Upadhyaya Shruti (Orcid ID: 0000-0002-1266-263X)

Towards Improved Precipitation Estimation with the GOES-16 Advanced Baseline Imager: Algorithm and Evaluation

Shruti A. Upadhyaya[1,*], Pierre-Emmanuel Kirstetter[1,2,3,4,*], Robert J. Kuligowski[5], Maresa Searls[2]

[1] Advanced Radar Research Center, University of Oklahoma, Norman, Oklahoma

[2]School of Meteorology, University of Oklahoma, Norman, Oklahoma

[3]School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, Oklahoma

[4]NOAA/National Severe Storms Laboratory, Norman, Oklahoma

[5]NOAA/NESDIS/Center for Satellite Applications and Research, College Park, Maryland

[*]*Corresponding authors:* Pierre-Emmanuel Kirstetter (pierre.kirstetter@noaa.gov);

Shruti A. Upadhyaya (shruti.a.upadhyaya-1@ou.edu)

Funding Information:

Funding for this research was provided by the NOAA GOES-R Series Risk Reduction program, which provided support to the Cooperative Institute for Mesoscale Meteorological Studies at the University of Oklahoma under Grant NA16OAR4320115, and by NASA Earth Science Division Earth Science Research from Operational Geostationary Satellite Systems (ESROGSS) which provided support to the Advanced Radar Research Center at the University of Oklahoma under Grant NA16OAR4320115. P. Kirstetter acknowledges support from the NASA Global Precipitation Measurement Ground Validation program under Grant NNX16AL23G and the Precipitation Measurement Missions program under Grant 80NSSC19K0681.

Abstract:

The study introduces a new quantitative precipitation estimation (QPE) algorithm from Advanced Baseline Imager (ABI) observations from GOES-16 across the CONUS. It is developed and comprehensively evaluated using the Ground Validation Multi-Radar/Multi-Sensor (GV-MRMS) system as a benchmark, and features Random Forest (RF) machine learning-based QPE. The key innovations of the algorithm include a comprehensive set of satellite predictors derived from five infrared ABI channels, complemented by low-level environmental conditions from RAP Numerical Weather Prediction (NWP) model, and outputs of probability of precipitation type for seamless integration of varying precipitation rates across types. A systematic analysis of the predictors is performed. The analysis reveals that satellite predictors contribute more to high-intensity precipitation estimates, whereas environmental predictors primarily condition low-intensity precipitation. Combining both categories of predictors improve scores (correlation coefficient, CC=0.41) overall, with the greatest improvement in the Warm Stratiform precipitation type. Introducing precipitation type information through probabilities further improves correlation (0.46). An inter-comparison of the RF model with the Self-Calibrating Multivariate Precipitation Retrieval (SCaMPR) shows that RF has better detection and quantification scores than SCaMPR (Heidke Skill Score, HSS=0.91 vs. 0.19; CC=0.44 vs. 0.26). Both retrievals display similar performance patterns across different regions of the CONUS. For example, skill degrades in complex terrain. Precipitation processes in complex terrain and their variability may not be well captured, especially with the degraded ABI resolution in the Western CONUS. It is recommended to complement the 11.2μm legacy channel with at least the 6.2μm channel for global precipitation retrievals using GEO sensors.

Keywords: GOES-16, Numerical Weather Prediction, Precipitation, Machine Learning, Classification, Geostationary Satellites

1. Introduction:

Precipitation is highly variable in space and time, and its estimation at fine scale and with low latency is pivotal for advancing our understanding of precipitation formation and evolution. As extreme rainfall events intensify under a changing climate (IPCC, 2021; Fowler et al., 2021), precipitation-related disasters such as flash floods have aggravated societal and economic consequences. Fine-scale and timely observations are needed to monitor rapidly-evolving high-impact precipitation, extract critical information from rapidly developing storms, and provide improved forecasts which can increase warning lead time for the associated hazards. As many regions around the world lack sufficiently dense surface-based observation networks, global observations from satellites are a highly attractive alternative (van Emmerik et al., 2015). Satellites in low-Earth orbit (LEO) offer global coverage but not frequently enough to capture the rapidly-evolving features of hazardous precipitation. Geostationary satellite (GEO) sensors have been used for decades to retrieve precipitation at sub-hourly time scales and at lower latency than other space-based precipitation products used for near-real time applications. For example, the Integrated Multisatellite Retrievals for GPM-Early (IMERG-E; Huffman et al. 2015) combines multiple satellite sensors but has a latency of around 4 hours. On the other hand, the NOAA's operational GOES-16/17 Self-Calibrating Multivariate Precipitation Retrieval (SCaMPR; Kuligowski 2002; Kuligowski et al., 2016) has a latency lower than 5 min, making it suitable for monitoring and forecasting fast-paced hydrologic processes such as flash floods. However, because challenges exist due to the indirectness of the relationship between cloud top observations from GEO sensors and surface precipitation (e.g., Kirstetter et al., 2018) further developments are required to advance GEO satellite precipitation estimates. While GEO sensors are critical for precipitation science and applications, the information content in IR observations may remain not well understood and is probably underutilized. Progress can be made by investigating the IR information content and by quantifying its benefit in terms of QPE accuracy.

