



Calhoun: The NPS Institutional Archive
DSpace Repository

NPS Scholarship

Theses

2022-03

**EMULATING PASSIVE MICROWAVE
OBSERVATIONS WITH PATCH-TO-PIXEL
CONVOLUTIONAL NEURAL NETWORKS**

Hall, Micky S.

Monterey, CA; Naval Postgraduate School

<https://hdl.handle.net/10945/69649>

This publication is a work of the U.S. Government as defined in Title 17, United States Code, Section 101. Copyright protection is not available for this work in the United States.

Downloaded from NPS Archive: Calhoun



Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

<http://www.nps.edu/library>



**NAVAL
POSTGRADUATE
SCHOOL**

MONTEREY, CALIFORNIA

THESIS

**EMULATING PASSIVE MICROWAVE OBSERVATIONS
WITH PATCH-TO-PIXEL CONVOLUTIONAL
NEURAL NETWORKS**

by

Micky S. Hall

March 2022

Thesis Advisor:
Second Reader:

Marko Orescanin
Scott Powell

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.			
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE March 2022	3. REPORT TYPE AND DATES COVERED Master's thesis	
4. TITLE AND SUBTITLE EMULATING PASSIVE MICROWAVE OBSERVATIONS WITH PATCH-TO-PIXEL CONVOLUTIONAL NEURAL NETWORKS			5. FUNDING NUMBERS RCLJN
6. AUTHOR(S) Micky S. Hall			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Office of Naval Research, One Liberty Center, 875 N. Randolph Street, Suite 1425 Arlington, VA 22203-1995			10. SPONSORING / MONITORING AGENCY REPORT NUMBER
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			12b. DISTRIBUTION CODE A
13. ABSTRACT (maximum 200 words) Geostationary (GEO) satellites such as the GOES constellation are equipped with Advanced Baseline Imager (ABI) sensors that have a very high temporal resolution with a very low spatial resolution and provide visible through infrared data every 15 minutes. In contrast, Low Earth Orbit (LEO) satellites with Global Precipitation Measurement Microwave Imager (GMI) sensors have very high spatial resolution with a low temporal resolution that provide data as infrequently as every 15 hours. The purpose of this research is to study the viability of using the ABI data to regress to a synthetic GMI dataset. Specifically, the focus is on improving the ability to make predictions on the under-represented data points within our dataset and being able to generalize well to future distributions of data. This thesis has created a sampling technique that combines over and under sampling in conjunction with a purpose-built Residual Neural Network to perform regression from multi-spectral ABI data to a single GMI channel. In doing so, we prove that it is possible to predict under-represented values more accurately in datasets when using our sampling method and to generalize well to future data. Using our approach, we predict within 5 Kelvin for 34.5% of the tail of the test data compared to only 24.4% when we used an unsampled dataset. We also are able to prevent our mean absolute error from rising by 1 Kelvin when measured across three test datasets that span a timeframe of five months.			
14. SUBJECT TERMS neural network, GMI, ABI, CNN, satellite			15. NUMBER OF PAGES 65
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

**EMULATING PASSIVE MICROWAVE OBSERVATIONS
WITH PATCH-TO-PIXEL CONVOLUTIONAL NEURAL NETWORKS**

Micky S. Hall
Lieutenant, United States Navy
BS, United States Naval Academy, 2017

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

**NAVAL POSTGRADUATE SCHOOL
March 2022**

Approved by: Marko Orescanin
Advisor

Scott Powell
Second Reader

Gurminder Singh
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Geostationary (GEO) satellites such as the GOES constellation are equipped with Advanced Baseline Imager (ABI) sensors that have a very high temporal resolution with a very low spatial resolution and provide visible through infrared data every 15 minutes. In contrast, Low Earth Orbit (LEO) satellites with Global Precipitation Measurement Microwave Imager (GMI) sensors have very high spatial resolution with a low temporal resolution that provide data as infrequently as every 15 hours. The purpose of this research is to study the viability of using the ABI data to regress to a synthetic GMI dataset. Specifically, the focus is on improving the ability to make predictions on the under-represented data points within our dataset and being able to generalize well to future distributions of data. This thesis has created a sampling technique that combines over and under sampling in conjunction with a purpose-built Residual Neural Network to perform regression from multi-spectral ABI data to a single GMI channel. In doing so, we prove that it is possible to predict under-represented values more accurately in datasets when using our sampling method and to generalize well to future data. Using our approach, we predict within 5 Kelvin for 34.5% of the tail of the test data compared to only 24.4% when we used an unsampled dataset. We also are able to prevent our mean absolute error from rising by 1 Kelvin when measured across three test datasets that span a timeframe of five months.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1 Introduction	1
1.1 Research Objectives and Contributions	3
1.2 Organization	3
2 Background	5
2.1 Neural Networks	5
2.2 Sampling for Regression	7
2.3 Capabilities and Limitations Imposed by our Sensors	11
3 Methodology	15
3.1 Data Overview	15
3.2 Initial Dataset	15
3.3 Initial Dataset Munging	16
3.4 Sampling Training Datasets	16
3.5 Test Datasets	19
3.6 Experiment Methodology	20
4 Results	23
4.1 Training Results	23
4.2 Analysis of Test Data Results	24
5 Conclusion	39
5.1 Future Work	40
List of References	43
Initial Distribution List	47

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

Figure 2.1	ResNet Residual Block. Source: [8].	6
Figure 2.2	Example of Dataset that has not Been Sampled	8
Figure 2.3	Example of Dataset Figure 2.2 Being Downsampled	9
Figure 2.4	Example of Dataset Figure 2.2 Being Upsampled	9
Figure 2.5	Example of Grouping Figure 2.2 into a Minority and Majority Class	10
Figure 2.6	Example of Oversampling Minority Class and Undersampling Majority Class in Figure 2.5	11
Figure 3.1	Band 13 with No Sampling	17
Figure 3.2	Band 13 Downsampled 35%	18
Figure 3.3	Band 13 Fully Sampled	18

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Table 2.1	Classification Error of ResNets with Differing Number of Layers on the CIFAR-10 Test Set. Source: [8].	7
Table 2.2	GMI Technical Summary. Source: [18]	12
Table 2.3	ABI Technical Summary. Source: [20]	13
Table 3.1	Validation and Test Days by Month	16
Table 3.2	Records in Each Training Dataset	19
Table 3.3	Records in Test Days	20
Table 4.1	Lowest Validation Loss by Dataset for Each Band.	24
Table 4.2	Metrics for Datasets Sampled to 65% in January.	25
Table 4.3	Metrics for Datasets Sampled to 50% in January.	26
Table 4.4	Metrics for Datasets Sampled to 35% in January.	26
Table 4.5	Metrics for Unsampled Datasets in January.	26
Table 4.6	Percent of Predictions within a Given Range for Datasets Sampled to 65% in January.	27
Table 4.7	Percent of Predictions within a Given Range for Datasets Sampled to 50% in January.	28
Table 4.8	Percent of Predictions within a Given Range for Datasets Sampled to 35% in January.	28
Table 4.9	Percent of Predictions within a Given Range for Unsampled Datasets in January.	29
Table 4.10	Metrics for Datasets Sampled to 65% in February.	30
Table 4.11	Metrics for Datasets Sampled to 50% in February.	30

Table 4.12	Metrics for Datasets Sampled to 35% in February.	30
Table 4.13	Metrics for Unsampled Datasets in February.	31
Table 4.14	Percent of Predictions within a Given Range for Datasets Sampled to 65% in February.	32
Table 4.15	Percent of Predictions within a Given Range for Datasets Sampled to 50% in February.	32
Table 4.16	Percent of Predictions within a Given Range for Datasets Sampled to 35% in February.	33
Table 4.17	Percent of Predictions within a Given Range for Unsampled Datasets in February.	33
Table 4.18	Metrics for Datasets Sampled to 65% in May.	34
Table 4.19	Metrics for Datasets Sampled to 50% in May.	34
Table 4.20	Metrics for Datasets Sampled to 35% in May.	35
Table 4.21	Metrics for Unsampled Datasets in May.	35
Table 4.22	Percent of Predictions within a Given Range for Datasets Sampled to 65% in May.	36
Table 4.23	Percent of Predictions within a Given Range for Datasets Sampled to 50% in May.	36
Table 4.24	Percent of Predictions within a Given Range for Datasets Sampled to 35% in May.	37
Table 4.25	Percent of Predictions within a Given Range for Unsampled Datasets in May.	37

List of Acronyms and Abbreviations

ABI	Advanced Baseline Imager
AI	artificial intelligence
CNN	convolutional neural network
GEO	Geostationary Equatorial Orbit
GMI	Global Precipitation Measurement Microwave Imager
GOES	Geostationary Operational Environmental Satellite
GPM	Global Precipitation Measurement
LEO	Low Earth Orbit
MAE	Mean Absolute Error
MSE	Mean Squared Error
ML	machine learning
NAVGEN	Navy Global Environment Model
NN	neural network
P2P-CNN	Patch-to-Pixel Convolutional Neural Networks
ReLU	Rectified Linear Unit
ResNet	residual network

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

First of all, I would like to express my gratitude to Dr. Marko Orescanin, my thesis advisor, for his advice, mentoring, and encouragement throughout the thesis process.

Thank you to my wife, Skyelar Hall, for constantly listening to me rant and complain, for enduring the smell of my coffee at midnight and beyond, and for insisting on our weekly date night so that I would take a break.

Finally, a special thanks to my children, Trey and Rowan, for always reminding me that there was a whole world outside of my office just waiting for us to go play in.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:

Introduction

This thesis will be examining the full pipeline of retrieving satellite imagery, parameterizing it for use in a neural network, sampling it, and training it on a Patch-to-Pixel Convolutional Neural Networks (P2P-CNN) for synthetic data generation. The aim will be to create a neural network model that can ingest satellite sensor data from a Geostationary Equatorial Orbit (GEO) weather satellite as an input feature space, and within a certain degree of accuracy generate the values that a Low Earth Orbit (LEO) satellite is able to natively observe. The driving force behind this research is that the best predictive weather models currently used today by both military and civilian meteorologists are driven by data that can only be collected by a LEO satellite. GEO satellites can provide frequent imagery, but at relatively coarse spatial resolution. Polar orbiting satellites, which are a subcategory of LEO satellites, can provide higher spatial resolution imagery, but only a few times per day. Therefore, if we can reproduce the data that LEO satellites generate synthetically using a GEO satellite that would enable weather forecasting models to ingest information at much higher temporal and spatial scales which we speculate will result in increased forecasting accuracy of extreme events.

