	73.2
	2-layer (5 × 5)	59.9	93.7	73.1
	MistNet	72.6	96.1	82.7
Contemporary	2-layer (1 × 1)	96.2	66.5	78.6
	2-layer (5 × 5)	96.0	88.6	92.1
	MistNet	96.4	99.1	97.7

Table 3: Performance comparison of MistNet to two different 2-layer convolutional neural networks on historical and contemporary data sets.

Data set	Method	Precision	Recall	F-score
Historical (all)	DZ+adjacent sweeps	79.8	99.3	88.5
	DZ+VR+SW	79.4	99.3	88.3
	MistNet	93.5	99.0	96.2
Historical (weather)	DZ+adjacent sweeps	42.2	97.6	59.0
	DZ+VR+SW	41.7	97.8	58.5
	MistNet	72.6	96.1	82.7
Contemporary	DZ+adjacent sweeps	93.7	99.5	96.5
	DZ+VR+SW	94.2	99.4	96.7
	MistNet	96.5	99.1	97.8

Table 4: Comparing different approaches to predicting at multiple elevations. MistNet predicts at 5 elevations using all 15 channels as input. DZ+adjacent sweeps uses only reflectivity information from the target elevation and adjacent sweeps as input. DZ+VR+SW uses three products at the target elevation as input.

- *Predicting at 5 elevations using 15 channels.* Table 4 compares MistNet to the two baselines that predict each sweep separately using three selected input channels. MistNet achieves higher F-scores than either baseline. The baselines misclassify rain as biology substantially more than MistNet: they have lower precision on each data set, and the difference is more pronounced on historical data, which has a higher percentage of rain than the contemporary data, and especially on the weather subset, where MistNet has F-score 82.7 % compared to 58.9 % for the better baseline model.

- *Size of training set.* Figure 3 shows the results of increasing training set size. Performance increases significantly from 100 to 1000 training scans. In both historical and

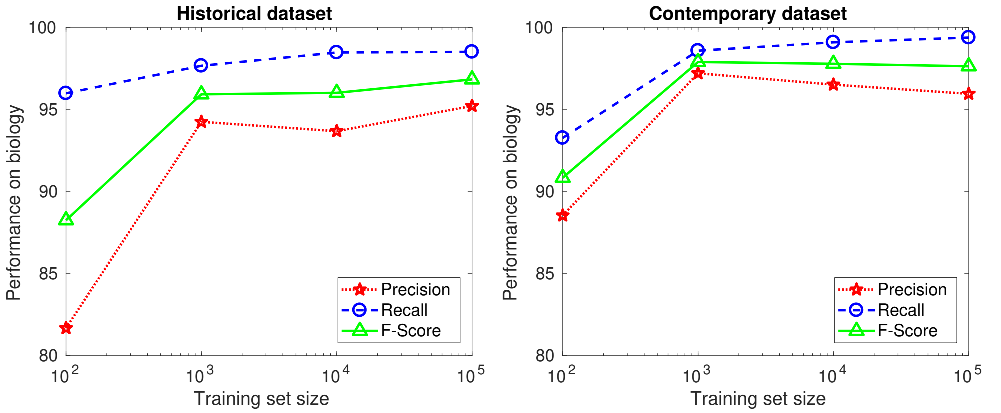


Figure 3: Performance of MistNet as a function of the size of training data. The performance improves significantly from 100 to 1000 training examples. The recall of biology prediction continues to increase with more training data. This suggests more data is useful for reducing the confusion of recognizing biology as precipitation.

Data set	Post-processing?	Precision	Recall	F-score
Historical (all)	no	93.5	99.0	96.2
	yes	98.7	95.9	97.3
Historical (weather)	no	72.6	96.1	82.7
	yes	92.7	82.8	87.5
Contemporary	no	96.4	99.1	97.7
	yes	99.1	96.7	97.9

Table 5: Performance of MistNet with and without post-processing.

contemporary data, recall continues to increase with more training data, but precision may stay the same or even decrease after 1000 scans. Improvements in recall suggest that bigger training sets allow the model to better recognize cases of biology that can be confused with precipitation.

- *Post-processing.* Table 5 compares results with and without post-processing. Post-processing always predicts fewer pixels to be biology—and hence will have higher precision and lower recall—than the standard prediction rule. Post-processing improves F-score on each data set; again, the difference is most pronounced on the weather subset of historical data (4.8 %), and is very slight on contemporary data (0.2 %).
- *Pre-training.* Table C.2 compares models with and without pre-training. The pre-

trained models outperform the randomly initialized ones, but with a modest overall increase in F-score. The difference is most pronounced on the weather subset (2.7 %), and very slight on contemporary data (0.1 %). On all evaluation sets precision improves and recall is nearly unchanged, which indicates that pre-training helps recognize some cases of rain that would otherwise be misclassified as biology.

- *Low-resolution augmentation.* Table C.3 compares MistNet with and without training data augmentation by low-resolution rendering. Augmentation yields slight F-score improvements (0.9 % historical data, 0.2 % contemporary).

3.1 Comparison with Whole-Scan Classification

	Pixel-level			VPRs (RMSE)		
Method	Precision	Recall	F-score	Clear	Weather	All
MistNet	98.7	95.9	97.3	40.2	262.3	155.3
Oracle scan-level	100	81.2	89.6	0.0	655.2	379.2

Table 6: Pixel-level classification performance and per-height-bin root mean-squared error (η, $\mathrm{cm}^2\,\mathrm{km}^{-3}$) for MistNet and an oracle scan-level classifier on historical data. "Clear", "weather", and "all" refer to different subsets of historical evaluation set.