Significant advances in satellite and sensor technology have led to the development of new-generation GEO sensors with improved spatial, temporal, and spectral information. The latest GEO satellite sensors such as the Advanced Baseline Imager (ABI) onboard the GOES-R satellites, the Advanced Meteorological Imager (AMI) onboard GEO-KOMPSAT-2A, and the Advanced Himawari Imager (AHI) onboard Himawari-8/9, capture observations at 0.5-2km spatial resolution every 10min or less across 16 spectral channels. Compared to the five channels on the previous GOES imager, the additional spectral information from the ABI improves the capability to detect very thin cirrus clouds, discriminate snow cover from clouds,

estimate cloud particle size and phase, and detect and track low- and mid-tropospheric water vapor (Schmit et al., 2005). Over the last few decades, several indices have been derived from one or several channels to identify cloud properties and precipitation microphysics. Recently, increased efforts have focused on effectively utilizing this big data at high spatial, temporal, and spectral resolution along with physical information for improving quantitative precipitation estimation (QPE) from space (Hirose et al., 2019; Kuligowski et al., 2016; Kirstetter et al., 2018; Lazri et al., 2020; Ouallouche et al., 2018).

As an example, Hirose et al. (2019) developed a QPE algorithm for the AHI sensor onboard the Himawari-8 satellite that uses Brightness Temperatures (BT) from 9 Infrared (IR) channels and combinations of Brightness Temperature Differences (BTD) as predictors. BTDs have been one of the earliest indices that combine information from two channels, and have been widely used for precipitation detection, classification, and quantification (Ba and Gruber, 2001; Kuligowski et al., 2016; Tjemkes et al., 1997; Upadhyaya and Ramsankaran, 2014, 2016). Min et al. (2018) also developed a similar algorithm for AHI by using additional information from the Global Forecast System (GFS) and surface elevation as environmental predictors. Recently, Ouallouche et al. (2018) and Lazri et al. (2020) developed a separate day and night-time QPE for SEVIRI (Spinning Enhanced Visible and Infrared Imager, onboard the MSG (Meteosat Second Generation) satellite. In addition to BTs and BTDs, spatial predictors such as the 10.8μm BT mean and variance across 5 x 5 grids were used along with a temporal predictor computed as the 10.8μm BT difference between two consecutive 15-min SEVIRI observations. For daytime, the retrieval used Visible (VIS) channels, whereas only IR channels are used for nighttime. Meyer et al. (2017) specifically focused on the significance of textural/spatial predictors on QPE from SEVIRI. With ABI, recent work on precipitation quantification involved the advanced machine learning (ML) techniques such as Deep Neural Networks (Tao et al., 2018) and Convolutional Neural Networks (Sadeghi et al., 2019), yet these studies used only two out of the 16 ABI channels. All of the above studies utilised one or a combination of ML techniques such as Random Forest, Artificial Neural Networks, Support Vector Machines, Deep Neural Networks and Convolutional Neural Networks. Meyer et al. (2016) evaluated some of the commonly used ML techniques and concluded that no single technique overperformed compared to the others. They recommended focusing further research on developing suitable predictors. In contrast with the above studies, Upadhyaya et al. (2021a, b) examined a large number of predictors (260) compared to using only infrared channels from ABI onboard GOES-16 to systematically analyse the detection of precipitation occurrence and types. The studies showed the significance of (1) deriving new satellite-based predictors and

(2) combining satellite information with Numerical Weather Prediction (NWP) model based environmental predictors.

The overarching goal of this study is to develop a low latency precipitation retrieval that provides "explainable QPE" for GOES-16 satellite observations across CONUS. Predictors discussed in Upadhyaya et al. (2021a, b) for precipitation typology are evaluated for precipitation quantification using precipitation type probabilities and the Random Forest (RF) ML technique. The QPE is designed to estimate instantaneous precipitation rates every 10 min (which is the temporal resolution of ABI full-disk imagery) at the full spatial resolution of the ABI infrared channels.

It is evaluated by comparing the retrievals to an external reference to assess the overall QPE accuracy, and also to the operational SCaMPR product to establish a baseline comparison with the state-of-the-art. Both SCaMPR and the retrievals developed in this study are designed at the ABI spatial-temporal resolution and use the same five ABI infrared channels; in this sense, the primary GEO information content is the same for both retrievals. The twofold comparison aims at depicting the opportunities and the challenges on the way to improving GEO precipitation estimation, specifically how much can be gained by deriving additional predictors from the same data, and the benefit of combining more model-predicted parameters than SCaMPR.

It is crucial to analyse and report the progress obtained with the new generation GEO QPEs for identified challenges such as:

(1) the impact of precipitation types on quantification (Upadhyaya et al., 2020);

(2) the impact of degraded ABI spatial resolution at higher zenith angle (Yu et al., 2008) on QPE;

(3) QPE in traditionally challenging regions such as complex terrain (Kuligowski 2011; Upadhyaya and Ramsankaran, 2018), and coastal areas (Nieman et al., 2004; Zou et al., 2011).

A detailed evaluation of QPE performance is expected to benefit precipitation science and users for precipitation applications, as well as provide guidance to algorithm developers for further algorithm improvement. Therefore, a comprehensive evaluation of QPE is performed across different climate regions, surfaces (ocean, coast, land) and regions (complex/flat terrain).

Following the Introduction, Section 2 of this paper describes the data sets and their processing, outlines the RF model, and introduces the evaluation scores for performance

assessment. Section 3 describes the results from various designed experiments. It is followed by concluding remarks in Section 4.

