CLOUD IMAGE CLASSIFICATION USING
MACHINE LEARNING

By
MARCUS PETER COTE

A thesis submitted to the Department of Ocean Engineering and Marine Sciences of Florida Institute of Technology in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
METEOROLOGY

Melbourne, Florida
December 2022
CLOUD IMAGE CLASSIFICATION USING

MACHINE LEARNING

A THESIS

By

MARCUS PETER COTE

Approved as to style and content by:

______________________________
Steven Lazarus, Ph.D.
Professor
Ocean Engineering and Marine Sciences
Major Advisor

______________________________
Michael E. Splitt
Assistant Professor
College of Aeronautics

______________________________
Pallav Ray, Ph.D.
Associate Professor
Ocean Engineering and Marine Sciences

______________________________
Eraldo Ribeiro, Ph.D.
Associate Professor
Computer Engineering and Sciences

______________________________
Richard B. Aronson, Ph.D.
Professor and Department Head
Ocean Engineering and Marine Sciences

December 2022
ABSTRACT

CLOUD IMAGE CLASSIFICATION USING MACHINE LEARNING

by Marcus Peter Cote

Chairperson of Advisory Committee: Steven Lazarus, Ph.D.

Machine learning is a rapidly expanding technology that has proven to be highly useful for image classification. Ground-based camera networks are an emerging resource for aviation weather information with near real-time imagery available online for public viewing and download. While raw web camera imagery can be analyzed by aviators, high pilot workload motivates the use of machine learning to autonomously interpret cloud type information from images that is relevant to aviation weather hazards. In particular, transfer learning is a machine learning approach by which elements of a pre-trained machine learning model are refitted for new tasks. By employing transfer learning for image classification, the outermost layers of complex convolutional neural networks (CNNs) can be quickly retrained for new classes of imagery as organized by the user. In this research, a tiered methodology is employed whereby machine transfer learning is used to develop cloud image classification schemes of increasing complexity. To achieve accurate results using transfer learning, a large and diverse training dataset is required. During May of 2022, four publicly accessible ground-based web cameras were installed at FIT (Florida Institute of Technology) Aviation – located at the Melbourne International Airport (KMLB). These data are used to develop an extensive cloud image archive.
Image categories of interest include a variety of cloud types and sky conditions, prioritizing those relevant to aviation safety hazards such as towering cumulus, cumulonimbus, and precipitation. Utilizing Google’s TensorFlow machine learning platform in Python, transfer learning was conducted with their Inception v3 convolutional neural network for deep learning. Several iterations of models were developed and tested for accuracy to assess the impacts of training data organization, application to different camera sites, and time of day. As proof of concept, the models were used to classify the FIT Aviation web camera imagery in real time to supplement weather information for potential users while simultaneously allowing for the near real-time tracking of model performance. Results reveal that cloud-type image classification using transfer learning is a viable method for extracting high-temporal-resolution information from growing web camera resources while minimizing the human component of weather information processing. Classification output demonstrates that the model correctly identifies hazard-related image content for a large percentage of cases, especially when raw model output is adapted to optimize model hit rates while minimizing false alarms and missed events.
ACKNOWLEDGEMENTS

I would like to thank Dr. Steven Lazarus and Professor Michael Splitt for continued opportunities and for making this research possible with partial funding from the Federal Aviation Administration PEGASAS grant, General Dynamics Radionuclide Aerosol Sampler/Analyzer (RASA) program, and Schnoor endowment. I would also like to thank my close friends, family, teachers, and committee members who have been incredibly helpful and supportive along the way.
# TABLE OF CONTENTS

ABSTRACT................................................................................................................................. iii
ACKNOWLEDGEMENTS........................................................................................................... v
TABLE OF CONTENTS................................................................................................................ vi
LIST OF FIGURES ..................................................................................................................... vii
LIST OF TABLES ......................................................................................................................... x
INTRODUCTION ............................................................................................................................. 1
   LITERATURE REVIEW .............................................................................................................. 3
      AVIATION WEATHER ACCIDENTS...................................................................................... 3
      ARTIFICIAL INTELLIGENCE APPLICATIONS.................................................................. 6
DATA AND METHODS .................................................................................................................. 18
   DATASETS: WEB CAMERA RESOURCES ........................................................................... 21
METHODOLOGY .......................................................................................................................... 25
   METHODS: STAGE 1 ................................................................................................................ 28
   METHODS: STAGE 2 ................................................................................................................ 31
   METHODS: STAGE 3 ................................................................................................................ 37
   METHODS: STAGE 4 ................................................................................................................ 48
RESULTS AND DISCUSSION ....................................................................................................... 49
   RESULTS: STAGE 1 ................................................................................................................ 49
   RESULTS: STAGE 2 ................................................................................................................ 50
   RESULTS: STAGE 3 ................................................................................................................ 61
CONCLUSIONS ............................................................................................................................. 75
LITERATURE CITED ..................................................................................................................... 77
LIST OF FIGURES

Figure 1. Diagram depicting the workflow of supervised machine learning (modified from Zaccone, 2016) ........................................................................................................................................................................ 8
Figure 2. Graphical depiction of feature extraction in layers during the deep learning process (Zaccone, 2016) ........................................................................................................................................................................ 10
Figure 3. Illustration of Inception v3 model structure and modification for transfer learning applications (Intel, 2019) .................................................................................................................................................... 14
Figure 4. Illustration of feature recognition elements that are created by convolutional layers of deep learning for cloud images (from Zhang et al., 2018) ........................................ 17
Figure 5. The FAA WeatherCams website interface (FAA, n.d.) ............................................................................................................................ 22
Figure 6. The Alert Wildfire website interface (Alert Wildfire, n.d.) .................................................................................................................... 23
Figure 7. The NPS Air Quality Webcams website interface (NPS, n.d. b) ........................................................................................................ 24
Figure 8. Images from four AirportView.net cameras installed at FIT Aviation (KMLB) on 22 August 2022 at 1:00 p.m. and 10:00 p.m. ET ................................................................................................. 26
Figure 9. Four-stage graphic describing the methodological approach to construct a cloud image classifier for this study .......................................................................................................................... 27
Figure 10. Map of the Great Smoky Mountains National Park region .......................................................... 29
Figure 11. Supplementary information for the Look Rock camera site provided on the NPS Air Quality Webcams site (NPS, n.d. a) .......................................................................................................................... 30
Figure 12. Digital images produced by the Look Rock NPS Air Quality Webcam site before and after the 2018 camera upgrade ........................................................................................................ 31
Figure 13. Metadata contained within the Look Rock image files before and after the 2018 camera upgrade ................................................................................................................................. 33
Figure 14. Training images and independent test images from pre-2018 web camera data at the Look Rock NPS Air Quality Webcam site located in the Great Smoky Mountains National Park ......................................................................................................................... 34
Figure 15. The three stages of convective clouds (Learnweather, 2021) ....................................................... 36
Figure 16. Location of four AirportView.net cameras stationed at FIT Aviation at the Melbourne International Airport (KMLB) in Melbourne, Florida ........................................................................ 38
Figure 17. Representative images for the seven categories used to create the stage 3 training dataset ................................................................. 41

Figure 18. A screenshot of the near real-time, publicly accessible Google Sheets document used to log image classification output for the AirportView.net west-facing camera site at FIT Aviation ......................................................................................................................... 43

Figure 19. NOAA Sky Watcher Chart with information regarding the visual characteristics and cloud base heights of various cloud types (NOAA, n.d. b) .......... 46

Figure 20. A stage 3 training data image captured by the AirportView.net west-facing camera at FIT Aviation (KMLB) that was categorized as precipitation and archived NEXRAD mosaic reflectivity imagery from the NOAA National Centers of Environmental Information (NCEI) on 6 June 2022, at 3:20 p.m. ET ........................................ 47

Figure 21. An independent test image from stage 3 that was labeled ‘clear sky’. Clouds partially visible along the distant horizon were determined to be an insignificant element of the image based on the stage 3 methodology ................................................................. 47

Figure 22. Three examples of independent test images: cumulus, clear sky, and obscured view that have been inferenced by a cloud-image classifier that was trained on these image categories .................................................................................................................. 49

Figure 23. Cumulative performance of cloud image classifier results for all training data and independent test image resolutions (as described by Tables 6-9) ................................................................. 56

Figure 24. The image from the Look Rock camera site in the Great Smoky Mountains National Park taken on 12 May 2016 at 9:15 a.m. ET and responsible for the ‘incorrect’ identification of Cu as TCu .................................................................................................................. 58

Figure 25. Training/independent test image candidate taken on 4 May 2012 at 1:45 p.m. ET that was rejected during quality control as the image was determined to contain mixed cloud types, including clouds existing in a state between fair-weather cumulus and towering cumulus .................................................................................................................. 61

Figure 26. Percentage of classified independent test images that exhibit a level of confidence above the prescribed thresholds that corresponds to the correct image category .................................................................................................................. 67
Figure 27. Distribution of top-1 results for the west-facing camera AirportView.net located at FIT Aviation (KMLB) for images captured ................................................................. 69
Figure 28. Nighttime IR mode image from the west-facing Airportview.net camera at FIT Aviation........................................................................................................................................................................ 70
Figure 29. Fair-weather cumulus cloud-only image from the west-facing Airportview.net camera at FIT Aviation ........................................................................................................................................................................ 71
Figure 30. Distribution of top-1 results for all AirportView.net cameras located at FIT Aviation (KMLB) for all available times of day (4 a.m. to 10 p.m. ET)........................................ 72
Figure 31. Images captured by the AirportView.net cameras at FIT Aviation (KMLB) on 27 October 2022 at 9:00 a.m. ET and labeled as clear sky ................................................................. 74
LIST OF TABLES

Table 1. List of settings that are available in the retrain.py algorithm ........................................ 20
Table 2. Breakdown of all image data used in the stage 2 ................................................................. 35
Table 3. Counts for each of the image categories that comprise the stage 3 training dataset .......................................................... 40
Table 4. Counts for each of the image categories that comprise the stage 3 test image dataset .............................................................................................................................. 44
Table 5. Classification results from the three independent test images in Figure 22 ........ 50
Table 6. Image classifier performance for an algorithm trained using original (high) resolution images. 25 independent high resolution test images each of Cu and TCu were classified and assigned a cloud type based on whether or not they exceeded the selected thresholds (0.70, 0.80, and 0.90) ........................................................................................................ 52
Table 7. Image classifier performance for an algorithm trained using original (high) resolution images. 25 independent medium resolution test images each of Cu and TCu were classified and assigned a cloud type based on whether or not they exceeded the selected thresholds (0.70, 0.80, and 0.90) ........................................................................................................ 53
Table 8. Image classifier performance for an algorithm trained using medium resolution images. 25 independent medium resolution test images each of Cu and TCu were classified and assigned a cloud type based on whether or not they exceeded the selected thresholds (0.70, 0.80, and 0.90) ........................................................................................................ 54
Table 9. Image classifier performance for an algorithm trained using medium resolution images. 25 independent high resolution test images each of Cu and TCu were classified and assigned a cloud type based on whether or not they exceeded the selected thresholds (0.70, 0.80, and 0.90) ........................................................................................................ 55
Table 10. Stage 3 matrix of the image classification scheme .................................................. 63
Table 11. Hazard detection matrix utilizing the top-1 accuracy data from Table 10 .... 65
INTRODUCTION