Table 6 compares the performance of MistNet to an oracle whole-scan classifier on the historical evaluation set. Henceforth, MistNet is trained using the full data set of 239 128 scans, with pre-training, with low-resolution augmentation, and post-processing is applied to predictions. The oracle whole-scan classifier eliminates *all* rain and therefore has pixel-level precision of 100 %. However, its recall is only 81.2 %, meaning it excludes about 19 % of biology. In contrast, MistNet retains an additional 14.7 % of the total biology for a recall of 95.9 %, and still has an excellent precision of 98.7 %, leading to a significantly better F-score. For the task of computing VPRs, MistNet has slightly higher error on the clear data but substantially lower error on the weather data, leading to a lower overall error.

3.2 Biology Case Studies

Figure 4 shows the biological case studies. Panel (a) illustrates the cumulative migration traffic across the US from 1999-2018. An examination of latitudinal cross-sections of the US reveals passage of upwards of 3 billion birds, assuming a radar cross-section of 11 cm^2 (mass: 26 g–36 g). The significance of the midwest as a migration corridor is apparent. Panel (b) shows the cumulative spring migration traffic over one station for 20 years, and can be used to examine year-to-year consistency and variability in timing of spring migration. For example, the date by which 50% of total spring migration occurred over KHGX varied by only 11 days (mean: April 28$^{\text{th}}$), and there was no difference across years ($F_{1,18} = 0.481$, $p = 0.497$). Panel (c) further zooms in and shows the nightly migration over KHGX during 2018. From this we can see that migratory activity isn't uniform night-to-night, but occurs in bursts. For example, 51.3% of migration occurs on the top 10 nights. Panel (d) further zooms in to show the migration intensity at different heights during the night of April 29, 2018. The ascent behavior at the onset of migration and altitude distribution during later part of migrants is apparent.

4 Discussion

Discriminating biology from precipitation has been a long-standing challenge in radar aeroecology that has substantially limited accessing the full biological potential of historical weather radar data. MistNet provides a fully automated method for extracting biological signals from historical WSR-88D weather radar data and opens the entirety of the more-than-25-year archive of US weather radar data for long-term and large-scale biological studies. The high resolution of MistNet retains 15 % more biology than previously used whole-scan methods, providing a more complete data set that includes previously-absent information on biological targets interacting with weather systems.

MistNet can help address contemporary and pressing ecological challenges. Our case studies highlight the temporal and spatial dexterity of data products enabled by MistNet, spanning from multiple decades to single nights and from continental scales to

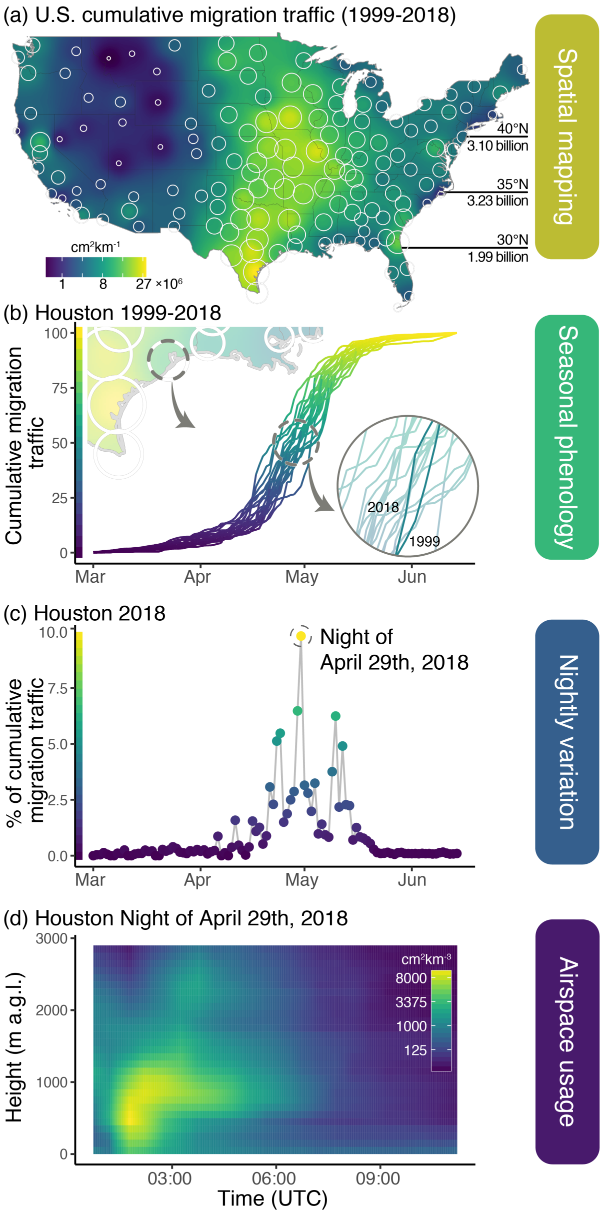


Figure 4: Case study. (a) Average cumulative migration traffic across the continental United States from 1999-2018. White circles show radar locations and are scaled to the cube root of cumulative migration traffic. Site estimates are interpolated using inverse distance weighting. Latitude lines show estimated passage numbers across different latitudinal cross-sections. (b) Site-level cumulative traffic curves shown for Houston, Texas (KHGX) from 1999-2018. Inset shows the period when approximately 50 % of migratory passage occurred. (c) Seasonal timing of migration showing the nightly variation in cumulative migration passage. The peak night of activity is highlighted with a gray circle. (d) Aerial migratory activity for the night of April 29th, 2018 (i.e., peak night of activity) from sunset to sunrise.

individual parcels of airspace. At the largest spatial and longest temporal extents, we can examine cumulative migration traffic across the continental US to identify flyways, critical hot spots, and estimate long-term changes in total biomass passing broad ecoregions for conservation and ecological applications. Beyond multi-decadal summaries, these data can be used to quantify yearly phenology at specific points of coverage and examine how migration timing may be changing (or not) with modified habitats and climates. At the site level, we can examine the progression of migration within a single season to identify peak nights of movements, with the potential to relate environmental conditions and motivate conservation action (e.g., halting wind turbines) and civic engagement (e.g., Sullivan *et al.*, 2009). During single nights, we can examine which regions of the atmosphere migrants are using to investigate details of their flight behaviors, such as speed and direction in the different altitude layers.