The first step will be deciding on the architecture for the Patch-to-Pixel Convolutional Neural Network. In our P2P-CNN we will be taking in a patch of Advanced Baseline Imager (ABI) data and attempting to regress to a single pixel located in the center of this patch for a given band of Global Precipitation Measurement Microwave Imager (GMI) data. It is possible to create a model that ingests every ABI band and produces a prediction for every GMI band, but this creates problems when attempting to sample the data for greater outlier prediction performance. In the paper “A Patch-to-Pixel Convolutional Neural Network for Small Ship Detection with PolSAR Images” the authors performed a similar experiment [1]. These authors took polarimetric synthetic aperture radar (PolSAR) images and attempted to locate ships. These images have a resolution of 5 meters per pixel, and there is only a fraction of a percent of pixels that contain a ship. They designed their experiment to take in a patch of 126x126 pixels and output information about all the pixels and attempt to classify if each pixel contained a ship. After completing their model and adjusting their hyperparameter

they set up a ROC curve for their model and showed that their method outperformed all the preexisting models of ship classification in PolSAR imagery [1]. We hope to use this same type of CNN and patch structure to produce a new dataset that will closely represent our LEO data.

The next challenge will be figuring out how to sample the dataset to ensure we have task-appropriate and balanced feature space to build a model from. In general, physical data is imbalanced in nature and special care has to be taken in data preparation to avoid overfitting the model to underlying distributions in training. Since we are attempting to build a discriminative model for meteorological data, most of the data will be homogeneous most of the time, but that is not the data that we most care about. We want to ensure that we can predict the outliers in our dataset which will correlate to irregular weather patterns and changes of water content and type in the atmosphere. To do this we must ensure that our training dataset holds enough examples of these outliers so that it can recognize their pattern. In the book *Progress in Artificial Intelligence* [2] there is a chapter entitled “SMOTE for regression” in which the authors discuss adapting the well-known classification sampling technique SMOTE to be able to handle regression problems. After they developed their sampling technique, they then proceeded to test it on 17 different well-known machine learning datasets using 20 different learning approaches each using no sampling for their training sets, and then again using their SMOTE for regression sampling method. They then computed the F-1 score for each of the datasets and found that the F-1 score increased with their sampling method until the F-1 score was already above a certain range, usually around 50%, without sampling.

By deciding on a proper architecture for our network, studying the effect that sampling has on our outlier predictions, and building a pipeline that can seamlessly progress from raw data to proper predictions, we hope to be able to build a tool that can be used by both government and civilian meteorologists to more effectively fuse available information with their numerical weather prediction models. In doing so, we will have proven that it is possible to synthetically generate data from different satellites, which could open the path to future research by examining if other space-based sensors can extract more information content from currently available observations.

1.1 Research Objectives and Contributions

In this thesis, the focus is on discovering the most reliable and efficient method of generating synthetic GMI data from authentic ABI data. GMI data is an inherently imbalanced dataset, so the main objective is to make accurate predictions for the less represented data points. This is accomplished by utilizing differing sampling methods and the degrees to which each dataset is sampled. Once differing datasets are generated, benchmarking will be performed using a single configuration of a residual network (ResNet) to determine the effectiveness of the sampling techniques used. Specific research questions are:

- Is it possible to make accurate predictions on the least represented data points?
- Is there an optimal sampling method and size for such a large dataset?
- Is it possible to make predictions for each GMI band, or will only some of them be able to be regressed on?
- How quickly does time degrade the reliability of the model? That is to say if a model is trained on data from January, will it be able to make similarly accurate predictions on data from the following months?
- Is there a difference in the reliability of predictions of each extreme, both high and low?

Key contributions of this research include the first instance, to the author's knowledge, of a project that has attempted to regress on every GMI channel assimilated by the Navy's global model, the Navy Global Environment Model (NAVGEM), by sampling a dataset for each separate channel.

1.2 Organization

This thesis is organized into four additional chapters. Chapter 2 introduces Neural Networks, sampling techniques, and inherent properties of both GMI and ABI sensors. Chapter 3 discusses the datasets and their creation, ResNet models used in this thesis, and the research methodology. Chapter 4 examines the study's outcomes, including metrics, observations, efficiency, and accuracy. Chapter 5 discusses the research conclusions, future work, and final observations.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2: Background

This chapter explores several vital technical concepts relevant to neural networks, methods for sampling a dataset for use in regression-based tasks, and the capabilities and limitations of ABI and GMI sensors. First, there is a brief discussion of deterministic neural networks and their limitations. Following that, an overview of different sampling techniques and how they can be applied to regression-based tasks will be explored. After that, we will examine the technical capabilities of the sensors that were used and the effects that they had on the experiment.

2.1 Neural Networks

Neural networks are becoming ever more popular for tasks ranging from classifying cat pictures [3], playing video games [4], and even helping to diagnose diseases [5]. At the core, every neural network (NN) is a contained feedback loop made up of neurons and connection weights. Neurons must be arranged into at least two layers to make the most basic NN, the input layer and an output layer. It is much more common to add hidden layers made up of additional neurons and connections between the input layer and output layer to make deeper neural networks. Once the NN has been created each neuron in the input layer must receive a value. These values are passed onto the following layers, and based on the weights of the connections between layers may be adjusted before they are delivered to the next layer. Once the values have made it to the output layer a decision is made on how to validate the results depending on if the NN is using a supervised, unsupervised, or reinforcement-based system. After a determination has been made, the weights of the connections are adjusted to decrease the error on the next epoch of training, and the process begins again.

As mentioned previously, three major types of training can be performed using NN: supervised, unsupervised, and reward based [6]. When performing supervised training the data that is being trained on has a known answer that the model can reference at the end of every training epoch. This allows the model to more accurately update the weights for

each connection because it can easily calculate the error between the known answer and the output that is produced. Unsupervised learning is similar to supervised learning, but there is not a known answer for every input. This makes calculating the error for the output more difficult because a custom learning method must be created to update the weights for each connection. Finally, reinforcement learning is conducted by performing a task and receiving a score based on the performance of the task. After the task has been completed the score that was recorded is used to determine how much to change the weights of the connections. If the task was performed well then the weights do not need to be adjusted by a great deal, but if the task was performed poorly then significant changes can be made to increase performance. For this thesis, only supervised learning will be used.

Two subsets of NNs are convolutional neural network (CNN)s and ResNets. A CNN can take in multiple layers of data, including height, width, and even depth to create a 3D matrix of input data. This data is then passed through multiple hidden layers as in a traditional NN, but the idea of the initial matrix is preserved throughout the layers. This is important because it brings relevance to data that is stacked in meaningful layers such as much of traditional imagery [7]. A ResNet is a NN or any subset of NN that contains at least one Rectified Linear Unit (ReLU) activation layer. A ReLU layer allows other layers to be skipped to prevent every layer from being adjusted at the beginning of training. This can be seen in Figure 2.1. This helps to prevent the problem of vanishing gradients within a model. Once the model has trained through a few epochs the skipped layers will begin to make adjustments as well to increase the feature space that the model is exploring to continuously increase the effectiveness of the model [8]. This thesis will utilize a CNN with ReLU activation layers and will be referred to from here on out as a ResNet.

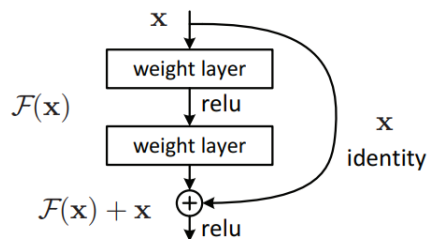


Figure 2.1. ResNet Residual Block. Source: [8].

After the decision to use a ResNet was made the network’s architecture had to be decided upon. Instead of creating a custom model, it was instead decided to adapt the method of He et al. [8]. They developed six different ResNets, each with a differing number of layers. The number of layers they used was 20, 32, 44, 56, 110, and 1202. Their analysis showed that ResNet110 has the lowest classification error on their designated test set, but ResNet56 was within .2% and trained faster. Due to the similarity in error rate and the speed up in training, ResNet56 was chosen as the architecture to be used for this research.

# Layers	# params	error(%)
20	0.27M	8.75
32	0.46M	7.51
44	0.66M	7.17
56	0.85M	6.97
110	1.7M	6.43
1202	19.4M	7.93

Table 2.1. Classification Error of ResNets with Differing Number of Layers on the CIFAR-10 Test Set. Source: [8].

2.2 Sampling for Regression

When presented with an imbalanced dataset and the task to make predictions based on the data given, and future similar data that will be generated, the question of predicting outliers will arise [9]. Some tasks, such as predicting the median housing cost [10] in a new neighborhood, may not require or expect outliers to be predicted with any level of accuracy. However, tasks such as the detection of cancer in an MRI scan rely heavily on accurately predicting outliers consistently [5]. A widely used technique to increase the accuracy of making predictions on such outliers is resampling the dataset. Two of the most common forms of sampling are downsampling and upsampling.

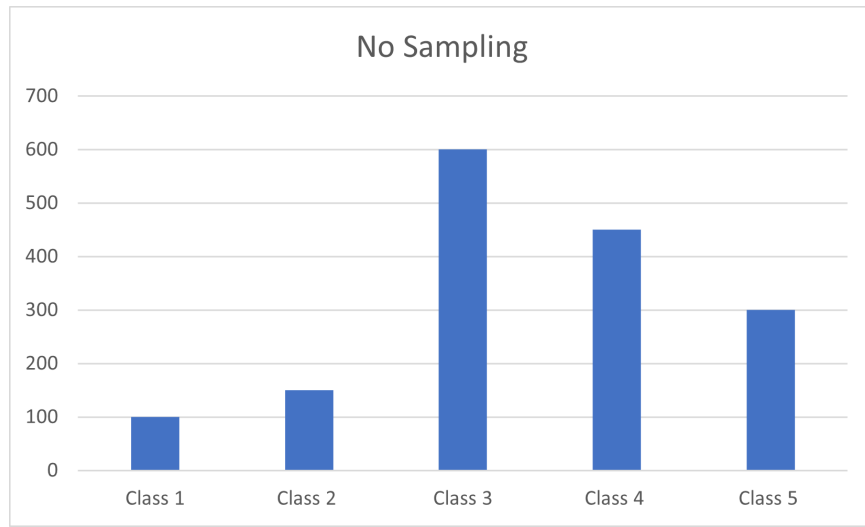


Figure 2.2. Example of Dataset that has not Been Sampled

Downsampling, also known as undersampling, can best be described as taking instances of the majority classes out of the dataset to create a more balanced distribution between the minority and majority classes. Upsampling, also known as oversampling, is accomplished by taking instances of the minority class and copying them or creating similar synthetic instances to increase the representation of the minority class in the distribution. Both of these methods lend themselves well to classification tasks that have a set number of classes to sample their distribution around, but using them for datasets focused on regression is more difficult because there are no set classes that each instance can fall in [11].