Weather information resources available to aviators have become increasingly numerous in the modern technological era. Pilots have access to formal aviation weather observations such as METARs and pilot reports (PIREPs). Terminal Aerodrome Forecasts (TAFs) are official FAA aviation weather forecasts issued for U.S. Airports (NOAA, n.d. a). Advisories are issued within the contiguous United States including AIRMETs, SIGMETs, convective SIGMETs, and CWAs to warn aviators of potential flight hazards of varying severity such as icing, turbulence, dust/sandstorms, thunderstorms, winds, hail, and tornadoes (NOAA, n.d. a). SIGMETs issued outside of the CONUS follow international coding standards to warn of hazards such as tropical cyclones, radiological clouds, as well as the aforementioned dangers (NOAA, n.d. a). Additionally, satellite imagery and aircraft/ground-based Doppler radar are valuable sources of weather information to supplement a pilot’s planning and in-flight workflows. A growing source of weather information for aviation exists in ground-based camera networks from which near real-time digital imagery is available for analysis. Ground-based camera sites can provide additional information to aviators about cloud type and other fast evolving weather hazards. Ground-based cameras may also be used to supplement gaps in weather information in regions where official NWS/FAA observation stations (ASOS/AWOS) are lacking.
The motivation for this research is to use ground-based web camera networks as a potential resource to pilots while minimizing the workload associated with weather information processing. Specifically, this research seeks to develop a product that gives pilots access to processed weather information that is near real-time, reflects changing weather conditions, and can be quickly integrated with existing information. As introduced above, ground-based camera sites can provide cloud-type and other sky condition information relevant to aviation and supplement current gaps in observations. Because pilots already face a burden from information overload, it is unrealistic for a pilot to check dozens of camera sites before and during a flight to obtain valuable weather data. This problem motivates the use of machine learning to autonomously interpret cloud information from imagery and present concise, near real-time hazard information to a common dashboard.

The literature is first explored to examine aviation incidents where additional weather data (i.e., cloud information from ground-based camera sites) might have been lifesaving to pilots. Also, the topics of artificial intelligence and machine/transfer learning are explored in the context of image classification and meteorological applications. Next, the methodology and datasets for a tiered research approach are described in which machine transfer learning is used to classify increasingly complex cloud types and sky conditions present in webcam imagery. Results of the tiered research cases are discussed in detail, and aviation hazard detection applications are evaluated for their viability in potential
operational use. Closing remarks and conclusions are presented as well as potential follow-up research opportunities.

LITERATURE REVIEW

AVIATION WEATHER ACCIDENTS. A consensus among studies in the peer-reviewed literature is that most aviation accidents caused by weather reside in the general aviation community (Chamberlain and Latorella, 2001; DeFilippis et al., 2018; Fultz and Walker, 2016; Johnson et al., 2019; King et al., 2017; NTSB, 2005). Research also reveals that within the general aviation sector, these incidents related to weather result in the highest rate of fatalities among all general aviation accident types (Chamberlain and Latorella, 2001; DeFilippis et al., 2018; Fultz and Walker, 2016). Furthermore, the studies demonstrate that the most common cause of weather-related aviation accidents is the unplanned encounter of Instrument Meteorological Conditions (IMC) during Visual Flight Rules (VFR) conditions (Chamberlain and Latorella, 2001; Goh and Wiegmann, 2001; NTSB, 2005). While ceiling and visibility reduction are predominant accident-driven factors as expressed in DeFilippis et al. (2018), Chamberlain and Latorella (2001) highlight the threat of convective weather to the general aviation population. In a study conducted from 1982-1993, data revealed that thunderstorms were a factor in less than five percent of general aviation incidents (Chamberlain and Latorella, 2001). However, of these small percentage of accidents, the majority (66%) of cases involved fatalities (Chamberlain and Latorella, 2001). Issues cited within their study include the unique challenge of convective weather for pre-flight planning.
and in-flight analysis because of its fast evolution and sparse coverage (Chamberlain and Latorella, 2001). While weather related events such as icing, hail, lightning, heavy rain, and winds/turbulence are of great interest to general aviation safety, flights near thunderstorms can also lead to ceiling and visibility reductions (Chamberlain and Latorella, 2001). In a more recent study by Fultz and Walker (2016), general aviation was also reported as the deadliest type of air travel with costs of accidents in the United States in the range of one to four billion dollars per year (Fultz and Walker, 2016). In their study conducted from 1982-2013, 60,000 general aviation accidents occurred in the United States resulting in 20,000 fatalities and comprising nearly 80% of all aviation fatalities (Fultz and Walker, 2016). Weather was found to be an element or principal factor in nearly a quarter of all cases with a small, but fatal, percentage of accidents related to convective weather (Fultz and Walker, 2016). Additionally, the data revealed that weather was a factor in around 35% of fatal general aviation accidents during the period of study.

Several issues pertaining to the reception of weather information during general aviation flights are discussed in the literature. In Johnson et al. (2019), the authors indicate that in-flight weather decisions made by general aviation pilots can be supplemented by radar images, pilot reports (PIREPS), and other information populated on their available electronic displays. However, much of this information, such as ground-based NEXRAD radar imagery, is received up to tens of minutes after the reported weather conditions were observed (Johnson et al.,
Chamberlain and Latorella (2001) emphasize the lack of in-flight information that exists for rapidly changing convective weather. While aircraft-based Doppler weather radar is a valuable tool, most general aviation aircraft are not equipped with this type of technology due to high costs and lengthy training procedures (Johnson et al., 2019). Pilot reports are typically verified, reputable sources of weather information pertaining to conditions such as icing and turbulence, but statistics reveal that general aviators do not submit many reports (Johnson et al., 2019). Automated weather information sources are available to pilots such as those from ASOS, AWOS, and ATIS, but this information can sometimes lack the spatial and temporal resolution critical to weather phenomena such as convective weather (Chamberlain and Latorella, 2001). Chamberlain and Latorella (2001) propose that pilots’ decision-making, awareness, and underlying safety could be improved by weather information that is more available, complete, and user-focused. Of particular interest is information that is real-time and displayed graphically such that it can be overlaid with other information that is already consumed by pilots.

Although additional information about aviation weather hazards could be beneficial to pilots, the literature reveals that too much weather information or a lack of interpretive knowledge of weather information is of equal concern. As stated in a 2005 NTSB article on aviation safety hazards, each time pilots fly, weather must be continually evaluated to identify conditions that may affect flight safety. This entails assessing all pertinent information such as physical views from
the cockpit as well as the weather observation, advisory, and, forecast resources discussed previously (NTSB, 2005). In a study by DeFilippis et al. (2018), their results showed that some pilots have trouble accurately interpreting current operational forms of graphical aviation weather information products. Of the various information-based products, pilots scored the lowest on their assessment of surface station plots as produced by ASOS/AWOS sites (DeFilippis et al., 2018). Goh and Wiegmann (2001) report similar findings that suggest many of these deadly general aviation incidents, such as unexpected flight into IMC, may be ascribable to a lack of situational awareness. In summary, an analysis of the literature reinforces the notion that weather can be a potentially deadly general aviation safety hazard and thus pilots could benefit from higher quality and more timely information. However, the literature also suggests that information overload and the potential for misinterpretation, perhaps resulting from excessive information, can be problematic. As such, supplementary weather information should be delivered in a way that is concise, readily available, and straightforward to understand.

**ARTIFICIAL INTELLIGENCE APPLICATIONS.** Although there is extensive literature on the subject, a brief introduction of artificial intelligence algorithms is presented here – particularly as it relates to the machine learning framework, transfer learning, and applications to meteorological image classification. Applications of artificial intelligence have increased dramatically over recent years – driven by the big data needs of large corporations such as
Google, Microsoft, Facebook, Apple, as well as the healthcare, automotive, and gaming industries (Abu et al., 2019; Kothari, 2018). Artificial intelligence encompasses a broad sector of the computer sciences that is interested in replicating complex thinking and decision-making processes conducted by the human brain (Zaccone, 2016). Of particular interest is machine learning which, according to Zaccone (2016), is based upon the study of algorithms and processes that are capable of interpreting data and gaining knowledge from them. A quickly emerging subset of machine learning is image classification, and a popular approach to this problem involves the use of deep learning through artificial neural networks (ANNs, Zaccone, 2016). The most common form of image classification involves ‘supervised learning’ for which a set of training images organized by desired output category are submitted to the algorithm for training (Zaccone, 2016). Prior to algorithm training, a portion of the full dataset is typically set aside for validation. Processes, such as deep learning through artificial neural networking, are then conducted in an effort to recognize patterns and features of images within the training set. Numerous model parameters (perhaps millions) are assigned weights that, in theory, constitute image features which allow for accurate distinguishing between user-organized image categories. The training process is iterative in the sense that the validation image set aids in the tuning of algorithm parameters in order to maximize performance and/or to allow the training process to terminate when validation accuracy has reached a desired threshold. Since the validation images have been withheld from the training processes, bias (i.e., overfitting)
during parameter tuning is minimized (Giorgos, 2021). Once training has commenced, classification of the validation set can be used to give estimates of model skill which, in turn, can aid in the selection of the right model architecture for the desired task (Giorgos, 2021). The completed system should be tested for accuracy using an independent dataset (often called a test set, and not to be confused with the validation set defined above). If the final validation accuracy indicates that the model parameters have been skillfully tuned and rigorous independent testing demonstrates stable classification performance for additional imagery, then model development can be advanced towards implementation for operational use (see flowchart, Figure 1). However, if the final validation accuracy significantly exceeds the independent testing performance, then overfitting has likely occurred during training (Giorgos, 2021).

Figure 1. Diagram depicting the workflow of supervised machine learning (modified from Zaccone, 2016).
As previously mentioned, artificial neural networking is a data processing method that is inspired by the structure of neurological circuits (Zaccone, 2016). For deep learning, the system is comprised of multiple layers of computational nodes, or units, which perform tasks and transformations in order to recognize and amplify characteristics of the training data (Zaccone, 2016). Convolutional neural networks (CNNs) have become a “dominant approach” for image classification (Zaccone, 2016). The pixels of a digital image become a three-dimensional input to CNNs (e.g., length, width, and color channels), and individual segments of the image undergo mathematical (e.g., matrix) transformations during a process called convolution (Zaccone, 2016). Additional layers, such as pooling, concatenation, and dropout layers work in tandem with convolutional layers to construct feature maps of increasingly higher detail (e.g., see Figure 2 in Zaccone, 2016). The CNN learns which features to recognize and how to recognize them during training (Pritt & Chern, 2017). In the final stages of the network, dense, or fully connected, layers and activation function layers are used. These layers convert the output information from previous layers, which vary in dimension, to a dimension which corresponds to the image categories that were organized by the user (i.e., the number of categories with their respective labels). Using the training data that has been organized into user-desired categories for supervised machine learning, the deep learning process acts to develop a scheme which uses recognized features to distinguish between these categories.
TensorFlow is a popular tool that is used for artificial intelligence and machine learning applications such as artificial neural networking for image classification. TensorFlow is a machine learning framework developed by Google which can be used for training and classification of images by employing a variety of machine learning architectures. CNN architectures available for use with TensorFlow include those such as Xception, VGG, ResNet, MobileNet, and Inception (Howard et al., 2017; Karim, 2019). These families of CNN deep learning architectures vary in algorithm construction, including the number and type of internal layers and resultant parameters used for feature extraction (Karim, 2019). For example, the MobileNet algorithms (e.g., MobileNet, MobileNetV2, and
MobileNetV3) are less computationally expensive and have been optimized to operate efficiently on less powerful devices (Howard et. al, 2017). Similar to the MobileNet family, Inception v3 is a multilayer convolutional neural network that makes use of more complex convolutional methods which result in a higher number of trainable parameters and larger computational expense (Howard et al., 2017; Karim, 2019). Despite the increased computational requirements, albeit less than some other architectures, Inception v3 can be applied to a variety of image classification tasks with minimal user manipulation, and it has been found to be quite accurate for many image classification applications.