Our results show that deep learning is an effective tool for discriminating rain from biology in radar data, and is likely to be successful for other recognition tasks in radar data. Key ingredients to MistNet's success are a large enough training set, which is enabled by gathering labels automatically from dual polarization data, and an architecture that is able to use all available information—from all products across multiple elevations—while making predictions. An interesting technical aspect of MistNet's architecture is the fact that information is compressed down from 15 to 3 channels at the first layer, but MistNet is later able to make predictions at 5 separate elevations. The exact mechanisms by which the model compresses and retrieves information from these channels is an interesting topic of future research.

There are several promising research directions for future applications of deep learning to radar tasks. One direction is to improve performance by tracking recent progress in deep learning for images, for example, to adopt architectures such as residual networks (He *et al.*, 2016) instead of the VGG-16 architecture used in MistNet. A more substantial change would be to explore novel architectures that are completely customized for radar data, which would necessitate training models from scratch. We observed in every model we tested that pre-training improves performance, but the gains were less than 1 % in MistNet's final architecture, so pre-training may not be essential. Simple two-layer networks trained from scratch are about 5 % worse than MistNet. An intermediate architecture may achieve a good trade-off between size and accuracy; or, a different deep architecture tailored to radar data may outperform ImageNet-based models. Because radar data is sampled from a three-dimensional volume, a volumetric approach (Wu *et al.*, 2015; Maturana & Scherer, 2015) or point-based approach (Qi *et al.*, 2017; Su *et al.*, 2018), may be more appropriate. Finally, the predictions can be improved by taking temporal information into account, for example, to discriminate between weather and biology based on different patterns of motion within the radar domain.

Although we focused on historical data, our results also provide several insights about the use of dual polarization data. First, we provide a comprehensive empirical validation of simple thresholding rules for discriminating precipitation from biology. The common practice of thresholding correlation coefficient is effective; we also confirm the observation of Kilambi *et al.* (2018) that thresholding depolarization ratio is slightly more effective, and recommend this to practitioners of radar aeroecology.

We observed that the pixel-level classification performance of MistNet, which uses only legacy data products, is *better* than simple thresholding rules using dual-pol products. A deep learning model that uses both legacy and dual-pol data products is an obvious candidate to achieve the best possible classifications using dual-pol data. We are unsure how this would compare with existing hydrometeor classification algorithms from the meteorology community (Lim *et al.*, 2005; Park *et al.*, 2009). It is likely that deep neural networks can learn to detect spatial patterns and textures that complement the pixel-level information used by hydrometeor classification algorithms. Hydrometeor classification algorithms, which discern 10 different classes, can potentially benefit future applications of neural networks by providing better sources of training labels.

A tantalizing possibility is to use deep learning with dual polarization data to make finer-grained classifications of biological scatterers, for example, to discriminate birds, bats, and insects, or more specific groups such as size classes of birds or finer taxonomic groups (Stepanian *et al.*, 2016; Bauer *et al.*, 2018). While this is exciting, it is unclear what distinctions are possible using weather radar data alone. Any research in this direction must begin by assembling ground-truth data sets to evaluate and train algorithms. We believe the most promising near-term applications will be recognition of specific patterns in radar data such as bat and bird roosts or mayfly hatches, where humans can judge with reasonable certainty the identity of the scatterers and therefore assemble evaluation and training sets using available radar and geospatial data. For example, Chilson *et al.* (2018) recently trained a deep learning model using large human-labeled data sets to find radar scans containing swallow roosts (Bridge *et al.*, 2016; Laughlin *et al.*, 2016; Kelly & Pletschet, 2017). Detailed analyses of other specific patterns in radar data (Van Den Broeke, 2019) and cross-calibration with other sensors (Nilsson *et al.*, 2018a; Liechti *et al.*, 2018) may reveal over time the ability to distinguish other biological phenomena.

Radar aeroecology is advancing rapidly and leading to significant discoveries about continent-scale patterns of migration (Bauer *et al.*, 2018). To overcome big data challenges, we are relying increasingly on algorithms for all parts of the analysis. As the field moves quickly in this direction, we believe it is critical to advance methodological foundations including software, data, and empirical benchmarks to validate individual components of the analysis. MistNet is a general-purpose and empirically validated method to discriminate precipitation from biology, and can enable large-scale, reproducible measurements of whole migration systems. MistNet is available in the open-source WSRLIB software package (Sheldon, 2017) and is part of the `vol2bird` algorithm in bioRad (Dokter *et al.*, 2018a).[6]

Acknowledgments

This work was supported by the US NSF Advances in Biological Informatics program (ABI-1661259 to DS). CN, KH and ADs work was supported through the Edward W. Rose Postdoctoral Fellowship.

[6] It will be added per the terms of the data accessibility agreement prior to publication.

Data Accessibility

References

Ansari, S., Del Greco, S., Kearns, E., Brown, O., Wilkins, S., Ramamurthy, M., Weber, J., May, R., Sundwall, J., Layton, J., Gold, A., Pasch, A. & Lakshmanan, V. (2018) Unlocking the potential of NEXRAD data through NOAAs big data partnership. *Bulletin of the American Meteorological Society*, **99**, 189–204.

Bauer, S., Chapman, J.W., Reynolds, D.R., Alves, J.A., Dokter, A.M., Menz, M.M.H., Sapir, N., Ciach, M., Pettersson, L.B., Kelly, J.F., Leijnse, H. & Shamoun-Baranes, J. (2017) From agricultural benefits to aviation safety: Realizing the potential of continent-wide radar networks. *BioScience*, **67**, 912–918.