2. Data and Methods

2.1. Satellite and Environmental Predictors

Conventionally, most global operational precipitation algorithms such MERG, Global Satellite Mapping of Precipitation (GSMaP; Aonashi et al. 2009), the Climate Prediction Center morphing algorithm (CMORPH; Joyce et al. 2004), and Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks - Cloud Classification System (PERSIANN-CCS; Hong et al. 2004) utilize the 11.2 μm channel that is available globally from the Climate Prediction Center (CPC) at the National Oceanic and Atmospheric Administration (NOAA) National Weather Service (Janowiak et al. 2001). However, in the new era of GEO sensors, more channels are available and have been shown to provide additional information on precipitation (Hirose et al., 2019; Tao et al., 2018; Upadhyaya et al., 2021a, b).

The categories of predictors used in this study are listed in Table 1A. In total, seven categories of predictors are used: the first four consist of satellite predictors derived from the BT fields of five ABI IR channels at wavelengths 6.2μm (referred as T6.2), 7.3μm, 8.5μm, 11.2 μm, and 12.3μm; the fifth is the ABI zenith angle; the sixth consists of environmental predictors derived from the zero-hour analysis of the Rapid Refresh model (RAP; Benjamin et al., 2016); and the seventh category is MRMS precipitation type probabilities . Table 1A also provides an example for each category of satellite predictor along with its acronyms and Table 1B lists all of the environmental predictors. The number and significance of the predictors in each category discussed below.

Table 1. A. Categories of predictors used in this study. **B.** The Environmental predictors are further detailed in section B of the table.

Category	***Predictor Type***
A. Predictors Category (Acronym: Example)	
I	Brightness temperature (BT: T6.2*)
II	Brightness temperature difference (BTD: T8.5 – T11.2)
III	Difference of BTDs (D-BTD: (T6.2 – T7.3) – (T8.5 – T11.2))

IV	GLCM textures: mean, variance, homogeneity, contrast, and entropy (Te: T11.2 mean)
V	Satellite zenith angle (Ze)
VI	Environmental predictors (NWP: Listed below)
VII	Precipitation type classification probabilities from MRMS
B. Environmental Predictors	
VI:1	Vertically integrated precipitable water (PWV) (kg/m2)
VI:2	1000-700-hPa mean relative humidity (%)
VI:3–6	Relative humidity (%) at 900, 850, 700 and 500 hPa
VI:7	Surface equivalent potential temperature (K)
VI:8	Surface-based convective available potential energy (CAPE) (J/kg)
VI:9	Surface temperature (C)
VI:10–12	Temperature (K) at 850, 700 and 500 hPa
VI:13	Height of 0C isotherm (km)
VI:14 – 16	**Wind shear (m/s) between the surface and 850 hPa, 700 hPa and 500 hPa**
VI:17– 18	**Lapse rate (K/km) at 850-500-hPa and 850-700-hPa**
VI:19	**Wet bulb temperature**

*T6.2 is read as the brightness temperature of ABI channel 6.2μm; the **bold rows** in environmental predictors are derived from RAP output fields and other predictors are directly available from RAP output.

The first predictor category is BTs sampled at all five channels that are available during both day and night (5 BTs). Following Kuligowski et al. (2016), these IR observations are parallax-adjusted. T11.2 provides cloud top brightness temperature that is a proxy indicator for cloud top height (Ba and Gruber, 2001). The water vapor (WV) absorption channels (T6.2 and T7.3) are sensitive to different levels of tropospheric WV (upper and lower, respectively for

cloud-free regions). Other IR window channels (T8.5 and T12.3) display slightly greater absorption due to WV than T11.2 and are commonly known as "dirty" IR bands. Note that, following the design of several operational products such as PERSIANN-CCS and SCaMPR, the scope of this study is limited to only IR channels to maintain consistency across day/night retrievals. GEO VIS and Short-Wave IR (SWIR) channels will be considered in future analyses to more fully exploit GEO observations.

The second category of predictors consists of BTDs of all channel combinations (10 BTD predictors). The difference between two channels highlights cloud and precipitation properties unavailable from single channels. For example, the difference between T6.2 – T7.3 is used to separate low-level clouds from high-level clouds (Giannakos and Feidas, 2013; Kuligowski, 2011), and the T6.2 – T11.2 difference is commonly used to separate thin cirrus clouds from overshooting cloud tops (Upadhyaya et al., 2014, 2016; Ouallouche et al., 2018).

The third category of predictor is the difference of BTDs (D-BTD: 25 predictors). The difference between T8.4 – T11.2 and T11.2–T12.3 is used to identify cloud phase (So and Shin, 2018). Upadhyaya et al. (2021a, b) demonstrated the significance of additional D-BTDs combinations in detecting precipitation occurrence and convective precipitation.

The fourth category consists of texture predictors, which encode spatial information as an index. Texture indices, namely mean, variance, homogeneity, contrast, and entropy, are derived using the Grey Level Co-occurrence Matrix (GLCM; Haralick et al., 1973) for all possible combinations of BTs, BTDs and D-BTDs across 5 x 5 grids of ABI images. These texture predictors contribute the greatest number of predictors of any category (200 total). T11.2 spatial textures have long been recognised for detecting precipitation (Adler and Negri, 1988; Ba and Gruber, 2001; Hsu et al., 1997). Upadhyaya et al. (2021a, b) showed for the first time that spatial textures derived from BTDs and D-BTDs contribute considerably to identifying precipitation types. In this study, their relevance for quantification is evaluated.

