THESIS

ASSESSING THE STATE-DEPENDENCY OF INFRARED SATELLITE PRECIPITATION ERRORS

Submitted by

Eric Goldenstern

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2022

Master's Committee:

Advisor: Christian Kummerow

Christine Chiu
Imme Ebert-Uphoff

ABSTRACT

ASSESSING THE STATE-DEPENDENCY OF INFRARED SATELLITE PRECIPITATION ERRORS

The sensing and prediction of precipitation remains at the forefront of weather forecasting, building upon centuries of measurement and study. While in-situ and ground-based methodologies such as rain gauges and weather radars provide the best assessments of precipitation, they are prone to sampling issues and coverage gaps both over challenging terrain and in developing areas of the world. As a result, the use of remote sensing methodologies, namely satellites, have allowed for the expansion of precipitation measurement to encompass nearly the entire Earth. However, unlike rain gauges, satellites are incapable of directly sensing precipitation; rather, they must infer it from the spectral information that can be captured from space through a mathematical framework known as a retrieval.

While satellite precipitation retrievals are a boon to the meteorological community due to their ability to fill in these coverage gaps, their indirect nature inevitably gives rise to errors in the measurements themselves. Furthermore, these errors have historically been specific to their training area and are not directly comparable to the errors in other areas. Therefore, this thesis aims to begin disentangling these errors into more generalizable metrics through known information about the measurements themselves and the environmental state being observed. To do this, a neural-network style retrieval algorithm was developed using infrared and lightning data from the Geostationary Operational Environmental Satellite – 16 (GOES-16) to create a

validation statistics study. The error from this retrieval, selected to be its bias statistic, was then analyzed both in the context of the satellite data and ancillary meteorological data. From these analyses, it was shown that an understanding of the satellite data allows for limited reproducibility of the retrieval bias tendencies across multiple areas of study, and that ancillary environmental information can shed additional light on how these errors are influenced by the underlying meteorological state. Though this thesis does not create an exact, quantitative methodology for such an assessment, it does provide a direction in which a framework can be established to predict precipitation uncertainties for a more global perspective.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Christian Kummerow, for his continued guidance and patience while developing this thesis. I would also like to thank my committee members, Dr. Christine Chiu and Dr. Imme Ebert-Uphoff, for all of their help, both directly and indirectly, as this thesis progressed. A special thank you to Kyle Hilburn, both for providing me with an exceptional dataset to work with and for answering my many machine-learning related questions. A thank you as well to the Kummerow research group members, past and present; having you all around to talk about nearly any topic has made my time here memorable.

To my family and friends, I cannot thank you all enough. I could not have become the person I am today without all of you in my life. To all of the teachers and mentors I have had in the past, I am truly indebted to you all. Your guidance and knowledge cultivated my desire to learn, and is what allowed me to reach this point.

Lastly, a thank you to the whole of the CSU Atmospheric Science department. The community that has been cultivated here is second to none and has allowed me to flourish as a scientist and as a person. A special thank you to all those CSU staff working in the front office, as custodians, and elsewise. Our community continues to run smoothly thanks to you all, and I could not be more grateful for that.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iv
CHAPTER 1: INTRODUCTION 1
CHAPTER 2: DATA AND ALGORITHM DESCRIPTION 10
2.1 ABI and GLM Description 10
2.2 MRMS System Description 12
2.3 ERA5 Ancillary Meteorological Data Description 12
2.4 Data Pre-Processing 15
2.5 GPE-CNN Precipitation Retrieval Description 19
2.6 GPE-CNN Validation Performance 24
CHAPTER 3: ANALYSES AND RESULTS 29
3.1 Retrieval Performance: Sector Intercomparisons 29
3.2 Retrieval Performance: Training Area Analysis 37
3.3 Satellite Data Partitioning 44
3.4 Environmental State Influences 52
CHAPTER 4: DISCUSSION AND SUMMARY 61
REFERENCES 66
APPENDIX A: IMPACTS IN REMAINING SECTORS 70

CHAPTER 1: INTRODUCTION

Water is one of the most ubiquitous materials on Earth, making up roughly 70 percent of the Earth's surface and being present as one of the most universal and influential atmospheric constituent gases. Because of this prevalence and the importance of water for sustaining life, the transferal of water between its atmospheric and terrestrial sources has been documented and studied for centuries. Known as the hydrological cycle, this continual transfer of water through the Earth system has been categorized into several processes, including evaporation, vapor transport, and precipitation (Chahine 1992; Chen and Pfaendtner 1993). Of these processes, precipitation is likely the most influential on everyday life. Changes in precipitation can result in droughts and flooding, both of which present unique and potentially severe challenges to the native flora and fauna. Specifically, drought-stricken areas become more prone to wildfires, as the fuel loading in the area is increased due to a lack of moisture (Bailing et al. 1992), while flooding events are widely considered one of the most impactful atmospheric phenomena in terms of death and property destruction (Pielke and Downton 2000). Given these high-impact consequences, it is important now more than ever to understand precipitation processes. In doing this, the quantification of uncertainties is a central requirement, as it is this knowledge that influences our confidence in precipitation measurements.

The need for actionable precipitation estimates has led to the development of many different technologies capable of directly and indirectly assessing precipitation both before and during its occurrence. Initially, precipitation data was largely collected using ground-based in-situ measurement devices, known as rain gauges. Even in modern scientific applications, rain gauges are generally considered the most reliable means of assessing precipitation. These

gauges, however, are not perfect, as the measurements are highly affected by the wind, which can cause two gauges in close proximity to report substantially different precipitation amounts (Habib et al. 1999). Also, rain gauge networks tend to be concentrated in wealthier and more highly-populated areas, causing large amounts of the Earth to be nearly unobserved.

With the development of further indirect measurement technologies, two methods have become widely accepted by the scientific community: RAdio Detection And Ranging (RADAR) and satellite instrument analyses. RADAR is a generally ground-based technique in which pulses of microwave energy are directed at a precipitating system. This energy is then scattered by precipitation-sized hydrometeors and the portion of said radiation which is directed back at the sensor is collected and analyzed. This is what creates the RADAR "echoes" that most are familiar with. While these devices have the advantage of better spatial coverage in comparison with rain gauges, RADAR also presents some unique challenges. Firstly, these microwave pulses tend to spread horizontally and vertically with distance traveled, meaning that the returned energy from a distant system may not fully capture the structure or intensity of the identified precipitation. Also, these devices are subject to many different types of interference that can be terrestrial in nature, such as beam blocking due to nearby orographic features, or based in atmospheric conditions, such as from anomalous propagation related to static stability.

Analyses of precipitation from satellites have seen an increase in use in recent years. Beginning with the first meteorological satellite image received from the Television Infrared Observation Satellite 1 (TIROS-1) in 1960, satellite instrumentation has developed rapidly, allowing for observations of atmospheric and terrestrial features across much of the electromagnetic spectrum. Modern environmental satellites are generally identified as Low-Earth Orbiting (LEO) or Geostationary Orbiting (GEO) sensors, with each orbit strategy having certain

advantages and drawbacks. LEO systems generally orbit within a few hundred kilometers above Earth's surface, allowing for high spatial resolution imaging along the satellite's orbital path. Also, LEO satellites typically carry microwave sensors, which are uniquely suited for precipitation assessment due to the ability of microwave energy to penetrate cloud features. These sensors, however, have poor temporal resolution, often only seeing a given location on Earth twice a day. As such, though LEO satellites often perform very well in understanding precipitation due to the fine-scale features they can see and the high information content in their spectral data, they are often not as useful for real-time assessment of precipitating systems. GEO satellites, on the other hand, orbit at roughly 36,000 kilometers above Earth's surface. This orbital strategy allows these sensors to view the same area of Earth continuously, as at this orbit, the satellite is able to match the speed of Earth's rotation. This distance from Earth, however, degrades the spatial and spectral resolution of the sensor, resulting in data with good temporal coverage and very little temporal latency but decreased information content regarding precipitating systems.

Despite the inherent limitations of the sensors, satellites are becoming increasingly favored for precipitation assessment, as they are able to provide data coverage where rain gauges and RADAR are unable to. These satellites, however, do not measure precipitation itself, but rather infer it through the spectral information gathered. As such, relating the information obtained by the satellite to the occurrence and properties of precipitation is necessary. This relationship is developed through what is known as a retrieval. Satellite retrievals can come in several varieties, each with their own optimal uses. Interestingly, many of these algorithms use either LEO or GEO satellite data primarily, with the other type playing a much lesser or non-existent role in the retrieval. This thesis focuses on GEO infrared algorithms, as their use is

preferred for applications regarding situations that require fast temporal refresh at high spatial resolutions. GEO precipitation retrieval algorithms have existed for many decades, first beginning with the climatological threshold-regression technique of the Geostationary Operational Environmental Satellite (GOES) Precipitation Index (GPI; Arkin and Meisner 1987, Joyce and Arkin 1997), advancing to the multispectral approach of the GOES Multispectral Rainfall Algorithm (GMSRA; Ba and Gruber 2001), and progressing into the hybrid LEO-GEO strategy of the current National Oceanic and Atmospheric Administration (NOAA) operational algorithm, the Self-Calibrating Multivariate Precipitation Retrieval (SCaMPR; Kuligowski 2002, Kuligowski et al. 2016) and the neural network approach pioneered by the Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks (PERSIANN; Hsu et al. 1997, Hong et al. 2004, Sadeghi et al. 2019) algorithm.

While such retrievals allow us to better assess precipitation characteristics with near-global coverage, such algorithms are not exact representations of the current state, and therefore have inherent errors associated with them. Such errors are quantified through the use of validation analyses, which use a 'ground truth' dataset to allow the retrieval algorithm to be tested on data which it has not previously seen. Such activities provide actionable information on the behaviors of a given retrieval algorithm over the validation area. These methods, however, often lack the ability to fully explain the retrieval errors over areas where the data necessary for validation does not exist. Also, these analyses are often only related to the environment anecdotally, and therefore do not present a concrete interpretation of which set of conditions create trustworthy estimations and which do not. An example of this discrepancy can be seen through the evaluation of an improved PERSIANN architecture, the

PERSIANN Cloud Classification System (PERSIANN-CCS; Hong et al. 2004), over different geographic areas. The performance of this algorithm was first highlighted through a case study of a flash-flooding event near Las Vegas, Nevada on 8-9 June 1999. For this event, PERSIANN-CCS achieved a correlation with radar observations of about 0.45 with a 4km, hourly data resolution. However, another study performed by Mastrantonas et al. 2019 examined the performance of PERSIANN-CCS, among other retrievals, over the Kinu Basin in Japan with data extending from 2015-2016 and found that this correlation was around 0.22 at the same spatiotemporal resolution. In fact, PERSIANN-CCS was concluded to be the worst-performing retrieval of the five algorithms tested in this region. While the greater data volume in Mastrantonas et al. 2019 likely played a role in this difference, it can be seen that once an algorithm is removed from its calibration domain, its 'trustworthiness' can fluctuate, meaning that the retrieval's validation statistics are not generalized to cases outside of its training area. As such, being able to understand the error characteristics of these algorithms is essential to understanding their best use scenarios.

A more nuanced interpretation of retrieval errors can come from two perspectives: through the optical properties and inherent limitations of the satellite data utilized to classify performance, or through ancillary information regarding the meteorological state. To date, most focus has been on assessing the satellite sensor properties to understand a retrieval's error characteristics; this is a logical first step, as it is these data that generate the retrieval in the first place. Such attempts include the Precipitation Uncertainties for Satellite Hydrology (PUSH; Maggioni et al. 2016) algorithm, which performs a pixel classification based on a minimum precipitation threshold and approximates errors through probability density function (PDF) matching. This methodology allows for the retrieval itself to be compared with its 'ground truth'

in a way that provides context to its systematic and random errors. Few studies attempt to quantify errors from a 'regime' perspective, likely owing to the ambiguity by which a 'regime' can be defined. Still, Henderson et al. 2017 identified precipitation regimes via cluster analysis of derived satellite precipitation products that produced similar results across multiple regions and seasons, creating a basis by which errors within oceanic precipitation retrievals can be assessed without input from additional ground validation data.

Such studies have provided frameworks by which satellite retrieval errors can be related to properties captured by the satellite data, as they hypothesized that these data can be used to develop self-similar regimes even in geographically distinct regions. These methods, however, do not explicitly consider the effects of the meteorological state on precipitating systems. This is an important aspect of assessing precipitation, as the environment can modify the larger-scale cloud features of precipitating systems, thereby affecting the intensity and type of precipitation that occurs. As such, this thesis aims to explore the precipitation characteristics and meteorological state dependency of satellite precipitation retrieval errors. To date, such a methodology has not been deeply explored, despite its potential usefulness in explaining why precipitation retrievals perform well in some contexts and poorly in others. In doing so, several assessments of how this state dependency can be characterized were developed, combining information from satellite radiance data and meteorological state variables to describe retrieval errors without the use of a 'ground truth'.

To accomplish this task, a simple retrieval algorithm was built. Since the algorithm was not intended to be operational, it was left mostly unoptimized, meaning that it develops its predictions through the relationships between precipitation and satellite information without the input of any prior rain/no-rain discrimination or seasonal variations. The retrieval system which

this framework was designed around is an encoder-decoder style neural network retrieval similar in design to the GOES Radar Estimation via Machine Learning to Inform NWP (GREMLIN) algorithm described by Hilburn et al. (2020); this neural network is hereafter referred to as the GOES Precipitation Estimator using Convolutional Neural Networks (GPE-CNN). The predictor data came from the GOES-16 satellite, utilizing normalized data from channels 08, 09, 10, 13, and 15 along with a derived group-extent density from GLM developed by Hilburn et al. (2020). These channels were chosen such that GPE-CNN would only operate with infrared and lightning data, which circumvents the disruption in data that occurs in the visible and shortwave infrared channels due to the diurnal cycle. The 'ground truth' which the algorithm was trained to estimate came from the Multi-Radar Multi-Sensor (MRMS; Zhang et al. 2016) PrecipRate product. MRMS is a system of algorithms developed by the National Severe Storms Laboratory (NSSL) which were designed to integrate information from RADAR, surface observations, lightning networks, satellite observations, and numerical models to develop radar-based products over the continental United States (CONUS). Both the GOES-16 and MRMS data were resampled to the 3km High Resolution Rapid Refresh (HRRR) model mass grid via group averaging, such that all data types could be adequately compared.

In combination with this simple retrieval and post-processing framework, analyses of both the satellite-based and ancillary environment information were used to develop links between the environment and the validation bias statistic. Such an analysis would allow for certain bias regimes present within the retrieval to be attributed to more specific conditions regarding the atmospheric state, thus allowing for a more holistic assessment of retrieval performance. This was important for increasing confidence in areas where the retrieval algorithm was not validated. An example of this is displayed in Figure 1, in which Panel a represents the

retrieval's performance over its validation area in the Central Plains and Panel b represents the retrieval's performance over the Northwestern CONUS. From this, one can see that the retrieval has very different representations between the two regions, despite them being located within CONUS. These differences were reflected in the validation statistics, as the Central Plains retrieval was much more closely aligned with MRMS than the Northwest. Since the main difference between these regions is the environment in which the respective precipitation developed, determining those conditions as they relate to the precipitation would allow for one to better characterize these performances, thus aiding in the interpretation of the retrieval precipitation. In essence, the purpose of this thesis is to assess the ability to generalize the validation statistics taken over the Central Plains, where said statistics are fairly well-known, and using information contained both within the satellite data and other ancillary environmental data, to predict the validation performance in other areas, the Northwest CONUS standing as one such example. With careful interpretation, such a relationship between validation statistics and external information would allow for substantially more confident applications of these retrievals in areas where proper validation data is sparse or nonexistent.

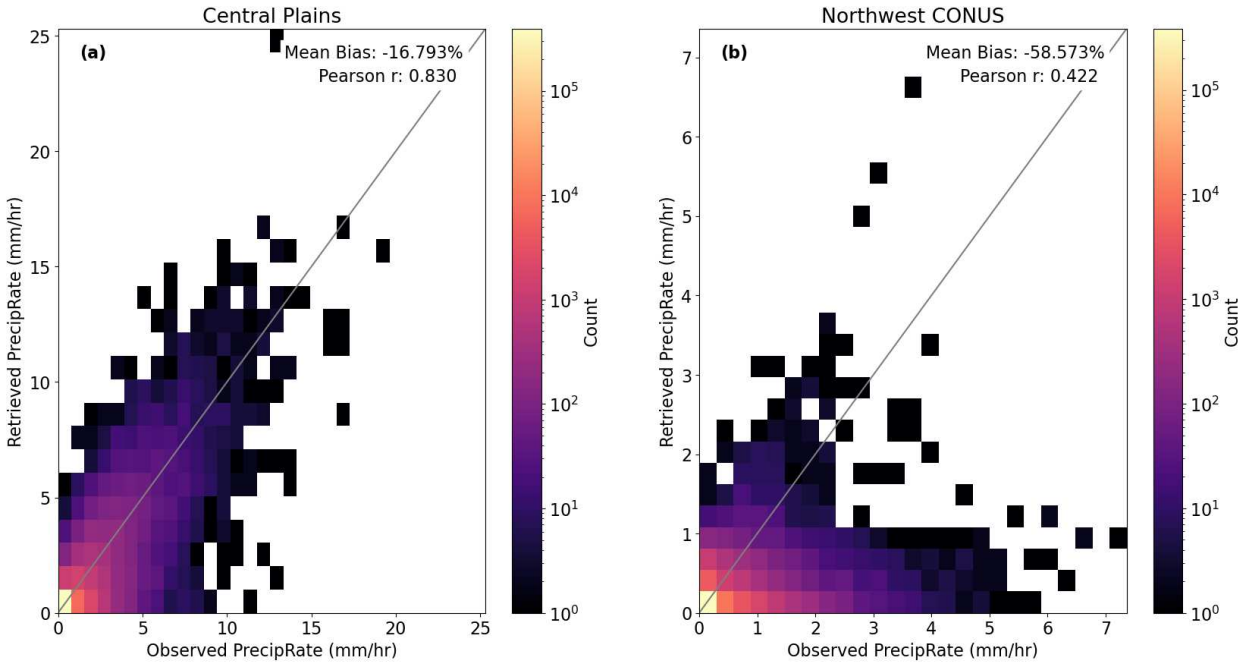

Figure 1: 2D Histograms of retrieved vs observed PrecipRate for the (a) Central Plains and (b) Northwest CONUS regions. By comparing the two panels with each other, conclusions regarding the agreement of the two regions can be made, both through their qualitative and quantitative differences.

The remainder of this thesis is structured as follows. Chapter 2 details the data and the retrieval algorithm structure utilized for this thesis and provides insight as to the rationale behind the construction of both aspects. Chapter 3 discusses the performance of the retrieval algorithm, both within its given training area and in other areas over which the retrieval was not trained. These areas include four similarly-sized domains over different areas of the CONUS. Also, details regarding the efforts taken to understand the ability of the satellite, environmental, and combined information sets to explain the bias tendencies noted within the retrieval are included. Chapter 4 will summarize these findings, as well as provide avenues of approach for future development.

CHAPTER 2: DATA AND ALGORITHM DESCRIPTION

Throughout this thesis, multiple data sources with differing spatial scales were utilized to explore the ability of ancillary data to assess the bias tendencies of a simple, geostationary precipitation retrieval. These data included observations from the Earth-observing sensors aboard the Geostationary Operation Environmental Satellite 16 (GOES-16), known as the Advanced Baseline Imager (ABI) and Geostationary Lightning Mapper (GLM), the PrecipRate product developed by the Multi Radar Multi Sensor (MRMS) system, and meteorological state variables from the European Centre for Meteorological Weather Forecasting (ECMWF) Reanalysis, 5th Generation (ERA5). The GOES and MRMS data were utilized for the development and training of the precipitation retrieval algorithm, GPE-CNN, described in greater detail in subsequent sections. The ERA5 data was utilized outside of the retrieval algorithm development; instead, its purpose was to assess the ability of the environmental state to assess the regional biases present in the retrieval estimates.

2.1 ABI and GLM Description

The primary data used in this thesis were from the GOES-16 ABI and GLM sensors. These sensors have been operational since December 2017, and have been important in increasing the relevant information about atmospheric processes from geostationary orbit. Though mounted on the same satellite bus, ABI and GLM have different functions. ABI observes electromagnetic radiation in 16 spectral channels that span the visible-infrared (VIS-IR) spectrum and includes an onboard blackbody calibration device to stabilize the radiances observed by the sensor (Schmidt et al. 2017). The 16 spectral channels are a marked improvement in spectral resolution over the five spectral channels of the legacy GOES sensors.