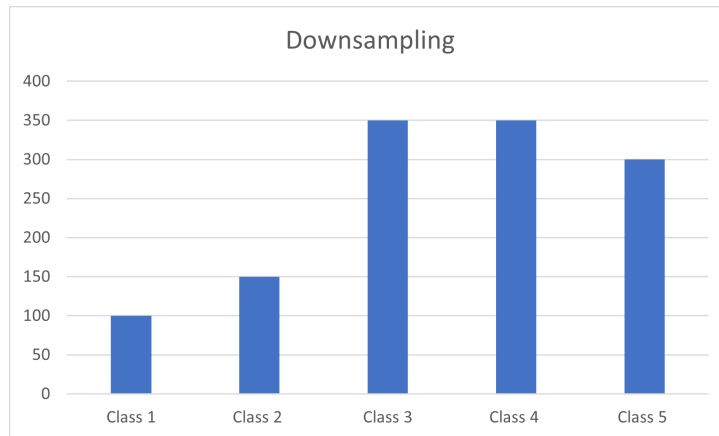


Figure 2.3. Example of Dataset Figure 2.2 Being Downsampled

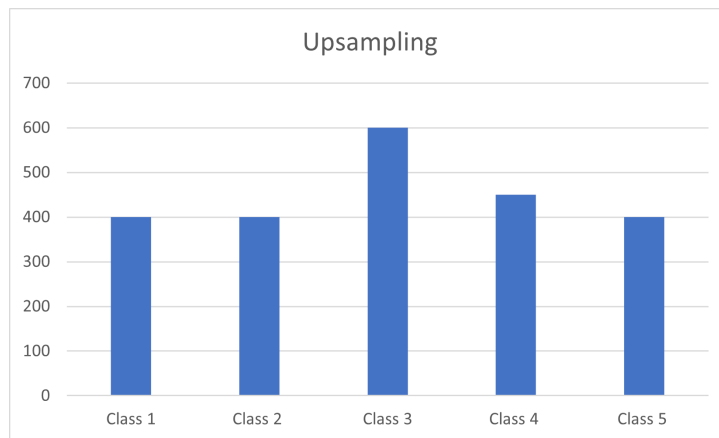


Figure 2.4. Example of Dataset Figure 2.2 Being Upsampled

There has been a great deal of research done in identifying the best techniques to use when sampling a dataset for use with classification-based tasks in which classification of rare cases is desired, but there has been a remarkable lack of work done when it comes to regression-based tasks [12]. A few notable contributions to this problem include the creation of the REBAGG technique for use with ensemble training [13], the work done by Torgo et al. in adapting traditional classification based sampling techniques for regression [14], and the continuation of that work to develop the SMOTE for Regression technique [2]

and the SMOGN technique [15]. It was proposed by Torgo et al. [2] that it is possible to perform traditional undersampling and oversampling on datasets where the target variable is a continuous value by setting a threshold for the rare cases in the dataset that will represent the minority class and then allowing all other instances to fall into the majority class. By setting this threshold and separating the data into two distinct classes Torgo et al. [2] was also able to combine oversampling and undersampling similarly to the traditional SMOTE technique used for classification and called it SMOTE for Regression or SMOTE-R.

SMOTE-R begins by determining the threshold for the minority class and then separating the data into their separate classes. The majority class is then downsampled by randomly removing samples until it has been reduced to a predetermined level. The minority class is subsequently upsampled by synthetically generating new samples by interpolating similar preexisting records. Once both tasks are complete the distributions of the dataset between the two classes are more balanced to increase the accuracy of predicting the minority class [2]. Figure 2.5 is an example of how Figure 2.2 would be distributed if Class 1 and Class 2 were assimilated into the minority class with all other classes being assigned to the majority class. Figure 2.6 is then a representation of oversampling the minority class and undersampling the majority class to form a more balanced dataset.

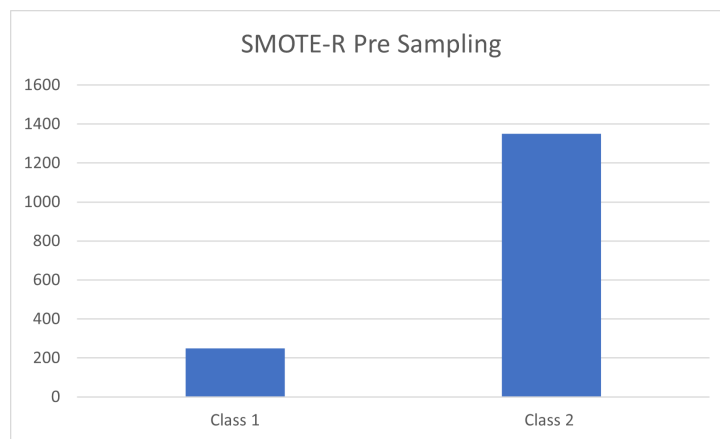


Figure 2.5. Example of Grouping Figure 2.2 into a Minority and Majority Class

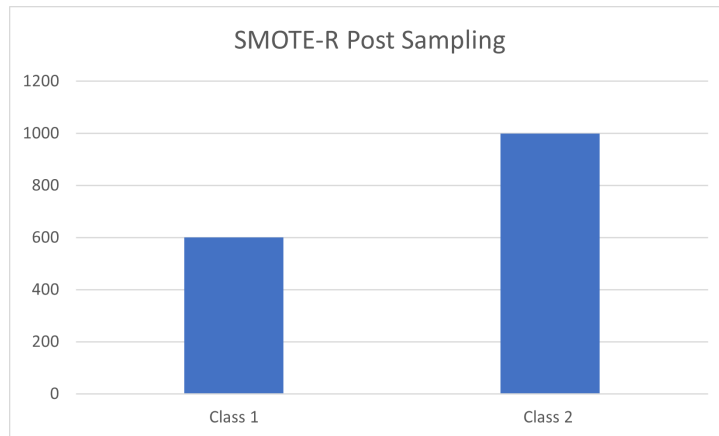


Figure 2.6. Example of Oversampling Minority Class and Undersampling Majority Class in Figure 2.5

All of these methods were considered when deciding the best solution for the sampling of the dataset presented in this thesis. In the end, a combination of several sampling techniques was merged to create a dataset that better represented the types of predictions the model that was created was intended to make.

2.3 Capabilities and Limitations Imposed by our Sensors

LEO satellites that are equipped with passive microwave sensors have the intended use to collect data for use in meteorology applications, specifically, they are commonly used for weather forecasting applications. GMI sensors have excellent spatial resolution but do not continuously observe the same area and suffer from poor temporal resolution as a result. These sensors operate by collecting microwave energy, processing the frequency bands of interest, and then delivering the data to the ground station [16]. The newest GMI sensor arrays are equipped to capture thirteen different channels worth of data as can be seen in Table 2.2. Microwave radiation tends to be less sensitive to absorption by water vapor and scattering by liquid water and ice at lower frequencies; however, some channels, such as the 23 GHz channel, have a heightened sensitivity to water vapor. Research has been performed to use data from GMI sensors to classify convective classes within the data [17], but no work has been done to make predictions of GMI data using other sources to the author's knowledge.

Band	Frequency	Polarization
1	10.6	Vertically
2	10.6	Horizontally
3	18.7	Vertically
4	18.7	Horizontally
5	23	Vertically
6	37	Vertically
7	37	Horizontally
8	89	Vertically
9	89	Horizontally
10	166	Vertically
11	166	Horizontally
12	183±2	Vertically
13	183±7	Vertically

Table 2.2. GMI Technical Summary. Source: [18]

The Geostationary Operational Environmental Satellite (GOES) platforms, which are equipped with ABI sensors, create a much larger array of products from the data that they collect. ABI sensors have excellent temporal resolution since they are only equipped to GEO satellites, but because they are so far away from Earth's surface they have poor spatial resolution. The sensors on GOES-16 and GOES-17 each have sixteen different bands, compared to only five bands on previous generations of GOES. These bands have been summarized in Table 2.3. This data is used operationally for tracking cloud formations, fires, smoke, volcanic ash plumes, aerosols, and air quality [19].

ABI Band	Wavelength(μm)	Type	Spatial Resolution (km^2)
1	0.47	Visible	1
2	0.64	Visible	0.5
3	0.86	Near-Infrared	1
4	1.37	Near-Infrared	2
5	1.6	Near-Infrared	1
6	2.2	Near-Infrared	2
7	3.9	Infrared	2
8	6.2	Infrared	2
9	6.9	Infrared	2
10	7.3	Infrared	2
11	8.4	Infrared	2
12	9.6	Infrared	2
13	10.3	Infrared	2
14	11.2	Infrared	2
15	12.3	Infrared	2
16	13.3	Infrared	2

Table 2.3. ABI Technical Summary. Source: [20]

Even though it is clear that GMI and ABI sensors have some overlap in functionality, it must be stressed that their data is not interchangeable due to the vast difference in distance of their orbits and the collection properties that limit each array. The largest limitation that was realized during this research is the difficulty in making predictions about GMI data over land. GMI sensors are affected heavily by the emissive properties of the surface that they are over. The ocean has emissive properties that are dependent upon its surface roughness, temperature, and salinity, which are much more horizontally homogeneous than soil moisture or soil type, which affect emissivity over land [21]. Because of this property, information about the soil type and moisture content would need to be known by any model trying to use ABI data to make accurate predictions on corresponding GMI data over land, which falls outside of the scope of what this thesis is researching.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3: Methodology

This chapter discusses the process of correlating ABI to GMI data, the initial “munging” of the dataset, the sampling technique used to create training datasets, the technique used to create testing datasets, and the research methodology.

3.1 Data Overview

Throughout this thesis, multiple datasets have been created to decide the best method to increase our performance on under-represented or rare data points. The process begins by correlating GMI and ABI data to make the main dataset from which we can pull records for later datasets. This dataset is then divided between observations that take place over water and observations that take place over land. This data is then downsampled and subsequently upsampled based on the distribution of a single GMI band. Once it is known which records are needed for the final dataset, they are retrieved and put into TensorFlow Record files for faster data consumption by the ResNet.

3.2 Initial Dataset

The initial ABI data was collected using sensors aboard the GOES-16 and GOES-17 satellites. The initial GMI data was collected using various LEO satellites that were equipped with the proper GMI sensors. As discussed in Chapter 2 and seen in Table 2.2 and Table 2.3, GMI has 13 different bands while ABI has 16. It was decided that for this research band 8 and band 9 were dropped from all GMI records because they are not assimilated in the Navy global model system. Bands 1–6 from ABI were not used because they detect reflected shortwave radiation, and we chose to use only terrestrial infrared data so that we could execute our methods during both day and night. This data was then organized according to the time it was collected and then by the coordinates of the data collected. Once organized, individual pixel values from the GMI data were correlated with a patch of ABI pixel values and stored along with other corresponding metadata. Such metadata included the latitude of collection, longitude of collection, the time that the collection was made, the viewing

angle of the satellites, if the value is over land or sea, and a unique key id. This process was performed on a tenth of the data spanning January 01, 2020, to July 01, 2020, which resulted in 23,053,000 individual correlated records.