The fifth and last category in the satellite predictors is the satellite zenith angle, which accounts for the varying spatial resolution of the ABI sensor which is around 2 km on the East Coast to 3-4km in Central CONUS, and around 12-16 km on the West Coast, and also accounts for the effects of limb darkening; i.e., increased attenuation of radiation as the length of the path increases with satellite zenith angle. Thus, a total of 241 satellite predictors are used for QPE.

Environmental predictors from NWP models complement the cloud-top information from GEO satellite sensors by providing low level environmental conditions that contribute to improving precipitation detection and quantification. The environmental predictors listed in

second part of Table 1 (19 predictors) include the vertically integrated precipitable water, surface temperature, wet bulb temperature, surface-based Convective Available Potential Energy (CAPE), height of the 0°C isotherm, air temperature, relative humidity, lapse rate and wind shear at different levels of atmosphere. In the preceding study on precipitation detection and classification (Upadhyaya et al., 2021a, b), it is found that atmospheric moisture content predictors such as relative humidity and precipitable water have the highest contribution in detecting stratiform types, while predictors such as lapse-rate and CAPE consistently have the highest contribution for convective precipitation.

The last category of predictors used to develop the quantification model includes probabilities of precipitation types (listed later in this section) that are predicted by the classification model (Upadhyaya et al., 2021a, b). Upadhyaya et al. (2021a) showed the value of probabilistic classification over deterministic separation of precipitation types. The probabilistic classification accounts for the under constrained information coming from cloud-top satellite observations and transition challenges from one type to another (e.g. non-uniform beam filling: NUBF reported in Kirstetter et al., 2012, and Upadhyaya et al., 2020). The full information content provided by precipitation type probabilities is thus more suitable than a deterministic classification.

The approach explored in this study differs from the commonly used two-step approach of classifying precipitation types first before quantifying precipitation for each type separately (e.g., Kuligowski et al., 2016; Hirose et al., 2019; Ouallouche et al., 2018). The model developed herein seamlessly accounts for the varying relationships between passive satellite sensor observations and surface precipitation of different types through probabilities of precipitation types. Hereafter, the impact of developing different models for each type is contrasted with incorporating classification probabilities into a single model.

2.2. Precipitation products: Ground Validation-Multi-Radar/Multi-Sensor (GV-MRMS) and SCaMPR

In this study, the Ground Validation-Multi-Radar/Multi-Sensor (GV-MRMS) product is used to develop and train the QPE model. GV-MRMS is the surface reference precipitation developed by Kirstetter et al. (2012; 2014; 2018) using MRMS (Zhang et al., 2016) to support the Global Precipitation Measurement mission for ground validation of satellite sensors and combined precipitation products across CONUS. GV-MRMS maximizes the accuracy of combined radar-gauge surface precipitation estimates through adjustments and conservative quality and quantity controls. The GV-MRMS precipitation types are used to train satellite-

based probabilistic precipitation type estimates (Upadhyaya et al., 2021a, b) and to evaluate satellite-derived QPE across different GV-MRMS detected precipitation types. Seven precipitation types from GV-MRMS are: Hail, Convective (Conv), Tropical Convective/Mix (Trp_ConvMix), Tropical StratiformMix (Trp_StratMix), Warm Stratiform (WarmStart), Cool Stratiform (CoolStrat) and Snow.

This study is carried out across the CONUS for the period of summer 2018 (June through September). Note the snow precipitation type is not considered given the low sample size across the summer season. GV-MRMS products are spatially aggregated to match the spatial resolution of ABI and are temporally aggregated to a scale of 30 min to mitigate uncertainty due to temporal matching and to the indirect link between cloud-top observations and precipitation processes. Regarding precipitation type, the aggregated product represents the highest proportion of occurrence of precipitation type within the considered space-time window. Conservative quality controls are applied to the resampled GV-MRMS data to derive the precipitation reference. Details on data preparation, matching, and thresholds used for quality control are provided in Upadhyaya et al. (2021a, b).

The RF precipitation model is inter-compared with the latest version of SCaMPR (Kuligowski et al. 2016). SCaMPR (abbreviated as SC) uses the same five ABI spectral channels described in Section 2.1. It derives various indices to detect and quantify precipitation using discriminant and multiple linear regression techniques, respectively. SC uses environmental predictors of relative humidity to adjust for reduced surface-level precipitation due to subcloud evaporation. For details of the algorithm, the reader is referred to Kuligowski et al. (2016).