ABI also greatly improves upon the legacy sensors' spatiotemporal resolution. The spectral channels observed by ABI have varying spatial resolutions, with the red-visible channel at a resolution of 0.5km, the near-infrared channels at 1km, and the infrared channels at 2km; legacy GOES sensors had spatial resolutions at least twice as coarse. ABI is also capable of delivering observation of the CONUS with a 5-minute temporal resolution, which is much finer than the 15-minute resolution of legacy sensors.

GLM differs from ABI both in its spectral viewing and its operating style. While ABI utilizes a suite of channels spanning over the VIS-IR spectrum, GLM uses a single spectral band focused in the near-infrared to detect lightning flashes. It does so through its scanning strategy as a high-speed event detector, utilizing a wide field of view (FOV) and narrow band filtering to detect the minute radiance changes indicative of lightning (Goodman et al. 2013). Because of this strategy, GLM's spatial resolution varies widely depending on proximity to nadir, or the center of the sensor's FOV; at nadir, resolution is roughly 8km, but degrades to about 14km at the limbs. This viewing strategy does allow for near-hemispheric coverage, meaning that GLM observations are taken alongside the 15-minute full-disk temporal strategy. The inclusion of lightning data in the retrieval allows for a more informed assessment of the convective capability of a given scene; the presence of lightning is a good proxy for convective strength, as systems with significant convective activity are able to suspend much more ice within their updrafts, and thus increase the lightning flash activity within the storm. For use as a predictor in this thesis, the GLM data was processed identically to Hilburn et al. (2020), where the group data were taken to create group-event density (GED) maps. This was done by accumulating group events using the group area over a 15-minute timespan.

2.2 MRMS System Description

When developing a retrieval algorithm, it is necessary to consider the desired output of the algorithm. In this case, the PrecipRate product from the National Oceanic and Atmospheric Administration (NOAA) National Severe Storms Laboratory (NSSL) MRMS system was utilized as the target and ground truth for the retrieval algorithm. MRMS is a blended data system that uses the United States Weather Surveillance Radar-1988 Doppler (WSR-88D) radar network and some weather radars operated by Environment Canada as its base data, with observations from CONUS rain gauge networks, Rapid Refresh (RAP) model data, and satellite observations blended into the dataset. By combining these data types, MRMS is able to generate composite radar variables alongside several quantitative precipitation forecasting (QPF) and severe weather forecasting products (Zhang et al. 2016). The precipitation products are delivered on a 1km, 2-minute spatiotemporal resolution, providing information regarding fine-scale atmospheric processes. MRMS precipitation products also have a rich history in satellite precipitation estimation, as with the Global Precipitation Mission (GPM; Kirstetter 2012, 2014) as well as GOES-16 (Upadhyaya et al. 2018, Sun et al. 2020). The MRMS data utilized in this thesis was taken from the standard data processing of MRMS, which are located in the Iowa State University meteorological data archive (https://mtarchive.geol.iastate.edu).

2.3 ERA5 Ancillary Meteorological Data Description

To understand the implications which the environment has on the retrieval, ancillary meteorological data was compiled to assess the large-scale atmospheric state. For the purposes of this thesis, the ERA5 model was utilized. While not a real-time model, ERA5 is able to produce model observations on a global scale at fine spatiotemporal resolution. This iteration of the

ECMWF reanalysis was implemented as the improved replacement of the ERA-Interim model, described by Dee et al. (2011). These improvements involved multiple aspects of the model, including its spatiotemporal resolution and its data assimilation methodology. More specifically, the ERA5 model delivers estimates at roughly quarter-degree resolution for its surface products and half-degree resolution for its atmospheric level products, as opposed to the roughly three-quarter-degree resolution of ERA-Interim. This increase in resolution allows for more precise assessment of finer-scale features, somewhat decreasing the burden on the parameterization schemes used in the model. The temporal resolution of ERA5 was also increased to hourly from ERA-Interim's maximum 3-hourly resolution. These improvements in spatiotemporal resolution altogether have been successful in adding roughly a day in skill to the reanalysis forecasts, thus improving the overall dependability of ERA5 in comparison with its predecessors. Data assimilation in ERA5 is done using the Integrated Forecast System (IFS) cycle 41r2. This cycle utilizes a four-dimensional variation (4DVAR) scheme to assimilate a wide variety of data into the reanalysis model. The ERA5 meteorological information used in this thesis is a combination of both surface and atmospheric level data.

Given the large number of state variables available in ERA5, a selection of those available were chosen for further analysis based on their potential relationship with precipitation. These included both native ERA5 variables and derived quantities from said native variables. Because the full combination of all remaining state variables that could be examined from ERA5 was also deemed too large for the problem at hand, a reduction was performed to determine which variables would be included in the environmental analysis. This analysis involved two methods by which the variable selection occurred: correlation with bias and independence of information. For the correlation analysis, the Pearson correlations for the relevant ERA5

variables were taken against the GPE-CNN bias statistic. From this analysis, only those variables with a correlation of at least 0.09 were retained; this correlation value was chosen based on the observed correlations of the relevant variables to highlight those with the strongest linear relationship with the retrieval bias. This left seven variables for further analysis. For the independence test, a correlation matrix was constructed for investigation. From here, the variable suite was reduced until no two variables were correlated greater than 0.5. This added two more variables, 2-meter dewpoint (d2m) and 700hPa wind shear (WSS700), to the original seven. This resulted in nine total variables for analysis that satisfied at least one of the inclusion tests, those being the 700hPa relative humidity (RH700), the difference between RH700 and the surface relative humidity (dRH700), the 500hPa relative humidity (RH500), the difference between RH500 and the surface relative humidity (dRH500), the 500hPa specific humidity (q500), the 500hPa and 700hPa meridional wind speed (v500 and v700), and the 2-meter dewpoint and 700hPa speed shear previously identified. These variables are described in Table 1.

Table 1: The selected ERA5 variables and their bias correlations. These variables were selected for further analysis owing to their relationship with this thesis's retrieval bias and the independence of the information they were expected to provide.

	RH700	dRH700	RH500	dRH500	q500	v500	v700	d2m	WSS700
r	-0.18	-0.15	-0.13	-0.12	-0.12	-0.1	-0.09	-0.03	-0.03

With this group of variables, there were a couple of things to notice. Most notably, six of the nine variables used here were related to moisture, while the remaining three were wind variables. This set of variables did make physical sense, as available moisture and wind shear both are central to the types of precipitation that the environment can support. These variables were also consistent with other works that explored the relationship between meteorological

variables and precipitation (Merenti-Välimäki and Laininen 2002, Loriaux et al. 2016, Petkovic et al. 2018). Another notable aspect of this group was that two of its members, d2m and WSS700, were only weakly correlated with bias. Despite their weak relationship with the retrieval bias, they do carry physical significance with precipitation, even if the relationship is somewhat tedious. Dewpoint temperature is necessary for understanding if the near-surface environment is saturated, which would support the ability for precipitation to reach the surface. The wind shear at 700hPa has traditionally been used by weather forecasters as a proxy for convective organization, since stronger vertical wind shear often indicates that the environment is conducive to the development of more robust convective systems, which can produce enhanced precipitation. As such, though they appear unimportant in a statistical sense, these two variables were retained due to their relationship with precipitation as used by forecasters.

2.4 Data Pre-Processing

Before the data described above can be utilized in either the retrieval algorithm or bias estimator, they must undergo a requisite amount of preprocessing. The first step of doing this was to determine which of the available GOES-16 channels would be used as predictors for the retrieval. For the purposes of this thesis, it was decided that GOES water vapor absorption channels (ABI Channels 08; 6.2 microns, 09; 6.9 microns, 10; 7.3 microns), the 'clean' longwave channel (ABI Channel 13; 10.3 microns), the 'dirty' longwave channel (ABI Channel 15; 12.3 microns) and the GLM GED derived data. The ABI channels were chosen all record observations within the infrared portion of the electromagnetic spectrum; this choice was made to avoid the issues with the diurnal cycle that arise with the use of visible and shortwave infrared radiance data. The water vapor channels were chosen because they respond directly to

atmospheric moisture. These channels exist within the water vapor absorption window, meaning that in areas where the sensor picks up less infrared radiation from the Earth, there was likely a notable presence of water vapor. The choice to use all three available water vapor channels was made with the understanding that these channels each interact with different levels of the atmosphere, which would provide for a more comprehensive interpretation of atmospheric moisture. The two longwave infrared channels were chosen due to their utility as split-window difference inputs in assessing whether a given cloud feature was likely to be precipitating. Though the split-window difference itself was not used in the retrieval, the inclusion of both infrared channels was hoped to provide the same level of information. The difference between the two channels lies in where each one is centered. The 'clean' infrared channel has a central wavelength of 10.3 microns, placing it at the upper bound of the infrared atmospheric window. As such, this channel is less attenuated by the atmosphere, thus providing a cleaner view of the Earth. The 'dirty' channel, however, centers at a wavelength around 12.3 microns. At this wavelength, atmospheric absorption is more pronounced, resulting in greater contamination of a given scene by several atmospheric constituents, namely clouds. Though infrared radiative properties are only indirectly linked to precipitation, it was hoped that these predictors would create an adequate representation of precipitation process owing to their overlapping information content. These data were also normalized via min-max scaling, such that no one input variable gains predominance solely due to scale. This scaling operates as such:

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X_n is the normalized data, X is the raw data, X_{min} is the minimum scaling value, and X_{max} is the maximum scaling value. Table 2 details the retrieval predictors and target further.

The datasets described above also exist at differing spatiotemporal resolutions, and therefore must be resampled to ensure proper comparisons. In the case of the GOES-16 and MRMS data, this resampling fulfilled two objectives: for use in GPE-CNN and for use in the bias analyses. For GPE-CNN, these data were resampled to the 3km spatial resolution High Resolution Rapid Refresh (HRRR) model mass grid from their native resolutions shown in Table 2 and were matched with the GOES-16 Full-Disk temporal resolution of 15 minutes. This matching resulted in the CONUS3 dataset, as developed by Hilburn et al. 2020. The temporal resolution for this thesis was designed to be hourly; this was done by taking only the top-of-hour data. This was done for all available data in 2018 and for data from January, April, July, and October 2019. The data were also broken into five sectors, each representing a different geographic region of the CONUS; these sectors include the Central Plains, Northeast, Southeast, Northwest, and Southwest (Figure 2). Each sector encompasses a 384x384 pixel area, which equates to a roughly 10.5x10.5 latitude/longitude degree area.

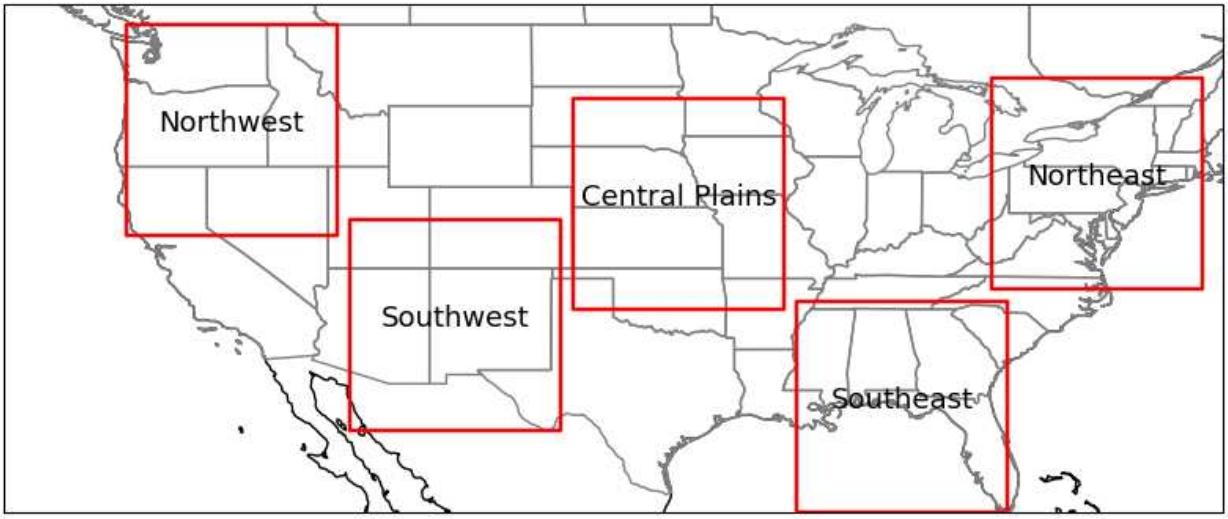

Figure 2: A representation of the five sectors utilized in this thesis. Each sector is to be analyzed by a retrieval algorithm developed specifically over the Central Plains sector.

Table 2: Predictor and target data information for GPE-CNN. Both the predictors, which are the channel and GLM data, and the target MRMS data are shown here with their native resolutions and with the minimum (X_{min}) and maximum (X_{max}) values by which they were normalized.

	CH08	CH09	CH10	CH13	CH15	GLMGED	MRMS
Native Resolution	2km Geostationary	2km Geostationary	2km Geostationary	2km Geostationary	2km Geostationary	~8km Geostationary	1km Lambert Conformal
X_{min}	200 K	200 K	200 K	200 K	200 K	0	0 mm/hr
X_{max}	250 K	250 K	250 K	300 K	300 K	50	50 mm/hr

For use in analyzing the large-scale patterns that influence retrieval bias, the data were coarsened to a 96x96km spatial resolution, with each coarsened gridpoint representing a subdomain of the given sector. This resolution was chosen to most closely align with a 1x1 latitude/longitude degree area where a round number of said subdomains can be created within the given sector. Figure 3 displays an example of how these subdomains are aligned within a sector. In the case of ERA5, the data are first resampled to the 3km HRRR grid via nearest neighbors, then are coarsened by averaging the 3km resampled data to the subdomain resolution. For the GOES data, the resolution coarsening also occurred from the 3km to subdomain resolution, but were done for several descriptive statistics, namely means, standard deviations, as well as binary groupings in the case of the lightning data. The effects of these aggregations will be discussed further in Chapter 3.

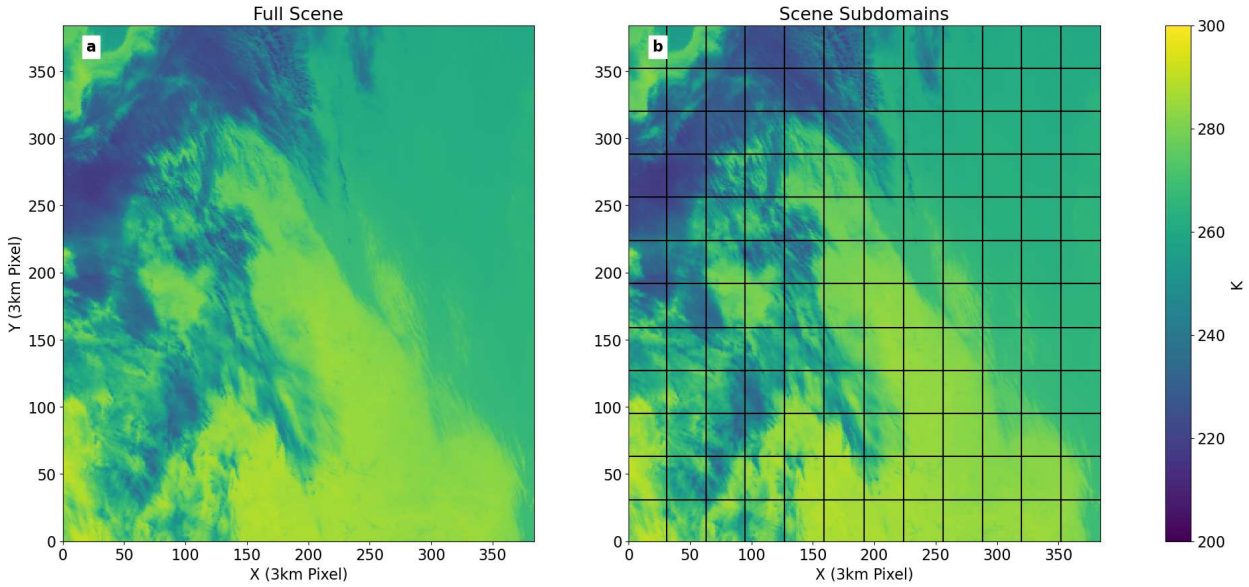

Figure 3. An example of (a) a sector scene and of (b) its subdomain partitioning. The subdomains are displayed as the gridded area bounded by the black lines in panel b, and will be utilized primarily in subsequent analyses.

2.5 GPE-CNN Precipitation Retrieval Description

As described in Chapter 1, the retrieval developed for this thesis was an encoder-decoder style convolutional neural network, named GPE-CNN, that was similar in design to the GREMLIN algorithm (Hilburn et al. 2020). GPE-CNN's architecture utilizes two-layer, double-convolution encoder and decoder blocks, with maxpooling, upsampling, and Rectified Linear Unit (ReLU) activation between layers (Figure 4). The use of an encoder-decoder style network was chosen due to its utility in mapping a set of images to another target image, known as image-to-image translation. This is done by considering the input images in fragments, known as kernels, from which information is extracted through the use of filters that adapt to the data as it is seen. In doing this, the network then learns patterns within the image based on the points themselves as well as the spatial context provided by the kernels.

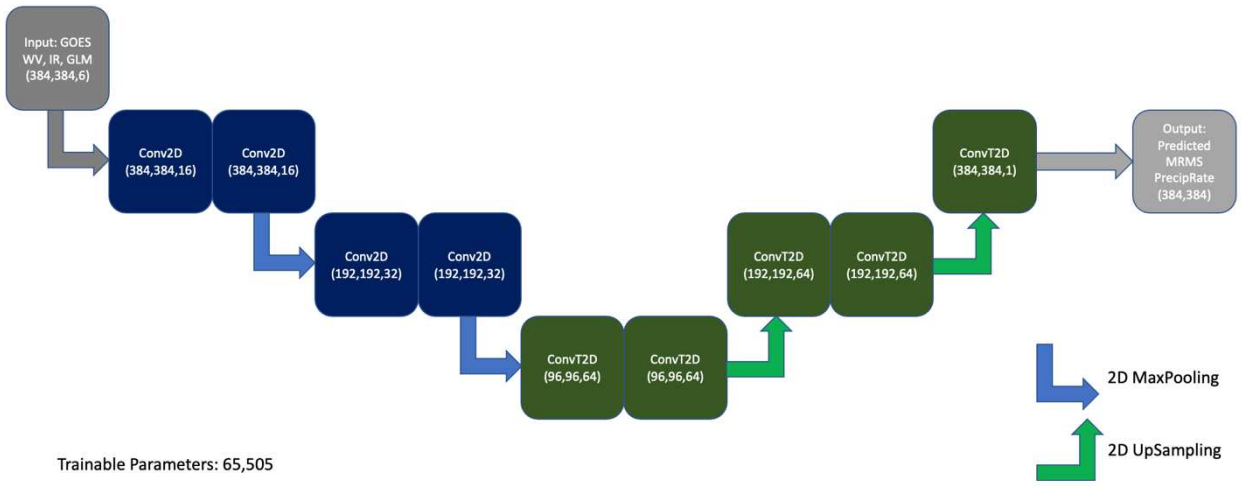

Figure 4: A graphical representation of the GPE-CNN architecture. The network is a two-layer, double-convolution encoder-decoder network, similarly modeled after the GREMLIN architecture (Hilburn et al. 2020).

For this network, a convolution kernel size of 2x2 was chosen; though typically convolution kernels are 3x3, limited hyperparameter tuning showed that this size did not allow the network to learn from its given dataset, while a 2x2 kernel size did. ReLU was chosen as the activation function both due to the prevalence of its use in other machine learning algorithms as well as its ability to constrain the predictions made by the network to only positive values. ReLU is represented mathematically as:

$$ReLU = \begin{cases} 0 \; if \; x < 0 \\ x \; if \; x \geq 0 \end{cases} \tag{2}$$

The loss function by which the network was trained was chosen to be unweighted mean squared error (MSE), calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{truth} - Y_{pred}\right)^2 \tag{3}$$

where n is the number of points, Y_{truth} are the values from the truth dataset, and Y_{pred} are the values from the prediction dataset. Though a weighted version of the MSE calculation such as that utilized in GREMLIN can be applied to better constrain the loss function, several

experiments failed to yield a weighting scheme which allowed the retrieval to properly learn and predict precipitation with physical relevance. This was believed to be related to the extreme class imbalance that is inherent to precipitation data, since at least 80 percent of the MRMS PrecipRate dataset recorded zero precipitation. Because of the lack of loss function and hyperparameter tuning performed, GPE-CNN remained mostly unoptimized; this was done intentionally, as the goal of this thesis was not to present another retrieval algorithm, but to highlight the types and sources of error within similar retrievals. This network was trained over 100 epochs using the Adam optimizer with a batch size of 64 and a learning rate initialized at 10^{-4}. The training data included all top-of-hour imagery available in the Central Plains sector throughout the entirety of 2018.