Inception v3 was first proposed in a paper by Szegedy et al. (2015) and developed by Google, Inc. before its release in 2015. The CNN has over 40 layers with some comprised of many units (i.e., computational nodes) which allow for a parameter count of nearly 24,000,000 (Szegedy et al., 2015). As with many of the CNN architectures including those mentioned above, Inception v3 was tested for accuracy through training on the ImageNet dataset from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 (Szegedy et al., 2015). In this annual challenge, software developers use their machine learning algorithms to compete in the training/classification of a dataset containing 1,000 image classes comprised of a total of over 1,000,000 digital images. Inception v3 achieved a high level of accuracy when trained on this dataset (Szegedy et al., 2015), leading to widespread use of the model for diverse image classification purposes (e.g., Chowdary, Punn, Sonbhadra, & Agarwal, 2020; Guan et al., 2017; Lin, Li, Luo,
Many models that have been pre-trained on the ImageNet database are available for direct implementation in TensorFlow. These models can be used to classify images for up to 1,000 image classes that are represented by the ImageNet dataset without any additional training. However, to create a customized image classification scheme for specific image categories that are desired by the user (and perhaps not contained within the ImageNet dataset), the user can leverage the pre-trained model using a popular and effective machine learning technique referred to as transfer learning.

Transfer learning is a machine learning approach where previously trained models are refitted for new tasks (TensorFlow, 2022). As described above, CNNs typically involve many layers of transformations, each leading to extracted image features of increasingly higher detail. However, much of the ‘knowledge’ gained during the deep learning process is applicable to image categories that may vary in scope and type from the original dataset (TensorFlow, 2022). For example, parameter weights associated with image features such as edges, shading, curves, lines, textures, and shapes exist in all types of images (TensorFlow, 2022). While this explanation is an oversimplification, transfer learning allows the user to ‘retrain’ the pre-trained model by manipulating just the outermost layers of complex CNNs (i.e., the input data and final fully connected/activation function layers). One of the benefits of this approach is that it allows for relatively quick model development with minimal computational power. For the Inception v3 architecture, the TensorFlow authors created publicly-available Python code files...
named ‘retrain.py’ and ‘label_image.py’ which allow users to employ transfer learning with their own organized and labeled image datasets. The software is licensed under the Apache License, Version 2.0, where users are granted unrestricted copyright and patent licenses to reproduce, create derivative works, and sell/distribute any works that originate from these algorithms (The Apache Software Foundation, 2004). In the documentation for the retrain.py code file, authors share that the pre-trained Inception v3 model is unmodified besides replacement of a fully connected layer and ‘softmax’ layer (i.e., a type of activation function layer) at the end of the model which allows previous layer data to be applied to new input data (see Figure 3). The retrain.py file allows the user to initialize retraining (i.e., transfer learning) on user-selected image data that is organized in simple file folder structures as well as adjust algorithm settings related to image preprocessing, training/validation processes, and file pathways. The label_image.py file allows the user to input images and then inference, or generate a ‘prediction’ (i.e., based on levels of confidence), that the image belongs to any of the trained image categories. Inferencing is completed according to TensorFlow graph (.pb) and text (.txt) files produced during retraining that contain information about trained parameter weights and image category labels, respectively. Guan et al. (2019) utilized the Inception v3 model with transfer learning to classify cytological (i.e., single cell type) images for lymph node diseases. Lymph node cells were medically tested and diagnosed for diseases, and their corresponding
images were sorted into appropriate classes for retraining. Results revealed nearly 90% accuracy of test image classification (Guan et al., 2019).

Figure 3. Illustration of Inception v3 model structure and modification for transfer learning applications. The black ‘x’ covers the layers of the pre-trained model (trained on ImageNet ILSVRC 2012 data) that are removed and replaced by softmax and fully connected layers for classification of new image categories (Intel, 2019).

As with many industries such as healthcare, the use of machine learning in image classification has been rapidly expanding in the field of meteorology. Buch and Sun (1995) discuss a variety of approaches for classifying cloud imagery including texture and topological feature extraction, clustering, thresholding, and neural networking. In their work, the authors use binary decision trees to distinguish cloud types in whole-sky imagery. More recent studies focus on cloud image classification using deep learning methods, such as Zhang et al. (2018), who state that despite the impressive performance of CNNs for image classification,
there are few studies which have evaluated their applicability to cloud classification. The authors also share that one of the fundamental issues of using deep learning (e.g., for cloud classification) is the availability of sufficient training data on which the model is initialized. Zhang et al. (2018) conducted their studies using the Singapore Whole-sky Imaging Categories (SWIMCAT) database and the Cirrus Cumulus Stratus Nimbus (CCSN) database hosted by Harvard Dataverse. However, both datasets have limited amounts of low-resolution imagery (e.g., 125x125 pixels and 256x256 pixels for SWIMCAT and CCSN, respectively). Additionally, the imagery is divided into broad categories including clear sky, patterned clouds, and thick dark clouds in SWIMCAT, and the CCSN is organized such that the cumulus cloud category represents multiple low-base cumuliform cloud types (e.g., fair-weather cumulus and towering cumulus). Much of the literature regarding cloud image classification applications of machine learning is associated with whole-sky imager (WSI) data and is related to studies of Earth’s radiation budget, general circulation models (GCMs), and the solar energy sector (Wan and Du, 2020; Ye, Cao, and Xiao, 2017; Zhang et al., 2018; Zhuo, Cao, and Xiao, 2014). Ye, Cao, and Xiao (2017) cite a lack of research on machine learning applications to cloud imagery, motivating their work on evaluating the challenges associated with the use of feature detection for cloud typing. They state that varying cloud types exhibit different distinguishing features. For example, cumulus and altocumulus are well defined by their shape while cirrus, stratocumulus, and stratus are better defined by their texture (Ye, Cao, and Xiao, 2017). Interestingly,
the literature contains few, if any, examples of cloud image classification research that utilizes ground-based web camera imagery available from several sources to be described by this research. Additionally, there have been few studies whereby a machine learning approach is applied to cloud-image classification as it relates to aviation weather hazards. Yet, some ground-based web camera networks have archived data that are highly suitable for machine learning.

Pritt and Chern (2017) used deep learning techniques to develop an image classification scheme for object recognition in satellite imagery. Their work strives to enhance the use of machine learning for law enforcement and other entities by aiding in the identification and location of objects in satellite imagery such as seaports, boats, and many complex features of the built environment. To increase available data for the algorithm training process, duplicate (preprocessed) satellite images were used that were rotated 90, 180, and 270 degrees from their original format (Pritt and Chern, 2017). While this technique is applicable to satellite imagery with views that face directly downward (nadir) towards earth’s surface, it may not be beneficial to apply this style of preprocessing to ground-based webcam imagery where the sky, foreground, and horizon have distinct directionality in the images. However, the use of other preprocessing techniques (e.g., random image cropping) could potentially reduce model overfitting to webcam site-specific image contents such as image foreground. Results from numerous studies (including those referenced above) indicate that the use of machine learning for complex object and cloud recognition is certainly feasible. The classification of clouds and other
complex objects was found to be quite accurate, although challenges still exist. Heinle, Macke, and Srivastav (2010) applied masking techniques to training and validation images from whole-sky imager data to minimize errors related to washed-out pixels due to the sun and foreground structures. They state that cumuliform clouds of varying altitudes (e.g., cumulus versus altocumulus) and types of expansive cloud cover (e.g., stratocumulus versus stratus) were sometimes misinterpreted by the image classifier due to similarities in appearance. They also cite that the dynamic spatial-temporal nature of clouds is especially problematic with respect to automating the classification. Ultimately, much remains to be learned regarding the applications of these techniques as an aviation weather resource.

Figure 4. Illustration of feature recognition elements that are created by convolutional layers of deep learning for cloud images (from Zhang et al., 2018). The naming scheme at the top describes the convolutional layer number (e.g., Conv1, Conv2, etc.) and the digital media channel number from the layer (e.g., 12, 255, 384, etc., Zhang et al., 2018).
A large quantity of training data is needed to successfully develop an accurate machine learning classification scheme for digital imagery – especially images that contain highly variable image content such as clouds. While a transfer learning approach may mitigate some of this data burden, it does not circumvent the need for a large and diverse training dataset which is necessary to maximize model performance. As previously mentioned, if the user initializes an off-the-shelf machine transfer learning algorithm for image classification, the software typically has full control of the training processes and the resultant image features that are extracted to predict image contents. For some off-the-shelf algorithms, the user may have the ability to adjust basic settings for dataset preprocessing and training/validation. In the retrain.py code file for Inception v3 transfer learning, the TensorFlow authors provide code that allows the user to randomly preprocess images. Image preprocessing acts to augment images before the training process commences to improve image diversity and potential feature enhancement. In the code, the user has options to randomly crop, scale, flip, and change pixel brightness in training images. Additionally, the user can adjust model settings such as the number of training steps, the learning rate (i.e., part of parameter tuning), and the percentage of images that are designated as part of the validation set. The code also allows for designation of a percentage of images as part of a test set. These images are used to provide an independent final test of model accuracy after
training/validation completes. However, most users will require additional, much more rigorous independent testing and evaluation for robustness, especially before utilizing the model for operational purposes. In this research, all independent testing datasets to be described in the methodology are not associated with the test set feature available in the retrain.py algorithm (i.e., testing_percentage and test_batch_size in Table 1). Any validation and testing metrics provided automatically by the retrain.py algorithm have been used only for initial assessments, and therefore these details have been withheld from the discussion. For a full list of settings and their default values, refer to Table 1. As part of the inherent internal layer structure of the pre-trained Inception v3 model, all input images for training and inferencing are automatically resized (scaled) to a resolution of 299x299 pixels.
Table 1. List of settings that are available in the retrain.py algorithm. The parameter names as they appear in the code, a brief description of their function, and their input type/default values are provided.