2.3. Random Forest precipitation retrieval

This study implements the RF technique (Breiman 2001) with the open source scikit-learn (Pedregosa et al. 2011) package. RF builds on the concept of decision trees where a single tree is considered to be brittle and sensitive to small changes in the input parameter space (McGovern et al., 2019). To reduce overfitting and get unbiased estimates, RF uses ensembles of trees by introducing the "Randomness" in two forms; i.e., by training each tree in the ensemble (1) on a random subset of samples called "Bagging" (Breiman, 1996) and (2) on a subset of the predictor space. RF is found to be very effective at building non-linear relationships between predictors and the predictand. It has been used for different meteorological applications such as for hail forecasting probability (Gagne et al., 2017), precipitation type forecasting from ground radars (Elmore and Grams, 2016), and satellite

precipitation detection and quantification (Kuhnlein et al., 2014; Lazri and Ameur, 2018; Ouallouche et al., 2018; Upadhyaya et al., 2021a, b). In the present study, the RF model is fed with the predictors listed in Table 1, and it is trained against the quality-controlled GV-MRMS precipitation reference across the CONUS. The most commonly tuned RF parameters are (1) the number of bootstrap samples (represented as n) to develop n number of trees and (2) the number of randomly selected features (represented as m) for each RF tree. In this study, 500 trees with m = √(no. of features) are used (see Section 3.1 for a discussion of features). A sensitivity analysis was performed to fine-tune these parameters (not shown) but did not indicate significant impact on the performance accuracy. Thus, for all experiments, these two parameters were kept as indicated above.

Regarding dataset preparation, the entire summer 2018 (June to September) dataset is broken down into 3 equal datasets, each containing 10 consecutive days from each of the four months). The first two datasets are used for precipitation type classification and regression training, respectively, and the remaining dataset is used as independent testing data. The dataset is separated this way for each month in order to provide representative samples from the entire summer season and ensure that events in the training and testing datasets have good spatial and temporal representations. The independence between the training and testing datasets caused by some of the same storms ending up in both datasets was checked, and its impact on the results was found to be negligible. Since the training data is highly imbalanced across precipitation types, it is balanced by using random sampling (for details refer to Upadhyaya et al., 2021a). Note that the validation data distribution remains unchanged to evaluate the retrievals on the natural distribution of precipitation rates and types. To address the issues with aggregated evaluation scores caused by skewed distribution of data, a comprehensive validation analysis is performed by conditioning on precipitation types, precipitation rates, and predictors to assess their impact (details in Section 2.4).

2.4. Evaluation

The RF model is evaluated for different scenarios listed below with the independent validation dataset. The quantification scores listed in Table A1 are used to compare different models, along with the relative bias conditioned by the reference precipitation rates and the predictors. In a second stage of performance evaluation, the RF detection and quantification capabilities are compared with SC for the exact same time period and area. Note that since the spatial resolution of SC is the same as RF retrieval, a direct comparison is performed without any spatial aggregation. The inter-comparison is performed across four different conditions:

1. Precipitation types from GV-MRMS;
2. ABI Zenith angle bins (10° bins) as shown in Figure 1a;
3. Transition zones from ocean to coast to land, where "coast" is defined as the region within 0.1° inland from shoreline (NOAA Medium Resolution Shoreline; Graham et al., 2003). Note that the shoreline of the Great Lakes is also considered as a coastal region;
4. Flat and complex terrain (Figure 1b). Following Daly et al. (2002), elevation gradients are calculated between the four adjacent cells of the digital elevation model (ASTER GDEM V3, 2019) by using a moving window technique. Elevation gradients less than a threshold of 0.015m/m are considered flat. This resulting terrain mask is at the resolution of DEM (1 arc second ~ 30 m) and therefore resampled to ABI grid using a nearest neighbour approach.

The last two analyses further decompose the evaluation across different U.S. climate regions defined by Karl and Koss (1984; Figure 1c). The Ocean region is also further divided into: East, West, South, and Great Lakes as shown in Figure 1c.

Bulk error metrics such as correlation or bias (Table A1) depict averaged performances because they are computed over samples that gather a variety of precipitation situations (types, rates) and sampling conditions (Ze) for which the retrievals are likely to behave differently. Conditional biases are also computed to provide an in-depth assessment of the precipitation retrievals.