After training on the 2018 dataset, GPE-CNN's performance was first assessed qualitatively. Figures 5 and 6 show examples where the retrieval performed well and poorly, respectively. In this case, relative performance was determined solely by the scene R^2, rather than through the other statistics utilized. From these examples, it became clear that there are certain precipitation structures, with notable satellite-identified conditions, which are favored by GPE-CNN over others.

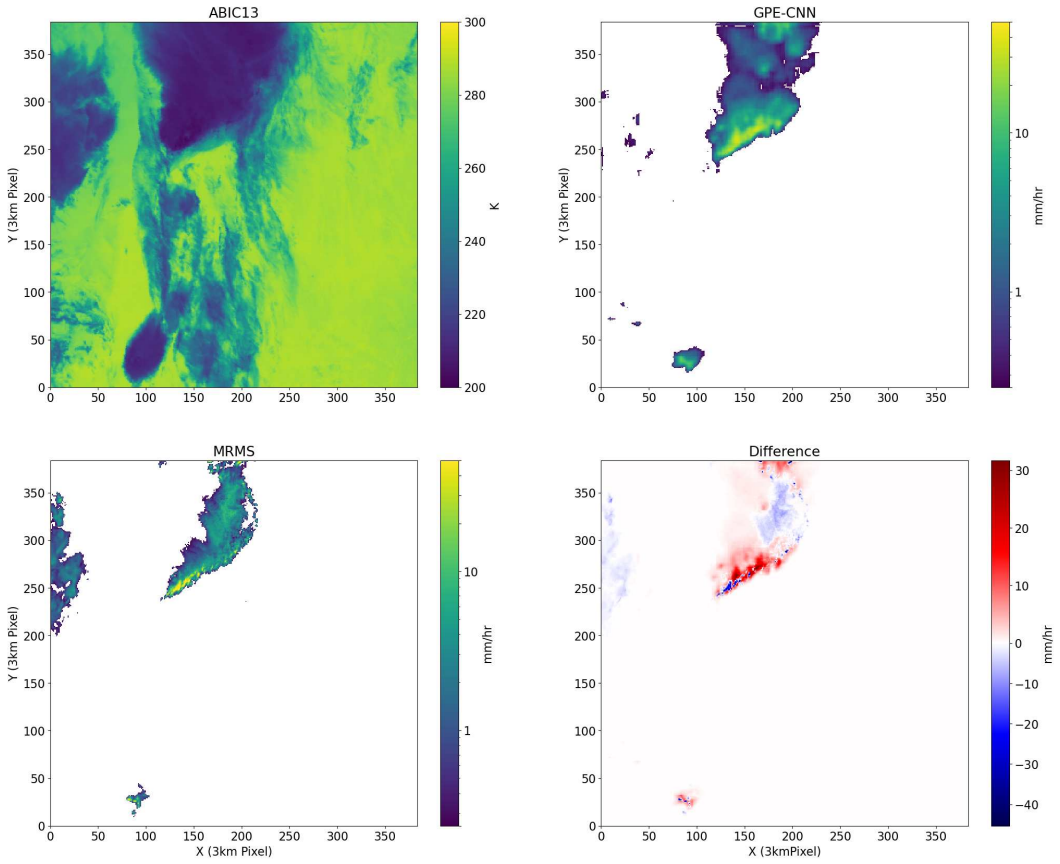

Figure 5: A well-performing sample from the GPE-CNN testing dataset, having a Pearson correlation of 0.6. Such examples tend to be convective, and as such have coherent cloud features that also exhibit overall cold brightness temperatures and tend to have lightning present within the scene.

First considering Figure 5, one notes that the MRMS and GOES-16 observations depicted a mesoscale convective system (MCS) precipitation structure, with enhanced precipitation located near regions of more vigorous convection. GPE-CNN seemed to understand this phenomenon, coming into good agreement with MRMS in terms of precipitation location and intensity. It was noted, however, that GPE-CNN overestimated the coverage of precipitation within the scene alongside an overall underestimation of the precipitation intensity. These discrepancies likely came about from two key features of the retrieval. Firstly, the retrieval is almost entirely based on infrared information. This type of information is only attainable at cloud top, rather than from within the column where the precipitation is occurring. As a result, an

association is made between the cloud top features, namely cloud top temperature, and precipitation intensity. While this relationship loosely represents the relationship between brightness temperatures and precipitation intensity, it does result in an excessive portioning of precipitation over sufficiently cold non-precipitating clouds, such as cirrus clouds. Also, a limitation of using a convolutional neural network for this task is that the use of kernel information results in the prediction being smoother than the truth data. This smoothing may also be in part responsible for the misattribution of finer-scale features in the truth data.

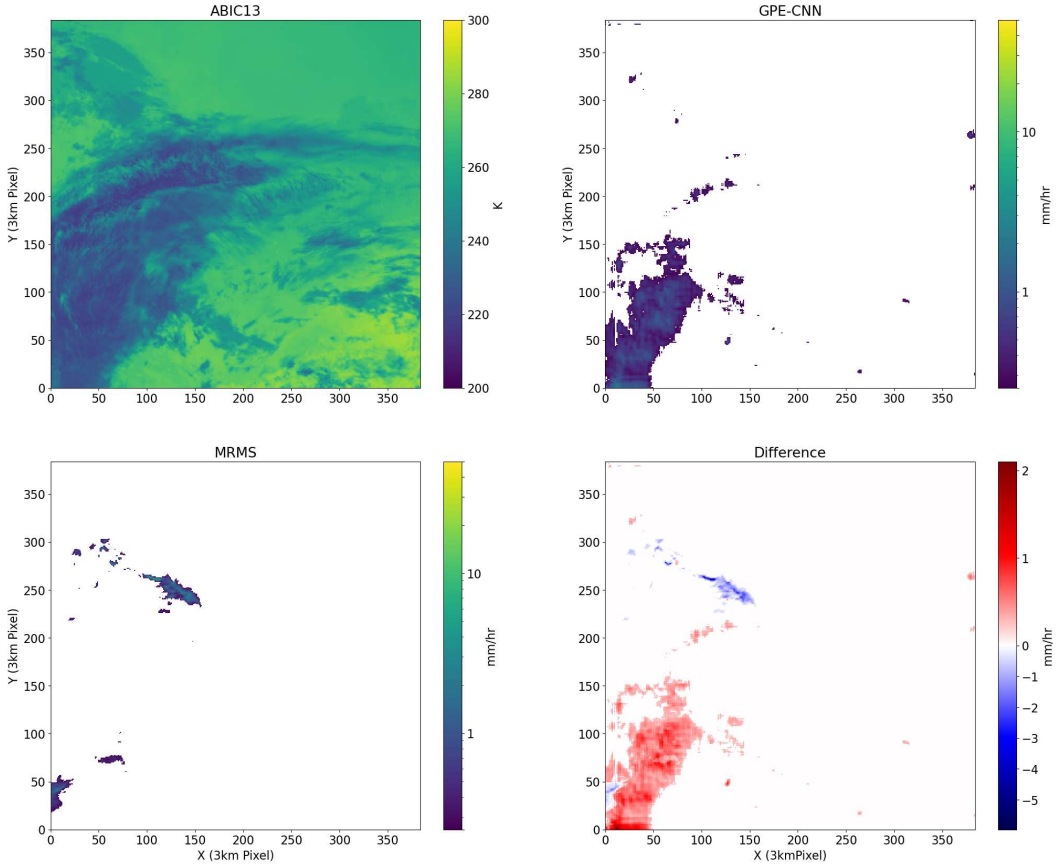

Figure 6: An ill-performing sample from the GPE-CNN testing dataset, with a Pearson correlation of 0.06. In most cases, poor performance is related to the inability of GPE-CNN to determine which cloud features should be precipitating, either through difficult to identify cloud features and/or the absence of lightning information.

Considering those cases where the retrieval tended to perform poorly, such as the example provided in Figure 6, one can begin to understand the incidences in which the retrieval was unable to properly assess the precipitation structure. From Figure 6, one can see that the MRMS observed precipitation was slight and did not cover a substantial area; the coverage and intensity of the precipitation suggested that this scene represented sparse stratiform precipitation. Even to the human eye, discerning noteworthy cloud features in such regimes can be difficult, thus making our own ability to locate and structure precipitation in these cases less accurate. GPE-CNN behaved much the same, substantially misplacing and overestimating the intensity of the precipitation here. This caused major errors in the retrieval, resulting in its poor performance. This example highlighted the limitations of the infrared imagery which were being used as predictors. In cases of sufficiently cold non-precipitating clouds, the retrieval believes that these clouds should be precipitating since similar clouds did precipitate. As a result, the retrieval erroneously assigns precipitation, and often produces too much.

2.6 GPE-CNN Validation Performance

The GPE-CNN algorithm was also validated using MRMS data over the Central Plains sector collected during the months of January, April, July, and October of 2019. These months were chosen as representative months for each season, such that the validation dataset remained separate from but temporally consistent with the training dataset. To assess the overall performance of the algorithm, the coefficient of determination (R^2), MSE, and bias statistic were primarily utilized. R^2 ranges between 0 and 1 normally, and is calculated as:

$$R^2 = \left(\frac{n \sum_{i=1}^{n}(xy) - (\sum_{i=1}^{n} x) * (\sum_{i=1}^{n} y)}{\sqrt{[n \sum_{i=1}^{n}(x^2) - (\sum_{i=1}^{n} x)^2] * [n \sum_{i=1}^{n}(y^2) - (\sum_{i=1}^{n} y)^2]}} \right)^2 \quad (4)$$

where n is the number of data points, x is the observed dataset, and y is the predicted dataset. For ease of interpretation, the coefficient of determination is also known as the square of the correlation coefficient. The bias statistic is calculated as a point statistic of the deviation of the retrieval from its corresponding truth.

$$Bias = RR_{pred} - RR_{truth} \tag{5}$$

where RR_{pred} is the retrieved rain rate at a given gridpoint and RR_{truth} is the corresponding truth rain rate. For use as a comparison metric between the various sectors, this bias was utilized as a percent difference between the mean values of the observed and predicted precipitation rates by sector, which was termed the Mean Bias Error (MBE):

$$MBE = \frac{\left(RR_{pred,av} - RR_{truth,av}\right)}{RR_{truth,av}} * 100 \tag{6}$$

where $RR_{pred,av}$ is the mean predicted precipitation rate and $RR_{truth,av}$ is the mean observed precipitation rate. MSE is calculated in the same manner as the algorithm's loss function.

Using these performance metrics, samples from the testing dataset were analyzed. The overall performance of GPE-CNN is illustrated through the convergence plots shown in Figure 7. From these values, it was shown that the algorithm displayed overall modest performance, with the average R^2 for the model being around 0.36 after training is complete. Though not particularly impressive, the application of this algorithm as an infrared and lightning-based precipitation retrieval likely led to this being the upper bound of retrieval performance. In fact, a comparison of this and other validation statistics in Table 3 showed that GPE-CNN performed quite similarly in its validation area as other established infrared precipitation retrieval algorithms have in theirs. The average testing R^2 also closely matched this value, showing that the network was able to both learn and converge to a solution. The MSE of the network also showed these characteristics, with these values trending toward a minimum of around $3e^{-4}$ by

the conclusion of training. Figure 7 shows that by the end of the training period, the algorithm virtually ceased to improve in both the training and testing datasets. This was encouraging, as the tendency of the algorithm to converge supported the assumption that, despite minimal optimization, the algorithm was able to both learn and develop predictions that were robust across the examples within the training and testing datasets.

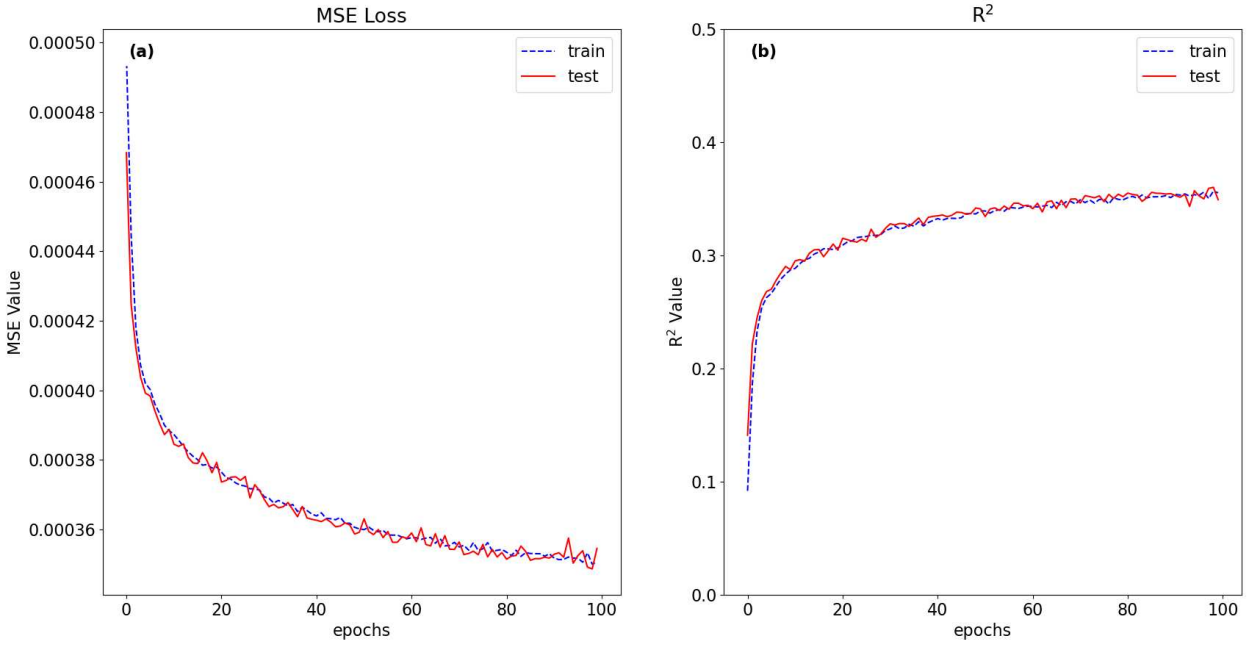

Figure 7: Convergence plots describing GPE-CNN training and testing performance in terms of (a) MSE and (b) R^2. Such plots allow for the building of confidence that the algorithm was able to converge and was able to perform robustly over the wide range of scenarios it encountered.

Table 3: Performance Statistics of GPE-CNN compared to SCaMPR and PERSIANN-CNN for their respective training areas and datasets. From these performance statistics, it can be seen that GPE-CNN performs comparably with other operational and research-based retrieval algorithms utilizing similar inputs.

	RMSE	CC	POD	FAR	CSI
SCaMPR	0.7	0.44	-	-	-
PERSIANN-CNN	0.88	0.41	0.67	0.56	0.37
GPE-CNN	0.65	0.43	0.52	0.65	0.31

Further evaluation was performed by examining the overall error tendencies of GPE-CNN. Figure 8 shows this tendency in the form of a 2D-histogram of the predicted vs observed MRMS PrecipRate for the Central Plains training data. Since this thesis is concerned with the large-scale error tendencies, Figure 8 was developed using data at the subdomain level. From this figure, one can see that the overall performance of GPE-CNN was quite good. Here, the data spread remained fairly close to the identity line, indicating that the retrieval was able to represent the 1x1 degree PrecipRate product with good accuracy. The high Pearson correlation and low mean bias error (MBE) for this sector also showed that GPE-CNN was able to properly assess the structure of the data and create reasonable predictions in most cases.

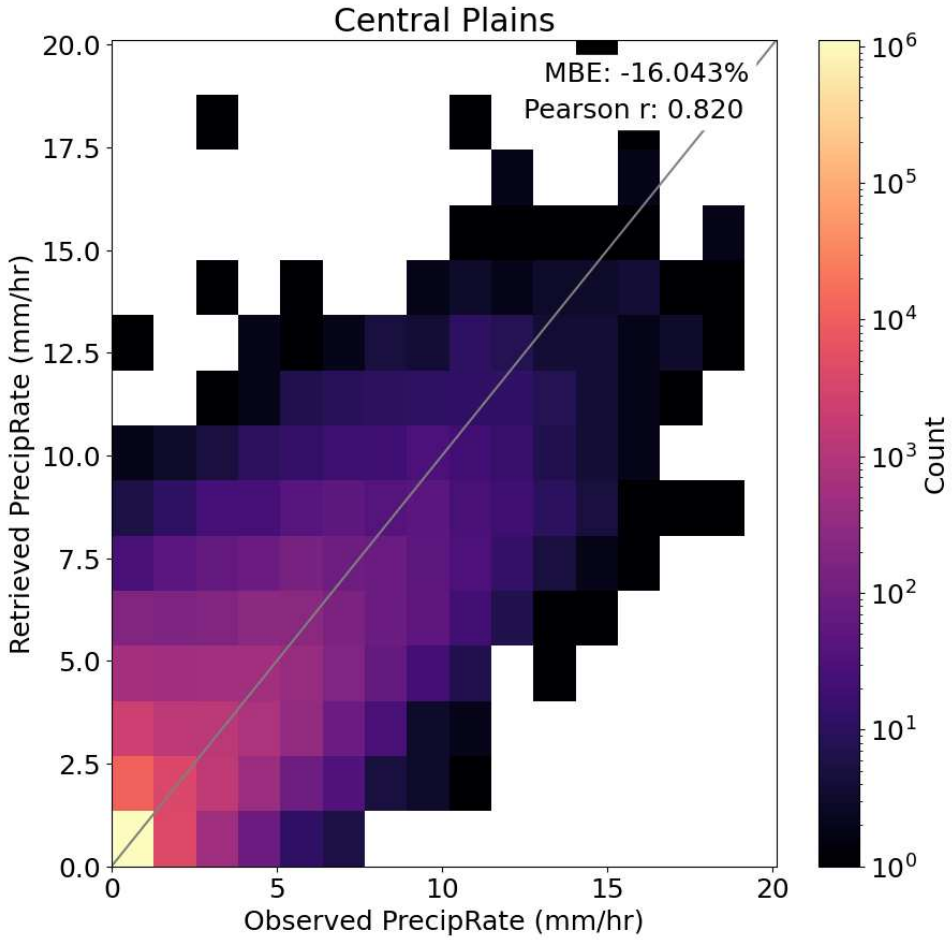

Figure 8: A 2D Histogram of GPE-CNN Central Plains subdomain performance. From this figure, it can be seen that GPE-CNN is able to generalize to the precipitation seen in the Central Plains, with a high correlation (0.82) and relatively low MBE (-16%).

CHAPTER 3: ANALYSES AND RESULTS

The focus of this chapter is on elucidating the nuances of the GPE-CNN retrieval performance both within and outside of its training area. In the case of infrared precipitation retrieval, the information content of the predictors is low, and as such, assumptions become necessary aspects of the retrieval. This leads to the assumptions becoming important identifiers in the retrieval performance, which is a trait that this set of analyses aims to exploit. These analyses were designed to assess the reasons why strong performance in a given region does not necessitate similar performance in another. To do this, the five sectors discussed in Chapter 2 will be utilized; this will allow for comparisons of retrieval performance in areas that are in close geographic proximity, and therefore should be similar in terms of sensor viewing angle and solar zenith angle, but remain meteorologically distinct enough to elucidate differences in model performance. The differences in performance highlighted here will then be explored both in the context of the satellite data and environmental data, allowing for an assessment of how such information provides improved explainability of performance under certain regimes.

3.1 Retrieval Performance: Sector Intercomparisons

While GPE-CNN was shown to perform satisfactorily within the Central Plains training sector, it was not immediately clear if said performance would translate to other, similarly-sized sectors. As such, further performance analysis of GPE-CNN was carried out over the five sectors described in Chapter 2. This stage of the analysis was based mainly in the overall behaviors observed in each sector and how they compare with the Central Plains. Figure 9 shows the general estimation behaviors of all five sectors as two-dimensional histograms of retrieved versus observed precipitation rates at subdomain resolution for each sector. The averaged

subdomain MRMS and GPE-CNN precipitation rates were used, as the original resolution of the datasets was deemed too noisy for this and subsequent analyses. This and subsequent statistical comparisons were performed using the MBE and Pearson correlation; these indicators were chosen to provide consistent comparisons between the sectors.