3.3 Initial Dataset Munging

As discussed in Chapter 2, all land data must be removed from our dataset due to the difficulties imposed by land emissivity. This results in a remainder of 15,188,000 records. To further pare this down the data was divided into months. Three days from each month are then set aside for a validation dataset and three additional days were set aside for a test dataset. The days selected for each month can be seen in Table 3.1. Training was only conducted on the January training set, which was comprised of 2,839,500 records before sampling.

Month	Validation Days	Test Days
January	6, 13, 19	8, 21, 26
February	4, 11, 27	7, 16, 24
March	5, 14, 22	8, 16, 28
April	5, 16, 27	3, 15, 20
May	7, 13, 24	4, 16, 25

Table 3.1. Validation and Test Days by Month

3.4 Sampling Training Datasets

After the land records were removed, and the data was further divided into months, data sampling began with a focus on the January training set. We took an approach similar to that of SMOTE-R. The main difference between our approach and SMOTE-R is that for our distributions we binned each record into multiple classes by taking the floor of each GMI pixel instead of creating a smaller group of classes. We also did not synthetically

generate new records for the minority classes by interpolating multiple records together. Instead, once we had a distribution for each GMI band using our bins we found the largest bin and reduced it by a given percentage by randomly removing records from its bin. Once the largest bin was reduced by that percentage all other bins were reduced to having at most the same number of records as that bin. If the bin already had that many records or fewer then no records were removed. Once all records were finished being removed then the upsampling process began. Upsampling was accomplished by taking all bins that had fewer records than the bin with the most records and randomly duplicating records within the bin until a flat distribution of data was created. For this thesis, each GMI band had three different datasets created by sampling down the largest bin by 35%, 50%, and 65%, resulting in the creation of 44 different sampled datasets, for a total of 45 datasets when the original non-sampled dataset is included. The change in the distribution of GMI's band 13 when the largest bin was downsampled by 35% can be seen in Figures 3.1 through 3.3, and the number of records in each sampled dataset can be viewed in Table 3.2.

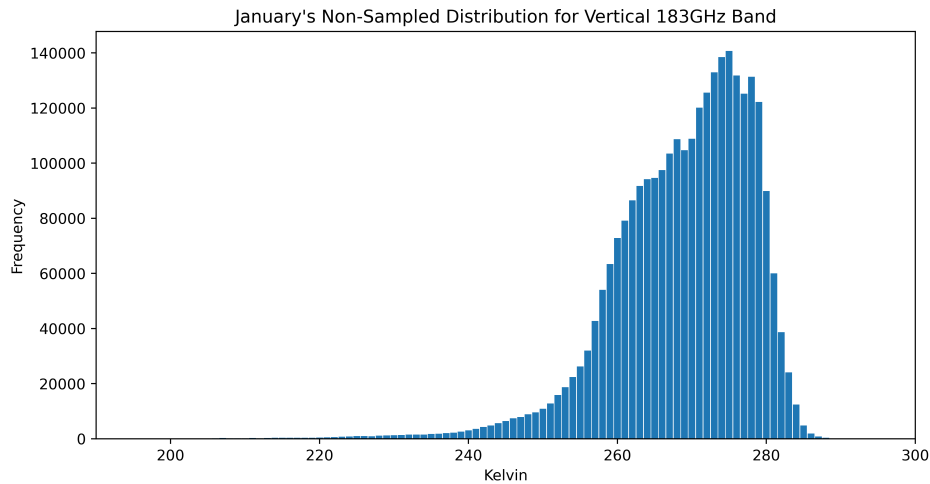


Figure 3.1. Band 13 with No Sampling

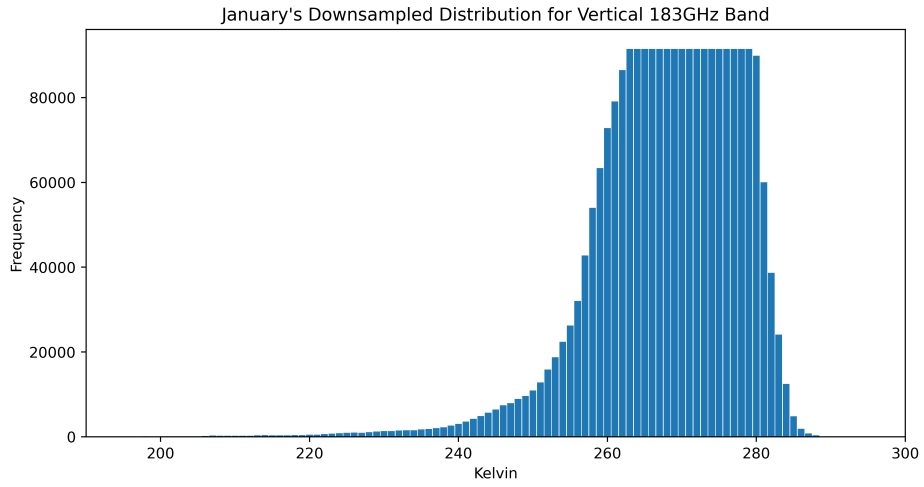


Figure 3.2. Band 13 Downsampled 35%

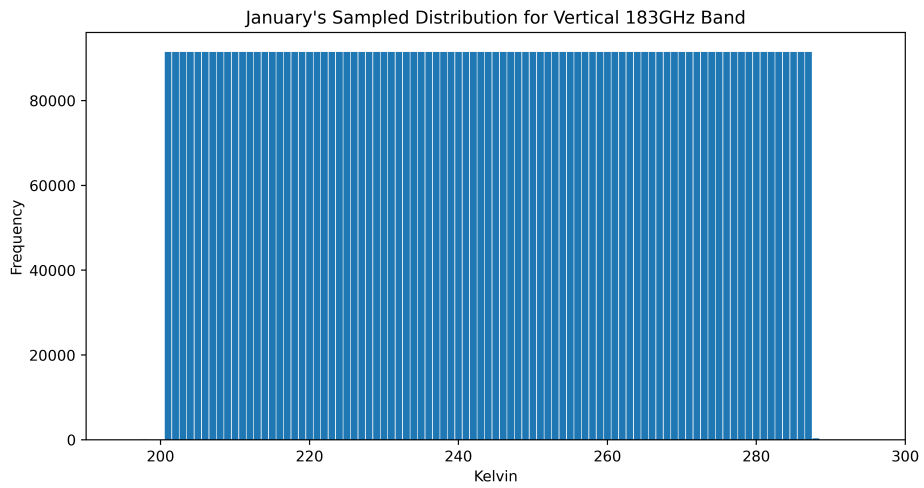


Figure 3.3. Band 13 Fully Sampled

GMI Band	35%	50%	65%	00%
1	1.52M	1.18M	0.83M	2.83M
2	1.38M	1.07M	0.75M	2.83M
3	0.82M	0.63M	0.45M	2.83M
4	1.3M	1.01M	0.71M	2.83M
5	0.69M	0.53M	0.37M	2.83M
6	1.22M	0.94M	0.66M	2.83M
7	1.02M	0.83M	.058M	2.83M
10	2.38M	1.93M	1.35M	2.83M
11	2.11M	1.62M	1.14M	2.83M
12	1.00M	0.77M	0.54M	2.83M
13	1.59M	1.22M	0.85M	2.83M

Table 3.2. Records in Each Training Dataset

3.5 Test Datasets

Once the 44 sampled training datasets were created various test sets needed to be created to test the properties of the models that were to be trained. Some of the properties that we wanted to test were the temporal resilience of our models and how well our models performed in different parts of the ocean. Rather than making a new dataset each time a new property was to be tested, it was decided that we would separate the records by days and then combine the days into a single large dataset when they were needed. The test datasets that yielded us the greatest insight were the original three days removed from January, the first seven days of February, and the first seven days of May. The days from January allowed us to test our model on data that it should be able to make accurate predictions on since it was data that was extremely similar to what was seen in the training and validation datasets. The February dataset allowed us to test to see if there was any immediate drop-off in the effectiveness of our models once the data was no longer being affected by possible lingering weather events between days in the training and test dataset. Finally, the May dataset showed if time was going to have any great impact on our model's results, or if the

model had learned enough to still be able to make decent predictions and demonstrate the persistence of skill by generalizing on unseen distributions. The number of records per day of test data can be seen in the following table.

Date	Number of Records
08 January	228,100
21 January	286,500
26 January	208,600
01 February	156,600
02 February	125,700
03 February	180,100
04 February	68,000
05 February	65,400
06 February	137,900
07 February	124,800
01 May	118,600
02 May	172,700
03 May	86,300
04 May	97,900
05 May	146,600
06 May	124,300
07 May	180,800

Table 3.3. Records in Test Days

3.6 Experiment Methodology

Experiments on the datasets were conducted using Python 3.6 and TensorFlow 2.4. Models were trained using a cluster of NVIDIA Quadro RTX 8000s.

3.6.1 Training

Models were trained using the non-sampled dataset to use as a baseline for comparison with models to later be trained on the sampled datasets. Each model was trained using four GPUs, sixteen CPUs, 100GB of memory, an Adam optimizer [22], a batch size of 1024, a base learning rate of 0.001 with a reduction of the learning rate on a plateau of seven epochs, for a total training cycle of 45 epochs. Initial tests were also performed on the 35% sampled datasets for bands 5, 11, and 13 with all of the previous parameters remaining the same except training was performed for 200 epochs. A negligible difference was observed with the added training time. Performance for all bands seemed to reach the maximum between epochs 25 and 35, so to drastically reduce the training time the number of epochs was lowered to 45.

3.6.2 Testing

After the training for all 55 models were completed they were tasked with making predictions on the data listed in Table 3.3 above. These predictions were then analyzed to test how well the R2 score, mean absolute error, and mean squared error turned out. Each of these measurements was taken for the full day's predictions, the middle 90% of data of each day, and the top and bottom 5% of data for each day. How many predictions were within 1 K, 3 K, and 5 K were also measured for the full day's predictions, the middle 90%, and then the top and bottom 5% of data for each day. Using these metrics we can determine if sampling had any major impact on the prediction of outliers in our dataset, or if sampling could have hurt the accuracy of our predictions overall.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4: Results

This chapter presents the results of our work when attempting to create a model that will reliably predict each GMI band using an unsampled dataset and three sampled datasets. The results of each model will be examined based on how well the model trained in regards to their validation loss during training, their R2, Mean Absolute Error (MAE), and Mean Squared Error (MSE) scores on test datasets, and what percentage of predictions in each test dataset was within an acceptable range of error.