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Description</th>
<th>Type</th>
<th>Default value</th>
</tr>
</thead>
<tbody>
<tr>
<td>flip_left_right</td>
<td>Whether to randomly flip half of training images horizontally</td>
<td>Boolean</td>
<td>False</td>
</tr>
<tr>
<td>random_crop</td>
<td>Percentage determining how much of a margin to randomly crop off of training images</td>
<td>integer</td>
<td>0</td>
</tr>
<tr>
<td>random_scale</td>
<td>Percentage determining how much to randomly scale up the size of the training images by</td>
<td>integer</td>
<td>0</td>
</tr>
<tr>
<td>random_brightness</td>
<td>Percentage determining how much to randomly multiply the training image input pixels up or down by</td>
<td>integer</td>
<td>0</td>
</tr>
<tr>
<td>architecture</td>
<td>Which model architecture to use (MobileNet or Inception v3)</td>
<td>string</td>
<td>‘inception_v3’</td>
</tr>
<tr>
<td>how_many_training_steps</td>
<td>How many training steps to run before ending</td>
<td>integer</td>
<td>4000</td>
</tr>
<tr>
<td>learning_rate</td>
<td>How large a learning rate to use when training</td>
<td>float</td>
<td>0.01</td>
</tr>
<tr>
<td>validation_percentage</td>
<td>Percentage of images to use as a validation set</td>
<td>integer</td>
<td>10</td>
</tr>
<tr>
<td>testing_percentage</td>
<td>Percentage of images to use as a test set</td>
<td>integer</td>
<td>10</td>
</tr>
<tr>
<td>eval_step_interval</td>
<td>How often (in steps) to evaluate the training results</td>
<td>integer</td>
<td>10</td>
</tr>
<tr>
<td>train_batch_size</td>
<td>How many images to train on at a time</td>
<td>integer</td>
<td>100</td>
</tr>
<tr>
<td>test_batch_size</td>
<td>How many images to test on (value of -1 causes entire test set to be used)</td>
<td>integer</td>
<td>-1</td>
</tr>
<tr>
<td>validation_batch_size</td>
<td>How many images to use in an evaluation batch for each iteration (value of -1 causes entire validation set to be used)</td>
<td>integer</td>
<td>100</td>
</tr>
</tbody>
</table>
Generally, as more training images are provided for each image category (e.g., cloud type), the classification scheme can better account for natural variability among the images. However, as the user is generally unable to control predictors and image elements recognized by the algorithm without deeper manipulation of the model architecture, accuracy of the classification scheme must be tested on a case-by-case basis and with a diverse variety of independent test images. The research in this study has been conducted using the retrain.py and label_image.py code files created by the TensorFlow authors. The settings in Table 1 were unchanged from their default values. The preliminary research was conducted by executing unmodified versions of the aforementioned Python code files (.py) within a Docker virtual container. While successful, this approach was somewhat limiting. As a result, the code was modified to run in a Jupyter Notebook Python environment (.ipynb) and subsequently integrated with other Jupyter Notebooks for automation and batch/real-time test image classification. Training/independent test image datasets and other methods for this research are described in detail below.

DATASETS: WEB CAMERA RESOURCES

There are several ground-based meteorological webcam networks that exist across the United States and North America. These webcam networks typically record and upload near real-time imagery (stills) for public viewing and download from various online resources. The United States Federal Aviation Administration (FAA) has a network of ground-based webcams that are prominent in regions such
as Colorado, Alaska, and parts of Canada. The FAA Aviation Weather Camera program is intended to provide aviators with supplementary information for current weather conditions and planning purposes (FAA, n.d.).

![Map of FAA WeatherCams](image)

Figure 5. The FAA WeatherCams website interface (FAA, n.d.). Green dots denote active camera sites with imagery available for viewing, and yellow dots indicate camera sites that are currently in maintenance mode.

A ground-based camera network is also operated by ALERT Wildfire. ALERT Wildfire has an abundance of webcam sites across the Western United States. This camera network was created to monitor wildfires in regions of complex terrain and sparse population. However, because many of these camera sites produce images that also consist of land and sky elements, secondary cloud information can also be gleaned from this imagery.
Figure 6. The Alert Wildfire website interface (Alert Wildfire, n.d.). Blue arrows indicate the location and orientation of available camera sites. The yellow arrow denotes the camera site/orientation that has been selected for viewing. Some camera sites pan/tilt and do not have fixed viewing angles.

The National Park Service (NPS) maintains a network of Air Quality Webcams that capture imagery from approximately twenty national parks. The cameras coincide with air quality and other meteorological instruments. Similarly, the imagery from the NPS Air Quality Webcams contains both land and sky elements, and thus cloud information can be extracted readily. The NPS Air Quality Webcams upload images once every fifteen minutes during the daytime at
most sites, and archived imagery is available for public download dating back to the time of installation for all camera sites.

Figure 7. The NPS Air Quality Webcams website interface (NPS, n.d. b). The webpage features a map of air quality web camera sites that are located at national parks around the United States. In the Southeast United States, the Great Smoky Mountains National Park region is comprised of three camera sites: Look Rock (facing east), Clingmans Dome (facing west), and Purchase Knob (facing northeast).

Airportview.net provides images from public and private airports in the United States of current weather conditions alongside text and graphical aviation weather information. A recent addition to this webcam network includes four cameras that were installed in May of 2022 at Florida Institute of Technology (FIT) Aviation at the Melbourne International Airport (KMLB, see Figure 8). The FIT
Aviation cameras produce real time high-resolution images (1280x720 pixels) of the sky and limited foreground in the north, east, south, and west-facing directions at a rate of one per minute. The cameras produce this imagery from 4:00 a.m. to 10:00 p.m. Eastern Time (ET) and switch to an IR nighttime mode when light levels darken below a threshold value. Because of this, limited amounts of nighttime sky imagery are available for archive and analysis. As part of this research, Jupyter Notebook Python code was written that automatically archives the images. Since its inception, the archives have grown to more than 325,000 images.

METHODOLOGY

As indicated, the methodology in this study uses an off-the-shelf machine transfer learning algorithm (Inception v3 architecture) to distinguish between different cloud types in webcam imagery. Ultimately, this research seeks to identify various cloud types relevant to aviation weather hazards (e.g., turbulence, low visibility, etc.). A tiered approach, comprised of four stages (Figure 9), was designed to limit the degrees of freedom associated with the training of an image classification scheme with respect to various challenging environmental factors including time of day, season, camera location, cloud types, etc. These constraints are systematically evaluated as part of the algorithm development process. In following sections, I provide a description of the four stages in Figure 9.
Figure 8. Images from four AirportView.net cameras installed at FIT Aviation (KMLB) on 22 August 2022 at 1:00 p.m. and 10:00 p.m. ET (left and right columns, respectively). The images on the right were captured using the nighttime IR mode. The rows are segregated based on the camera facing direction: a) north, b) east, c) south, and d) west.
Figure 9. Four-stage graphic describing the methodological approach to construct a cloud image classifier for this study.

The preliminary methodology (described in stages 1 and 2 below) has been conducted using algorithm training and independent test image data from the National Park Service Air Quality Webcams network. The availability of downloadable archived imagery is highly desirable for constructing training datasets critical to the machine learning process (the AirportView.net archives for FIT Aviation had not yet been developed for this research). Using archived imagery, images can be mined for cloud types of interest. As previously mentioned, training requires large data sets as they are more likely to capture the inherent variabilities in the images such as those related to physical cloud appearance, proximity to the camera, time of day, and season. The Look Rock camera site within the Great Smoky Mountain National Park was chosen for the preliminary
study presented here. The region was chosen because of its location in the Southeastern United States where summertime cumulus fields, towering cumulus, and thunderstorms commonly occur. Cloud types such as towering cumulus and cumulonimbus are often associated with turbulence and other aviation weather hazards – making this area an attractive location to test the transfer learning approach. The Look Rock camera site produces images with a large portion of sky visible within the camera’s field of view and minimal foreground obstructions. Interestingly, the Smoky Mountains National Park region is among areas in the United States where official NWS/FAA surface observations are lacking. As a result, this research also serves to better understand how webcam imagery in regions like these could be used to supplement gaps in observational coverage. Figure 10 illustrates the lack of official surface observation sites within the Great Smoky Mountains National Park region and the complementary gap-filling nature of the ground-based camera stations. For supplementary information about the Look Rock camera site, see Figure 11.

**METHODS: STAGE 1.** Determine whether or not it is feasible to use an off-the-shelf transfer learning framework like Google’s TensorFlow and Inception v3 architecture to identify cloud types. Initially, images from the Look Rock camera site were used to test the transfer learning capabilities. The images were collected and organized into three categories – each containing approximately several dozen images including cumulus clouds, obscured view, and clear sky. At this stage, the cumulus cloud category represented all low-base cumuliform clouds
of any vertical extent (e.g., fair-weather cumulus and towering cumulus). An Inception v3 transfer learning model was initiated to retrain on these three image categories. The accuracy of the algorithm was then tested using several independent images of each type. As the first stage was only used to conduct an initial feasibility check, additional details about the methodology are not included here. An example and short discussion of results are presented in the stage 1 results section below.

![Map of the Great Smoky Mountains National Park region. The solid black dots denote official NWS/FAA ASOS and AWOS sites, some of which display wind barb and surface temperature observations for 21:48 UTC on 21 January 2022. The shaded ellipse depicts the camera sites that are located within the Great Smoky Mountains National Park. The three labeled camera icons denote the air quality web camera locations that are available for public viewing and download via the NPS Air Quality Webcams website.](image-url)
Figure 11. Supplementary information for the Look Rock camera site provided on the NPS Air Quality Webcams site (NPS, n.d. a): a) geographical landmarks visible in the camera’s field of view, b) annotated map depicting the camera orientation and other park information, c) comparison of good and bad visibility days due to differences in air quality.
**METHODS: STAGE 2.** Determine whether or not transfer learning with Inception v3 can be used to distinguish between two similar types of clouds at a single camera site, similar time of day (i.e., solar angle), and similar season (to avoid foreground inconsistencies). As part of this stage, 100 images of both fair-weather cumulus and towering cumulus were selected from the NPS Air Quality Web Camera archives at the Look Rock site. During the data collection process, it was discovered that the Look Rock camera underwent an upgrade in 2018 that resulted in a change in the exposure level, aspect ratio, and resolution of uploaded images (Figure 12).

![Figure 12](image). Digital images produced by the Look Rock NPS Air Quality Webcam site before the 2018 camera upgrade (left) and after the 2018 camera upgrade (right), taken at similar times of day.

To circumvent any problems and/or issues that might be associated with the change in the image characteristics, and to maximize data availability, images were selected from pre-2018 archives for a nine-year period (2010-2018). Images were selected from the late spring and summer seasons only – so as to ensure that the
foreground contained foliage of a consistent appearance. However, the impact of foreground appearance on algorithm training is not entirely known and is thus an element of this research. To minimize impacts on the algorithm training process that are related to time of day, image selection was limited from mid-morning to late afternoon (approximately 9 a.m.-5 p.m.). Because sunlit elements within the images are also a function of camera orientation (e.g., direction of lens in the vertical and horizontal planes) as well as clouds or other shadow producing features, lighting induced image variations could not be entirely avoided (see findings in stage 2 results section). Two image resolutions are available (pre-2018) on the National Park Service Air Quality Webcams webpage: 1600x1200 pixels (original) and 1000x750 pixels (medium). Additional image-related information is provided in Figure 13. Training datasets of 100 images for fair-weather cumulus and towering cumulus clouds as well as 25 independent test images for each cloud type were collected at both original (i.e. high) and medium resolutions (e.g., see Figure 14). With the exception of resolution, both training and test image sets for each cloud type are identical. This isolates the impact (if any) of the user-input image resolution on the training and classification processes. Using the identical training sets of 100 fair-weather cumulus and 100 towering cumulus images at high and medium resolutions, transfer learning was conducted to create two image classifiers. The two image classifiers were then tested for accuracy using the 25 independent test images of each cloud type in both high and medium resolutions (see Table 2 for a visual breakdown of training and testing datasets). In doing so,
models were used to classify images matching their respective training image resolution (e.g., the model trained on high resolution images classifying high resolution test images) and of opposite training image resolution (e.g., the model trained on high resolution images classifying medium resolution test images).