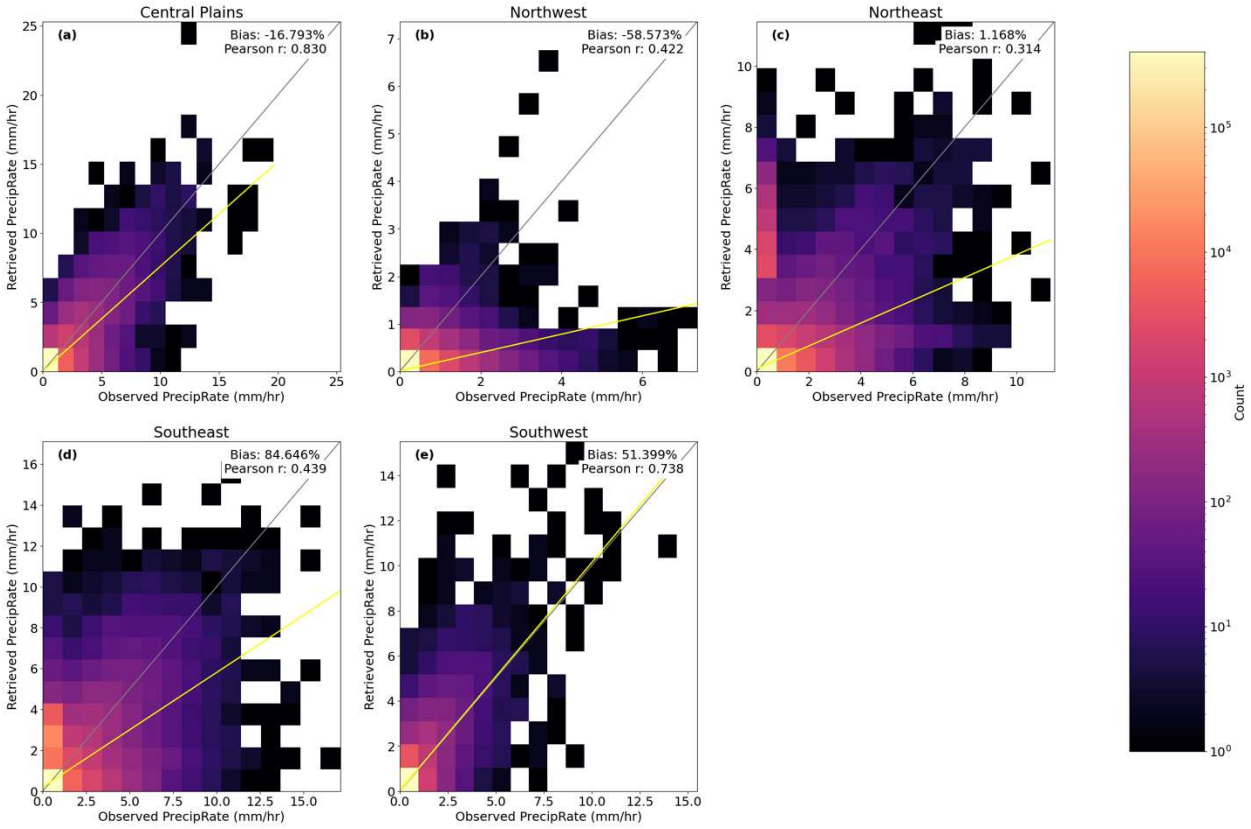

Figure 9: 2D histograms of retrieved and observed PrecipRate by GPE-CNN across all sectors. The grey lines represent a theoretical 1:1 relationship between GPE-CNN and MRMS, while the yellow lines represent the line of best fit for the data. Comparing each of these sectors to the Central Plains training area allows for qualitative and quantitative analysis of the change in performance of GPE-CNN.

Beginning first with the behavior of the Central Plains (Figure 9a), it was clear at GPE-CNN did well. The data scatter fell closely to the identity line (grey line), and the Pearson correlation of 0.83 showed that at this scale, GPE-CNN was reliable. The MBE was found to be about -16%, indicating that in the Central Plains, GPE-CNN had a tendency to slightly

underestimate the precipitation rate. While this represented the bias of the mean rain rates here, it was still worth noting that substantially better and worse biases did occur, indicated by the spread in the histogram in Figure 8. Still, a low MBE highlighted the ability of GPE-CNN to generalize to many different scenarios in the Central Plains.

From this same figure, however, it was apparent that the remaining four sectors did not perform similarly to the training sector. The most similar sector to the Central Plains was the Southwest (Figure 9d); here, it can be seen that the Pearson correlations of these sectors were comparable, being only about 11% different. However, while GPE-CNN displayed an MBE of -16%, the Southwest exhibited an MBE of 51%, representing a 67-point difference between these sectors. This extends largely from the shift in the estimation tendencies observed; the Central Plains tended to exhibit underestimation while the Southwest exhibited an overestimation tendency. This indicated that GPE-CNN had a very different bias tendency in the Southwest, despite their initial appearance of agreement. Given that both sectors tend to be dominated by deep convective systems, this discrepancy suggested that the underlying environmental conditions may play a role in causing the systematic biases between regions. The southwestern CONUS is much more arid than the Central Plains, and as such, there is likely a greater tendency for evaporation to occur, resulting in GPE-CNN overestimating precipitation here.

The most visually different sector from the Central Plains in terms of performance was the northwestern sector (Figure 9b). Here, it was apparent that GPE-CNN had a strong underestimation tendency, which can be highlighted with an MBE of -58%, representing a 42-point difference from the Central Plains MBE. This does seem to underrepresent the frequency and intensity with which the Northwest underestimated rain rate, which could be due to the amount of near-zero precipitation values here. The Pearson correlation for the Northwest was

0.42, a roughly 50% decrease from the Central Plains. These results suggested that GPE-CNN was not able to adequately assess the precipitating systems that existed in the Northwest sector using information from the Central Plains. Like the Southwest sector, these differences are most likely tied to differences in the overall environmental states in these sectors. In the Northwest sector, the environment has strong maritime influences that are not present in the Central Plains. As a result, the precipitating systems may "look and behave" differently from those seen in GPE-CNN's training area. This effect may also extend to how these systems appear to the satellite; since maritime environments are typically warmer and more moist than continental ones, precipitating systems in these environments may also be warmer and less lightning-prone than their landlocked counterparts.

Finally, the northeastern (Figure 9c) and southeastern (Figure 9e) sectors displayed a much wider variance in behaviors when compared to the Central Plains. In the Northeast sector, the Pearson correlation of 0.31 was the lowest of all sectors, representing a 62% difference from the Central Plains, alongside an MBE of about 1%. The MBE for the Northeast was particularly interesting, as on its own, it signals that the sector actually performed better than the Central Plains, which has an MBE of -16%. However, considering the substantial scatter and the high concentration of points for lightly precipitating samples noted in the Northeast histogram, the MBE was likely influenced by these factors to be much a much smaller value. The Southeast sector performed better in terms of correlation at 0.44, or a 47% difference, but performed the worst of the sectors in terms of MBE, which was about 84%, or a 100-point difference from the Central Plains MBE. The large scatter in possibilities denoted in these two sectors' histograms seemed to speak to a much more variable presentation than was captured in the training sector. Though these eastern sectors are most often influenced by continental airmasses, the presence of

the Gulf of Mexico and the Gulf Stream provide a maritime component to the climatology of these sectors. This combination of continental and maritime influences can lead to many environments that are not captured in the training data, and therefore create challenges for GPE-CNN. Because of this, the performance of the retrieval in these sectors becomes unclear.

From this intercomparison it was apparent that while GPE-CNN did well in the Central Plains, this behavior was not transferrable to the other sectors considered. This resulted in noticeably different performances in each sector, especially when considering MBE. Since the version of GPE-CNN used in each of these sectors was the same throughout, these differences could be attributable to information not explicitly included in the retrieval. From this, it was hypothesized that the differences in meteorological states between sectors could explain these differences in performance. Since clouds and precipitation are responses to their environment, this aspect was therefore hypothesized to be the most influential in determining the retrieval biases given certain meteorological states.

In pursuit of this hypothesis, several techniques were explored to develop a relationship by which GPE-CNN's bias could be robustly predicted given the ancillary information provided by ERA5. These techniques are listed in Table 4. These techniques were all attempted with various combinations of the available ERA5 data, including some variables like convective available potential energy (CAPE) and maximum vertical velocity, among others, which were not included in Table 1. Each attempt represented a different perspective by which this relationship was explored for. First, the more heuristic techniques, being thresholding and PDF matching, were tested, as it was believed that these methods would be the simplest and most explainable. This involved developing thresholds for the environmental data that, when implemented alongside the bias information, would potentially yield a useful relationship. This

thresholding was done on a univariate basis, meaning that only one variable at a time was tested using this method. Despite the perceived relationships between the ERA5 data and the retrieval bias, this method was unable to discern a relationship of any kind between the two. The PDF matching technique was also used in conjunction with this, with the belief that information about the distribution of the biases when coupled with the ERA5 data would provide a direction from which a relationship could be developed. In this case, the PDFs developed were too similar to one another, indicating that no real separation was achieved. From this test of methods, it was decided that the relationship being sought was much more complex than heuristic methods could interpret.

Understanding this greater complexity and following the rationale presented in Henderson et al. (2017) and Elsaesser et al. (2010), different methods of classification, namely K-means and hierarchical clustering, were attempted to create environmental regimes with distinct bias distributions. These methods were explored with varying numbers of clusters and combinations of ERA5 variables. In all instances of these clustering algorithms, groups with distinct bias characteristics were not attained. Instead, the biases in each respective group displayed large amounts of scatter, which indicated that the developed clusters would result in no better than a random guess as to predicting the bias of a given area with respect to the environmental data. Empirical Orthogonal Function (EOF) analysis was also used to establish the different modes in the environmental data by which the retrieval bias could be assessed. This resulted in a similar conclusion as the clustering algorithms. Various regression techniques were also attempted, believing that the relationship between the retrieval bias and the environmental data was not an appropriate task for classification. These methods were able to develop relationships between the ERA5 data and GPE-CNN's biases. These relationships, however,

were incredibly weak, with correlations well below 0.1 once the non-raining areas were excluded; with the non-raining areas included, correlations were incredibly high, but the Critical Success Index (CSI) values were very low. This seemed to suggest that these methods were unable to develop an actionable relationship both due to the abundance of non-raining areas in all possible environmental states as well as an overall inadequacy of simpler regressions to adequately assess the problem.

With the inability of the afore-mentioned techniques to adequately characterize retrieval bias with environmental data, more complex methods were explored. With this, different machine learning-based techniques were investigated, those being random forests, support vector machines, and Naïve-Bayes classifiers. These methods explored both nonlinear classification and regression, with both styles being attempted with the random forests. These algorithms, with the exception of the Naïve-Bayes classifiers, also included information about the retrieval precipitation intensity, in the hopes that information would more closely constrain the relationship. In all cases, these more complex methods did not exactly reveal a usable relationship between the ERA5 data and the retrieval biases either. These methods still seemed to be unable to perform robustly when confronted with the data imbalance that is typical of precipitation data, as the non-raining areas always dominated the performance of the methods. However, while testing these methods with varying interpretations of the retrieval bias, the Naïve-Bayes classifier did show some promise in determining if a given area should be biased or not. This was done by considering GPE-CNN's bias as a binary class, those classes being areas that displayed biases below +-1 mm/hr, and those with biases outside of that range. This gave the Naïve-Bayes classifier a correlation around 0.87 with a CSI of 0.21. This was by far the best performance achieved by any of the methods tested. However, with the low CSI displayed by

this classifier, it was deemed to not be usable, as a high correlation with a low CSI meant that this method was heavily favoring the prediction of a specific class, namely the unbiased class. When the dataset was balanced, this method's CSI increased to 0.7, but when this model was tested on an imbalanced dataset, the CSI dropped back to 0.2. This indicated that the Naïve-Bayes classifier, while it showed promise, was too brittle in the face of class imbalances to be usable.

Despite the apparent relationships between the environmental data used and precipitation, none of the methods attempted were successful in relating the environmental data to the retrieval biases. Upon determining that the development of a relationship between environmental data alone and GPE-CNN's biases was not feasible, further consideration was given as to why this was the case. When looking back at the performance of GPE-CNN as determined by its validation in the Central Plains, the observation that GPE-CNN appeared to favor certain precipitation regimes; this was shown in Figures 5 and 6, which displayed the tendency for GPE-CNN to better assess convective scenes than stratiform scenes. Those results led to a modification of the original hypothesis that environmental data alone is sufficient for predicting retrieval biases. It was then decided that in order for a usable relationship between biases and the environment to be established, one must also understand those scenarios in which the retrieval is likely to perform well or poorly. This assumption then led to the need to understand those circumstances within GPE-CNN, which reintroduced the use of satellite data to determine those characteristics that allowed the retrieval to do well and those which prevented it. The remainder of this thesis was therefore dedicated to understanding each of these components, beginning first with an assessment of which satellite characteristics enabled or prevented good performance in

GPE-CNN and followed by how this understanding affected one's ability to assess retrieval bias using environmental information.

Table 4: A list of the methods used when attempting to create a robust relationship between the ERA5 data and GPE-CNN biases. In each of these attempts, the performance of these given methods resulted in failure of these methods to develop a usable connection between the two quantities. These methods were also attempted with various sets of ERA5 variables, some of which were not included in the subsequent analyses using ERA5 data.

ERA5-Bias Relation Methods Attempted
Hierarchical Clustering
K-Means Clustering
PDF Matching
Heuristic Thresholding
Random Forests
Support Vector Machines (SVMs)
Naïve-Bayesian Classifiers
Empirical Orthogonal Functions (EOFs)
Logistic Regression
Multi-linear Regression

3.2 Retrieval Performance: Training Area Analysis

To better understand what may be causing the large discrepancies in GPE-CNN's performance, an examination of the large-scale correlations between the retrieval and MRMS precipitation was performed. To do this, the data within each subdomain were identified at their native 3km resolution and correlated with their corresponding MRMS data. The use of these subdomains allowed for the characterization of the large-scale behaviors in the retrieval and provided a scale for direct comparison against the ERA5 data. Figure 10 shows the distribution

of these correlations within the Central Plains testing data. From this partitioning, it was clear that there were a large number of negatively and positively correlated subdomains. It is also worth noting that a large proportion of these subdomains have incalculable correlations, shown in Figure 10 as the green bar. These subdomains were found to be cases where no precipitation was recorded in the MRMS and/or GPE-CNN outputs; given the inability to calculate their correlations and their representation of clear-sky conditions, these subdomains were not explicitly considered in the subsequent correlation analysis.

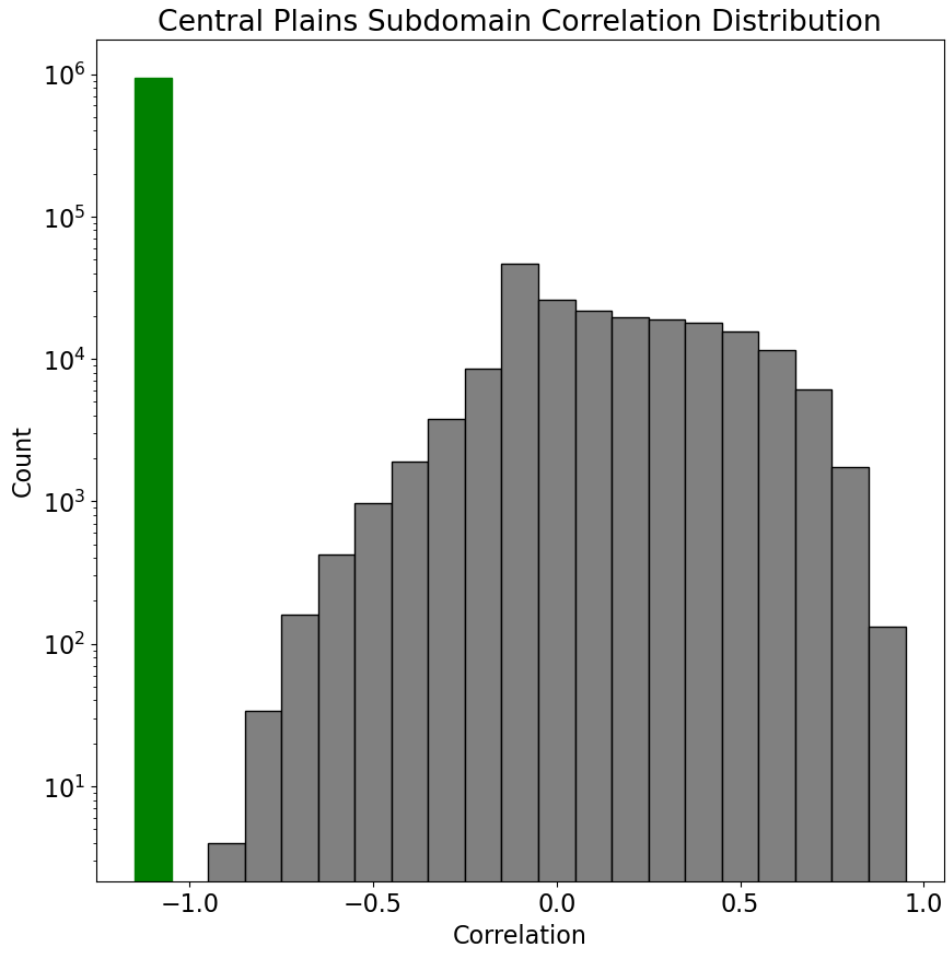

Figure 10: The distribution of the Central Plains subdomain correlations. The correlations were binned every 0.1 correlation value. The green bar represents those subdomains where the correlation was incalculable. Here, it is shown that there is a large possibility of subdomain correlations that can be observed, with some cases being strongly negative and others strongly positive.

To better understand the reasoning behind the occurrence of negatively and positively correlated subdomains, a random sample of 10 subdomains from each correlation range bin was taken and qualitatively assessed, resulting in roughly 200 total sample subdomains. This number of subdomains was chosen to balance the analysis, as there was a massive data imbalance issue noted previously. Figure 11 shows an example of one of these subdomains with a strong positive correlation. From this example, the most apparent aspect that likely contributed to its strong performance came from the presence of lightning. In Figure 11f, there were two 'hotspots' of lightning present near the center of the subdomain, along with a much broader area of weaker lightning surrounding these areas. Comparing these locations with the MRMS and GPE-CNN precipitation data, it was clear that the retrieval heavily favored the presence of lightning in assigning enhanced precipitation. This lightning signature was also coincident with a coherent area of cold brightness temperatures; this was noted in all of the infrared channels. As such, GPE-CNN appeared to perform best with convective scenes, since these scenes often had strong lightning information as well as obvious and cold clouds.

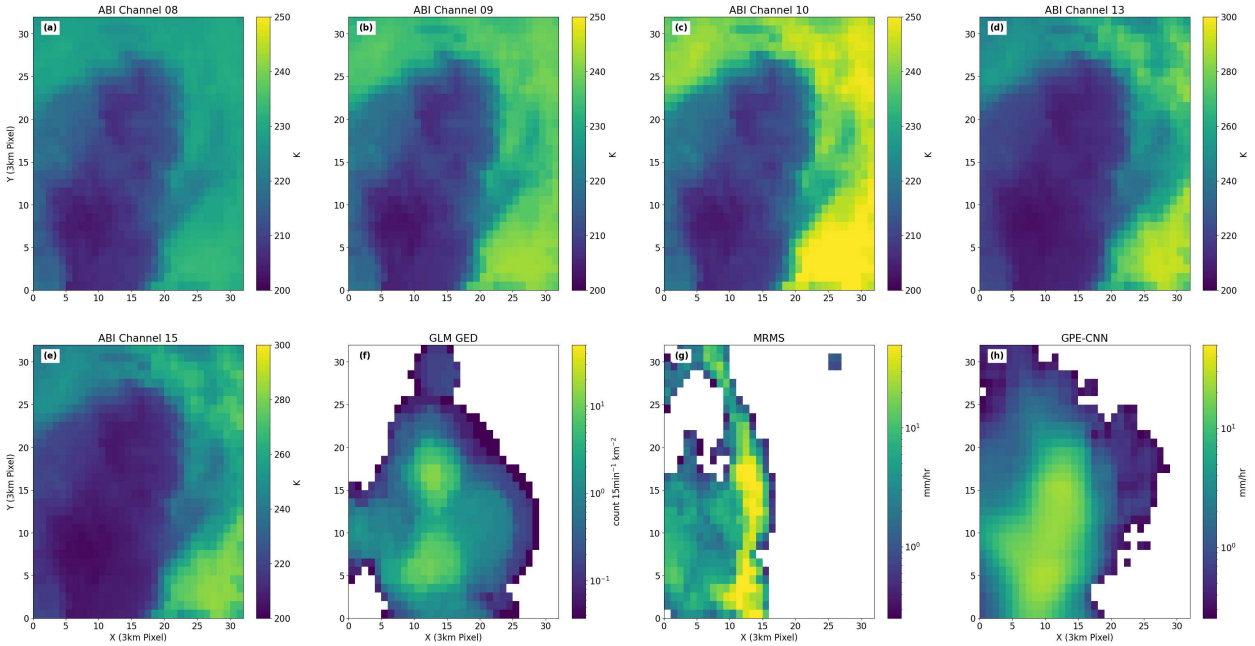

Figure 11: An example of a highly positively correlated subdomain in the predictor, target, and prediction outputs. In these cases, the subdomains display some combination of lightning presence, cold brightness temperatures, and/or strong brightness temperature heterogeneity.