4.1 Training Results

After creating three separate datasets for all eleven GMI bands that we have chosen to model, the same ResNet architecture was trained on each dataset to determine if there would be an appreciable difference in the results between the bands themselves and also between the amount of sampling that was done for each dataset. The results in Table 4.1 show the best epoch in each band's datasets. When referencing Table 4.1 to Table 2.2 it becomes immediately evident that horizontally polarized bands performed better during training than their vertical counterparts. GMI bands ten through thirteen, which are more sensitive than lower frequencies to scattering by liquid water and ice in clouds, had the best results during training. We theorize that this is the case since ABI bands eight through ten are also sensitive to water vapor in the atmosphere, so a strong correlation between the two can be made. It can also be seen that no level of sampling caused an overall performance increase during training. However, this alone can not be used to state that sampling was ineffective at increasing the ability of the model to predict outliers. That must be examined in greater detail using the test datasets results.

GMI Band	65%	50%	35%	Unsampled
01	394.00	401.24	408.43	319.74
02	164.33	183.39	190.26	148.34
03	516.72	486.62	444.146	409.39
04	206.31	198.60	177.34	167.52
05	199.10	202.49	185.50	170.26
06	440.95	432.66	383.69	361.02
07	113.31	119.91	109.93	101.25
10	110.05	115.97	114.97	97.64
11	82.37	79.12	84.30	68.40
12	15.89	15.95	15.45	17.30
13	39.16	39.07	37.77	36.18

Table 4.1. Lowest Validation Loss by Dataset for Each Band.

4.2 Analysis of Test Data Results

After a model was created for each dataset we wanted to see how well it would do on test data from within the same time frame that the model was trained, how well it would do on test data immediately following the training data, and finally on how well it would do on test data removed from the training data by a few months. As discussed in chapter three, we created all of these test data sets by using data set aside from January, the first week of February, and then the first week of May respectively. Using this test data we determined how well our models make predictions on all of the data as a whole, and also how well it makes predictions on rare values.

4.2.1 January Test Data

By performing predictions on test datasets built from data that is at most 24 hours removed from data that is seen in the training dataset, we can establish a good baseline of how well our model is performing. To do this we examine the results as three separate entities. The first is as a whole so that we can determine how well predictions are performing overall.

Next, we sort the predictions in ascending order based on the true values and pull out the top five percent and bottom five percent of the predictions, related to the predicted microwave brightness temperature (K). Here, the underlying understanding is that majority of values in nature are following unsampled data distribution and the tail ends are of interest in this experiment. By examining these smaller sets of predictions we can judge how well our models can perform on rare data and if sampling did increase our performance in this area.

In Tables 4.2 through Tables 4.5 you can see the R2, MAE, and MSE for all of the test data, for the bottom five percent of data, and the top five percent of data for each GMI band's dataset. Two main conclusions can be drawn from these tables. The first is that the overall accuracy of predictions does not seem to be affected by sampling. When comparing the model trained on the unsampled dataset to the models trained on the rest of the datasets, the model on the unsampled data had a better R2 score for 24 out of 33 comparisons, a lower MAE for 18 out of 33 comparisons, and a lower for MSE 27 out of 33 comparisons. Similar statistics hold for both the bottom and the top percentages as well, but the results are dispersed enough that a definitive statement based on these metrics alone can not be made. The second is that bands 10 through 13 were much easier to make predictions on regardless of the level of sampling performed. This was expected due to the strong correlation between what these GMI bands are used to observe, specifically water vapor, and what the upper ABI bands can observe.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	0.38	0.26	-0.11	-0.27	0.10	-0.42	-0.17	0.68	0.71	0.88	0.77
Bottom R2	-262.56	-463.53	-512.63	-664.16	-106.65	-229.51	-565.91	-2.74	-1.60	-0.18	-0.70
Top R2	-3.12	-9.26	-7.46	-35.84	-31.60	-16.77	-17.41	-54.13	-12.46	-4.26	-5.38
MAE	7.85	6.69	14.48	10.67	10.76	16.87	8.05	5.78	3.96	1.89	2.63
Bottom MAE	7.54	11.93	20.05	22.72	10.26	13.30	13.44	20.30	21.48	5.20	12.57
Top MAE	44.80	40.97	31.93	46.48	21.56	24.23	22.52	2.48	1.43	2.16	1.71
MSE	248.14	118.73	465.24	193.08	212.38	544.47	112.75	83.08	46.01	9.37	23.38
Bottom MSE	130.88	228.96	1080.52	533.67	295.07	427.11	240.68	851.04	789.82	122.20	383.44
Top MSE	2865.40	2169.39	1998.72	2226.29	947.89	1255.39	569.88	19.35	6.06	6.35	4.44

Table 4.2. Metrics for Datasets Sampled to 65% in January.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	0.36	0.31	-0.07	-0.26	0.04	-0.13	-0.31	0.69	0.72	0.88	0.78
Bottom R2	-349.25	-545.15	-568.55	-1 247.40	-325.58	-224.07	-317.35	-2.30	-1.36	-0.25	-0.70
Top R2	-2.74	-9.41	-4.93	-5.72	-8.47	-16.74	-11.31	-26.44	-19.41	-6.16	-1.24
MAE	7.79	5.79	13.72	8.48	10.77	14.54	7.84	5.52	3.83	1.88	2.55
Bottom MAE	8.68	11.77	21.31	21.05	20.14	13.85	6.85	18.28	18.93	5.37	12.47
Top MAE	41.09	42.80	25.32	13.10	11.35	24.05	13.93	2.05	1.95	2.61	0.98
MSE	257.18	109.60	450.36	191.13	226.21	432.65	126.44	78.53	44.13	9.21	22.31
Bottom MSE	173.92	269.19	1 198.15	1 001.62	895.16	417.03	135.15	751.21	718.37	129.37	383.01
Top MSE	2 601.37	2 201.07	1 400.68	406.04	275.45	1 253.24	381.09	9.63	9.18	8.65	1.56

Table 4.3. Metrics for Datasets Sampled to 50% in January.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	0.32	0.11	0.19	-0.05	0.12	-0.06	-0.10	0.66	0.73	0.90	0.79
Bottom R2	-406.57	-384.21	-388.97	-1 002.02	-328.27	-165.10	-307.62	-2.36	-1.58	-0.31	-0.99
Top R2	-1.82	-20.29	-10.32	-8.86	-8.78	-18.60	-10.72	-79.16	-17.36	-5.79	-3.24
MAE	6.93	4.86	12.10	7.50	10.26	13.45	7.03	5.79	3.62	1.72	2.31
Bottom MAE	8.05	7.45	20.22	18.37	22.43	12.03	6.69	18.68	20.44	6.00	13.98
Top MAE	30.38	63.80	41.45	16.09	11.11	24.68	13.57	2.43	1.75	2.51	1.45
MSE	272.90	142.27	341.36	159.76	207.25	407.14	106.12	88.30	43.04	8.29	21.42
Bottom MSE	202.38	189.86	820.36	804.74	902.55	307.77	131.02	765.97	783.34	135.93	449.62
Top MSE	1 961.08	4 502.83	2 673.15	596.08	284.34	1 384.65	362.80	28.14	8.26	8.20	2.95

Table 4.4. Metrics for Datasets Sampled to 35% in January.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	0.36	0.20	0.16	0.17	0.24	0.02	0.02	0.69	0.73	0.91	0.83
Bottom R2	-623.25	-1 686.50	-383.89	-548.64	-205.30	-168.64	-181.28	-2.58	-1.09	-0.39	-0.68
Top R2	-2.93	-6.09	-10.55	-21.40	-11.25	-23.92	-13.66	-96.59	-33.47	-2.01	-2.46
MAE	7.62	5.06	12.78	8.28	9.49	12.92	6.60	5.88	3.95	1.55	2.09
Bottom MAE	11.93	19.36	20.60	18.23	16.82	13.41	5.92	18.84	16.67	5.77	12.56
Top MAE	45.07	32.30	46.64	36.00	14.07	30.36	16.20	3.46	3.40	1.60	1.27
MSE	257.12	128.14	351.12	126.33	178.63	376.58	94.03	78.80	43.39	7.41	17.46
Bottom MSE	309.98	831.73	809.69	440.99	565.47	314.32	77.39	814.38	635.91	143.56	380.48
Top MSE	2 732.08	1 500.76	2 727.92	1 354.01	356.08	1 760.11	453.73	34.26	15.51	3.64	2.40

Table 4.5. Metrics for Unsampled Datasets in January.

We can also see how sampling affects the percentage of predictions that we can make within 1, 3, and 5 K in Tables 4.6 through 4.9. This observation is important because it shows that while some of our predictions are very far off, the majority of our results are within a 5 K range of error. This continues to hold especially true for bands 10 through 13. Once again, we can see that sampling does not seem to have a great effect on the effectiveness of our predictions as a whole. When looking at all of the predictions made on each dataset that were within 5 K of the true value, we see that the unsampled dataset had a higher percentage in 21 of the 33 comparisons. However, once we begin to look at the top and bottom five percent of the data separately we begin to see a much greater difference. The unsampled dataset had a higher percentage in only 10 of 33 comparisons of the bottom five percent and only 8 of 33 comparisons in the top five percent. The greatest disparity of results can be seen in the bottom seven bands which have already been shown to be the most difficult to make accurate predictions on. Specifically, with the dataset that has been downsampled by 35%, we can see differences of up to 46% in the number of predictions in the bottom of our dataset and differences up to 37% in the top portion of the dataset.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.18	0.23	0.07	0.12	0.09	0.07	0.13	0.19	0.24	0.42	0.39
All < 3.0	0.46	0.59	0.20	0.33	0.24	0.20	0.34	0.47	0.62	0.85	0.80
All < 5.0	0.65	0.80	0.32	0.51	0.37	0.31	0.52	0.62	0.81	0.96	0.90
Bottom < 1.0	0.09	0.11	0.01	0.06	0.04	0.02	0.12	0.04	0.03	0.12	0.05
Bottom < 3.0	0.30	0.32	0.04	0.16	0.14	0.10	0.47	0.13	0.09	0.46	0.14
Bottom < 5.0	0.47	0.58	0.09	0.26	0.23	0.22	0.67	0.21	0.15	0.69	0.22
Top < 1.0	0.05	0.00	0.04	0.09	0.11	0.06	0.10	0.57	0.55	0.18	0.35
Top < 3.0	0.14	0.00	0.09	0.25	0.30	0.21	0.24	0.83	0.83	0.63	0.94
Top < 5.0	0.20	0.00	0.12	0.37	0.44	0.33	0.37	0.88	0.89	0.97	1.00

Table 4.6. Percent of Predictions within a Given Range for Datasets Sampled to 65% in January.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.14	0.17	0.06	0.11	0.08	0.06	0.11	0.16	0.23	0.39	0.34
All < 3.0	0.37	0.47	0.18	0.31	0.23	0.18	0.31	0.44	0.58	0.82	0.77
All < 5.0	0.55	0.66	0.30	0.46	0.35	0.28	0.47	0.63	0.77	0.95	0.88
Bottom < 1.0	0.09	0.07	0.03	0.04	0.08	0.05	0.16	0.05	0.04	0.18	0.06
Bottom < 3.0	0.24	0.20	0.12	0.16	0.23	0.15	0.46	0.14	0.11	0.55	0.17
Bottom < 5.0	0.33	0.35	0.23	0.28	0.37	0.27	0.65	0.23	0.19	0.76	0.29
Top < 1.0	0.04	0.00	0.06	0.10	0.08	0.08	0.07	0.47	0.48	0.15	0.58
Top < 3.0	0.11	0.01	0.17	0.23	0.24	0.20	0.23	0.77	0.80	0.60	0.98
Top < 5.0	0.15	0.03	0.24	0.36	0.39	0.30	0.39	0.88	0.89	0.97	1.00