Bauer, S., Shamoun-Baranes, J., Nilsson, C., Farnsworth, A., Kelly, J.F., Reynolds, D.R., Dokter, A.M., Krauel, J.F., Petterson, L.B., Horton, K.G. & Chapman, J.W. (2018) The grand challenges of migration ecology that radar aeroecology can help answer. *Ecography*.

Bridge, E.S., Pletschet, S.M., Fagin, T., Chilson, P.B., Horton, K.G., Broadfoot, K.R. & Kelly, J.F. (2016) Persistence and habitat associations of Purple Martin roosts quantified via weather surveillance radar. *Landscape Ecology*, **31**, 43–53.

Brooks, M. (1945) Electronics as a possible aid in the study of bird flight and migration. *Science*, **101**, 329.

Bruderer, B. (1997) The study of bird migration by radar part 1: The technical basis. *Naturwissenschaften*, **84**, 1–8.

Buler, J.J. & Dawson, D.K. (2014) Radar analysis of fall bird migration stopover sites in the northeastern US. *The Condor*, **116**, 357–370.

Buler, J.J. & Diehl, R.H. (2009) Quantifying bird density during migratory stopover using weather surveillance radar. *IEEE Transactions on Geoscience and Remote Sensing*, **47**, 2741–2751.

Buler, J.J., Randall, L.A., Fleskes, J.P., Barrow, Jr., W.C., Bogart, T. & Kluver, D. (2012) Mapping wintering waterfowl distributions using weather surveillance radar. *PloS one*, **7**, e41571.

Casement, M.B. (1966) Migration across the Mediterranean observed by radar. *Ibis*, **108**, 461–491.

Chilson, C., Avery, K., McGovern, A., Bridge, E., Sheldon, D. & Kelly, J. (2018) Automated detection of bird roosts using NEXRAD radar data and convolutional neural networks. *Remote Sensing in Ecology and Conservation.*

Chilson, P.B., Frick, W.F., Stepanian, P.M., Shipley, J.R., Kunz, T.H. & Kelly, J.F. (2012) Estimating animal densities in the aerosphere using weather radar: To Z or not to Z? *Ecosphere*, **3**.

Crum, T.D. & Alberty, R.L. (1993) The WSR-88D and the WSR-88D operational support facility. *Bulletin of the American Meteorological Society*, **74**, 1669–1687.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. & Fei-Fei, L. (2009) ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE.

Dokter, A.M., Desmet, P., Spaaks, J.H., van Hoey, S., Veen, L., Verlinden, L., Nilsson, C., Haase, G., Leijnse, H., Farnsworth, A., Bouten, W. & Shamoun-Baranes, J. (2018a) bioRad: biological analysis and visualization of weather radar data. *Ecography.*

Dokter, A.M., Farnsworth, A., Fink, D., Ruiz-Gutierrez, V., Hochachka, W.M., La Sorte, F.A., Robinson, O.J., Rosenberg, K.V. & Kelling, S. (2018b) Seasonal abundance and survival of North America's migratory avifauna determined by weather radar. *Nature ecology & evolution*, **2**, 1603.

Dokter, A.M., Liechti, F., Stark, H., Delobbe, L., Tabary, P. & Holleman, I. (2011) Bird migration flight altitudes studied by a network of operational weather radars. *Journal of The Royal Society Interface*, **8**, 30–43.

Dokter, A.M., Shamoun-Baranes, J., Kemp, M.U., Tijm, S. & Holleman, I. (2013) High altitude bird migration at temperate latitudes: a synoptic perspective on wind assistance. *PloS one*, **8**, e52300.

Doviak, R.J. & Zrnić, D.S. (1993) *Doppler radar and weather observations.* Courier Corporation.

Eastwood, E. (1967) *Radar ornithology.* Methuen.

Farnsworth, A., Van Doren, B.M., Hochachka, W.M., Sheldon, D., Winner, K., Irvine, J., Geevarghese, J. & Kelling, S. (2016) A characterization of autumn nocturnal migration detected by weather surveillance radars in the northeastern USA. *Ecological Applications*, **26**, 752–770.

Forsyth, D.A. & Ponce, J. (2003) *Computer vision: a modern approach.* Prentice Hall.

Gauthreaux, Jr., S.A. (1970) Weather radar quantification of bird migration. *BioScience*, **20**, 17–19.

Gauthreaux, Jr., S.A. & Belser, C.G. (1998) Displays of bird movements on the WSR-88D: patterns and quantification. *Weather and Forecasting*, **13**, 453–464.

Gauthreaux, Jr., S.A., Belser, C.G. & Van Blaricom, D. (2003) Using a network of WSR-88D weather surveillance radars to define patterns of bird migration at large spatial scales. *Avian migration*, pp. 335–346. Springer.

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep Learning.* MIT Press. `http://www.deeplearningbook.org`.

Graves, A., Mohamed, A.r. & Hinton, G. (2013) Speech recognition with deep recurrent neural networks. *Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on*, pp. 6645–6649. IEEE.

Harper, W.G. (1958) Detection of bird migration by centimetric radar—a cause of radar angels. *Proceedings of the Royal Society of London, Series B*, **149**, 484–502.

He, K., Zhang, X., Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. & Kingsbury, B. (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, **29**, 82–97.

Horton, K.G., Shriver, W.G. & Buler, J.J. (2015) A comparison of traffic estimates of nocturnal flying animals using radar, thermal imaging, and acoustic recording. *Ecological Applications*, **25**, 390–401.

Horton, K.G., Van Doren, B.M., La Sorte, F.A., Cohen, E.B., Clipp, H.L., Buler, J.J., Fink, D., Kelly, J.F. & Farnsworth, A. (2019) Holding steady: Little change in intensity or timing of bird migration over the Gulf of Mexico. *Global Change Biology.* In press.