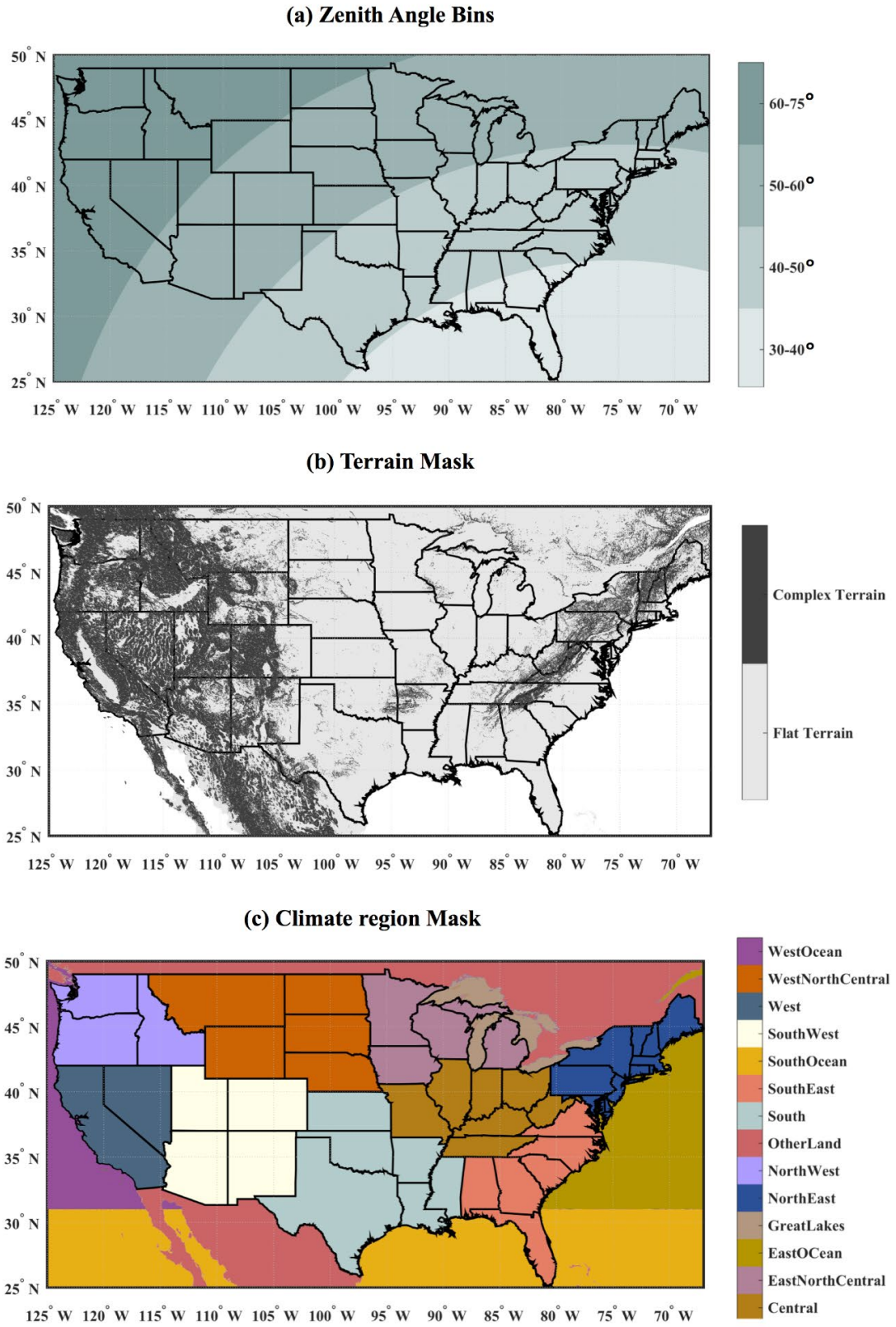


Figure 1. Depiction of regions for validation: (a) ABI zenith angle bins (b) flat and complex terrain and (c) climate regions

3. Results and Discussions

3.1. Predictor Importance and Selection

The most important predictors should be identified in order to make the model as computationally efficient as possible for real-time applications. Note that RF estimation is still computationally efficient even with a large number of predictors. However, the latency is impacted by the pre-processing of predictors; for example, texture predictors are computationally intensive, especially for large images (e.g., Tahir et al. 2004). The latency exponentially increases with the number of predictors; therefore, a trade-off needs to be reached between latency and accuracy by selecting only highly relevant predictors. A predictor selection exercise is therefore carried out to select the fewest possible predictors without significantly compromising the model accuracy. It is done in two stages. Initially, a preliminary RF model is trained with all 260 predictors listed in Table 1. Next, predictors are selected and ranked based on their importance according to the decrease in impurity score that is inherently computed by the RF (Louppe et al., 2013). Following a backward predictor elimination procedure, the least important predictors are sequentially eliminated to train the RF models with the remaining most important predictors (e.g., Genuer et al. 2010). Note that RF predictor importance has bias regarding correlated predictors (Strobl et al., 2008). Therefore, impurity score is only used as a relative indicator and the predictors are removed based on their impact on validation scores. In other words, the choice of the next predictor to test for removal is based on the impurity score, but the actual choice of whether or not to remove it is based on validation scores. This approach seems to reduce the risk of removing a predictor that should not be removed.

Figure 2 provides accuracy scores obtained with the RF models trained with various sets of predictors and validated against an independent validation dataset. Sets of predictors are defined by their threshold on cumulative importance and associated number of predictors given in the x-axis. Accuracy scores are the correlation coefficient (Table A1: Eq.1) and RMSE (Table A1: Eq. 2), shown on the primary and secondary y-axes, respectively, in Fig.2a. Because the proportion of precipitation types is uneven, the dominant precipitation types (e.g., warm stratiform) can influence the selection of predictors at the expense of under-represented precipitation types. Therefore, the impact of removing predictors was also evaluated for each precipitation type in Figure 2b, which provides a normalised RMSE (nRMSE) (Table A1: Eq.3) score for each precipitation type. Note that scores such as CC and RMSE can naturally vary with precipitation types (e.g., Convective type is associated with more quantitative variability than the Stratiform type

and therefore is associated with lower CC) and, and they are affected by heteroscedasticity within each type. Therefore, nRMSE is shown in Fig.2b to enable comparison among precipitation types.