Table 4.7. Percent of Predictions within a Given Range for Datasets Sampled to 50% in January.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.13	0.09	0.05	0.05	0.07	0.05	0.08	0.17	0.21	0.39	0.34
All < 3.0	0.36	0.33	0.16	0.16	0.20	0.16	0.25	0.44	0.55	0.82	0.75
All < 5.0	0.54	0.56	0.26	0.27	0.32	0.25	0.41	0.60	0.77	0.94	0.87
Bottom < 1.0	0.11	0.00	0.03	0.00	0.09	0.09	0.00	0.03	0.02	0.20	0.06
Bottom < 3.0	0.29	0.01	0.09	0.00	0.27	0.23	0.00	0.11	0.05	0.56	0.17
Bottom < 5.0	0.46	0.10	0.17	0.00	0.44	0.38	0.02	0.18	0.10	0.76	0.27
Top < 1.0	0.01	0.01	0.05	0.00	0.06	0.06	0.00	0.48	0.62	0.25	0.35
Top < 3.0	0.02	0.06	0.15	0.00	0.20	0.18	0.01	0.79	0.88	0.72	0.84
Top < 5.0	0.05	0.10	0.24	0.00	0.30	0.27	0.02	0.86	0.92	0.99	0.99

Table 4.8. Percent of Predictions within a Given Range for Datasets Sampled to 35% in January.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.13	0.20	0.06	0.07	0.08	0.07	0.13	0.15	0.20	0.45	0.41
All < 3.0	0.37	0.58	0.17	0.21	0.24	0.20	0.35	0.41	0.57	0.89	0.82
All < 5.0	0.59	0.79	0.28	0.36	0.38	0.32	0.54	0.57	0.76	0.97	0.92
Bottom < 1.0	0.00	0.00	0.00	0.00	0.05	0.01	0.08	0.03	0.05	0.16	0.05
Bottom < 3.0	0.00	0.03	0.01	0.00	0.15	0.04	0.31	0.10	0.14	0.50	0.15
Bottom < 5.0	0.01	0.19	0.04	0.00	0.26	0.13	0.62	0.19	0.24	0.73	0.28
Top < 1.0	0.02	0.03	0.00	0.00	0.05	0.05	0.04	0.23	0.01	0.31	0.45
Top < 3.0	0.04	0.08	0.00	0.00	0.15	0.15	0.15	0.72	0.56	0.91	0.96
Top < 5.0	0.07	0.11	0.01	0.00	0.27	0.22	0.24	0.84	0.83	0.99	1.00

Table 4.9. Percent of Predictions within a Given Range for Unsampled Datasets in January.

4.2.2 February Test Data

The goal of using test data from February is to test whether the models that were created in January could retain the skill that they had learned and adapt it to future data, or if the models had effectively become over-fit and learned the atmospheric conditions for January. If this was the case then a sharp deterioration should be seen in the February results, but they are not. As can be seen in Tables 4.10 through 4.13, the R², MAE, and MSE remain similar to that of the predictions made on the test data in January. All of the conclusions drawn from the January test data in regards to the efficacy of sampling still hold when comparing the February datasets.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	0.44	0.11	0.27	0.12	0.29	0.13	0.08	0.55	0.61	0.84	0.69
Bottom R2	-238.28	-330.32	-86.77	-176.23	-42.71	-65.31	-160.77	-3.91	-2.95	-0.56	-1.41
Top R2	-14.37	-72.39	-57.65	-56.66	-31.77	-36.10	-30.50	-26.77	-18.15	-2.52	-18.72
MAE	7.04	5.40	12.31	7.39	9.70	13.73	7.15	6.30	4.10	2.01	2.72
Bottom MAE	8.46	7.63	13.92	10.46	11.90	15.83	6.43	30.56	29.90	10.35	21.10
Top MAE	45.72	67.42	63.25	32.69	20.39	41.76	21.68	1.80	1.45	2.21	2.12
MSE	277.90	174.59	385.58	166.98	189.24	430.01	115.74	117.93	66.86	14.45	34.94
Bottom MSE	290.42	199.54	388.18	309.89	315.53	467.52	107.97	1462.14	1399.87	299.54	824.92
Top MSE	4111.48	5144.40	5134.95	1775.44	668.38	2686.38	690.17	8.08	5.28	6.43	7.32

Table 4.10. Metrics for Datasets Sampled to 65% in February.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	0.46	0.35	0.11	0.01	0.19	0.04	-0.03	0.54	0.61	0.83	0.67
Bottom R2	-360.12	-294.17	-121.92	-223.01	-55.45	-69.82	-180.03	-4.22	-2.47	-0.39	-1.37
Top R2	-20.67	-41.06	-55.21	-51.99	-29.80	-31.99	-31.90	-43.44	-27.04	-4.03	-14.83
MAE	7.47	6.12	13.43	8.08	10.30	14.79	7.77	6.58	4.33	2.18	2.98
Bottom MAE	10.92	8.94	16.60	12.06	13.33	16.42	6.81	31.78	27.37	8.81	20.97
Top MAE	61.47	50.23	56.42	32.55	19.21	41.16	22.61	2.21	1.86	2.71	1.83
MSE	268.88	128.80	469.51	187.94	217.11	478.39	129.63	118.74	65.66	15.09	36.79
Bottom MSE	438.30	177.77	543.61	391.70	407.49	499.31	120.82	1552.48	1228.00	266.77	810.86
Top MSE	5799.30	2948.13	4921.23	1631.66	628.15	2388.72	720.87	12.93	7.74	9.17	5.88

Table 4.11. Metrics for Datasets Sampled to 50% in February.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	0.44	0.24	0.06	-0.04	0.32	-0.07	-0.02	0.55	0.61	0.83	0.67
Bottom R2	-365.33	-403.53	-190.24	-287.56	-36.54	-106.14	-270.27	-4.44	-2.30	-0.30	-1.19
Top R2	-19.94	-38.41	-49.56	-78.96	-41.14	-32.49	-33.74	-32.86	-21.74	-1.15	-43.53
MAE	8.24	7.30	14.35	10.38	9.73	16.05	8.22	6.39	4.38	2.13	3.13
Bottom MAE	11.14	12.27	19.36	21.93	11.70	18.73	11.51	32.49	26.03	8.51	20.04
Top MAE	62.49	47.15	54.21	49.11	21.14	40.80	26.32	2.07	1.66	1.73	3.28
MSE	281.74	149.97	497.05	198.35	182.34	531.34	128.90	117.37	65.72	14.91	36.95
Bottom MSE	444.61	243.63	845.78	504.57	271.02	755.37	181.05	1618.19	1168.07	249.62	749.13
Top MSE	5602.71	2762.89	4426.75	2462.00	859.43	2424.88	761.22	9.86	6.28	3.93	16.53

Table 4.12. Metrics for Datasets Sampled to 35% in February.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	0.37	0.30	0.25	0.27	0.40	0.18	0.11	0.60	0.64	0.85	0.68
Bottom R2	-204.51	-378.25	-71.03	-176.79	-28.46	-53.60	-159.82	-3.67	-1.55	-0.23	-1.07
Top R2	-29.30	-40.26	-62.03	-70.28	-28.60	-46.49	-35.64	-47.24	-46.28	-3.53	-16.00
MAE	7.76	5.11	12.50	7.68	9.05	13.41	7.01	6.10	4.48	1.89	2.73
Bottom MAE	12.28	10.08	15.26	16.09	10.54	16.71	7.53	28.58	20.72	8.99	19.81
Top MAE	83.30	48.35	69.81	46.37	19.80	50.90	24.44	3.01	3.33	2.49	2.21
MSE	312.17	137.16	395.99	139.65	159.43	408.63	111.95	103.58	60.64	13.13	35.56
Bottom MSE	249.43	228.41	318.57	310.88	212.65	384.98	107.33	1389.91	902.30	234.94	709.45
Top MSE	8107.06	2892.52	5518.21	2194.90	603.71	3438.24	802.70	14.04	13.05	8.26	6.31

Table 4.13. Metrics for Unsampled Datasets in February.

We are also able to see a similar relation in how many predictions fall within 1, 3, and 5 K across all of the data. This further reinforces the claim that small temporal differences do not cause great deterioration to the performance of the models. However, we do begin to see a slight deterioration in how well our models perform on the top and bottom 5% of the February data when compared to their January performance, but the models trained on the sampled datasets still significantly outperform the models that are trained on unsampled data when making predictions on these outliers.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.19	0.23	0.07	0.14	0.08	0.07	0.13	0.19	0.25	0.40	0.36
All < 3.0	0.49	0.59	0.22	0.38	0.25	0.20	0.37	0.47	0.61	0.81	0.75
All < 5.0	0.67	0.77	0.35	0.56	0.39	0.32	0.54	0.62	0.79	0.93	0.89
Bottom < 1.0	0.09	0.03	0.03	0.07	0.10	0.03	0.17	0.02	0.01	0.06	0.02
Bottom < 3.0	0.29	0.21	0.11	0.23	0.25	0.09	0.43	0.06	0.04	0.22	0.07
Bottom < 5.0	0.51	0.59	0.21	0.40	0.36	0.17	0.59	0.09	0.07	0.39	0.12
Top < 1.0	0.02	0.00	0.00	0.02	0.01	0.02	0.03	0.47	0.58	0.20	0.23
Top < 3.0	0.07	0.01	0.01	0.08	0.05	0.08	0.08	0.84	0.86	0.72	0.81
Top < 5.0	0.12	0.01	0.02	0.14	0.11	0.12	0.15	0.91	0.95	0.99	0.98

Table 4.14. Percent of Predictions within a Given Range for Datasets Sampled to 65% in February.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.16	0.16	0.07	0.12	0.08	0.07	0.12	0.15	0.23	0.35	0.31
All < 3.0	0.43	0.44	0.21	0.34	0.24	0.19	0.34	0.41	0.56	0.78	0.72
All < 5.0	0.61	0.63	0.34	0.52	0.38	0.30	0.51	0.58	0.76	0.92	0.87
Bottom < 1.0	0.10	0.06	0.03	0.06	0.11	0.02	0.16	0.02	0.02	0.13	0.02
Bottom < 3.0	0.30	0.25	0.09	0.20	0.27	0.09	0.43	0.06	0.06	0.36	0.06
Bottom < 5.0	0.47	0.50	0.15	0.34	0.39	0.17	0.61	0.11	0.10	0.54	0.11
Top < 1.0	0.01	0.00	0.02	0.02	0.02	0.02	0.02	0.48	0.48	0.12	0.31
Top < 3.0	0.04	0.01	0.06	0.06	0.06	0.05	0.05	0.80	0.78	0.56	0.85
Top < 5.0	0.07	0.02	0.10	0.10	0.12	0.08	0.11	0.85	0.91	0.96	0.98