Figure 13. Metadata contained within the Look Rock image files: a) prior to the 2018 camera upgrade and b) after the 2018 camera upgrade. The metadata provides information about image resolution for the original (i.e., high) and medium resolution images, camera type, and automatic exposure settings (e.g., exposure time and ISO speed). For a fair comparison of the exposure settings, the respective example images were taken from a similar month/time of day and during full-sunlight conditions. Some metadata fields not provided within the image files are left blank.
Figure 14. Training images and independent test images from pre-2018 web camera data at the Look Rock NPS Air Quality Webcam site located in the Great Smoky Mountains National Park. Training and testing images were selected from the late spring and summertime and range from mid-morning to late afternoon. Shown are cumulus clouds (left column) and towering cumulus clouds (right column).
Table 2. Breakdown of all image data used in stage 2: a) high and medium resolution image data (fair-weather cumulus and towering cumulus) that were used to train two image classifiers, b) resolution-based datasets of each cloud type used for classifier testing.

<table>
<thead>
<tr>
<th>a) Training</th>
<th>b) Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Classifier</td>
<td>Training data</td>
</tr>
<tr>
<td>Image classifier 1 (high resolution)</td>
<td>Fair-weather cumulus: high resolution</td>
</tr>
<tr>
<td></td>
<td>Towering cumulus: high resolution</td>
</tr>
<tr>
<td>Image classifier 2 (medium resolution)</td>
<td>Fair-weather cumulus: medium resolution</td>
</tr>
<tr>
<td></td>
<td>Towering cumulus: high resolution</td>
</tr>
<tr>
<td></td>
<td>Fair-weather cumulus: medium resolution</td>
</tr>
<tr>
<td></td>
<td>Towering cumulus: medium resolution</td>
</tr>
</tbody>
</table>

**Bias Minimization Practices: Stage 2.** For supervised machine learning to be possible, the user must provide the “truth” (i.e., the category that images correctly belong to) for both the training and testing images. While the truth can be fairly easily applied to an image classification scheme containing apples and oranges, for instance, challenges can arise when attempting to provide the truth for
images of clouds. Clouds, which are inherently chaotic in nature, can exist in mixed forms, and they may not always fit idealized meteorological groupings. To minimize subjectivity and bias related to the selection of both training and test images, a review session was conducted within the research advisory committee to quality-control these images for their respective cloud types (i.e., fair-weather cumulus versus towering cumulus). Fairly ‘pristine’ examples of each cloud type were chosen for this preliminary study such that images containing a mix of other cloud types not relevant to the study were, for the most part, not included. In addition, images containing cumulus clouds with slight vertical development (e.g., cumulus mediocris in Figure 15) were also filtered. However, a clear distinction between cloud types is not always possible and is one of several challenges of this study (see stage 2 results).

Figure 15. The three stages of convective (i.e., cumuliform) clouds (Learnweather, 2021). Fair-weather cumulus clouds (cumulus humilis) have little vertical development while towering cumulus (cumulus congestus) have deep vertical development and may begin to produce precipitation and/or lightning as they mature into cumulonimbus. As convective clouds develop vertically, they may exist in intermediate stages (cumulus mediocris) where cloud appearance cannot be classified definitively as fair-weather cumulus or towering cumulus.
METHODS: STAGE 3. Using the limited conditions described above for stage 2, the image classification framework is expanded to include the identification of additional cloud types and sky conditions: clear sky, cirrus clouds, fair-weather cumulus clouds, fair-weather cumulus clouds with the presence of middle/high clouds, towering cumulus clouds, cumulonimbus clouds, and precipitation. Using recent archived imagery from the AirportView.net cameras installed at FIT Aviation at the Melbourne International Airport (KMLB), the seven categories of images were selected for the west-facing camera site from 25 May 2022 to 9 July 2022 from approximately 9:00 a.m. to 4:00 p.m. Eastern Time (ET). Although images from all camera sites were used for preliminary testing of model universality (stage 3 results), the west-facing camera site was chosen, over that of the other three cameras, for model development because its field of view is oriented towards the interior of the Florida peninsula (as opposed to the nearby intracoastal waterway and Atlantic Ocean for the east-facing camera, see Figure 16). This decision was motivated by inland convective development, especially in the summer months. Hence, in addition to fair-weather cumulus, the west-facing camera captures developing towering cumulus, and cumulonimbus which are clouds of interest for aviation hazards. The east-facing camera view, for example, was found to be less desirable for model development because of prevailing clear skies (i.e., less image category diversity) that often resulted from an inland progressing sea breeze. Universal model development to achieve accurate classification results for all FIT Aviation camera orientations and even non-
collocated camera locations is of high interest (see fourth stage methods and conclusions sections).

Figure 16. Location of four AirportView.net cameras stationed at FIT Aviation at the Melbourne International Airport (KMLB) in Melbourne, Florida. The star indicates the location of the equipment tower where the four cameras are mounted, and transparent red triangles indicate the approximate field of view of each camera. The darker triangle indicates the approximate orientation of the west-facing camera.

Information about the seven image categories that were used to create the training dataset and respective image quantities are described in Table 3. Cloud
images were organized into two main classes: nonhazardous (i.e., clear sky, cirrus, fair-weather cumulus, and fair-weather cumulus with the presence of middle/high clouds) and those that may pose a hazard to aviation (i.e., towering cumulus, cumulonimbus, and precipitation). While fixed quantities of imagery per category were employed in the second stage methodology (100 images), this technique was not utilized in the third stage in order to maximize available training data for each image category. As a result, the counts of images in the seven classes represent the approximate prevalence of categories in the archived dataset. Altocumulus, stratus, and stratocumulus were not included in model development due to insufficient training data captured during the archived period. Convective (i.e., cumuliform) clouds of interest were divided into four categories: fair-weather cumulus, fair-weather cumulus with the presence of middle/high clouds (e.g., cirrus and altocumulus), towering cumulus, and cumulonimbus. Fair-weather cumulus clouds were broken into two sub-categories (i.e., with and without the presence of higher clouds) due to a high frequency of both cases within the archived data and noticeable differences in visual appearance. This organization serves to better discern whether model accuracy may be affected by the addition of image classes that reflect mixed cloud types. If cumulus clouds could not be classified as fair-weather cumulus or towering cumulus (e.g., as discussed in Figure 15 of stage 2), then the images were not included in the training dataset. With the exception of the two fair-weather cumulus cloud sub-categories, all other mixed cloud cases were organized in a manner that prioritized hazardous cloud types. For example, if
towering cumulus clouds were present within an image also containing higher clouds and/or fair-weather cumulus clouds, the image was labeled as the hazardous cloud type (e.g., towering cumulus). If towering cumulus and cumulonimbus clouds were both present in an image, the image was labeled as the more hazardous cloud type (e.g., cumulonimbus). Cumulonimbus clouds were labeled based upon the presence of an anvil and/or rain shaft extending from the cloud base to the ground/horizon. Images were labeled as precipitation if rain was falling at the camera location (e.g., falling raindrops visible in the image and raindrops on the camera lens) or if rain was close enough in proximity to the camera site that the rain comprised the entire field of view. All cirriform clouds types (e.g., cirrostratus and cirrocumulus) were included cirrus cloud category (see Figure 17). As in stage 2 and for all instances of supervised machine learning, the user is responsible for labeling (i.e., providing a ‘truth’ for) the image data. Practices and procedures used to reduce subjectivity and maintain a consistent organizational scheme are described in more detail in the stage 3 bias minimization practices section.

Table 3. Counts for each of the image categories that comprise the stage 3 training dataset (left to right): clear sky, cirrus clouds, fair-weather cumulus clouds, fair-weather cumulus clouds with the presence of higher clouds, towering cumulus clouds, cumulonimbus clouds, precipitation.

<table>
<thead>
<tr>
<th>Hazard</th>
<th>Nonhazardous</th>
<th></th>
<th>Hazardous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image type</td>
<td>Clear</td>
<td>Ci</td>
<td>Cu</td>
</tr>
<tr>
<td>Image count</td>
<td>255</td>
<td>1199</td>
<td>524</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 17. Representative images for the seven categories used to create the stage 3 training dataset. All images were collected from the AirportView.net KMLB west-facing camera from 25 May 2022 to 9 July 2022 from approximately 9:00 a.m. to 4:00 p.m. Eastern Time (ET). Left column from top to bottom: clear sky, fair weather cumulus, towering cumulus, precipitation. Right column from top to bottom: cirrus, fair-weather cumulus with the presence of higher clouds, cumulonimbus.
Following development of the stage 3 training dataset, transfer learning was conducted using the retrain.py algorithm. Additional (Python) code was developed to perform independent testing on a large dataset and to conduct near real-time model applications for assessment. The code, which was written as a Jupyter Notebook, automatically extracts images from the AirportView.net site (every minute, when available) from the west-facing camera at FIT Aviation. Using integrated code from label_image.py, inferencing is performed on these images, and the output is automatically logged to a live, publicly accessible Google Sheets document using a Google Cloud API service account and the gsread Python library (see Figure 18). Output levels of confidence (in decimal format) are presented for the seven image categories and always sum to 1.00 (100%). As a QC check, the algorithm also uses its image classification output to detect whether or not images are unique in order to avoid the presence of duplicate images in the archived dataset (e.g., pulling the same file when a new image is not available). The operational real-time image classification has been running nearly continuously, between the hours of 4:00 a.m. and 10:00 p.m., since August 2022. However, in order to keep the variations in lighting within the defined stage 3 restrictions, the training data is comprised only of images captured from approximately 9:00 a.m. to 4:00 p.m. Classification results for images taken outside the training data time restrictions are discussed in stage 3 results. The automatic, near real-time classification allows for the tracking of model performance. More than 50,000 images have been classified and archived for the west-facing camera site.
Figure 18. A screenshot of the near real-time, publicly accessible Google Sheets document used to log image classification output for the AirportView.net west-facing camera site at FIT Aviation. Shown is the most recent image (upper left) and the associated classification results for the seven image categories (bottom). As new images are processed, the classification percentages are appended at the top of the Google Doc (spreadsheet).