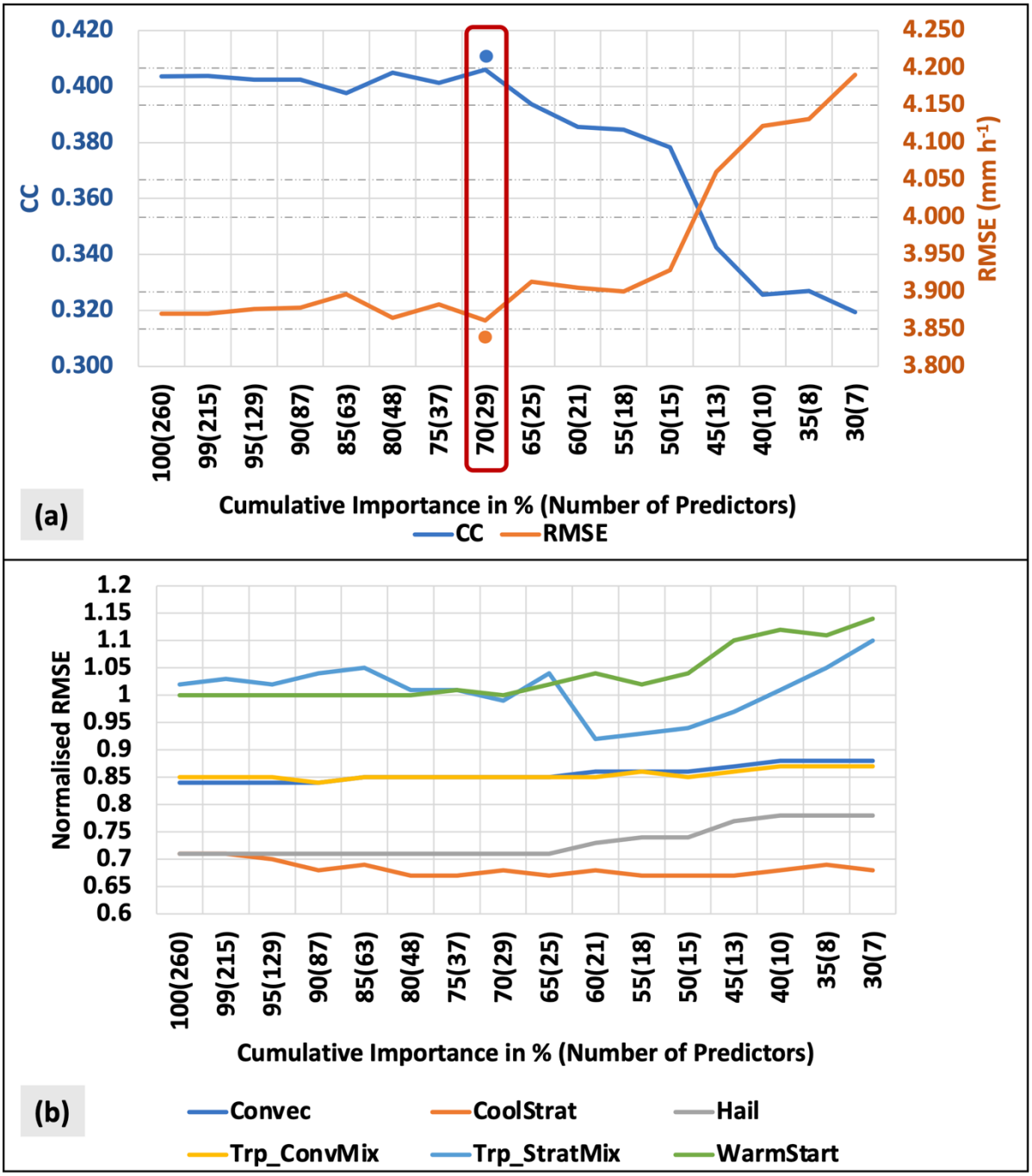


Figure 2. (a) Evaluation scores correlation coefficient (blue) and RMSE (orange) for the Stage-I predictor selection experiment, where the x-axis indicates the cumulative importance and number of the remaining predictors. The red box shows where the number of predictors

is further reduced by removing eight predictors which give similar information (as opposed to removing them based on cumulative importance) and the accuracy scores after removal are indicated by dots. (b) Same as (a) but Normalised RMSE is computed separately for each precipitation type.

From Fig. 2a, it can be observed that the accuracy scores remain almost the same until the combined importance of the selected predictors is reduced to 70% of the total importance (i.e., when number of predictors is reduced from 260 to 29). A consistent result is revealed in Fig. 2b, where nRMSE remains almost the same for each precipitation type, which confirms that the removed predictors are not significant for any precipitation type. Removing additional predictors as ranked by cumulative importance markedly degrades the accuracy scores, specifically for the Warm Stratiform precipitation type followed by Hail type. An exception is with Tropical Stratiform Mix where nRMSE improves by reducing predictors from 25 to 21 while it degrades for most other precipitation types. Therefore, 29 predictors are selected from this experiment. Within this set, removing highly correlated predictors (CC > 0.95) to keep only the predictors with the highest importance further reduces the number of selected predictors to 21. The resulting accuracy scores are slightly improved as shown in Fig. 2a (dots), in contrast to the degraded skill obtained when removing 8 additional predictors based on cumulative importance (Fig. 2a).

The technique used above is at a risk of keeping a few insignificant predictors in the model. Therefore, in the final stage of selection, each of the predictors is removed from the model one at a time (independently of its RF importance score) and the model is retrained. At the end of this cycle, the predictor whose removal produces the largest improvement in performance is removed. This process is repeated and an additional predictor is removed each time until it is observed that removing any more predictors reduces the validation accuracy, indicating that all the remaining predictors in the model are important. Figure 3 shows the final selected predictors arranged in decreasing order of their impact on validation score (CC and RMSE) when removed from the model. The predictor impact on validation score can be used as a proxy indicator of their importance and can be ranked based on the amount of degradation on performance metrics. Note that this final stage involves re-training the model several times and is computationally expensive; therefore, an initial selection based on RF importance is performed.