Table 4.15. Percent of Predictions within a Given Range for Datasets Sampled to 50% in February.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.12	0.12	0.06	0.07	0.08	0.06	0.10	0.18	0.23	0.37	0.29
All < 3.0	0.35	0.35	0.17	0.20	0.22	0.17	0.29	0.44	0.56	0.79	0.69
All < 5.0	0.53	0.52	0.29	0.32	0.36	0.27	0.45	0.60	0.75	0.92	0.85
Bottom < 1.0	0.09	0.00	0.02	0.00	0.08	0.04	0.00	0.01	0.02	0.11	0.02
Bottom < 3.0	0.30	0.01	0.07	0.00	0.22	0.13	0.00	0.04	0.06	0.36	0.07
Bottom < 5.0	0.47	0.05	0.13	0.00	0.31	0.22	0.03	0.07	0.10	0.55	0.12
Top < 1.0	0.01	0.01	0.02	0.00	0.06	0.02	0.00	0.43	0.47	0.26	0.17
Top < 3.0	0.02	0.02	0.05	0.00	0.18	0.05	0.00	0.78	0.85	0.90	0.50
Top < 5.0	0.03	0.04	0.08	0.00	0.28	0.09	0.01	0.90	0.94	1.00	0.76

Table 4.16. Percent of Predictions within a Given Range for Datasets Sampled to 35% in February.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.15	0.26	0.08	0.10	0.09	0.07	0.14	0.17	0.17	0.42	0.35
All < 3.0	0.43	0.64	0.22	0.32	0.27	0.21	0.38	0.43	0.52	0.84	0.77
All < 5.0	0.65	0.81	0.34	0.50	0.41	0.33	0.55	0.60	0.75	0.95	0.89
Bottom < 1.0	0.01	0.01	0.00	0.00	0.07	0.00	0.05	0.03	0.04	0.11	0.02
Bottom < 3.0	0.01	0.15	0.02	0.00	0.21	0.01	0.21	0.08	0.12	0.36	0.06
Bottom < 5.0	0.02	0.40	0.06	0.00	0.34	0.04	0.44	0.13	0.21	0.53	0.12
Top < 1.0	0.00	0.01	0.00	0.00	0.02	0.01	0.01	0.11	0.00	0.17	0.19
Top < 3.0	0.00	0.02	0.00	0.00	0.06	0.03	0.04	0.66	0.50	0.65	0.73
Top < 5.0	0.00	0.03	0.00	0.00	0.11	0.06	0.09	0.84	0.88	0.95	0.99

Table 4.17. Percent of Predictions within a Given Range for Unsampled Datasets in February.

4.2.3 May Test Data

The May test dataset was intended to be a stress test to see how much our models would deteriorate over four months. However, as evidenced by Tables 4.18 through 4.21, there was not a mass deterioration of skill between January and May. Because our models were trained on data that was collected from both the Northern and Southern Hemispheres they can experience a greater range of the seasons all within a single month. This range is what we theorize to be the key to maintaining our performance over such a long time frame.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	-0.06	-0.09	0.00	-0.27	0.22	-0.02	-0.08	0.50	0.57	0.80	0.66
Bottom R2	-166.66	-218.57	-71.74	-264.77	-141.28	-127.79	-18.14	-2.79	-2.84	-0.47	-0.85
Top R2	-334.50	-766.65	-250.34	-327.89	-88.40	-61.51	-47.85	-22.39	-6.38	-4.04	-7.94
MAE	10.55	6.59	13.77	9.28	11.52	14.46	7.37	6.92	4.38	2.05	2.50
Bottom MAE	8.35	6.00	9.49	8.58	14.88	14.19	7.17	28.80	31.96	8.50	18.28
Top MAE	132.62	91.91	115.83	67.30	34.13	66.97	34.79	3.53	2.36	1.98	2.00
MSE	777.43	320.04	664.79	310.80	271.03	497.83	135.53	130.83	71.35	14.84	32.24
Bottom MSE	119.23	65.33	126.89	166.53	330.19	293.85	111.71	1369.58	1445.34	226.47	705.68
Top MSE	18627.21	8539.93	13921.35	4949.58	1706.59	5226.29	1434.87	19.22	8.18	5.92	6.17

Table 4.18. Metrics for Datasets Sampled to 65% in May.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	-0.05	0.01	-0.15	-0.28	0.23	-0.02	-0.27	0.51	0.60	0.79	0.67
Bottom R2	-839.18	-448.71	-201.76	-336.75	-131.51	-168.20	-27.20	-2.78	-2.36	-0.37	-0.71
Top R2	-286.60	-636.96	-201.31	-293.08	-90.71	-50.15	-46.66	-15.19	-10.96	-5.65	-8.38
MAE	11.27	7.66	15.92	9.80	11.42	14.99	8.23	7.07	4.40	2.22	2.65
Bottom MAE	14.64	8.87	14.61	10.22	14.30	15.68	8.92	30.26	29.64	7.97	17.12
Top MAE	122.20	82.67	98.89	62.93	33.89	60.46	34.17	2.56	3.11	2.41	1.82
MSE	770.12	288.66	766.25	315.49	266.34	500.10	158.42	128.43	66.22	15.54	31.03
Bottom MSE	597.48	133.80	353.71	211.64	307.51	386.05	164.59	1363.03	1267.16	211.02	651.73
Top MSE	15967.86	7097.15	11205.31	4425.70	1750.66	4276.77	1400.01	13.30	13.26	7.80	6.48

Table 4.19. Metrics for Datasets Sampled to 50% in May.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	0.03	-0.05	-0.05	0.04	0.39	-0.20	0.09	0.53	0.59	0.80	0.68
Bottom R2	-1 097.08	-637.75	-144.60	-766.78	-82.08	-214.64	-28.75	-3.05	-2.34	-0.32	-0.63
Top R2	-229.79	-637.54	-204.00	-202.38	-102.90	-52.61	-36.56	-17.18	-6.91	-2.52	-12.79
MAE	11.85	8.62	15.32	11.19	10.04	16.59	7.70	6.84	4.47	2.16	2.69
Bottom MAE	17.14	11.01	13.27	21.71	12.11	17.00	10.87	29.99	28.91	7.64	16.71
Top MAE	106.27	82.24	100.37	55.02	35.97	60.73	30.94	3.22	2.33	1.59	2.50
MSE	712.72	307.41	701.40	236.26	211.72	586.66	113.75	123.39	68.43	14.67	30.08
Bottom MSE	780.87	190.04	254.01	481.10	192.80	492.00	173.62	1 462.36	1 258.54	203.32	622.54
Top MSE	12 813.63	7 103.64	11 354.42	3 060.77	1 983.48	4 481.99	1 103.20	14.93	8.77	4.13	9.53

Table 4.20. Metrics for Datasets Sampled to 35% in May.

Metric	01	02	03	04	05	06	07	10	11	12	13
R2	-0.09	-0.29	-0.01	0.11	0.41	0.08	0.03	0.54	0.64	0.81	0.67
Bottom R2	-212.68	-877.22	-165.07	-275.83	-82.05	-162.42	-22.30	-2.69	-1.86	-0.44	-0.65
Top R2	-350.01	-691.05	-245.09	-294.83	-87.68	-62.60	-42.69	-28.91	-18.87	-2.23	-8.96
MAE	11.39	8.12	14.56	8.59	9.66	13.76	7.14	6.67	4.38	1.81	2.40
Bottom MAE	10.74	11.87	15.37	12.38	10.92	16.63	8.59	27.75	25.27	8.32	16.85
Top MAE	138.35	86.83	115.25	66.35	33.85	68.56	32.84	4.45	4.47	1.52	2.03
MSE	796.93	378.45	676.01	217.56	205.15	450.73	121.35	120.77	60.06	14.45	31.69
Bottom MSE	151.95	261.29	289.70	173.46	192.74	372.86	135.96	1 332.58	1 075.84	222.46	629.71
Top MSE	19 488.56	7 698.92	13 630.46	4 452.11	1 692.98	5 317.21	1 283.44	24.57	22.04	3.80	6.88

Table 4.21. Metrics for Unsampled Datasets in May.

Once again, we do not see a striking dissimilarity between the results of the January, February, and May results when examining how many predictions were within five kelvin across the entire dataset. However, we do continue to see a noticeable deterioration among the top and bottom five percent of predictions when comparing the results to both January and February. The best performance is still being seen among the sampled datasets, and the unsampled dataset is still struggling to make these predictions on bands 1 through 7.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.16	0.24	0.07	0.11	0.07	0.07	0.13	0.16	0.25	0.37	0.36
All < 3.0	0.46	0.62	0.21	0.33	0.21	0.19	0.37	0.43	0.59	0.81	0.78
All < 5.0	0.65	0.79	0.35	0.51	0.33	0.31	0.55	0.59	0.74	0.94	0.91
Bottom < 1.0	0.03	0.02	0.02	0.01	0.03	0.03	0.05	0.02	0.00	0.14	0.03
Bottom < 3.0	0.13	0.15	0.08	0.08	0.08	0.06	0.26	0.07	0.01	0.40	0.10
Bottom < 5.0	0.31	0.50	0.21	0.27	0.12	0.11	0.49	0.12	0.02	0.59	0.19
Top < 1.0	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.08	0.20	0.28	0.31
Top < 3.0	0.00	0.00	0.00	0.01	0.05	0.01	0.03	0.47	0.71	0.79	0.75
Top < 5.0	0.00	0.00	0.00	0.02	0.09	0.02	0.06	0.83	0.95	0.96	0.96

Table 4.22. Percent of Predictions within a Given Range for Datasets Sampled to 65% in May.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.15	0.17	0.07	0.12	0.07	0.06	0.12	0.14	0.22	0.34	0.33
All < 3.0	0.42	0.46	0.20	0.34	0.21	0.18	0.34	0.39	0.56	0.78	0.75
All < 5.0	0.61	0.65	0.32	0.51	0.34	0.30	0.51	0.57	0.74	0.92	0.90
Bottom < 1.0	0.03	0.01	0.01	0.01	0.03	0.02	0.04	0.02	0.01	0.16	0.04
Bottom < 3.0	0.11	0.13	0.05	0.05	0.07	0.06	0.17	0.05	0.02	0.44	0.12
Bottom < 5.0	0.26	0.37	0.11	0.20	0.11	0.12	0.40	0.08	0.03	0.61	0.23
Top < 1.0	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.37	0.13	0.14	0.44
Top < 3.0	0.00	0.00	0.01	0.01	0.07	0.01	0.03	0.68	0.51	0.71	0.78
Top < 5.0	0.00	0.00	0.01	0.02	0.12	0.02	0.05	0.83	0.86	0.94	0.95

Table 4.23. Percent of Predictions within a Given Range for Datasets Sampled to 50% in May.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.13	0.12	0.06	0.06	0.07	0.06	0.09	0.15	0.23	0.36	0.32
All < 3.0	0.35	0.36	0.19	0.18	0.20	0.17	0.27	0.41	0.56	0.79	0.74
All < 5.0	0.52	0.56	0.31	0.31	0.34	0.27	0.43	0.58	0.74	0.92	0.89
Bottom < 1.0	0.03	0.00	0.01	0.00	0.03	0.02	0.00	0.02	0.01	0.18	0.04
Bottom < 3.0	0.12	0.00	0.04	0.00	0.08	0.07	0.00	0.06	0.03	0.47	0.13
Bottom < 5.0	0.22	0.03	0.09	0.00	0.13	0.15	0.01	0.10	0.04	0.64	0.23
Top < 1.0	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.12	0.25	0.38	0.21
Top < 3.0	0.00	0.00	0.00	0.00	0.10	0.01	0.01	0.51	0.71	0.88	0.68
Top < 5.0	0.00	0.00	0.01	0.00	0.15	0.02	0.02	0.85	0.93	0.98	0.89

Table 4.24. Percent of Predictions within a Given Range for Datasets Sampled to 35% in May.