Figure 12 provides an example of one of these random subdomains with a strong anticorrelation. From this example, it was clear that GPE-CNN misunderstood where the precipitation was occurring. One major difference between this example and that shown in Figure 12 was that there was no lightning present here. This further signaled the importance of lightning data in GPE-CNN, as these poorly-performing subdomains did not include any substantial lightning information. As a result, GPE-CNN was unable to use this information and was forced to use the indirect relationship between brightness temperature and precipitation. Another aspect of this scene that would complicate GPE-CNN's prediction was that this subdomain depicted a very homogeneous brightness temperature pattern. Looking specifically at ABI Channels 13 (Figure 12d) and 15 (Figure 12e), it was difficult to discern where in this subdomain precipitation should be expected, as there were no obvious cloud features by which to orient oneself. As a result, GPE-CNN was unable to locate the cloud areas that were

precipitating. Another interesting feature here is that the precipitation was occurring within the lower right quadrant of the subdomain, which contains slightly warmer brightness temperatures than the rest of the subdomain. Since GPE-CNN appeared to assume that precipitation becomes more likely with lower brightness temperatures, this subdomain would prove to be a particularly difficult challenge for the retrieval. In fact, given the predictors that GPE-CNN had, it is dubious if improvement can be achieved in these circumstances.

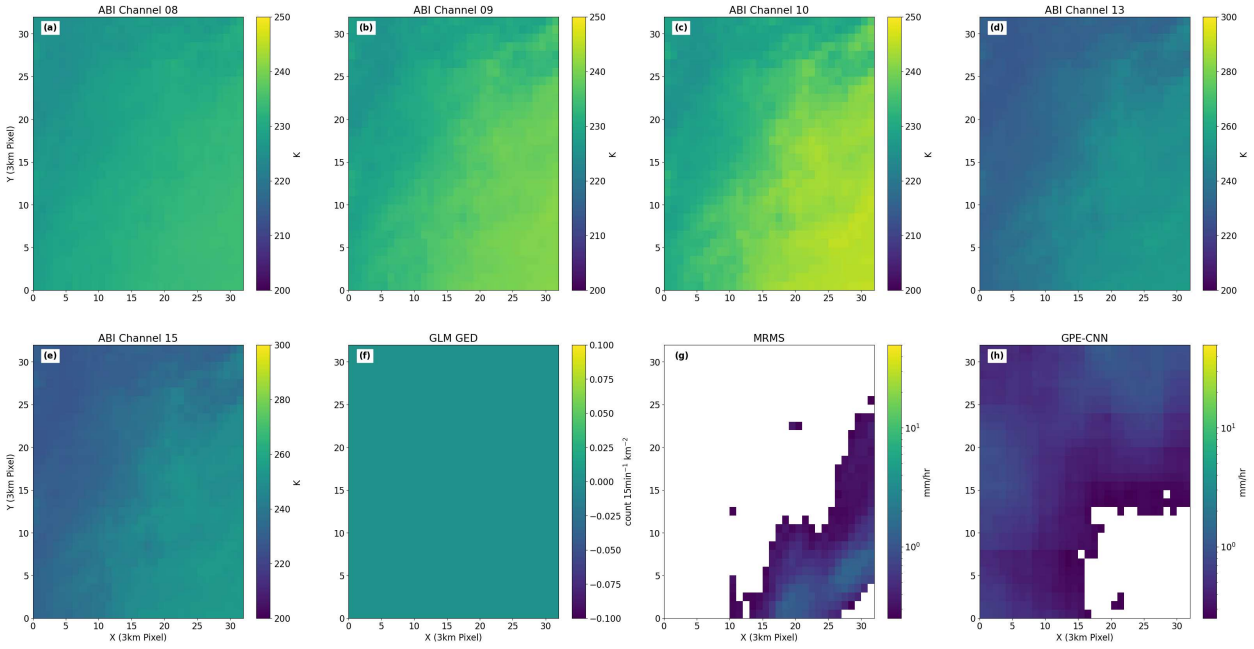

Figure 12: An example of a highly negatively correlated subdomain in the predictor, target, and predicted outputs. These subdomains most often exhibit a lack of lightning data, relatively cool brightness temperatures, and/or brightness temperature homogeneity.

Further analysis of these correlation bins was performed utilizing the descriptive statistics, namely the means and standard deviations, of the same randomly sampled subdomains in each bin. While this part of the analysis was performed with all of GPE-CNN's predictor variables, only the GLM data (Figure 13) and ABI Channel 13 (Figure 14) are included here. This choice was made for simplicity, as the remaining ABI channels showed similar but less

apparent behaviors as ABI Channel 13. Beginning with GLM, this analysis further reinforced how important the presence of lightning data was to this retrieval. Looking first at Figure 13a, it was clear that the anticorrelated subdomains almost never contained lightning, while in the positively correlated bins, at least a small fraction of the subdomains contained measurable lightning. Furthermore, there was a general positive trend between measured lightning and the strength of the positive correlation. This signified that the more lightning is present in a given area, the more likely that GPE-CNN will perform accurately. These conclusions were also reflected in the standard deviations shown in Figure 13b. In each case, the inclusion and prevalence of the GLM data in GPE-CNN was paramount to understanding how well the retrieval would perform.

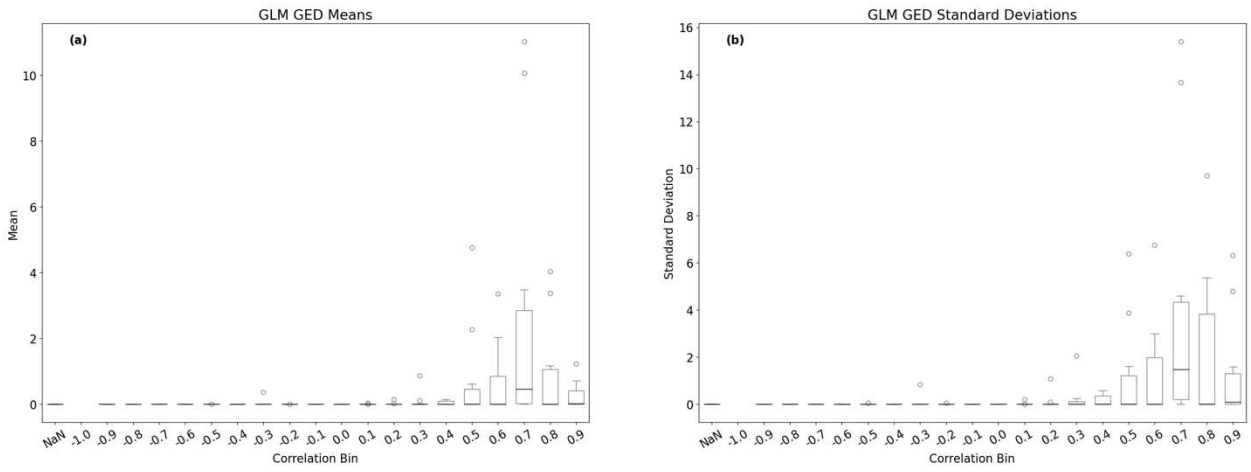

Figure 13: Boxplots of (a) means and (b) standard deviations by correlation bin for GLMGED. The presence of lightning is a major contributor to GPE-CNN's performance, as lightning information is disproportionately contained within the positively-correlated subdomains.

The other main conclusion drawn from this step of the analysis was that GPE-CNN tends to favor those areas with well-defined and cold cloud features. Figure 14 provides the same descriptive statistics as in Figure 13 but for the ABI Channel 13 brightness temperatures.

Initially, these results seemed to run counter to the qualitative analysis, as there did not appear to be a particularly strong trend in mean brightness temperature with correlation (Figure 14a). However, when considering the interquartile ranges (IQRs) of the mean brightness temperatures, it appeared that the positive correlations had a much wider spread in possible mean brightness temperatures than the negative correlation bins did. This assessment was amplified by the standard deviations of each correlation bin (Figure 14b), which showed that the positively correlated bins exhibited larger standard deviations, and therefore larger spread in potential variability, than the negatively correlated subdomains. This suggested that the ability to distinguish well-defined cloud features was rather consequential to GPE-CNN's performance, since greater heterogeneity in the subdomain suggests that more apparent clouds edges are present. Such features would allow for much easier determination of where to place precipitation. This delineation became very difficult in areas where the brightness temperature patten was homogeneous, such as that of stratiform and cirriform cloud cover, while cumuliform cloud cover tended to exhibit a more heterogeneous brightness temperature pattern. As a result, the effect on those positively correlated bins was that the mean values of brightness temperature were higher than expected, while the range of brightness temperatures within that same subdomain was wider.

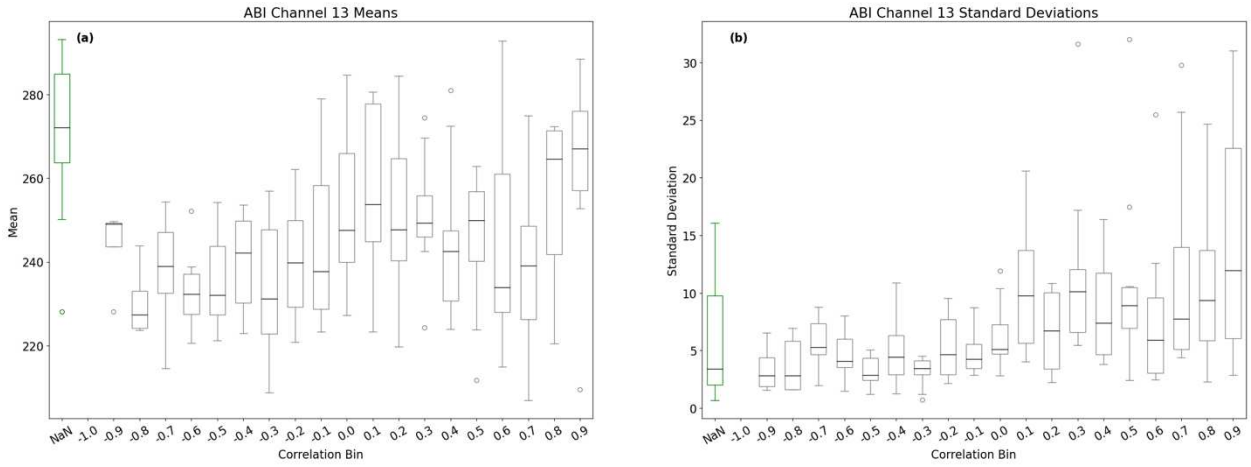

Figure 14: The same as Figure 11, but for ABI Channel 13. While exact trends in mean brightness temperature are not immediately clear, the brightness temperature standard deviations suggest that performance increases with large standard deviations (i.e.: more heterogeneous subdomains).

3.3 Satellite Data Partitioning

Given the impact of the subdomain-level satellite characteristics in the previous section on the performance of GPE-CNN, further scrutiny was given to these data. From the assumption that similar cloud types should represent similar precipitation regimes in the absence of environmental influence, manual grouping of the satellite inputs to separate these features was performed. In this case, not considering environmental impacts, each of these satellite data groups should appear nearly identical to each other within the different sectors. To do this, only the ABI Channel 13 and GLM data were used; while the other channels remain important contributors, the use of just these two was chosen for ease of interpretation.

From the analysis performed in Section 3.2, it was shown that the presence of lightning, subdomain brightness temperature heterogeneity, and cloud-top brightness temperature were most influential in determining large-scale performance. Therefore, the subdomains were classified into groups based on those features. In the case of mean brightness temperature in each subdomain, the thresholds used were fairly well-established in literature. From Arkin and Meisner 1987, it was determined that convective clouds tend to exhibit brightness temperatures

of around 235K or cooler. Work from Delgado et al. 2008 and Hanna et al. 2008 suggested that lightly-precipitating cloud features tended to exhibit infrared brightness temperature in the 250K-260K range; the warmer end was chosen to maximize this characterization. Any mean brightness temperatures warmer than 260K were considered to be representative of clear sky or non-raining conditions. The standard deviation of brightness temperature in each subdomain was used here as a proxy for both cloud edge detection and IR texture. While this interpretation was challenging, it appeared that larger standard deviations represented more coherent cloud features owing to the greater heterogeneity of the brightness temperatures in those areas. A threshold standard deviation of 10 Kelvin was set based on the standard deviation behaviors observed in several example scenes (not shown). In the case of lightning, the GLM data was taken and classified as a binary group. This was determined by the maximum value of the GLM data within the subdomain. As such, if the maximum amount of lightning detected was greater than zero, the subdomain is considered 'lightning-present', while the opposite is considered 'lightning-absent'. This thresholding methodology created 12 possible groups by which the satellite data can be classified, as described in Table 5.

From these groups, rough categorizations about the precipitation regimes they represent can be made. Here, these regimes were defined as "convective", "possibly convective", "stratiform", and "clear-sky". Groups 3, 4, 8, and 12 represented the "convective" groups. Groups 3 and 4 showed the most classic presentation of convection, owing to their cold brightness temperatures and lightning information. Groups 8 and 12 displayed much warmer brightness temperatures, but were still convective owing to their lightning information and stronger brightness temperature heterogeneities. Groups 1, 2, and 7 denoted the "possibly convective" groups. Oftentimes, these groups were located either within a weaker convective

portion of an MCS or along its edge, as determined through visual interpretation of several example scenes, and tended to meet at least some of the criteria that the "convective" groups presented. Groups 5, 6, and 10 were the "stratiform" groups. The satellite data in these cases most often showed very homogeneous cloud features that were slightly warmer than would be expected of convection. These groups also did not include lightning information. Group 9 was the "clear-sky" group; here, the warm brightness temperatures and lack of texture and lightning information indicated that these subdomains are mostly observing the Earth's surface. Group 11 was the only group that was tricky to classify. Its infrared depiction indicated it should be another 'clear-sky' example, but the presence of lightning information complicated this assessment. It is possible that Group 11 represented very small single-cell convection, though there were no apparent examples from this group that presented these conditions. As such, this group was deemed to represent those subdomains that contained more spurious cases of lightning propagation and were most likely unphysical.

Table 5: A description of the satellite data groups developed, with BT being the subdomain mean brightness temperature, std being the subdomain brightness temperature standard deviation, and GLM being the presence of lightning. The thresholds that these groups were developed from are related to the findings from Section 3.2.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
BT	T<=235	T<=235	T<=235	<=235	235<T<=260	235<T<=260	235<T<=260	235<T<=260	T>260	T>260	T>260	T>260
std	<=10	>10	<=10	>10	<=10	>10	<=10	>10	<=10	>10	<=10	>10
GLM	no	no	yes	yes	no	no	yes	yes	no	no	yes	yes

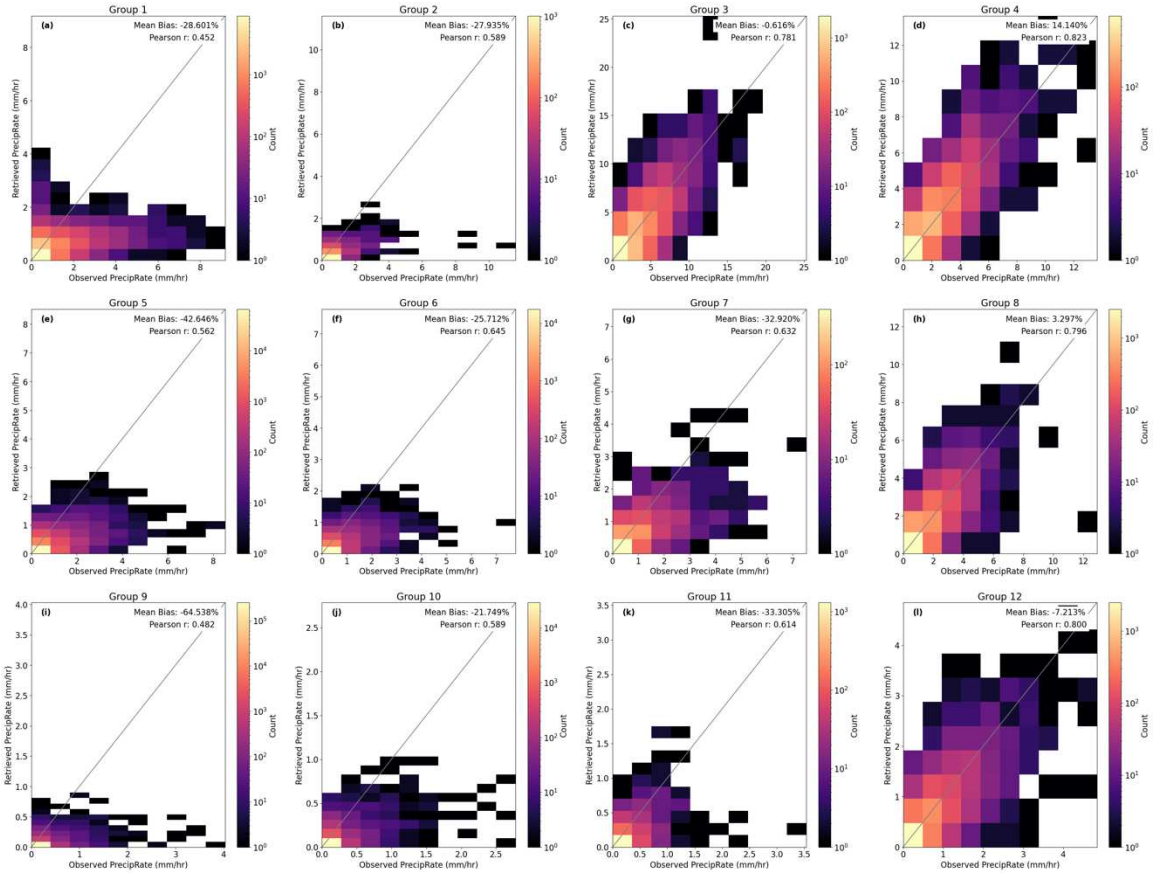

Figure 15: 2D histograms of observed and retrieved PrecipRate for the Central Plains test dataset partitioned by GOES group. Each of these groups displays different bias tendencies which can be compared to identical groups in the remaining sectors to determine if this partitioning can bring the sectors into better agreement with the Central Plains.

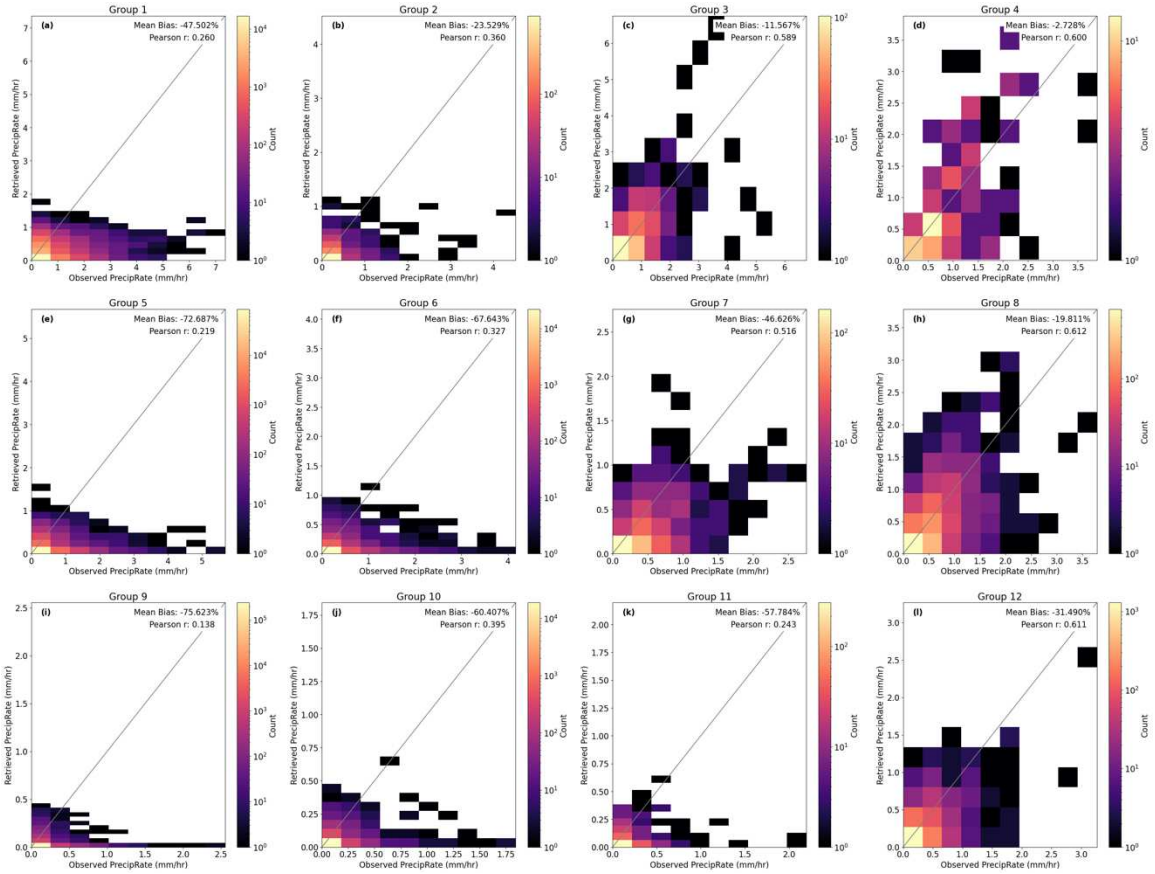

Figure 16: The same as Figure 15, but for the Northwest sector. When compared to the equivalent Central Plains groups, there is modest agreement in their general bias tendencies, but the greater specificity of the groups does not exactly align.