Horton, K.G., Van Doren, B.M., La Sorte, F.A., Fink, D., Sheldon, D., Farnsworth, A. & Kelly, J.F. (2018) Navigating north: how body mass and winds shape avian flight behaviours across a North American migratory flyway. *Ecology Letters*, **21**, 1055–1064.

Horton, K.G., Van Doren, B.M., Stepanian, P.M., Hochachka, W.M., Farnsworth, A. & Kelly, J.F. (2016) Nocturnally migrating songbirds drift when they can and compensate when they must. *Scientific Reports*, **6**, 21249.

Kelly, J.F. & Horton, K.G. (2016) Toward a predictive macrosystems framework for migration ecology. *Global Ecology and Biogeography*, **25**, 1159–1165.

Kelly, J.F. & Pletschet, S.M. (2017) Accuracy of swallow roost locations assigned using weather surveillance radar. *Remote Sensing in Ecology and Conservation*.

Kemp, M.U., ShamounBaranes, J., Dokter, A.M., van Loon, E. & Bouten, W. (2013) The influence of weather on the flight altitude of nocturnal migrants in midlatitudes. *Ibis*, **155**, 734–749.

Kilambi, A., Fabry, F. & Meunier, V. (2018) A simple and effective method for separating meteorological from nonmeteorological targets using dual-polarization data. *Journal of Atmospheric and Oceanic Technology*, **35**, 1415–1424.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012) ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pp. 1097–1105.

La Sorte, F.A., Hochachka, W.M., Farnsworth, A., Sheldon, D., Fink, D., Geevarghese, J., Winner, K., Van Doren, B.M. & Kelling, S. (2015a) Migration timing and its determinants for nocturnal migratory birds during autumn migration. *Journal of Animal Ecology*, **84**, 1202–1212.

La Sorte, F.A., Hochachka, W.M., Farnsworth, A., Sheldon, D., Van Doren, B.M., Fink, D. & Kelling, S. (2015b) Seasonal changes in the altitudinal distribution of nocturnally migrating birds during autumn migration. *Royal Society Open Science*, **2**, 150347.

La Sorte, F.A., Horton, K.G., Nilsson, C. & Dokter, A.M. (2019) Projected changes in wind assistance under climate change for nocturnally migrating bird populations. *Global Change Biology*, **25**.

Lack, D. & Varley, G.C. (1945) Detection of birds by radar. *Nature*, **156**, 446.

Laughlin, A.J., Sheldon, D.R., Winkler, D.W. & Taylor, C.M. (2016) Quantifying non-breeding season occupancy patterns and the timing and drivers of autumn migration for a migratory songbird using Doppler radar. *Ecography*, **39**, 1017–1024.

Laughlin, A.J., Taylor, C.M., Bradley, D.W., Leclair, D., Clark, R.G., Dawson, R.D., Dunn, P.O., Horn, A., Leonard, M., Sheldon, D., Shutler, D., Whittingham, L.A., Winkler, D.W. & Norris, D.R. (2013) Integrating information from geolocators, weather radar and citizen science to uncover a key stopover area for an aerial insectivore. *The Auk*, **130**, 230–239.

Liechti, F., Aschwanden, J., Blew, J., Boos, M., Brabant, R., Dokter, A.M., Kosarev, V., Lukach, M., Maruri, M., Reyniers, M., Schekler, I., Schmaljohann, H., Schmid, B., Weisshaupt, N. & Sapir, N. (2018) Cross-calibration of different radar systems for monitoring nocturnal bird migration across Europe and the Near East. *Ecography*.

Lim, S., Chandrasekar, V. & Bringi, V.N. (2005) Hydrometeor classification system using dual-polarization radar measurements: Model improvements and in situ verification. *IEEE transactions on geoscience and remote sensing*, **43**, 792–801.

Long, J., Shelhamer, E. & Darrell, T. (2015) Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3431–3440.

Maturana, D. & Scherer, S. (2015) VoxNet: A 3D convolutional neural network for real-time object recognition. *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pp. 922–928. IEEE.

McLaren, J.D., Buler, J.J., Schreckengost, T., Smolinsky, J.A., Boone, M., van Loon, E., Dawson, D.K. & Walters, E.L. (2018) Artificial light at night confounds broad-scale habitat use by migrating birds. *Ecology letters*, **21**, 356–364.

Nilsson, C., Dokter, A.M., Schmid, B., Scacco, M., Verlinden, L., Bäckman, J., Haase, G., Dell'Omo, G., Chapman, J.W., Leijnse, H. & Liechti, F. (2018a) Field validation of radar systems for monitoring bird migration. *Journal of Applied Ecology*, **55**, 2552–2564.

Nilsson, C., Dokter, A.M., Verlinden, L., Shamoun-Baranes, J., Schmid, B., Desmet, P., Bauer, S., Chapman, J., Alves, J.A., Stepanian, P.M., Sapir, N., Wainwright, C., Boos, M., Górska, A., Menz, M.H.M., Rodrigues, P., Leijnse, H., Zehtindjiev, P., Brabant, R., Haase, G., Weisshaupt, N., Ciach, M. & Liechti, F. (2018b) Revealing patterns of nocturnal migration using the European weather radar network. *Ecography.*

Park, H.S., Ryzhkov, A.V., Zrnić, D.S. & Kim, K.E. (2009) The hydrometeor classification algorithm for the polarimetric WSR-88D: Description and application to an MCS. *Weather and Forecasting*, **24**, 730–748.

Parker, J.A., Kenyon, R.V. & Troxel, D.E. (1983) Comparison of interpolating methods for image resampling. *IEEE Transactions on medical imaging*, **2**, 31–39.