The archived images were filtered to contain data from 9:00 a.m. to 4:00 p.m. ET only. After 4:00 p.m., the afternoon sun became visible in the field of view of the west-facing camera (for more on the solar impact on the images, see stage 3 results). The time-of-day filtered data consisted of approximately 15,000 images of which 1,000 were randomly selected for analysis. These images were categorized according to the same organizational procedures as described for the training dataset development, including bias minimization. Table 4 shows the distribution of cloud categories present in the randomly selected independent test images. Since the independent test images were randomly selected from all data available within the prescribed time restriction, the exact cloud type in each image was initially
unknown. Of the 1,000 images, 254 were not included in the analysis because they contained cumuliform clouds in an intermediate stage of convective development (Figure 15). As part of the methodology for collecting the training data, these intermediate convective clouds were not included as one of the seven image categories, and therefore they were also not included for testing. Additionally, 146 of the images contained cloud types that could not be distinguished or that contained clouds that were not included in the training dataset (e.g., altocumulus, stratus, stratocumulus, etc.). An analysis of classification accuracy for the remaining 600 test images is explored in the stage 3 results section.

Table 4. Counts for each of the image categories that comprise stage 3 the test image dataset. 1000 images were randomly selected from 4 August 2022 to 2 November 2022 from 9 a.m. to 4 p.m. ET. Cells shaded blue (orange) denote nonhazardous (hazardous) image categories (Table 3). The yellow-shaded cells denote images that could not be categorized including: Cu med. (e.g., cumulus mediocris, Figure 15) and N/A (e.g., cloud types not included in the training dataset).

<table>
<thead>
<tr>
<th>Image type</th>
<th>Clear</th>
<th>Ci</th>
<th>Cu</th>
<th>Cu w/ high clouds</th>
<th>TCu</th>
<th>Cb</th>
<th>Precip</th>
<th>Cu med.</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image count</td>
<td>49</td>
<td>54</td>
<td>166</td>
<td>161</td>
<td>59</td>
<td>103</td>
<td>8</td>
<td>254</td>
<td>146</td>
</tr>
<tr>
<td>Total: 600</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total: 400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bias Minimization Practices: Stage 3. To minimize subjectivity and bias while providing a “truth” (i.e., category) for training and independent images, techniques in stage 2 were adapted, and additional methods were established. A useful source of information for cloud typing is the NOAA Sky Watcher Chart (NOAA, n.d. b, Figure 19). The chart provides images and a brief description of the
visual properties for numerous cloud types. The chart also provides standardized cloud-base height thresholds for high clouds (e.g., cirrus), middle clouds (e.g., altocumulus), and low clouds (e.g., cumulus). If middle or high clouds (exact type uncertain) within the camera field of view were reported by nearby ASOS sites (e.g., Orlando International Airport (KMCO) to the northwest of Melbourne), then cloud base heights should be available in the ASOS data stream (the Iowa State University IOWA Environmental Mesonet (IEM) has an archive). These data can be used for additional verification. Similarly, available resources were used to minimize human error in the categorization of cumulonimbus cloud and precipitation images. To accomplish this, the NEXRAD reflectivity (0.5 degree elevation) from NOAA’s National Centers of Environmental Information (NCEI) archive was used to determine whether or not rainfall echoes existed in proximity of the west-facing camera. The volume scan nearest in time to the images was used (Figure 20). If clouds were vaguely visible on the distant horizon of a prospective image such that they occupied an insignificant portion of the image, the clouds were deemed as far enough away to have no immediate impact on aviation operations in the proximity of the camera site. These clouds were ignored during the image categorization process (e.g., see Figure 21). When organizing independent testing images into the seven image categories, output classification results (i.e., the model predictions for the cloud types corresponding to each image) were not visible to the human organizer. This ensured that classification output did not influence the human decisions made for image categorization. Ultimately, a
clear distinction between cloud types is not always concrete in nature, but human-supplied organization (i.e., truth) for training and testing images is a fundamental, unavoidable part of supervised machine learning. The aforementioned practices have been implemented to maximize reproducibility while minimizing potential human biases.

Figure 19. [NOAA Sky Watcher Chart](http://www.weather.gov/hq/NOACLOUDchart.pdf) with information regarding the visual characteristics and cloud base heights of various cloud types (NOAA, n.d. b).
Figure 20. A stage 3 training data image (left) captured by the AirportView.net west-facing camera at FIT Aviation (KMLB) on 6 June 2022, at 3:20 p.m. ET that was categorized as precipitation. The right panel shows archived NEXRAD mosaic reflectivity imagery from the NOAA National Centers of Environmental Information (NCEI) for 6 June 2022, at 3:20 p.m. ET. The star indicates the approximate position of the FIT Aviation camera site.

Figure 21. An independent test image from stage 3 that was labeled ‘clear sky’. Clouds (type unknown, outlined in red), partially visible along the distant horizon, were determined to be an insignificant element of the image based on the stage 3 methodology (i.e., far enough away to not impact aviation operations in proximity of the camera site location).
METHODS: STAGE 4. The goal is to determine whether a machine learning classification scheme can be developed to successfully distinguish between a variety of relevant cloud types at different camera sites. A single algorithm avoids the expensive overhead related to the training, development and testing on a site-by-site basis. In addition, the image classification is expanded to allow for varying time of day, different foreground/season, and changes in the camera field of view and/or resolution (for more details see the conclusions section).
RESULTS AND DISCUSSION

RESULTS: STAGE 1

The preliminary testing for stage 1 revealed that the image classifier was generally accurate, with correct/high output levels of confidence for the majority of independent test images (e.g., see Figure 22 and Table 5). Given this initial success, the next tier (stage 2) was then evaluated, i.e., “Can transfer learning be used to distinguish between two similar types of clouds under ideal constraints?” (Figure 9).

Figure 22. Three examples of independent test images: cumulus, clear sky, and obscured view (left-to-right, respectively) that have been inferred by a cloud-image classifier that was trained on these image categories. See Table 5 for classification results.
Table 5. Classification results from the three independent test images in Figure 22. The columns denote the type of independent test image classified by the trained algorithm. Image classifier output assigns a level of confidence to the test image (in decimal form, multiplied by one hundred for percent) for each of the image categories used in training. The confidence values in each column add up to 1.00, or 100%.

<table>
<thead>
<tr>
<th>Classified as:</th>
<th>Cumulus clouds</th>
<th>Clear Sky</th>
<th>Obscured view</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulus clouds</td>
<td>0.9966</td>
<td>0.0011</td>
<td>0.0015</td>
</tr>
<tr>
<td>Clear sky</td>
<td>0.0026</td>
<td>0.9913</td>
<td>0.0002</td>
</tr>
<tr>
<td>Obscured view</td>
<td>0.0008</td>
<td>0.0076</td>
<td>0.9983</td>
</tr>
</tbody>
</table>

RESULTS: STAGE 2

Image classifiers were trained to distinguish between fair-weather cumulus and towering cumulus clouds for high and medium resolution NPS Air Quality Webcam imagery. Results have been organized into a table format where the training/independent test image resolutions used to produce the statistics are given in the column headers of Tables 6-9. A threshold approach is used to convert confidence levels from the raw image classifier output into an identified cloud type. To accomplish this (and test the sensitivity of the algorithm), three threshold levels of confidence were selected, 0.70 (70%), 0.80 (80%), and 0.90 (90%) whereby a cloud type is assumed present in the image if the particular threshold is exceeded. The results are validated, separately for each threshold, by comparing the hits (i.e., the particular cloud type was identified by the classifier and present in the image), misses (i.e., the classifier identified the wrong cloud type) those that were classified as unidentified (i.e., the classifier was unable to assign a cloud type because the
output confidence levels fell below the specified confidence threshold). Figure 23 shows the cumulative performance of all image classifications that were conducted to produce Tables 6-9. To reiterate the methodology used at this stage, only ‘pristine’ example images of each cloud type were included in the training and test image datasets. Prospective training and test images that contained cumulus or towering cumulus and additional (i.e., mixed) cloud types (e.g., cirrus or altocumulus) and images containing less distinct examples of each cloud type were generally avoided.
Table 6. Image classifier performance for an algorithm trained using original (high) resolution images. 25 independent high resolution test images each of Cu and TCu were classified and assigned a cloud type based on whether or not they exceeded the selected thresholds (0.70, 0.80, and 0.90). The percentage of images identified correctly, incorrectly, and unidentified are given by the green, red, and gray shaded cells, respectively.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a. Level of confidence for identification: 0.70 or greater</strong></td>
<td>Classified as:</td>
</tr>
<tr>
<td></td>
<td>Cu</td>
</tr>
<tr>
<td>Input:</td>
<td>Cu</td>
</tr>
<tr>
<td></td>
<td>TCu</td>
</tr>
<tr>
<td><strong>b. Level of confidence for identification: 0.80 or greater</strong></td>
<td>Classified as:</td>
</tr>
<tr>
<td></td>
<td>Cu</td>
</tr>
<tr>
<td></td>
<td>TCu</td>
</tr>
<tr>
<td><strong>c. Level of confidence for identification: 0.90 or greater</strong></td>
<td>Classified as:</td>
</tr>
<tr>
<td></td>
<td>Cu</td>
</tr>
<tr>
<td></td>
<td>TCu</td>
</tr>
</tbody>
</table>
Table 7. Image classifier performance for an algorithm trained using original (high) resolution images. 25 independent medium resolution test images each of Cu and TCu were classified and assigned a cloud type based on whether or not they exceeded the selected thresholds (0.70, 0.80, and 0.90). The percentage of images identified correctly, incorrectly, and unidentified are given by the green, red, and gray shaded cells, respectively.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a. Level of confidence for identification: 0.70 or greater</strong></td>
<td>Classified as:</td>
</tr>
<tr>
<td></td>
<td>Cu</td>
</tr>
<tr>
<td>Input:</td>
<td></td>
</tr>
<tr>
<td>Cu</td>
<td>80%</td>
</tr>
<tr>
<td>TCu</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>b. Level of confidence for identification: 0.80 or greater</strong></th>
<th>Classified as:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input:</td>
<td>Cu</td>
</tr>
<tr>
<td>Cu</td>
<td>72%</td>
</tr>
<tr>
<td>TCu</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>c. Level of confidence for identification: 0.90 or greater</strong></th>
<th>Classified as:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input:</td>
<td>Cu</td>
</tr>
<tr>
<td>Cu</td>
<td>52%</td>
</tr>
<tr>
<td>TCu</td>
<td>0%</td>
</tr>
</tbody>
</table>
Table 8. Image classifier performance for an algorithm trained using medium resolution images. 25 independent medium resolution test images each of Cu and TCu were classified and assigned a cloud type based on whether or not they exceeded the selected thresholds (0.70, 0.80, and 0.90). The percentage of images identified correctly, incorrectly, and unidentified are given by the green, red, and gray shaded cells, respectively.