Using the information highlighted in Table 4, the MRMS and GPE-CNN data were partitioned by these groups in each of the five sectors. Figures 15 and 16 show how these groups appeared for Central Plains and Northwest sectors, respectively. For the sake of clarity, only these two sectors were shown, though the conclusions drawn from them are applicable to the remaining sectors, as outlined in Appendix A. Getting into these groups in greater detail, certain patterns of behavior can be noticed. Firstly, those groups that were more convective, Groups 3 (15c, 16c), 4 (15d, 16d), 8 (15h,16h), and 12 (15l,16l), were overall much better performers than the remaining groups. This was consistent with the general validation behavior of GPE-CNN, where the convective systems generally performed better than their stratiform counterparts. This

was further reflected in the validation statistics with these groups, as the Pearson correlations of these group ranged between 0.78 and 0.8 in the Central Plains. The Northwest sector was also similarly good in these groups, with correlations ranging between 0.59 and 0.61; these correlations showed a difference of about 25% compared to the Central Plains, which was half that of the overall difference between the two sectors for correlation. This strong correlation was mirrored in the histogram distributions, which showed that these groups trended near the identity line. In terms of MBE, there was a much greater discrepancy, with the Central Plains groups displaying values between -0.6% and 14% and the Northwest groups between -2% and -31%. These discrepancies amount to between differences in MBE between 1.4 and 45 points between these two sectors. Though these differences are substantial, this set of groups also showed the lowest MBE values in both sectors, indicating that GPE-CNN is still performing best here. Also, both sectors still displayed similar distributions in their histograms, which alongside the correlation results suggested that these groups better constrained the overall bias tendencies, but not exactly the biases themselves.

Outside of these groups, however, the bias tendencies exhibited consistent underestimates of precipitation intensity regardless of satellite presentation. Despite this similarity, there are different behaviors to be noted in the groups. For example, Groups 1 (Figure 15a, 16a), 2 (Figure 15b, 16b), and 7 (Figure 15h, 16h), which were most often associated with developed convection, did do better overall compared to these other groups. More specifically, these groups did show a greater spread in Pearson correlations than their more decidedly convective counterparts, between 0.45 and 0.64 for the Central Plains and 0.26 and 0.36 for the Northwest. The differences in correlation between these two sectors therefore ranged between 17% and 42%, which still showed some improvement from the overall correlation difference of 50%.

These groups, however, exhibited larger MBE values compared to the more convective groups, being between -27% and -32% in the Central Plains and between -24% and -46% in the Northwest. This amounted to differences in MBE between 3 and 19 points, indicating that similar bias responses were being captured, but full agreement was still not achieved. From this, it can be seen that there are still some differences between the two sectors in terms of MBE, though in terms of correlation, the two sectors agree fairly well.

Outside of the aforementioned groups, GPE-CNN's performance was noticeably degraded. Beginning with the stratiform groups, Groups 5 (Figures 15e, 16e), 6 (Figures 15f, 16f), and 10 (Figures 15j, 16j), the already known difficulty of GPE-CNN in predicting precipitation was further shown. Here, the Pearson correlations were between 0.56 and 0.65 in the Central Plains and between 0.22 and 0.4 for the Northwest. These ranges amounted to correlation differences between 32% and 61% between the sectors, which indicated that these groups were not as able to explain the bias tendencies. This was further reflected in the differences in MBE values between the Central Plains, which were between -21% and -42%, and the Northwest, which were between -60% and -72%. These ranges resulted in MBE differences between 39 and 51 points, which represented markedly worse agreement when compared to the previously discussed groups. These issues were not entirely surprising, as the stratiform precipitation regime was of particular difficulty for GPE-CNN, and from these results, it becomes difficult to determine how to better understand these biases. However, it was believed that environmental factors could contribute to this problem, as certain environments may be more conducive to producing precipitation that GPE-CNN can more easily recognize.

For the last set of groups, Groups 9 (Figures 15i, 16i) and 11 (Figures 15k, 16k), the interpretation of the biases and correlations becomes more difficult. Firstly, in the case of Group

9, its overall representation of non-raining conditions lends it to displaying both the lowest overall precipitation rates and the most substantial underestimation tendency of the groups. For this group, the MBE for the Central Plains was -64%, while it was -75% for the Northwest; this represented an 11-point disagreement in these values for the sectors, which indicated modest agreement. The Pearson correlations, however, were less supportive as the correlation for the Central Plains ($r = 0.48$) and the Northwest ($r = 0.14$) showed a 71% difference, much larger than the overall correlation difference. In terms of statistics, Group 11 behaved somewhat less favorably than Group 9, showing a 24-point difference between Central Plains (MBE = -33%) and Northwest (MBE = -57%) mean biases and a 61% difference between the Central Plains ($r = 0.61$) and Northwest ($r = 0.24$) correlations. While interpretation of Group 11's impacts was difficult due to a lack of understanding of what the group represented, Group 9 seemed to behave in a manner that signaled GPE-CNN's ability to recognize what non-raining scenes should look like, but also erroneously assigning that label to lightly precipitating regimes. This could also be in relation to the environment, as certain factors in the lower atmosphere, such as available moisture, may prevent some of these scenes from precipitating and not others.

Considering Figures 15 and 16 overall, it was shown that between these two sectors, each of the satellite groups displayed some self-similarity in their bias tendencies. This was most clear in correlation space, where the satellite groups generally displayed a decrease in correlation percent difference between the sectors. This was an encouraging result, as this showed that on a broader scale, partitioning the retrieval predictions by the information within the satellite data can draw the overall performance into better agreement within areas that were not trained upon. This improvement was most noticeable in the more convective groups, which showed a 25-point improvement from the overall correlation difference. Comparisons with the remaining sectors

further supported this determination, meaning that this method of separation was generalizable to other geographic locations. When considering MBE, however, this behavior became more difficult to interpret; in the groups more closely related to convection, similar MBE ranges were broadly being seen, while the more stratiform groups did considerably worse. Even then, the convective groups displayed different bias tendencies, with the MBE values in the Central Plains having both positive and negative values, while the Northwest only had negative values. These discrepancies were further highlighted in the actual bias values, which are described in greater detail in Appendix A. These results suggested that satellite data alone was unable to more closely constrain the individual bias values, but could constrain the overall bias tendencies. Because of this, it was believed that incorporating the ancillary environmental data into this analysis would allow for better characterization of the mean biases, owing to the varying effects of the different climatological regions represented in each sector. Also, since environmental modification that was not captured in the satellite data can affect precipitating systems, understanding the environment's ability to support precipitation should allow for a clearer understanding of the retrieval biases.

3.4 Environmental State Influences

While understanding the satellite perspective is important for assessing the effects of various external influences on the performance of GPE-CNN, understanding the large scale environment is also crucial. As described earlier, an identical cloud feature in two different environments does not necessarily lead to similar precipitation processes. Because of this, the group analysis described in the previous section can likely be augmented by the inclusion of information regarding the meteorological state.

For this analysis, the group of nine environmental variables described in Chapter 2 were utilized. These variables were then partitioned into the 12 satellite groups and analyzed as such. Figure 17 displays an example of this partitioning using 700hPa relative humidity (RH700) in the Central Plains. Here it can be seen that RH700 had different effects in each satellite group, though none of the groups displayed any clear relationships between RH700 and retrieval bias. However, some understanding of the overall bias behaviors can be gleaned here. More specifically, the non-lightning groups appeared to exhibit a weak nonlinear trend in bias. This can be highlighted using Groups 1 (Figure 17a, 18a), 2 (Figures 17b,18b), 6 (Figures 17f, 18f), 9 (Figures 17i, 18i), and 10(Figures 17j,18j), where the bias appeared to become more negative as RH700 increased. While this was a weak visual trend, its presence suggested that RH700 may be providing information that showed some promise in assessing the retrieval biases. This behavior was most strongly noted in the satellite groups that did not have lightning present in them. While no explicit analysis was carried out to investigate this reason, it was believed that RH700 and lightning data have similar information contents in this context, and as such, RH700 would be more impactful in those cases where lightning data was absent.

Though the aforementioned behavior regarding RH700 and retrieval bias was most noted in the non-lightning groups, some of the lightning groups did show this behavior as well. Group 7 (Figures 17g, 18g) is one such example; here there is still some visibility for this behavior, but it was much harder to determine. Groups 3 (Figures 17c, 18c) and 4 (Figures 17d, 18d) also appeared to show this behavior, though it was very modestly. This could be related to the lightning data, as discussed earlier. Since Group 7 does have lightning information, it could be that RH700 was providing overlapping information with the lightning data. However, since the behavior is still modestly visible here, RH700 still contributed some unique information,

indicating that it could be somewhat usable in these scenarios as well. From this, it seemed that RH700 would have its greatest contribution in the absence of lightning information, though it could still provide information when lightning was present.

The remaining groups, Groups 5 (Figures 17e, 18e), 8 (Figures 17h, 18h), 11 (Figures 17k, 18k), and 12 (Figures 17l, 18l), did not appear to experience any substantial influence from RH700 regarding their biases. The reasoning behind this was potentially linked to some of the defining characteristics for these groups. In the cases of Groups 5 and 11, GPE-CNN already struggled with their representation, and as such, additional information may not be useful in determining biases. Groups 8 and 12, however, are different in that they are among the cases where GPE-CNN showed better performance. As such, the inclusion of the lightning information was possibly masking a portion of the impact of RH700 here, like in the previous discussion involving the other convective groups, Groups 3 and 4. These groups also displayed warmer mean brightness temperatures, which may also be influencing this assessment. Since Groups 8 and 12 are overall warmer, yet include lightning, GPE-CNN may consider them weaker convection, which could bring about some greater uncertainty regarding precipitation intensity. As such, the overall greater uncertainty and the inclusion of lightning data may be compounding factors in these groups.

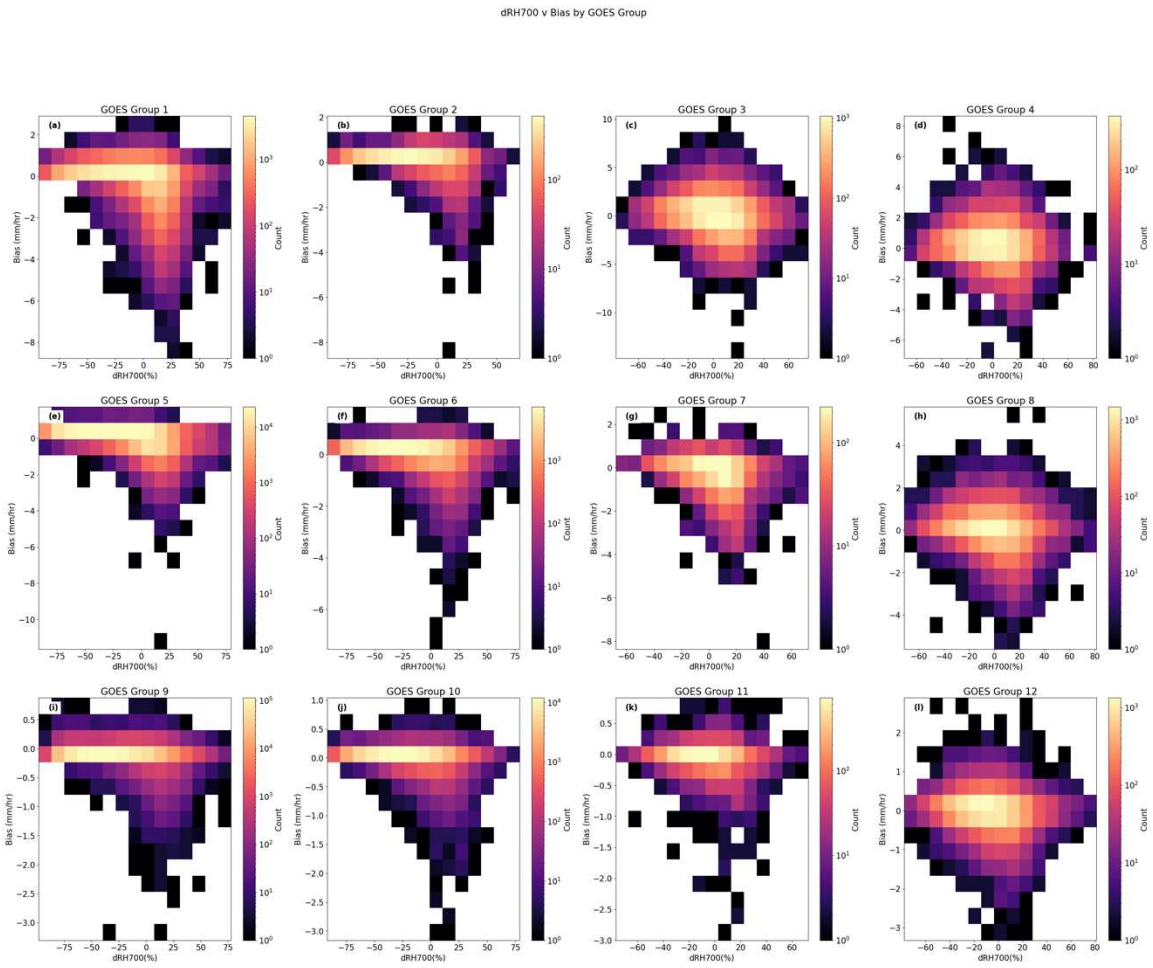

Figure 17: 2D histograms of bias and RH700 for the Central Plains partitioned by GOES group. In some of the groups, most notably those without lightning information, there appears to be a nonlinear trend between RH700 and retrieval bias. In lightning-present groups, this trend becomes weaker or virtually nonexistent.

When considering the effects of the wind variables selected for this analysis, the narrative changed. Figure 18 displays an identical analysis as done for RH700, but this time for the 500hPa meridional wind (v500). From this figure, it is apparent that there is no quantifiable trend between the retrieval biases and the wind. Though there was no immediate trend to assess, there were still interesting behaviors noted in these data. Most notably, it appeared that as the meridional wind speed increased, the range of possible biases decreased. A specific example of this can be seen in Group 8, where the spread in potential biases decreased from around 9 mm/hr at 0m/s to around 4 mm/hr at 30m/s; with both of these ranges centered at zero, it can then be

said that increasing wind speed was associated with greater certainty in GPE-CNN, and therefore more constrained biases. Such a behavior may result from the organization of the precipitation regime being considered, since midlevel winds play an important role in supporting the organization of various precipitating systems. This behavior was also noted in all of the satellite groups with the exception of Group 1, indicating that the information provided from this aspect of the environmental state was useful under most conditions. It is, however, worth noting that this behavior was not fully consistent among the remaining wind variables; while the 700hPa wind (v700) did show a similar but weaker behavior, the 700hPa speed shear (WSS700) did not. This was not entirely surprising, as the correlation with bias for this variable was the lowest of all variables tested, and as such, would have been the least likely to display a useful behavior regarding the retrieval biases. Overall, these results suggested that there was some useful information to be gained when using these wind variables in the cases where lightning was present. Furthermore, these results alongside those of RH700 suggested that a combination of variables could add value to constraining the biases of these groups. From this version of the environmental assessment, one can see that the interpretation of effects remained quite complex, but that there was information to be gained from this ancillary data. Because of this, it may be possible to decompose the effects of the environment on retrieval error utilizing methods that more carefully incorporate the environmental information into the analysis.

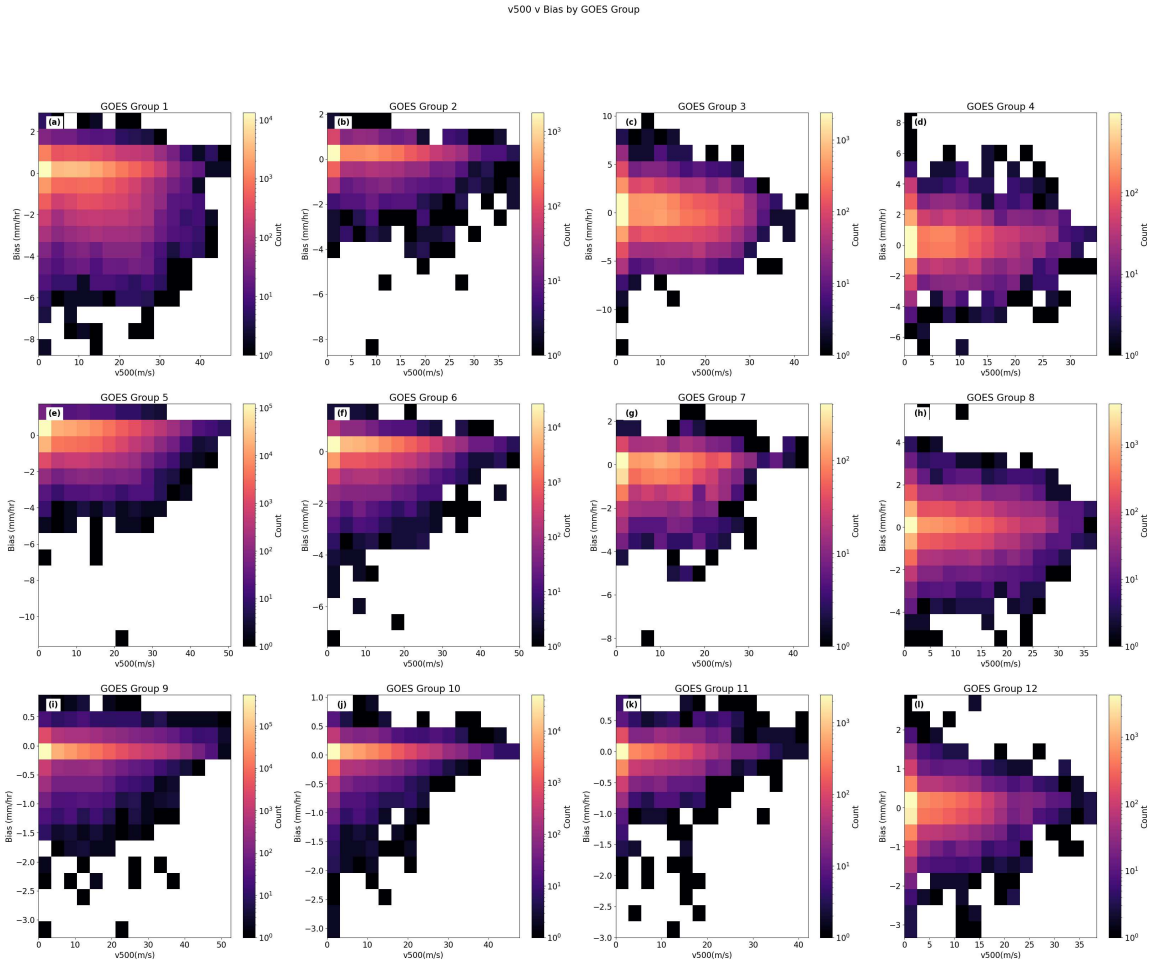

Figure 18: The same as Figure 17, but for bias and v500. While there are no immediately observable relationships between the retrieval bias and v500, a good portion of the groups show a notable decrease in bias with increasing wind speed.

Using the information gathered from these analyses, one can also determine how consistently these variables were able to influence the explainability of the retrieval biases when given a specific satellite group. To investigate this, a linear fit for the retrieved and observed precipitation rates were constructed for Group 8. This group was chosen to represent the more convective samples, which tended to have the best performance in GPE-CNN and were therefore assumed to be more likely to display usable behaviors. After performing the linear fit, the slope of the line of best fit was taken. This was used as a proxy for bias, as a slope less and 1 would indicate negative biases, while a slope greater than 1 would indicate positive biases. After getting

this slope for the full group, that same data was then partitioned into three categories, a low, medium, and high category, based in threshold values selected for each ERA5 variable used. Each of these categories then produced their own lines of best fit to be compared with the group's overall fit.