Qi, C.R., Su, H., Mo, K. & Guibas, L.J. (2017) PointNet: Deep learning on point sets for 3D classification and segmentation. *Proc Computer Vision and Pattern Recognition (CVPR), IEEE*, **1**, 4.

Razavian, A.S., Azizpour, H., Sullivan, J. & Carlsson, S. (2014) CNN features off-the-shelf: An astounding baseline for recognition. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '14, pp. 512–519. IEEE Computer Society, Washington, DC, USA.

RoyChowdhury, A., Sheldon, D., Maji, S. & Learned-Miller, E. (2016) Distinguishing weather phenomena from bird migration patterns in radar imagery. *CVPR workshop on Perception Beyond the Visual Spectrum (PBVS)*, pp. 1–8.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) Learning representations by back-propagating errors. *Nature*, **323**, 533.

Ryzhkov, A., Matrosov, S.Y., Melnikov, V., Zrnic, D., Zhang, P., Cao, Q., Knight, M., Simmer, C. & Troemel, S. (2017) Estimation of depolarization ratio using weather radars with simultaneous transmission/reception. *Journal of Applied Meteorology and Climatology*, **56**, 1797–1816.

Shamoun-Baranes, J., Dokter, A.M., van Gasteren, H., van Loon, E., Leijnse, H. & Bouten, W. (2011) Birds flee en mass from New Years Eve fireworks. *Behavioral Ecology*, **22**, 1173–1177.

Sheldon, D. (2017) WSRLIB: MATLAB toolbox for weather surveillance radar.

Simonyan, K. & Zisserman, A. (2014) Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, pp. 568–576.

Simonyan, K. & Zisserman, A. (2015) Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.

Stepanian, P.M., Horton, K.G., Melnikov, V.M., Zrnić, D.S. & Gauthreaux, Jr., S.A. (2016) Dual-polarization radar products for biological applications. *Ecosphere*, **7**.

Stepanian, P.M. (2015) *Radar Polarimetry for Biological Applications*. Ph.D. thesis, University of Oklahoma Norman, Oklahoma, USA.

Su, H., Jampani, V., Sun, D., Maji, S., Kalogerakis, E., Yang, M.H. & Kautz, J. (2018) SPLATNet: Sparse lattice networks for point cloud processing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2530–2539.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009) eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, **142**, 2282–2292.

Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. (2014) Deepface: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708.

Tangseng, P., Wu, Z. & Yamaguchi, K. (2017) Looking at outfit to parse clothing.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. (2015) Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497.

Van Den Broeke, M.S. (2019) Radar quantification, temporal analysis and influence of atmospheric conditions on a roost of American Robins (Turdus migratorius) in Oklahoma. *Remote Sensing in Ecology and Conservation*. In press.

Van Doren, B.M. & Horton, K.G. (2018) A continental system for forecasting bird migration. *Science*, **361**, 1115–1118.

Van Doren, B.M., Horton, K.G., Dokter, A.M., Klinck, H., Elbin, S.B. & Farnsworth, A. (2017) High-intensity urban light installation dramatically alters nocturnal bird migration. *Proceedings of the National Academy of Sciences*, **114**, 11175–11180.

Vedaldi, A. & Lenc, K. (2015) MatConvNet: Convolutional neural networks for MATLAB. *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 689–692. ACM.

Winkler, D.W. (2006) Roosts and migrations of swallows. *Hornero*, **21**, 85–97.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X. & Xiao, J. (2015) 3D ShapeNets: A deep representation for volumetric shapes. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920.

Zrnić, D.S. & Ryzhkov, A.V. (1998) Observations of insects and birds with a polarimetric radar. *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 661–668.

A Stations used in contemporary evaluation

The stations used in the contemporary data set are listed in Table A.1 and shown on a map in Figure A.1. One station was selected randomly from each of the 10° grid cells shown on the map. Three additional stations were added manually because they were known in some way to the researchers, either through prior radar analyses or availability of corroborating data from other sensors.

Station	City	Latitude	Longitude
KCBX	Boise, ID	43°29′27″ N	116°14′8″ W
KRIW	Riverton, WY	43°3′58″ N	108°28′38″ W
KOAX	Omaha, NE	41°19′13″ N	96°22′0″ W
KDLH	Duluth, MN	46°50′13″ N	92°12′35″ W
KBGM*	Binghamton, NY	42°11′59″ N	75°59′5″ W
KTYX	Montague, NY	43°45′21″ N	75°40′48″ W
KOKX	New York	40°51′56″ N	72°51′50″ W
KEYX	Edwards AFB, CA	35°5′52″ N	117°33′39″ W
KICX	Cedar City, UT	37°35′27″ N	112°51′44″ W
KEWX*	Austin, TX	29°42′14″ N	98°1′42″ W
KGRK	Fort Hood, TX	30°43′19″ N	97°22′59″ W
KTLX*	Oklahoma City, OK	35°19′59″ N	96°13′57″ W
KMOB	Mobile, AL	30°40′46″ N	88°14′23″ W
KJKL	Jackson, KY	37°35′27″ N	83°18′47″ W
KBRO	Brownsville, TX	25°54′58″ N	97°25′8″ W
KTBW	Tampa, FL	27°42′20″ N	82°24′6″ W

Table A.1: Stations selected for the contemporary data set. Stations marked by asterisks were selected manually; others were selected following a stratified random design.

B Additional details of MistNet and training

The full details of the MistNet architecture are shown in Figure B.2.[7] Yellow blocks indicate data, which includes input, output and intermediate activations; blue blocks indicate learnable parameters; red blocks indicate network operations such as pooling, up sampling, and convolutions. The backbone of the network (shared central pathway) consists of several 3×3 convolutional layers, ReLU non-linearities, and pooling blocks, following the VGG-16 architecture (Simonyan & Zisserman, 2015). The parameters of this pathway are initialized from the ImageNet dataset. We zero-pad the $15 \times 600 \times 600$ input to $15 \times 608 \times 608$ to make the spatial dimensions divisible by 32, the overall factor by which the input will be downsampled in the architecture. The first layer (adapter) maps the $15 \times 608 \times 608$ input to a $3 \times 608 \times 608$ using a 1×1 convolution making it compatible with the VGG-16 network that expects 3-channel images. Predictions for each

[7] We will include this image as a separate high-resolution supplementary file in the published version of the paper.