Range	01	02	03	04	05	06	07	10	11	12	13
All < 1.0	0.14	0.24	0.07	0.10	0.08	0.06	0.13	0.15	0.18	0.42	0.37
All < 3.0	0.38	0.58	0.20	0.29	0.25	0.19	0.37	0.42	0.53	0.87	0.80
All < 5.0	0.59	0.73	0.31	0.47	0.40	0.31	0.55	0.59	0.75	0.96	0.92
Bottom < 1.0	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.03	0.01	0.15	0.05
Bottom < 3.0	0.00	0.02	0.00	0.00	0.05	0.03	0.08	0.08	0.03	0.44	0.13
Bottom < 5.0	0.01	0.07	0.02	0.00	0.13	0.08	0.29	0.13	0.06	0.60	0.24
Top < 1.0	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.41	0.34
Top < 3.0	0.00	0.00	0.00	0.00	0.06	0.01	0.03	0.25	0.13	0.88	0.75
Top < 5.0	0.00	0.00	0.00	0.00	0.10	0.02	0.05	0.69	0.70	0.99	0.92

Table 4.25. Percent of Predictions within a Given Range for Unsampled Datasets in May.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5: Conclusion

Sampling large datasets for use in regression-based tasks is a difficult problem that does not have a single solution that can be used across all domains, and the more specific a domain becomes the more difficult sampling a dataset within that domain becomes. This research has been able to show that sampling can be used to increase the ability of neural networks to make predictions on outlier data in a complex dataset without degrading the overall performance of the model. However, the increase in performance was mainly seen in areas within the dataset that we had very little skill on before sampling. That is to say that areas that we can regress on accurately without sampling do not see a large performance increase from sampling, while areas that can not be regressed on accurately without sampling see a minor performance increase. It is also shown that it is possible to build a model that can overcome concept drift and generalize well on future distributions of data.

By testing the results of all of our models on test datasets from January, February, and May we were able to determine that our models retained their skill remarkably well. When taking the average MAE across all 11 GMI bands for our models that were trained on sampled datasets we see that January has an MAE of 8.14 K, February has an MAE of 7.07 Kelvin, and May has an MAE of 8.12 K. This proves that our models can generalize well to new data and have not overfitted on distributions represented in the training set.

The greatest benefit from sampling was seen when the datasets were downsampled by a smaller amount and more records were allowed to be upsampled to fill the dataset. While this method did not generate the best results in a comparison between every other generated dataset, it was never significantly behind any other dataset in any measure of performance either, and when it was ahead of the other datasets it was typically by the most appreciable amount.

When observing the January, February, and May test datasets we see that the models that were created using the smallest downsampling method outperformed the models that were created using the dataset that had no sampling performed on it when measuring the number of predictions that fall within a 5 K range that were made on the top and bottom five percent

of each dataset. When drawing these conclusions it is helpful to look at the predictions made across all 11 bands at once, at the bottom seven bands separately, and then at the top four bands. When we look at the bottom five percent of data across all 11 bands we can see that the sampled dataset outperformed the non-sampled dataset models by 10% in January, 11% in February, and 12% in May. If we continue to look at the bottom five percent of data, but only for the bottom seven GMI bands we see that the sampled dataset outperformed unsampled dataset models by 19% in January, 21% in February, and 21% in May. Next, if we continue to observe the bottom five percent of data but for only the top four GMI bands we see the only time that the unsampled models outperform the sampled models with the unsampled models outperforming by 4% in January, 7% in February, and 2% in May.

We can also observe the difference in how well the sampled models performed against the unsampled models when making predictions on the top five percent of data. Across all GMI bands the sampled models outperformed by 10% in January, 5% in February, and 4% in May. Across only the bottom seven GMI channels we see an increase in performance by 13% in January, 33% in February, and 1% in May. And finally on the top four GMI bands we see an increase in performance when using the sampled models by 2% in January, 4% in February, and 10% in May. This proves that using the developed sampling method does provide a significant advantage to predicting under-represented data points when compared to the results from models trained on non-sampled data.

This increase in performance is important because it shows that correlations can be made and that the performance is likely to be able to continue to increase with future work. With a continued focus on sampling techniques and feature engineering, it will likely be possible to create a model that can create a synthetic GMI dataset that is accurate enough to use within current meteorological models.

5.1 Future Work

There is still more work to be done when it comes to effectively trying to create a regression-based model for the synthetic generation of GMI data from authentic ABI data. Any future work done in this area should explore the addition of the ABI bands that we removed as well as including gathered metadata, such as the viewing angle of the satellite, to create a deep and wide neural network [23]. The inclusion of this data will likely incur a slower learning

rate due to an increase in the data that must be ingested by the model, but it should also allow the model to learn more about where rare data points are commonly found and lead to a more robust model.

Data augmentation and interpolation of data points when upsampling would also be worth exploring. During the creation of the sampled datasets for this research, data from bins that needed to be upsampled were copied and then added back to the bin that they were copied from. It would likely prove beneficial to interpolate data from within a single bin, or from surrounding bins, to create similar but not exact copies of preexisting data points to increase the skill of future models. Data augmentation could then be added to the data pipeline to further increase the diversity of what each model is encountering during training. This was previously demonstrated to benefit similar problems with GMI dataset [17]. Such diversity has traditionally led to more skilled models.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

- [1] K. Jin, Y. Chen, B. Xu, J. Yin, X. Wang, and J. Yang, “A patch-to-pixel convolutional neural network for small ship detection with polsar images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6623–6638, 2020 [Online].
- [2] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, “Smote for regression,” in *Progress in Artificial Intelligence [Online]* (Lecture Notes in Computer Science). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 378–389.
- [3] S. Tammina, “Transfer learning using VGG-16 with deep convolutional neural network for classifying images,” *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, pp. 143–150, 2019 [Online].
- [4] J. Togelius, S. Karakovskiy, J. Koutnik, and J. Schmidhuber, “Super mario evolution,” in *2009 IEEE Symposium on Computational Intelligence and Games*, 2009 [Online], pp. 156–161.
- [5] I. M. Nasser and S. S. Abu-Naser, “Lung cancer detection using artificial neural network,” *International Journal of Engineering and Information Systems (IJEAIS)*, vol. 3, no. 3, pp. 17–23, 2019 [Online].
- [6] A. Dongare, R. Kharde, A. D. Kachare *et al.*, “Introduction to artificial neural network,” *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 2, no. 1, pp. 189–194, 2012 [Online].
- [7] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in *2017 International Conference on Engineering and Technology (ICET)*. IEEE, 2017 [Online], pp. 1–6.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016 [Online], pp. 770–778.
- [9] P. Branco, L. Torgo, and R. P. Ribeiro, “Pre-processing approaches for imbalanced distributions in regression,” *Neurocomputing*, vol. 343, pp. 76–99, May 2019 [Online].
- [10] R. Timofeev, “Classification and regression trees (cart) theory and applications,” *Humboldt University, Berlin*, vol. 54, 2004 [Online].

- [11] T. R. Hoens and N. V. Chawla, “Imbalanced datasets: From sampling to classifiers,” *Imbalanced learning: Foundations, algorithms, and applications*, pp. 43–59, 2013 [Online].
- [12] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009 [Online].
- [13] P. Branco, L. Torgo, and R. P. Ribeiro, “REBAGG: REsampled BAGGING for Imbalanced Regression,” in *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*. PMLR, Nov. 2018 [Online], pp. 67–81. Available: <https://proceedings.mlr.press/v94/branco18a.html>
- [14] L. Torgo, P. Branco, R. P. Ribeiro, and B. Pfahringer, “Resampling strategies for regression,” *Expert Systems*, 2015 [Online].
- [15] P. Branco, L. Torgo, and R. P. Ribeiro, “SMOBN: A pre-processing approach for imbalanced regression,” in *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 2017 [Online].
- [16] T. Meissner, F. Wentz, and D. Draper, “Gmi calibration algorithm and analysis theoretical basis document,” *Remote Sensing System Technical Note*, 2012 [Online].
- [17] V. Petković, M. Orescanin, P. Kirstetter, C. Kummerow, and R. Ferraro, “Enhancing pmw satellite precipitation estimation: Detecting convective class,” *Journal of Atmospheric and Oceanic Technology*, vol. 36, no. 12, pp. 2349–2363, 2019.
- [18] “GPM Microwave Imager(GMI),” accessed Feb. 10, 2022 [Online]. Available: <https://gpm.nasa.gov/missions/GPM/GMI>
- [19] “Advanced Baseline Imager(ABI),” accessed Feb. 10, 2022 [Online]. Available: <https://www.goes-r.gov/spacesegment/abi.html>
- [20] “Advanced Baseline Imager(ABI) Bands,” accessed Feb. 10, 2022 [Online]. Available: <https://www.goes-r.gov/mission/ABI-bands-quick-info.html>
- [21] T. Jackson, D. Le Vine, A. Hsu, A. Oldak, P. Starks, C. Swift, J. Isham, and M. Haken, “Soil moisture mapping at regional scales using microwave radiometry: The southern great plains hydrology experiment,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 5, pp. 2136–2151, 1999 [Online].
- [22] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017 [Online].

- [23] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir *et al.*, “Wide & deep learning for recommender systems,” in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, 2016 [Online], pp. 7–10.

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
Ft. Belvoir, Virginia
2. Dudley Knox Library
Naval Postgraduate School
Monterey, California