Using this method, lines of best fit were developed for each of the nine ERA5 variables used previously for Group 8. These slopes, along with the slope for the entirety of Group 8 for each sector are shown in Figure 19. From this table, it can be seen that, while certain variables were able to develop more coherent relationships between the sectors, not all of the variables were able to perform this task. For example, v500 did quite a good job in developing more consistent slopes and slope behaviors between the sectors. In the Central Plains, the low, medium, and high categories for v500 displayed slopes of 0.908, 0.891, and 0.883 respectfully, indicating an increased tendency for underestimation as v500 increased, manifested as a decrease in slope of about 0.03. When compared with the remaining sectors, all sectors displayed increases in underestimation, those being about 0.25, 0.06, 0.06, and 0.12 for the Northwest, Northeast, Southeast, and Southwest respectively. It was also worth noting that for v500, both the Northwest and Southeast sector showed an increase in slope from the low to medium categories. This general consistency in slopes was also noted in v700, though there were more instances of mismatch with this variable; here, all but the Northeast sector displayed increases in overestimation, seen as an increase in slope, with increasing wind speed. These results suggested that the wind variables in particular could be used to enhance the explainability of the bias tendencies in GPE-CNN, therefore providing a pathway by which biases can be related to the environmental state.

While v500 and, to a lesser extent, v700 were capable of increasing the explainability of GPE-CNN's biases across the different sectors, there were several drawbacks noted. Firstly, these relationships still did not directly capture the actual bias values. Instead, it showed how the retrieved and observed precipitation rates can be related, which you could get the biases from. Also, the slope values themselves were not consistent across the sectors. In all cases, the Southwest sector had slope values greater than one, indicating overestimation, while the rest of the sectors had slopes below one, indicating underestimation. Even when not considering the Southwest sector, the remaining sectors did not have consistent slope magnitudes, meaning that while the bias tendencies were more generalizable, the individual bias values were not. This could potentially speak to the problem being overall ill-posed; there may just not be enough information included in this univariate analysis to determine a more robust relationship. As such, although considering the effects of the environment through the lens of the satellite groups did allow for more actionable information to be extracted regarding the biases, the problem remains complex and thus requires further investigation.

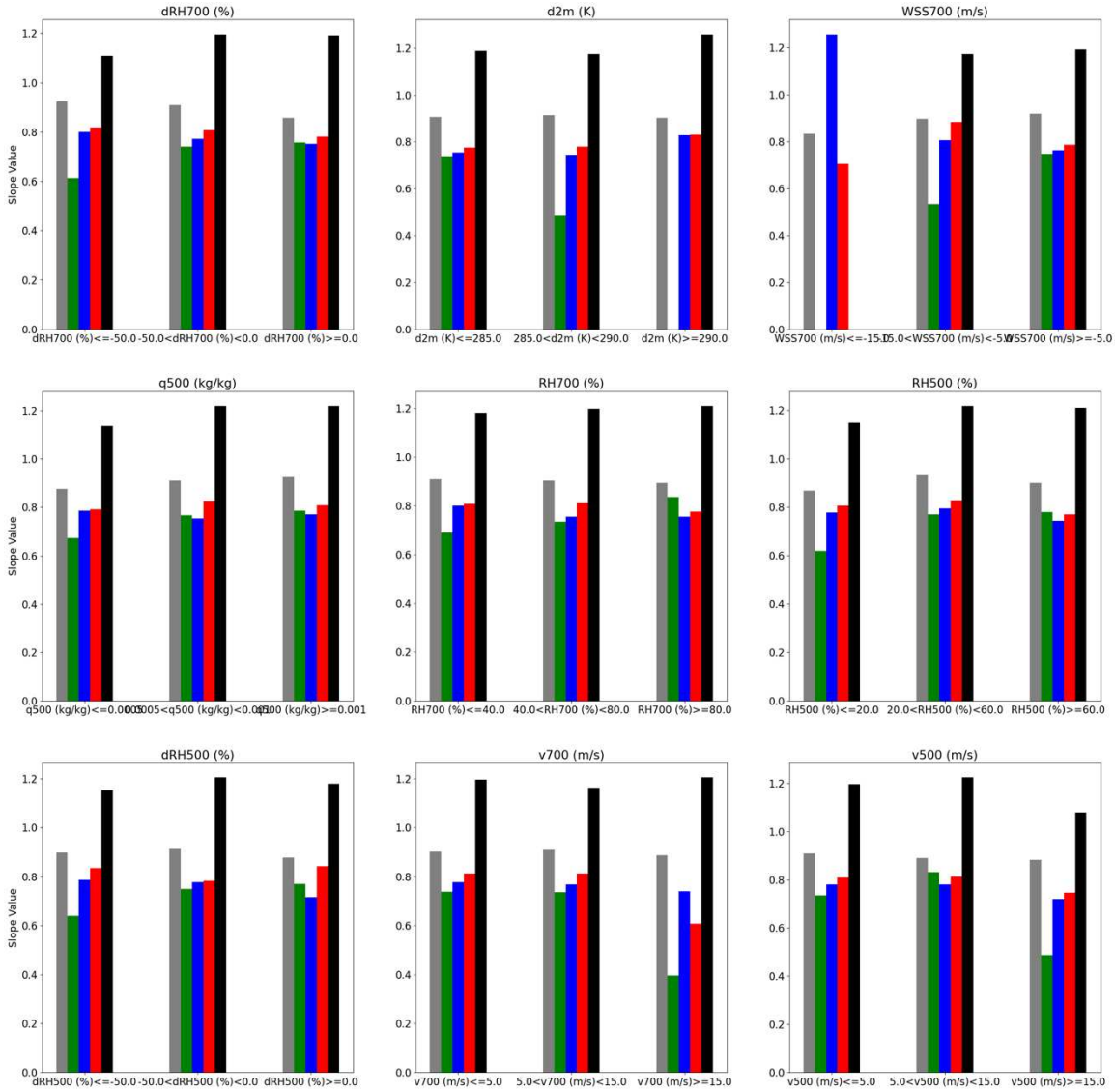

Figure 19: Bar charts of the best-fit slopes by threshold for each sector and ERA5 variable utilized in this thesis. The bar colors indicate the different sectors, those being the Central Plains (grey), the Northwest (green), the Northeast (blue), the Southeast (red), and the Southwest (black). The colored dashed lines indicate the overall best-fit slope for the like-colored sector. While the values of the slopes themselves were rarely in agreement, it was noted that in some of the variables, v700 and v500 being examples, the different sectors all saw similar changes in slopes between the groups.

CHAPTER 4: DISCUSSION AND SUMMARY

This thesis was developed to address the ability to understand the errors within a given infrared precipitation retrieval algorithm in terms of the environment in which the retrieval was performed. Since these algorithms are often used to provide precipitation estimates for areas where there is a lack of direct observations, understanding their performance in those areas is crucial to preserving confidence in these retrievals. Unfortunately, the lack of truth data to evaluate the retrieval against precluded this need. Thus, by attempting to diagnose the underlying error behaviors in a known training area and relating those behaviors to other areas outside of training, a framework was constructed to understand the general behaviors underlying these large-scale errors.

For simplicity and to avoid the broader assumptions associated with operational retrievals, a neural network-style retrieval algorithm, GPE-CNN, was developed for this thesis. This retrieval was purposefully left in a relatively simple state, taking only brightness temperature and lightning input from ABI and GLM to estimate MRMS precipitation rates over the Central Plains. Despite the relative lack of complexity in this retrieval, it was seen that not only did GPE-CNN often correctly locate and structure precipitation, but its performance statistics were comparable to those of previously-described retrieval algorithms. However, upon forcing GPE-CNN to predict precipitation in areas it was not trained in, substantial deviations in performance were noted. Specifically, while the Central Plains MBE as around -16%, the Northwest MBE was around -58%, indicating a 42-point difference between the two sectors. Other sectors displayed even greater bias differences, amounting to an up to 100-point difference in MBE in the Southeast sector. These percent biases for the minimum and maximum rain rates

also showed large discrepancies in MBE between the sectors, further signifying that GPE-CNN's performance can vary substantially in different geographic areas.

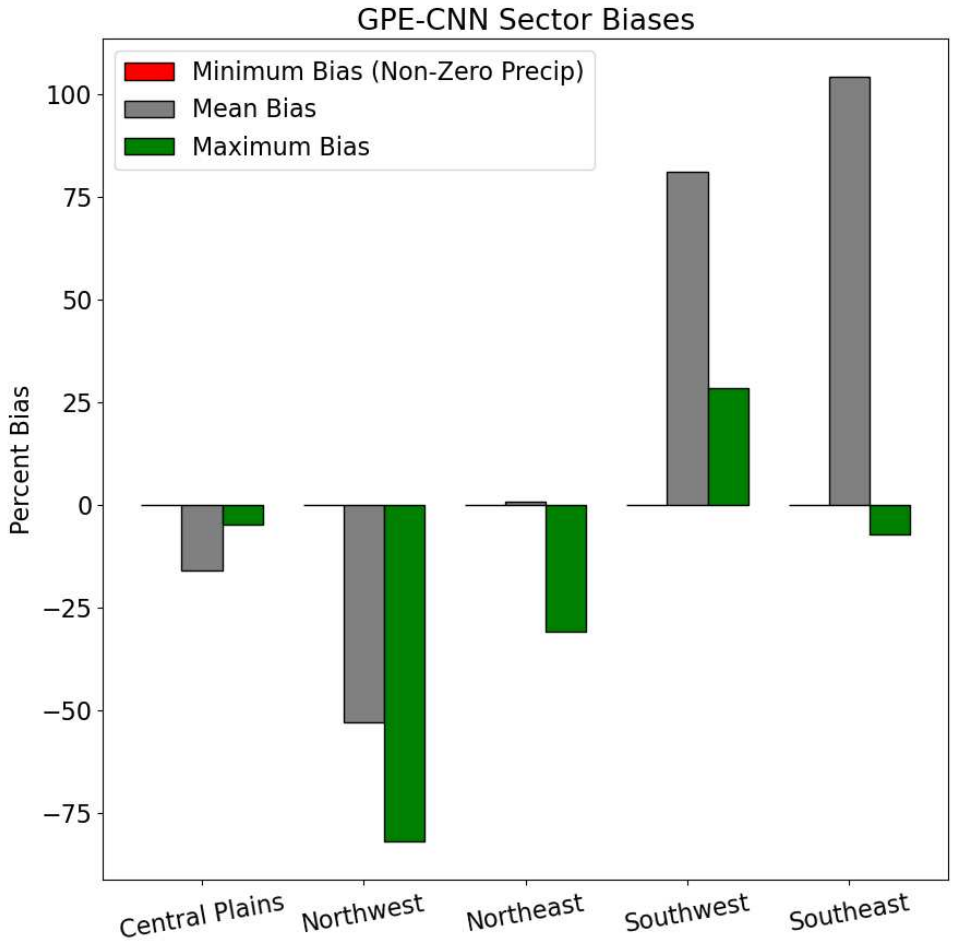

Figure 20: The minimum (red), mean (grey), and maximum (green) percent biases for each CONUS sector. The minimum and maximum biases are calculated using the minimum and maximum non-zero precipitation rates from the MRMS and GPE-CNN data, while the mean is the MBE. Ideally, these bias characteristics should match with the Central Plains, and any mismatch would indicate differences in GPE-CNN's performance when considered outside of the Central Plains training area.

To further identify the modes which contributed to the afore-mentioned deviations in performance, data analyses involving both the input satellite data and ancillary environmental data from ERA5 were performed to assess the impacts of both the sensor observations and the underlying meteorological state on these errors. Firstly, analysis of the satellite data suggested

that GPE-CNN was most impacted by both its ability to locate precipitating cloud features and the convective potential of said features; from this assessment, it was determined that more convective scenes generally performed better than their stratiform counterparts. These properties were further investigated through a correlation analysis relating the strength of the correlations of roughly one-degree areas within each given sector, termed subdomains, to the descriptive statistics of the satellite data. Using this analysis, it was shown that the lightning data was quite important in determining overall performance, with the presence of lightning almost exclusively occurring in the positively correlated subdomains. Furthermore, as the subdomain correlations increased, specifically from a correlation of 0.3 onward, the average lightning intensity also increased. This suggested that lightning intensity was also important in promoting the relationship between lightning and GPE-CNN large-scale performance.

Alongside the lightning data contributions, the brightness temperature standard deviation was also found to be a useful performance assessment metric. More specifically, there was a general increase in median ABI Channel 13 brightness temperature standard deviation from around 3K at a correlation of -0.7 to around 13K at a correlation of 0.8. This suggested that the amount of brightness temperature heterogeneity in a given subdomain was important in understanding GPE-CNN's performance. Understanding this, the coherence of cloud boundaries was then determined to be important in GPE-CNN's performance assessment alongside the prevalence of lightning. Inclusion of the mean brightness temperature, though difficult to interpret, was decided upon due to the indirect relationship between brightness temperature and precipitation. As cold cloud features were a sufficient condition for precipitation, colder average brightness temperatures can be associated with increasing precipitation intensity.

By partitioning the subdomains along these three variables, it was shown that the disagreement between sectors can begin to be characterized by these features. This was especially true when considering the Pearson correlations, which in some cases reduced the discrepancy between the Central Plains and Northwest by half. This, however, was not as clear in the MBE values, which often exhibited similar ranges of values but rarely agreed within the individual groups. These results indicated that the satellite group partitioning was able to discern the general behaviors of the retrieval biases, but was unable to more accurately depict the individual biases themselves. As such, parsing by the satellite features created a step by which these errors can be better understood, but this step required further information to properly assess bias in a quantitative sense.

Considering the environmental data analysis, it was shown that the problem did become even more complex. By first reducing the number of ERA5 variables considered to those that were most relevant for this problem, it was shown that moisture and wind properties were the most elucidative in terms of retrieval bias. When partitioned alongside the satellite data, it was also shown that while there were no clear relationships to be gathered from this data, the environmental inputs did have some limited explanatory power. However, this partitioning was found to not be universal, as the complexity of the analysis did not lend itself to strong generalizable relationships between the environment and the retrieval errors. In those groups where there was no lightning data, the environmental information appeared to have the greatest impact, especially considering the moisture variables, which seemingly made up for the lack of information instigated by the absence of lightning. In those groups with lightning present, the environmental data was still able to provide some unique contributions, namely through the wind

variables. In all cases, though, the environmental analysis did not provide specific information on how the retrieval biases and the environmental state were related.

Given the apparent complexity of this problem, further work is necessary to better understand the impact of this ancillary data in predicting retrieval errors. While this was attempted through the various algorithms discussed in Chapter 3, none of these methods were attempted in concert with the satellite group information. Knowing now that these groups have considerable explanatory power in the overall performance of GPE-CNN, using these groups alongside the environmental data with the methods listed in Table 4 may develop a usable relationship. Many of these algorithms operate in a multivariate space, which could aid in addressing the ill-posed nature of this problem. Other work may include the use of other environmental variables related to static stability and pressure perturbations, as studies have found that these two aspects of the environmental state also modulate precipitation. Finally, using satellite data that was more physically related to precipitation, such as passive microwave data, may make the desired relationships clearer. With this, it is worth noting that although the overall workflow of this thesis would remain identical, the satellite groups themselves would likely have to change, since passive microwave imagery does not have coincident lightning information. In any case, the work of this thesis clearly showed that this problem was incredibly complex and that universal statements regarding bias constraint were unlikely to exist in this style of retrieval.

REFERENCES

Arkin, P. A., & Meisner, B. N. (1987). The Relationship between Large-Scale Convective Rainfall and Cold Cloud over the Western Hemisphere during 1982-84. *Monthly Weather Review*, *115*(1), 51–74.

Ba, M. B., & Gruber, A. (2001). GOES Multispectral Rainfall Algorithm (GMSRA). *Journal of Applied Meteorology and Climatology*, *40*(8), 1500–1514.

Bailing, R. C., Meyer, G. A., & Wells, S. G. (1992). Relation of surface climate and burned area in Yellowstone National Park. *Agricultural and Forest Meteorology*, *60*, 285–293.

Chahine, M. T. (1992). The hydrological cycle and its influence on climate. *Nature*, *359*(6394), 373–380.

Chen, T.-C., & Pfaendtner, J. (1993). On the Atmospheric Branch of the Hydrological Cycle. *Journal of Climate*, *6*(1), 161–167.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., … Vitart, F. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Delgado, G., Machado, L. A. T., Angelis, C. F., Bottino, M. J., Redaño, Á., Lorente, J., Gimeno, L., & Nieto, R. (2008). Basis for a rainfall estimation technique using IR-VIS cloud classification and parameters over the life cycle of mesoscale convective systems. *Journal of Applied Meteorology and Climatology*, *47*(5), 1500–1517. https://doi.org/10.1175/2007JAMC1684.1

Elsaesser, G. S., Kummerow, C. D., L'Ecuyer, T. S., Takayabu, Y. N., & Shige, S. (2010). Observed Self-Similarities of Precipitation Regimes over the Tropical Oceans, *Journal of Climate*, *23*(10), 2686-2698. https://doi.org/10.1175/2010JCLI3330.1

Goodman, S. J., Blakeslee, R. J., Koshak, W. J., Mach, D., Bailey, J., Buechler, D., Carey, L., Schultz, C., Bateman, M., McCaul, E., & Stano, G. (2013). The GOES-R Geostationary Lightning Mapper (GLM). *Atmospheric Research*, *125–126*, 34–49. https://doi.org/10.1016/j.atmosres.2013.01.006

Habib, E., Krajewski, W. F., Nespor, V., & Kruger, A. (1999). Numerical simulation studies of rain gage data correction due to wind effect. *Journal of Geophysical Research Atmospheres*, *104*(D16), 19723–19733. https://doi.org/10.1029/1999JD900228

Hanna, J. W., Schultz, D. M., & Irving, A. R. (2008). Cloud-top temperatures for precipitating winter clouds. *Journal of Applied Meteorology and Climatology*, *47*(1), 351–359. https://doi.org/10.1175/2007JAMC1549.1

Henderson, D. S., Kummerow, C. D., Marks, D. A., & Berg, W. (2017). A regime-based evaluation of TRMM oceanic precipitation biases. *Journal of Atmospheric and Oceanic Technology*, *34*(12), 2613–2635. https://doi.org/10.1175/JTECH-D-16-0244.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hilburn, K. A., Ebert-Uphoff, I., & Miller, S. D. (2020). Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using goes-r satellite observations. *Journal of Applied Meteorology and Climatology*, *60*(1), 3–21. https://doi.org/10.1175/JAMC-D-20-0084.1

Hilburn, K., Lee, Y., Zupanski, M., & Wu, T. C. (2020). *Assimilating GOES-R Latent Heating in FV3 using Machine Learning*.

Hong, Y., Hsu, K.-L., Soroosh Sorooshian, & Xiaogang Gao. (2004). Precipitation Estimation from Remotely Sensed Imagery Using an Artificial Neural Network Cloud Classification System. *Journal of Applied Meteorology and Climatology*, *43*(12), 1834–1853.

Hsu, K.-L., Gao, X., Sorooshian, S., & Gupta, H. v. (1997). Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks. *Journal of Applied Meteorology and Climatology*, *36*(9), 1176–1190.

Joyce, R., & Arkin, P. A. (1997). Improved Estimates of Tropical and Subtropical Precipitation Using the GOES Precipitation Index. *Journal of Atmospheric and Oceanic Technology*, *14*(5), 997–1011.

Kirstetter, P. E., Hong, Y., Gourley, J. J., Chen, S., Flamig, Z., Zhang, J., Schwaller, M., Petersen, W., & Amitai, E. (2012). Toward a framework for systematic error modeling of spaceborne precipitation radar with NOAA/NSSL ground radar-based national mosaic QPE. *Journal of Hydrometeorology*, *13*(4), 1285–1300. https://doi.org/10.1175/JHM-D-11-0139.1

Kirstetter, P.-E., Hong, Y., Gourley, J., Cao, Q., Schwaller, M., & Petersen, W. (2014). Research Framework to Bridge from the Global Precipitation Measurement Mission Core Satellite to the Constellation Sensors Using Ground-Radar-Based National Mosaic QPE. In *Remote Sensing of the Terrestrial Water Cycle* (Vol. 206, pp. 61–79).

Kuligowski, R. J. (2002). A Self-Calibrating Real-Time GOES Rainfall Algorithm for Short-Term Rainfall Estimates. *Journal of Hydrometeorology*, *3*(2), 112–130.