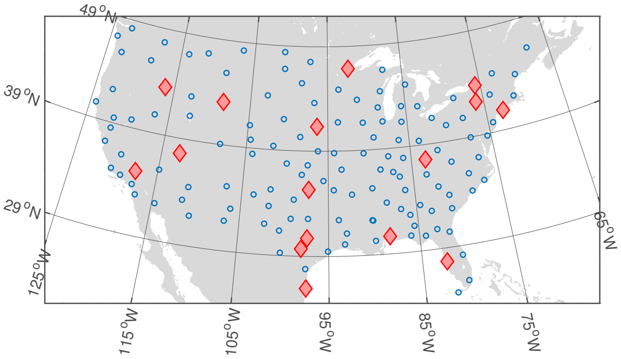


Figure A.1: Map of WSR-88D stations (small blue circles), and selected stations (red diamonds). One station was selected randomly from each 10° grid cell, and three additional stations were selected manually.

elevation are obtained by combining features from the pool3 (75×75), pool4 (36×36), and relu7 (19×19) layer outputs followed by several convolutional, ReLU and upsampling blocks to produce predictions of size 608×608, from which the central 600×600 portion is used. The 3-way softmax classifiers output the probability of the background, biology and rain classes at each elevation. The predictions pathways for each elevation are shown as separate pathways branching from the central shared backbone in the figure. Each pathway follows the architecture of the FCN8 network developed for image segmentation (Long *et al.*, 2015). All the parameters except for those in the central pathway are initialized randomly. We learn the model parameters using stochastic gradient descent at learning rate 10^{-4} and momentum 0.9 for 1.1 million iterations with batch size 64. The contemporary set was used for validation. At prediction time, we ignore the output probability for background class (since the identity of background pixels is known at prediction time). For non-background pixels, we renormalize the probabilities of precipitation and biology to sum to one and use these to make predictions in conjunction with the post-processing techniques described previously.

C Additional results

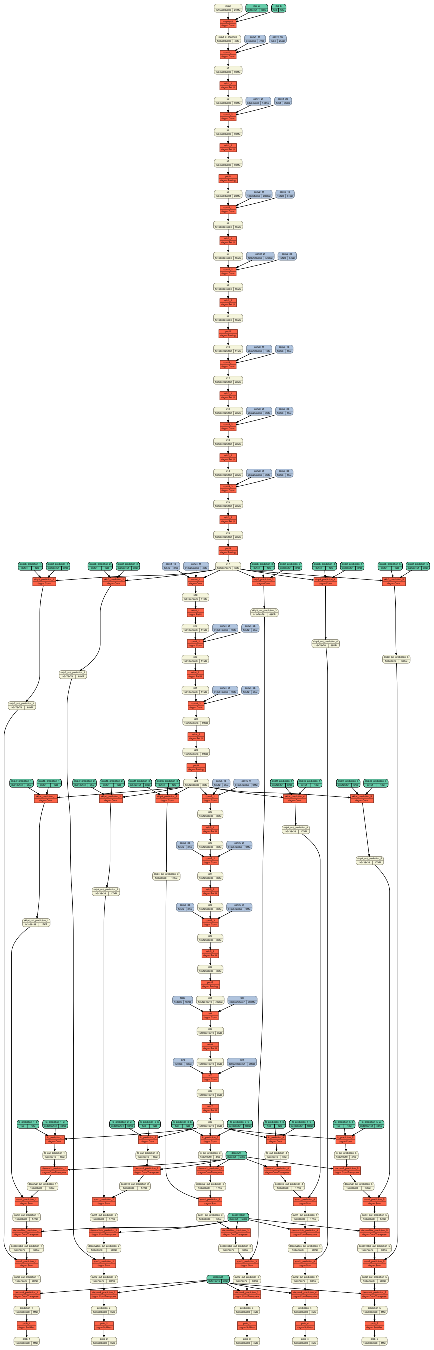
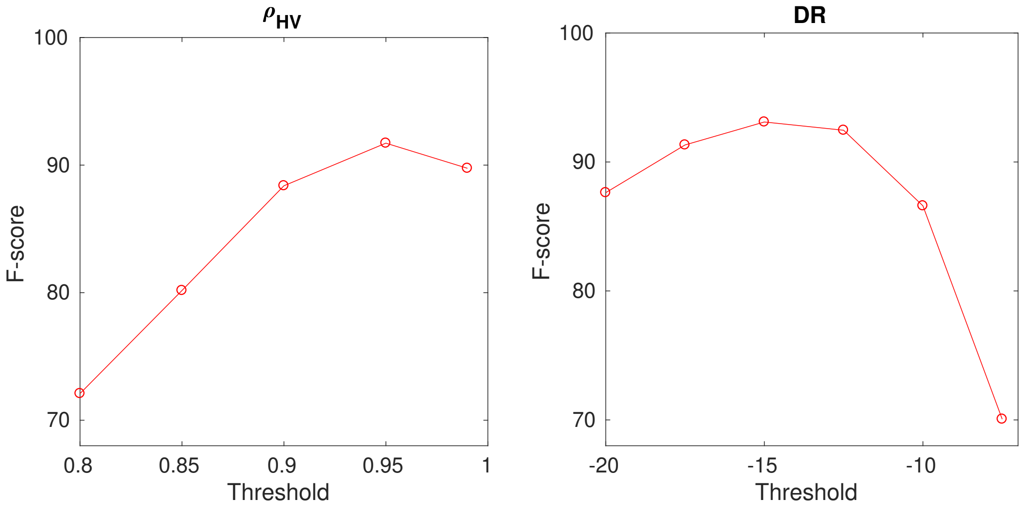