<table>
<thead>
<tr>
<th>Image Classifier Performance</th>
<th>Training resolution: Medium, Test image resolution: Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a. Level of confidence for identification: 0.70 or greater</strong></td>
<td></td>
</tr>
<tr>
<td>Classified as:</td>
<td>Cu</td>
</tr>
<tr>
<td>Input: Cu</td>
<td>80%</td>
</tr>
<tr>
<td>TCu</td>
<td>0%</td>
</tr>
<tr>
<td><strong>b. Level of confidence for identification: 0.80 or greater</strong></td>
<td></td>
</tr>
<tr>
<td>Classified as:</td>
<td>Cu</td>
</tr>
<tr>
<td>Input: Cu</td>
<td>68%</td>
</tr>
<tr>
<td>TCu</td>
<td>0%</td>
</tr>
<tr>
<td><strong>c. Level of confidence for identification: 0.90 or greater</strong></td>
<td></td>
</tr>
<tr>
<td>Classified as:</td>
<td>Cu</td>
</tr>
<tr>
<td>Input: Cu</td>
<td>52%</td>
</tr>
<tr>
<td>TCu</td>
<td>0%</td>
</tr>
</tbody>
</table>
Table 9. Image classifier performance for an algorithm trained using medium resolution images. 25 independent high resolution test images each of Cu and TCu were classified and assigned a cloud type based on whether or not they exceeded the selected thresholds (0.70, 0.80, and 0.90). The percentage of images identified correctly, incorrectly, and unidentified are given by the green, red, and gray shaded cells, respectively.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a. Level of confidence for identification: 0.70 or greater</strong></td>
<td></td>
</tr>
<tr>
<td>Classified as:</td>
<td>Cu</td>
</tr>
<tr>
<td>Input: Cu</td>
<td>72%</td>
</tr>
<tr>
<td>TCu</td>
<td>0%</td>
</tr>
</tbody>
</table>

| **b. Level of confidence for identification: 0.80 or greater** | | |
| Classified as:                  | Cu  | TCu | Unidentified |
| Input: Cu                       | 68% | 4%  | 28%          |
| TCu                             | 0%  | 96% | 4%           |

| **c. Level of confidence for identification: 0.90 or greater** | | |
| Classified as:                  | Cu  | TCu | Unidentified |
| Input: Cu                       | 60% | 4%  | 36%          |
| TCu                             | 0%  | 88% | 12%          |
The results suggest that image classifiers trained on both high and medium resolution imagery produced relatively consistent results when classifying both high and medium resolution test images. Independent test images containing fair-weather cumulus clouds were positively identified approximately 75% of the time at the 0.70 confidence threshold. Classification performance for fair-weather cumulus clouds diminished with increasingly strict confidence thresholds such that only approximately 55% of test imagery was positively identified at the 0.90 confidence threshold. The net result of an increasingly higher confidence threshold
is a systematic shifting of correct identifications to the unidentified category as shown in Figure 23.

Towerling cumulus cloud classification was very accurate among the resolution-dependent results shown by Tables 6-9. At the 0.70 confidence threshold value, all but one independent test image containing towering cumulus clouds were correctly identified. When both the training and test data were applied to medium resolution imagery, the classifier had a perfect (100%) correct identification rate at both the 0.70 and 0.80 confidence thresholds (Table 8, a and b). Cumulatively, this rate decreased slightly with increasingly strict confidence thresholds, but over 85% of towering cumulus test images were correctly identified at the 0.90 confidence threshold level.

Of the test images containing fair-weather cumulus clouds, 1 of 25 (4%) were incorrectly identified as towering cumulus – a trend that was consistent regardless of resolution or confidence threshold. Interestingly, a single image was responsible for the misidentification which is manifest at all resolutions and thresholds (i.e., the red shaded cells of 4.0% in Tables 6-9). The offending image and the average confidence levels for each cloud type are shown in Figure 24. There is a very high level of confidence (0.999) that the test image contains towering cumulus clouds. Of all 25 independent test images for fair-weather cumulus, Figure 24 was taken at the very beginning of the allowed window for time of day (9:00 a.m. to 5:00 p.m.). As the image was taken in the mid-morning (9:15
am), the east-facing camera site resulted in a scene that was backlit. Lens flares and other artifacts are also visible in the image as annotated on the figure.

![Image with annotations]

Figure 24. The image from the Look Rock camera site in the Great Smoky Mountains National Park taken on 12 May 2016 at 9:15 a.m. ET and responsible for the ‘incorrect’ identification of Cu as TCu. This single image comprises the 4.0% of cumulus imagery (1/25) that was incorrectly identified in Tables 6-9. The average level of confidence produced by the image classifiers for fair-weather cumulus and towering cumulus clouds are 0.001 and 0.999, respectively. See text for details.

Because fair weather cumulus clouds at the Look Rock site are typically most common during the afternoon (associated with diurnal heating), little in the way of training or test data was collected for the time of day shown in Figure 24. The lack of training data for cumulus clouds in these lighting conditions likely
resulted in trained model behavior that performs poorly for images taken at this time of day. In addition, it appears as if there might be cumuliform clouds in various stages of convective development in the background of the image. This issue highlights the challenges of quality control when selecting training and test images. Although some measures (e.g., a collaborative review sessions conducted by the research committee) have been implemented to reduce subjectivity and bias, quality control is challenging. The behavior motivates research on the addition of image categories that reflect mixed cloud types and/or convective (cumulus) clouds of intermediate vertical development.

An image that was not selected for algorithm training or testing is shown in Figure 25. Through analysis and quality control during the selection of training/test images, it was determined that this image contains cumuliform clouds in the cumulus mediocris stage (see Figure 15). For testing purposes, a high-resolution version of this image was classified using the high-resolution image classifier. As indicated in Figure 25, confidence was split relatively evenly between each cloud type. Like Figure 24, classification of this image illustrates some of the challenges of using machine learning for cloud type classification and has served as motivation for the additional research conducted herein.

In summary, an assessment of classifier performance reveals that results are generally accurate for this stage 2 case study, and no major performance differences exist among classification schemes that utilize different training/testing
image resolutions. Since the retrain.py and label_image.py code files automatically resize images to 299x299 pixels before training and test image classification, this behavior was generally expected. Despite automatic resizing of all images, the statistics presented in Tables 6-9 exhibit some sensitivity to image resolution of the training and test data sets. As briefly discussed above, the user does not have control of the image elements recognized by the algorithm during classifier training, and so testing on a large and diverse image data set is critical to understanding how the classifier responds and performs. This is especially true given the infinite degrees of freedom (i.e., each image is unique) such as cloud location (e.g., top of image, center, horizon), cloud appearance due to factors such as lighting and texture, etc. Additional research will be required to understand the impact of these and other relevant issues (e.g., lens artifacts, precipitation) that might affect classification accuracy.
Figure 25. Training/independent test image candidate taken on 4 May 2012 at 1:45 p.m. ET that was rejected during quality control as the image was determined to contain mixed cloud types, including clouds existing in a state between fair-weather cumulus and towering cumulus (see Figure 15 and stage 2 methods section) Confidence values are split relatively evenly between fair-weather cumulus and towering cumulus at 0.575 and 0.425, respectively.

RESULTS: STAGE 3

For image classifiers consisting of several or numerous image categories (e.g., the Inception v3 model retrained for stage 3 or the Inception v3 model trained on the ImageNet ILSVRC 2012 database), the top-1 and top-n accuracies (described below) are common metrics used to quantify model classification performance. Top-1 accuracy is defined as the percentage of test images that were
correctly assigned the highest level of confidence among all possible image
categories, i.e., an image which contains towering cumulus clouds that is input for
inferencing is assigned the highest output confidence level for towering cumulus. If
an image classifier was trained on hundreds of types of animals, for example, and
an image of a dog was input, a correct top-n identification (e.g., top 5) exists if the
dog category is among the n (e.g., five) highest levels of confidence output from
the model. The Inception v3 model trained and tested on the ImageNet ILSVRC
2012 database achieved a cumulative top-1 accuracy of 78.8% and top-5 accuracy
of 94.4% (Wolfram Research, Inc., 2022).

Table 10 displays the top-1 accuracies for each of seven cloud image
categories used in the stage 3 research as well as a cumulative top-1 accuracy for
all image categories (i.e., total number of correct top-1 identifications divided by
the total number of independent test images). As discussed in the stage 3
methodology (Tables 3-4), the image categories are divided into groups of
nonhazardous conditions and conditions that may pose a hazard to aviators. In
addition, the matrix-like table reveals all instances of an incorrect top-1
identification as well as cases where a misidentification resulted in classification of
the incorrect hazard type (known as false alarms and missed hazardous events).
Table 10. Stage 3 matrix of the image classification scheme. The columns are the categorized cloud type that are determined to be present in the image (i.e., the ‘truth’, see Table 4). The rows represent the image category predicted by the model using the top-1 criteria (i.e., the category of highest output confidence obtained from inferencing). Blue and orange columns represent cloud types deemed as nonhazardous and hazardous, respectively. Nonhazardous (hazardous) images classified as a hazardous (nonhazardous) denote false alarms (missed events) are colored in red.

<table>
<thead>
<tr>
<th>Categorized independent test images (truth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td><strong>Image total:</strong></td>
</tr>
<tr>
<td><strong>Classified as:</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Top-1 Accuracy</strong></td>
</tr>
</tbody>
</table>

**Cumulative top-1 accuracy: ** 361/600 = 60%

Table 10 indicates that the image classifier performed accurately (i.e., predicted the cloud type determined as the ‘truth’) for the majority (60%) of the independent test images inferenced. Cumulatively (i.e., the table diagonal), the image classification scheme obtained a top-1 accuracy that is approximately 19%
less than the Inception v3 model (78.8%) trained on the ImageNet ILSVRC 2012
database. However, it is important to note that the original Inception v3 model was
trained from scratch (i.e., did not employ transfer learning) and utilized far more
training data (i.e., 3 orders of magnitude), computational expense, and time during
its training. When assessing the top-1 accuracy of individual image categories, both
clear sky and precipitation categories carried a top-1 accuracy of over 80%.
Additionally, the cirrus, fair-weather cumulus with presence of higher clouds, and
towering cumulus cloud categories received top-1 accuracies of over 60%. When
analyzing the behavior of incorrect top-1 identifications, the cumuliform cloud
categories (e.g., fair-weather cumulus, fair-weather cumulus with the presence of
higher clouds, towering cumulus, and cumulonimbus) were especially problematic.
Of 166 fair-weather cumulus cloud images, 46 images (approximately 28%) were
classified as fair-weather cumulus with the presence of higher clouds. Additionally
of 103 cumulonimbus cloud images, 30 images (approximately 30%) were
classified as towering cumulus cloud images. Importantly, however, when
considering the identification of nonhazardous and hazardous conditions, 70% of
incorrect top-1 identifications did not result in a selection of the incorrect hazard
type. Table 11 depicts the accuracy of classification when output is utilized to
create a binary hazard detection system, rather than a system that recognizes
specific cloud-type categories.
Table 11. Hazard detection matrix utilizing the top-1 accuracy data from Table 10. The columns are the categorized hazard type determined to be present in the test images (i.e., ‘truth’, described in Table 4). The rows represent the hazard type predicted by the model. Correct negatives and hits represent the number of times the true hazard type in the image corresponds to that identified by the model even if the top-1 identification represents the incorrect cloud type/sky condition. Nonhazardous (hazardous) images classified as a hazardous (nonhazardous) denote false alarms (missed events) are colored in red.