Kuligowski, R. J., Li, Y., Hao, Y., & Zhang, Y. (2016). Improvements to the GOES-R rainfall rate algorithm. *Journal of Hydrometeorology*, *17*(6), 1693–1704. https://doi.org/10.1175/JHM-D-15-0186.1

Loriaux, J. M., Lenderink, G., & Pier Siebesma, A. (2016). Peak precipitation intensity in relation to atmospheric conditions and large-scale forcing at midlatitudes. *Journal of Geophysical Research*, *121*(10), 5471–5487. https://doi.org/10.1002/2015JD024274

Maggioni, V., Sapiano, M. R. P., & Adler, R. F. (2016). Estimating uncertainties in high-resolution satellite precipitation products: Systematic or Random Error? *Journal of Hydrometeorology*, *17*(4), 1119–1129. https://doi.org/10.1175/JHM-D-15-0094.1

Mastrantonas, N., Bhattacharya, B., Shibuo, Y., Rasmy, M., Espinoza-Dávalos, G., & Solomatine, D. (2019). Evaluating the benefits of merging near-real-time satellite precipitation products: A case study in the Kinu basin region, Japan. *Journal of Hydrometeorology*, *20*(6), 1213–1233. https://doi.org/10.1175/JHM-D-18-0190.1

Merenti-Välimäki, H. L., & Laininen, P. (2002). Analysing effects of meteorological variables on weather codes by logistic regression. *Meteorological Applications*, *9*(2), 191–197. https://doi.org/10.1017/S1350482702002049

Petković, V., Kummerow, C. D., Randel, D. L., Pierce, J. R., & Kodros, J. K. (2018). Improving the quality of heavy precipitation estimates from satellite passive microwave rainfall retrievals. *Journal of Hydrometeorology*, *19*(1), 69–85. https://doi.org/10.1175/JHM-D-17-0069.1

Pielke, R. A., And, J. R., & Downton, M. W. (2000). Precipitation and Damaging Floods: Trends in the United States, 1932-97. *Journal of Climate*, *13*(20), 3625–3637. http://nic.fb4.noaa.gov/products/

Sadeghi, M., Asanjan, A. A., Faridzad, M., Nguyen, P. H. U., Hsu, K., Sorooshian, S., & Braithwaite, D. A. N. (2019). PERSIANN-CNN: Precipitation estimation from remotely sensed information using artificial neural networks–convolutional neural networks. *Journal of Hydrometeorology*, *20*(12), 2273–2289. https://doi.org/10.1175/JHM-D-19-0110.1

Schmit, T. J., Griffith, P., Gunshor, M. M., Daniels, J. M., Goodman, S. J., & Lebair, W. J. (2017). A closer look at the ABI on the goes-r series. *Bulletin of the American Meteorological Society*, *98*(4), 681–698. https://doi.org/10.1175/BAMS-D-15-00230.1

Sun, L., Chen, H., Li, Z., & Han, L. (2021). Cross validation of GOES-16 and NOAA multi-radar multi-sensor (MRMS) QPE over the continental united states. *Remote Sensing*, *13*(20). https://doi.org/10.3390/rs13204030

Upadhyaya, S. A., Kirstetter, P. E., Kuligowski, R. J., & Searls, M. (2021). Classifying precipitation from GEO satellite observations: Diagnostic model. *Quarterly Journal of the Royal Meteorological Society*, *147*(739), 3318–3334. https://doi.org/10.1002/qj.4130

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., Grams, H., Wang, Y., Cockcks, S., Martinaitis, S., Arthur, A., Cooper, K., Brogden, J., & Kitzmillller, D. (2016). Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621–638. https://doi.org/10.1175/BAMS-D-14-00174.1

APPENDIX A: IMPACTS IN REMAINING SECTORS

In the development of the arguments of this thesis, analyses of all considered sectors were performed, though only the Central Plains and Northwest CONUS sectors are shown in detail. The use of only the Central Plains and Northwester sectors was chosen to maintain a more concise detailing of the effects of the satellite and environmental data in assessing the biases in GPE-CNN. This appendix is meant to display the results from the three remaining sectors specifically regarding the satellite data perspective and to show that the interpretations and conclusions from the thesis can in fact be applied to these sectors as well.

Figure A1 shows the satellite groups for the Northeast sector. When considering the groups in terms of the precipitation regimes they most often represented, the mostly convective groups are performing the best overall, as was highlighted in the body of the thesis. The main differences that begin to arise occur in the potentially convective and the stratiform groups, as several of these constituents exhibited a large overestimation spike at very small MRMS precipitation rates. While the reasoning behind these spikes is unclear, they could be related to the presence of a precipitation regime that GPE-CNN is unfamiliar with. In such a case, GPE-CNN would assume the precipitation characteristics of the most similar regime that it had seen, which could cause either severe over- or under-estimations to occur.

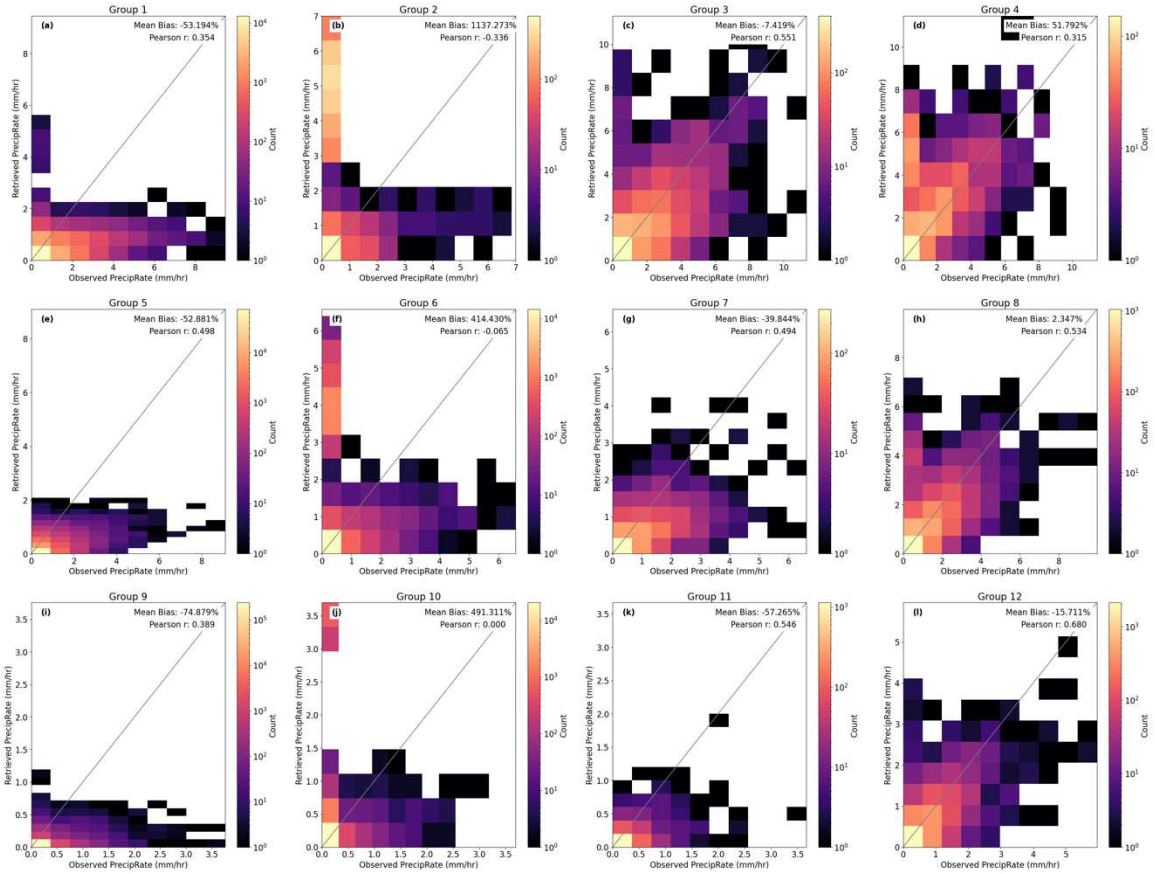

Figure A1: The Northeast sector testing data partitioned by the GOES satellite groups discussed in Chapter 3. While there are cases where certain groups experience a large degree of overestimation in their lightest precipitation rates, the overall behavior of the groups remained consistent with those in the Central Plains.

Figure A2 shows the satellite groups for the Southeast sector. Once the data from this sector had been divided into the satellite groups, one can see that its agreement with the Central Plains had greatly improved. Again, the best performance was noted in the mostly convective groups, though there was still a considerable amount of spread in these histograms. Also, in some of the potentially convective and stratiform groups, there were two maxima in the histograms: one with significant overestimation at small precipitation rates and one at the near-zero precipitation rates. This again could be related to the inclusion of a precipitation regime that was unknown to GPE-CNN in this sector. This does differ from the Northeast sector in that the

overestimations are not quite as egregious here, suggesting that this regime may be more similar to another that GPE-CNN had been trained on.

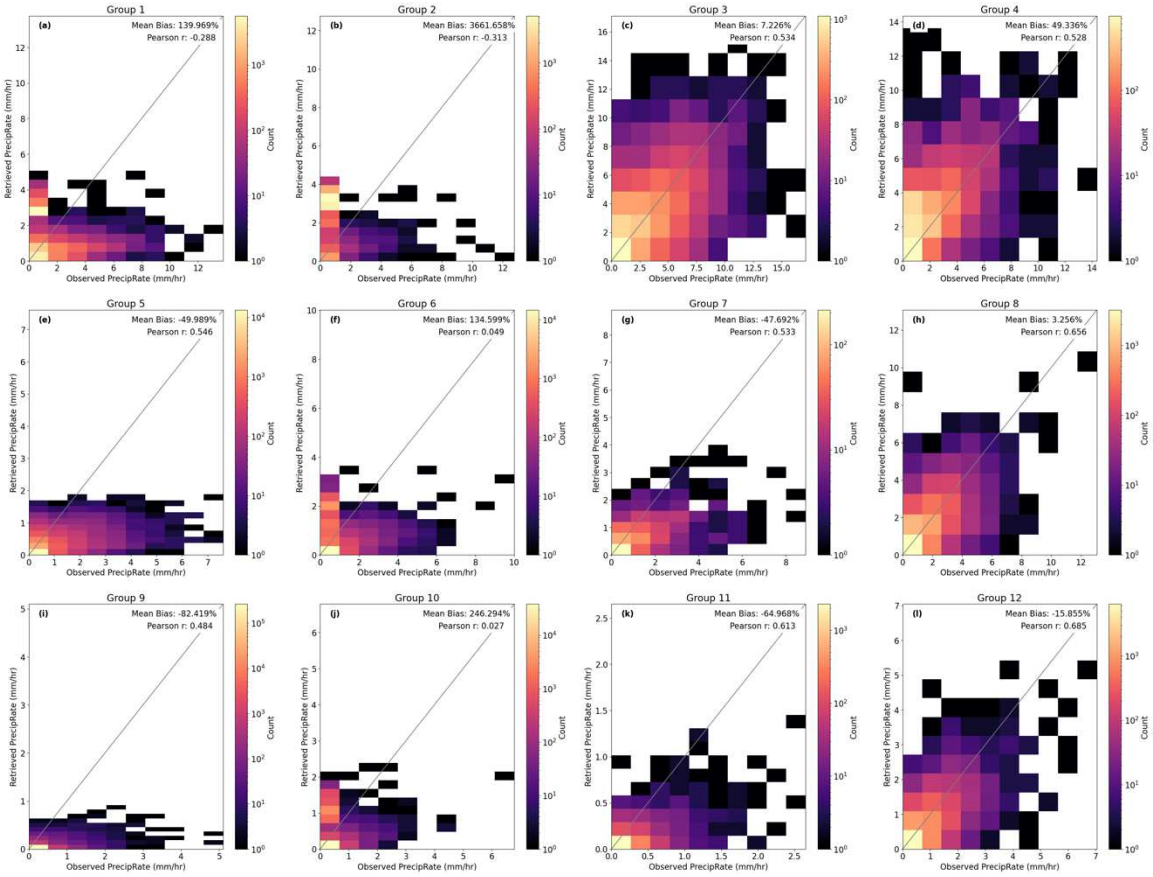

Figure A2: The same as Figure A1, but for the Southeast sector.

Figure A3 shows the satellite groups for the Southwest sector. This sector was the most similar to the Central Plains in the original intercomparison test, and this relationship mostly persisted through the satellite groups. The one exception to this were the overestimation spikes in Groups 2, 6, and 10. These overestimations may be related to the difference in environmental conditions between the Southwest and the Central Plains. Since the Southwest is more arid, the near-surface humidity is likely much lower than in the Central Plains. Because of this, the very

light precipitation in the MRMS data likely coincided with conditions that allowed for more precipitation to reach the surface in similar regimes in the Central Plains. As a result, the precipitation rates predicted by GPE-CNN would have been much greater than they should have been. This also likely explained the tendency of overestimation in this sector.

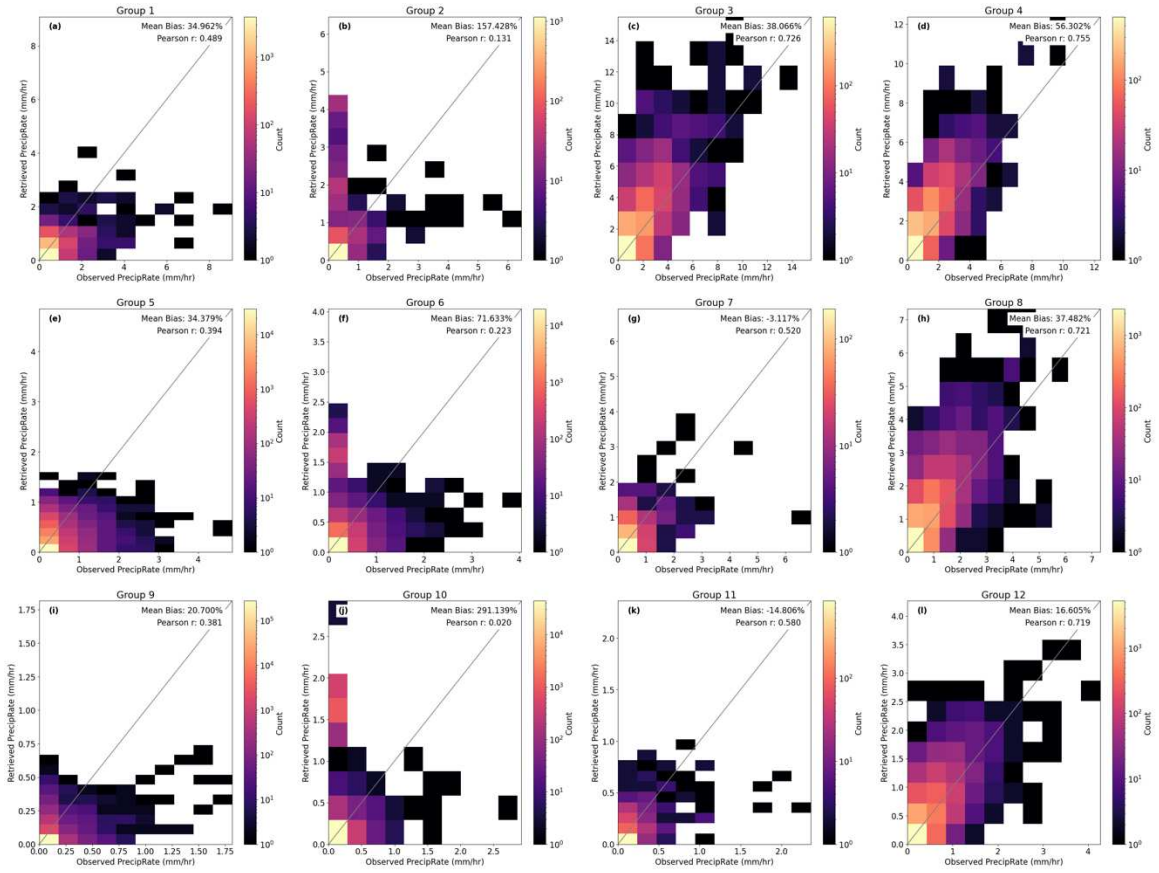

Figure A3: The same as Figure A1, but for the Southwest sector.

The mean bias and correlations can also be compared across all sectors to more comprehensively discuss the effects of the satellite groups on constraining GPE-CNN's performance. Tables AA1 and AA2 show this for mean bias and correlation respectively. From these tables, it was clear that in terms of correlation, dividing the data by these satellite groups provided improvement in most cases. In the more convective cases, this improvement was

universal, and in some cases reduced the differences between the Central Plains and the sector in question by 40%. In the more stratiform groups, however, the results were mixed, as the improvement shifted based on sector and group. This indicated that where GPE-CNN has the most information, i.e. in convective scenes, the satellite groups can go a long way in improving confidence, while this was not always the case in other groups. When considering mean bias, however, it became apparent that the satellite groups alone were inadequate in constraining biases. This was believed to be related to the information content of the satellite data itself, which was likely too low to assess the actual bias values, but adequate in assessing the overall bias tendencies. As such, the use of environmental data was believed to strengthen this relationship when considering the mean biases.

Table AA1: A comparison of the mean biases in mm/hr by satellite group across all sectors. Values in parentheses indicate the percent difference between the Central Plains and the given sector for that given group. The desired performance is that the percent differences in the satellite groups are lower than the percent differences shown in the 'Overall' row for a given sector. From this, it was shown that in terms of mean bias, these groups did not consistently reduce the bias difference between the Central Plains and the considered sector.

	Central Plains	Northwest	Northeast	Southeast	Southwest
Overall	-0.02	-0.03 (51%)	0.002 (107%)	0.133 (746%)	0.02 (206%)
Group 1	-0.147	-0.138 (6%)	-0.389 (165%)	0.87 (692%)	0.077 (153%)
Group 2	-0.091	-0.033 (63%)	2.884 (3283%)	2.699 (3079%)	0.244 (369%)
Group 3	-0.017	-0.101 (479%)	-0.159 (821%)	0.205 (1279%)	0.638 (3780%)
Group 4	0.302	-0.03 (109%)	1.002 (229%)	0.897 (195%)	0.716 (136%)
Group 5	-0.06	-0.066 (9%)	-0.111 (83%)	-0.19 (215%)	0.027 (144%)
Group 6	-0.029	-0.049 (67%)	0.647 (2308%)	0.255 (971%)	0.039 (232%)
Group 7	-0.326	-0.217 (34%)	-0.455 (39%)	-0.647 (98%)	-0.017 (95%)
Group 8	0.038	-0.098 (357%)	0.029 (24%)	0.039 (3%)	0.244 (537%)
Group 9	-0.004	-0.004 (0%)	-0.012 (196%)	-0.011 (189%)	0.001 (113%)
Group 10	-0.004	-0.012 (183%)	0.157 (2845%)	0.12 (2845%)	0.057 (1413%)
Group 11	-0.057	-0.084 (47%)	-0.124 (107%)	-0.118 (107%)	-0.015 (74%)
Group 12	-0.036	-0.079 (117%)	-0.093 (158%)	-0.089 (144%)	0.048 (232%)

Table AA2: The same as Table AA1, but for the Pearson correlations in each satellite group.

	Central Plains	Northwest	Northeast	Southeast	Southwest
Overall	0.83	0.42 (49%)	0.31 (62%)	0.44 (47%)	0.74 (11%)
Group 1	0.45	0.26 (42%)	0.36 (20%)	-0.28 (162%)	0.48 (7%)
Group 2	0.59	0.36 (39%)	-0.34 (158%)	-0.31 (153%)	0.13 (78%)
Group 3	0.78	0.59 (24%)	0.55 (29%)	0.53 (32%)	0.73 (6%)
Group 4	0.82	0.6 (26%)	0.32 (61%)	0.53 (35%)	0.76 (7%)
Group 5	0.56	0.22 (61%)	0.5 (11%)	0.55 (2%)	0.39 (30%)
Group 6	0.65	0.33 (49%)	-0.06 (109%)	0.05 (92%)	0.22 (66%)
Group 7	0.63	0.52 (17%)	0.49 (22%)	0.53 (16%)	0.52 (17%)
Group 8	0.8	0.61 (24%)	0.53 (34%)	0.66 (18%)	0.72 (10%)
Group 9	0.48	0.14 (71%)	0.39 (23%)	0.48 (0%)	0.38 (21%)
Group 10	0.59	0.4 (32%)	1e-6(100%)	0.03 (95%)	0.02 (97%)
Group 11	0.61	0.24 (61%)	0.55 (10%)	0.61 (0%)	0.58 (5%)
Group 12	0.8	0.61 (24%)	0.68 (15%)	0.69 (14%)	0.72 (10%)