Figure B.2: **Architecture of MistNet.** The network takes as input a 15 × 608 × 608 image and produces segmentation masks corresponding to 5 different elevations. Yellow blocks indicate data, blue blocks indicate parameters initialized with the ImageNet pre-trained model, green blocks indicate parameters initialized randomly, and red blocks indicate network layers. *(Best viewed digitally with zoom.)*



(a)

(b)

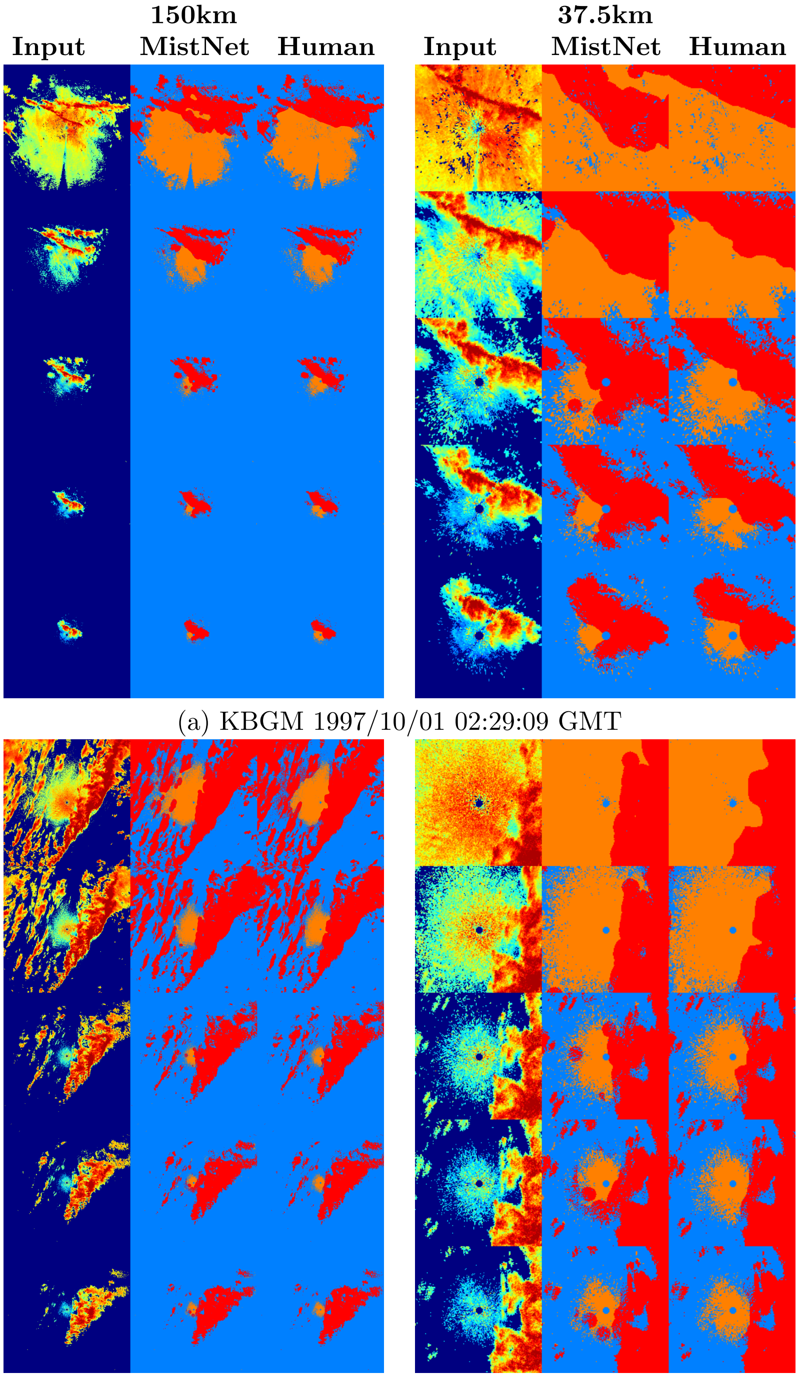
Figure C.3: Effectiveness of pixel-level classification using simple dual-pol thresholding rules. The plots show F-score on biology classification versus threshold values based on (a) ρ_{HV} and (b) DR.

Data set	Pre-training?	Precision	Recall	F-score
Historical (all)	no	92.0	99.2	95.5
	yes	93.5	99.0	96.2
Historical (weather)	no	68.0	97.0	80.0
	yes	72.6	96.1	82.7
Contemporary	no	96.1	99.3	97.7
	yes	96.5	99.1	97.8

Table C.2: Performance of MistNet with and without ImageNet pre-training.

Data set	Low-res. aug.?	Precision	Recall	F-score
Historical (all)	no	93.7	98.5	96.0
	yes	95.5	98.3	96.9
Historical (weather)	no	72.7	94.0	82.0
	yes	72.6	96.1	82.7
Contemporary	no	96.5	99.1	97.8
	yes	97.4	98.6	98.0

Table C.3: Performance of MistNet with and without data augmentation with low-resolution rendering.



(a) KBGM 1997/10/01 02:29:09 GMT

(b) KMOB 2015/11/01 02:50:41 GMT

Figure C.4: **MistNet Segmentation Results.** Segmentation results (red: rain, orange: biology, blue: background) predicted by MistNet are shown along with the human annotations in the ranges of 150km and 37.5km. Each example is shown as a stack of five rows from top to bottom corresponding to the elevation angles from 0.5 to 4.5 degrees.