<table>
<thead>
<tr>
<th>Classified as: (hazard type)</th>
<th>Nonhazardous</th>
<th>Hazardous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonhazardous</td>
<td>Correct negatives: 87% (375/430)</td>
<td>Missed events: 10% (17/170)</td>
</tr>
<tr>
<td>Hazardous</td>
<td>False alarms: 13% (55/430)</td>
<td>Hits: 90% (153/170)</td>
</tr>
</tbody>
</table>

When top-1 classification data (i.e., the rows/columns of Table 10) are referenced to create a binary (yes/no) hazard detection system, results show improved viability over the category-specific scheme in Table 10. Correctly identified nonhazardous and hazardous conditions (e.g., correct negatives and hits, respectively) comprise close to 90% of the independent test images. This performance is desirable as it minimizes false alarms and missed hazards. The results (i.e., large percentage of hits and correct negatives) also motivate the idea that cumuliform categories within the nonhazardous class (fair-weather cumulus with and without the presence of higher clouds) and within the hazardous class (towering cumulus and cumulonimbus) may exhibit the most similarities in appearance which may lead to the highest frequency of incorrect identifications and potential quality control issues during image labeling. Conversely, image categories
such as clear sky/cirrus (nonhazardous) and precipitation (hazardous) differ significantly in appearance from cumuliform categories within their respective hazard classes, thus fewer incorrect classifications exist.

Another way to assess classification performance is by using prescribed confidence thresholds (e.g., see stage 2 data analysis). Figure 26 shows the percentage of independent test data that meets varying confidence thresholds (e.g., 0.20, 0.30, 0.40, and 0.50) for the correct (i.e., ‘true’) image category. Data in this format serves to highlight model behavior that extends deeper than top-1 accuracy. In other words, the correct image may have been assigned a considerable (i.e., above-threshold) level of confidence despite not being the category with the highest output confidence level. However, this method of data manipulation is of less utility for classification schemes comprised of many (i.e., more than two) image categories, especially when the selected confidence threshold is below 0.50. This is because output confidence levels for multiple categories may simultaneously exceed the selected confidence threshold. For example, if a stage 3 confidence threshold is prescribed at 0.40, two-of-seven image categories may meet this threshold level of confidence for a single image. As the confidence threshold decreases, the number of image categories that may be satisfied simultaneously increases. In order to develop an operational model using thresholds below 0.5, additional algorithms (e.g., prioritized image categories or decision trees) would need to be developed, and these techniques are beyond the scope of current methodology.
Figure 26. Percentage of classified independent test images that exhibit a level of confidence above the prescribed thresholds that corresponds to the correct image category. It is important to note that individual images may simultaneously satisfy the thresholds for multiple image categories, although only one category can be true.

Figure 26 illustrates that instances of split confidence levels among image categories frequently exist in the image classification data. For example, while the top-1 accuracy for the fair-weather cumulus category is 49% (Table 10), approximately 60% of fair-weather cumulus imagery satisfies the 0.30 confidence threshold, and nearly 75% of imagery satisfies the 0.20 confidence threshold. Similar behaviors are seen for the other categories where increasingly strict confidence thresholds result in fewer images classified. For 8 total images in the
precipitation category which received a top-1 accuracy of 88% (Table 10), the confidence thresholds are exceeded by all 8 images (100%) when they are at or below 0.30. Ultimately, a threshold-based approach provides information for cases where classified images have received split levels of confidence among image categories. This information can offer insight into model behavior and supplement decisions made for the organizational schemes of future model iterations.

Although the model in stage 3 has been trained using image data from restricted times (9:00 am to 4:00 p.m. ET) and from one camera site only, it is of interest for future development to assess whether the model can be successfully applied to images that do not meet these restrictions. To understand more about model behavior related to time of day or camera location, the categorical distributions of top-1 results (i.e., the image category represented by the highest output level of confidence per data entry) are shown in Figure 27. It is important to understand top-1 results do not necessarily correspond to the correct (i.e., true) image category in the image. Images from the west-facing camera at FIT Aviation are broken down into two time groups (three time intervals, ET) – the first (group 1) is from 9:00 a.m. to 4:00 p.m. which is the same window that was used for the training and independent test images in stage 3. The second group (group 2) is comprised of images from 4:00 a.m. to 9:00 a.m. and 4:00 p.m. to 10:00 p.m. The latter interval includes times when images are captured using the camera's nighttime IR mode. For the second grouping (i.e., the two time intervals, Figure 27b), there are a large percentage of images that have been classified as
precipitation. Nearly all images classified as precipitation have been taken at nighttime while the camera's IR mode is enabled. An example of an IR nighttime image and the respective top-1 output level of confidence in shown in Figure 28. By visually assessing the image in Figure 28 and using archived NEXRAD imagery for verification, it was determined that precipitation was not present in the image. The uniform dark grey background of the IR image is similar in appearance and color to a daytime precipitation image, making it likely that these similarities are responsible for the poor model behavior.

**Time-dependent Distribution of Top-1 Results: West-Facing Camera**

<table>
<thead>
<tr>
<th>a) 9 a.m.–4 p.m. ET</th>
<th>b) 4 a.m.–9 a.m., 4 p.m.–10 p.m. ET</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clear</strong> 5%</td>
<td>Clear 5%</td>
</tr>
<tr>
<td><strong>Ci</strong> 10%</td>
<td>Ci 9%</td>
</tr>
<tr>
<td><strong>Cu</strong> 12%</td>
<td>Cu 6%</td>
</tr>
<tr>
<td><strong>Cu w/high clouds</strong> 26%</td>
<td>Cu w/high clouds 13%</td>
</tr>
<tr>
<td><strong>Tcu</strong> 26%</td>
<td>Tcu 13%</td>
</tr>
<tr>
<td><strong>Cb</strong> 16%</td>
<td>Cb 13%</td>
</tr>
<tr>
<td><strong>Precip</strong> 48%</td>
<td>Precip 4%</td>
</tr>
</tbody>
</table>

Figure 27. Distribution of top-1 results for the west-facing camera AirportView.net located at FIT Aviation (KMLB) for images captured. a) from 9:00 a.m. to 4:00 p.m. ET (group 1), and b) 4:00 a.m. to 9:00 a.m. and 4:00 p.m. to 10:00 p.m. ET (group 2). The training and test images are comprised of data from group 1 only. The distributions shown are for unverified top-1 results (i.e., they are not representative of the true distributions). Number of images per distribution: a) 14997, b) 25167.
Figure 28. Nighttime IR mode image from the west-facing Airportview.net camera at FIT Aviation. The image was taken on 1 November 2022 at 10:00 p.m. ET, and it received a top-1 level of confidence of 0.543 for the precipitation category. Additional information is provided in the text.

The model also performed poorly in the late afternoon, outside the training time window, as the afternoon sun becomes visible in the west-facing camera (Figure 29). In many instances, the model incorrectly identified the cloud type and sky conditions at this time. Similar to the image in Figure 24, it is possible that lens flares and other backlit elements (e.g., image exposure levels, the visible sun) are to blame for degraded model performance. In future iterations, it is likely that additional image categories and/or training data will be added to account for these types of issues.
Figure 29. Fair-weather cumulus cloud-only image from the west-facing Airportview.net camera at FIT Aviation. The image was taken on 1 November 2022 at 5:57 p.m. ET, and it received a top-1 level of confidence of 0.302 for the cumulonimbus category. See text for detail.

Although the model created for the stage 3 research was trained using images from the west-facing camera only, imagery from the other three camera angles were also classified by this model as a proxy for other locations. Similar to Figure 27, Figure 30 shows distributions of top-1 results for images from the west-facing camera (Figure 30a) as well as the east, south, and north-facing cameras (Figures 30b, c, and d, respectively). The distributions for each camera were composited using all available times (i.e., 4:00 a.m.-10:00 p.m. ET).
Figure 30. Distribution of top-1 results for all AirportView.net cameras located at FIT Aviation (KMLB) for all available times of day (4 a.m. to 10 p.m. ET). Images from all cameras have been classified using the stage 3 model trained only on images from the west-facing camera. The distributions shown are for unverified top-1 results (i.e., they are not representative of the true distributions. Number of images per distribution: a) 40164, b) 10241, c) 8492, d) 9729.
Figure 30 illustrates the presence of strong model behaviors that occur at the three proxy camera locations that were not included in the model training data. The true distributions of image categories are not assumed to be identical for all camera directions (i.e., the west-facing camera will always see the exact same cloud types as the east-facing camera). However, the distributions in Figures 30b, c, and d exhibit highly unrealistic distributions of the image categories. For example, images that were taken by the east-facing camera received top-1 classifications in the fair-weather cumulus category nearly 50% of the time, and the south and north-facing cameras showed similar classifications for towering cumulus and cumulonimbus clouds, respectively. Image categories including cirrus, fair-weather cumulus (with higher clouds), and clear sky make up only a fraction of a percent or less of the total distributions for these three cameras. Figure 31 shows examples of clear-sky images from the east, south, and north-facing cameras and their respective top-1 results. The classification results indicate that image foreground may be unique to the model's feature extraction scheme and thus any foreground changes in the test images may result in poor classification accuracy.
Figure 31. Images captured by the AirportView.net cameras at FIT Aviation (KMLB) on 27 October 2022 at 9:00 a.m. ET and labeled as clear sky. a) The east-facing camera which received a top-1 level of confidence of 0.866 for the fair-weather cumulus category, b) south-facing camera which received a top-1 level of confidence of 0.454 for the towering cumulus category, c) north-facing camera which received a top-1 level of confidence of 0.311 for the cumulonimbus category. See text for details.
CONCLUSIONS

The research conducted herein illustrates the viability of automated cloud identification and lays the groundwork for potential model improvement. Although not addressed directly herein, the motivation for this work is directed toward improving the identification of aviation weather hazards. This topic is currently not well-represented in the academic literature. A tiered approach was applied in stages where specific research questions were addressed. In stage 1, the feasibility of machine transfer learning, applied to cloud-image classification for ground-based web camera imagery, was demonstrated. Stages 1 and 2 illustrated the importance of a large and diverse image database from which thousands of images can be used in model training, validation, and testing. Stage 2 results highlighted the challenges related to quality control such as the organization of cumuliform cloud categories which are inherently non-discrete. These promising results motivated the development of an extensive cloud image database developed using four cameras located at FIT Aviation at the Melbourne International Airport. The stage 3 results revealed that a classification scheme could be trained and developed to successfully identify a variety of cloud types and sky conditions, and an accurate aviation hazard detection scheme could be organized by adapting model output. Lessons learned include issues such as limiting the training images to certain times of the day and training on single camera only. The latter results indicate that the foreground may be an intrinsic component of successful model development (i.e., dependent on the
camera field of view). A variety of avenues exist to learn more about cloud type classification from a combination of webcam imagery and machine learning. The results presented herein show an automated assessment of cloud types is possible and that there is potential to improve the model for future operational use in the aviation weather community. Future research opportunities include the following areas:

- Expand the classification scheme to include training data from additional times of day and for cloud types not currently represented (i.e., stage 4)
- Explore impacts of image foreground on universal model performance by including training images from multiple camera sites or by removing (cropping) image foreground from training and testing images
- Assess performance of alternative machine learning model architectures and preprocessing techniques for cloud image applications
- Evaluate training data organization methods, including handling of mixed-cloud cases and intermediate stages of convective (i.e., cumuliform) clouds of interest
- Expand the tiered methodology to include a 5th stage – application of the classification scheme for operational use as an aviation weather resource