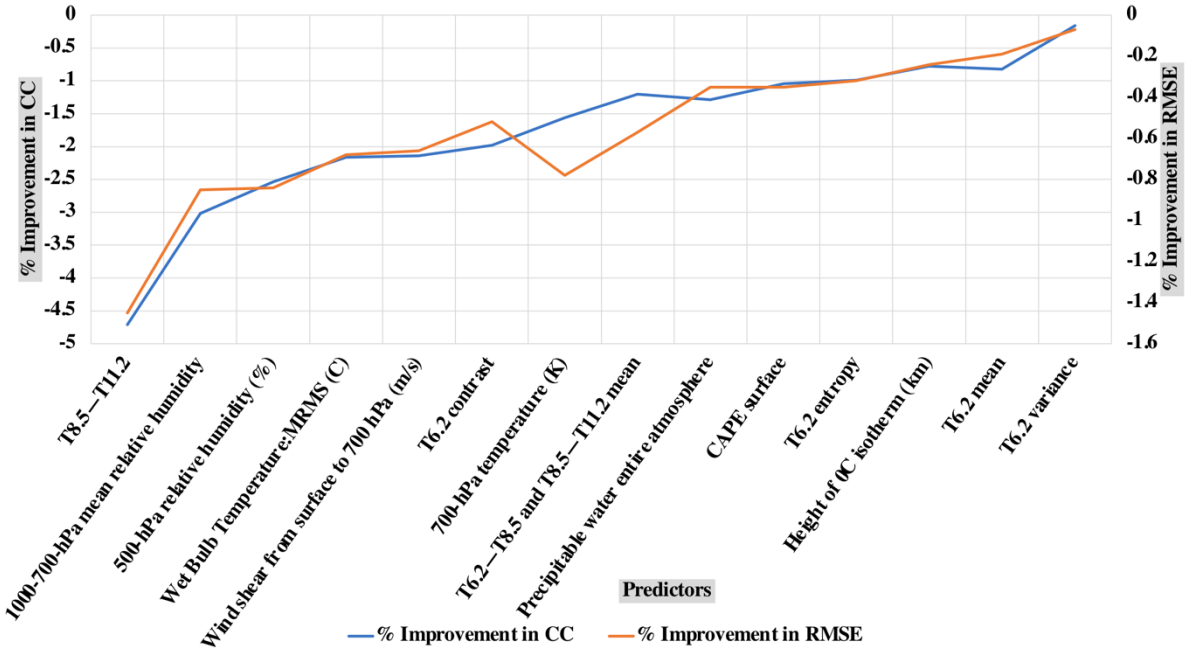


Figure 3. Final list of selected predictors at Stage-II. Blue (Orange) line represents % improvement correlation (RMSE) on the left (right) vertical axis. The predictors are arranged in the decreasing order of their impact on validation scores.

Figure 3 provides the most important predictors included in the identified parsimonious model, arranged in decreasing order of importance. From Fig. 3, the T8.5 – T11.2 satellite predictor is the most important predictor as it degrades considerably both CC and RMSE. The next predictors are environmental predictors such as relative humidity, temperature, and windshear. Out of fourteen selected predictors, six are satellite-based predictors and eight are environmental predictors, which suggest complementary information content brought by satellite observations and model parameters. The contribution of environmental predictors reflects the indirect relationship between the information provided by GEO observations and surface precipitation. Among satellite predictors, other than the BTD predictor T8.5 – T11.2, the texture features of BT at 6.2µm (T6.2) are confirmed to have the highest importance along with texture-based index derived from D-BTD: T6.2 – T8.5 and T8.5 – T11.2 mean. Note that out of six selected satellite predictors, five predictors are texture-based satellite predictors (T6.2 – T8.5 and T8.5 – T11.2 mean, T6.2 mean, T6.2 variance, T6.2 entropy, T6.2 contrast). This highlights the importance of these newly introduced texture-based satellite predictors for precipitation quantification. It is consistent with the finding that texture-based predictors lead

to the greatest improvement in precipitation type classification, especially for the detection of the WarmStratiform type (Upadhyaya et al., 2021b).

3.2. Benchmarking precipitation quantification: from legacy GEO sensors to the new generation GEO observations

Before evaluating the RF model, it can be helpful to take a broader look at the total value of the data from multiple IR bands. As noted in Section 2.1, most of the operational precipitation retrieval algorithms that operate at the global scale utilize only the 11.2 μm channel. Upadhyaya et al. (2021b) shows the significance of including additional channels for precipitation detection and classification. An experiment is conducted to understand the impact and highlight the need to incorporate multiple spectral channels from GEO sensors to improve the quantification of precipitation. Separate RF precipitation quantification models are trained with sets of BT predictors where BTs are successively introduced (Table 2). Following Upadhyaya et al. (2021b), the order of introduction is selected based on the historic availability of channels in GEO sensors.

Table 2. Quantification scores for experiments with different channels and their derived predictors.

Predictors with channel	*CC*	*RMSE (mm h^{-1})*	*Rbias (%)*
T11.2	0.20	4.01	-4.78
T11.2 + T6.2	0.28	3.94	-2.18
T11.2 + T6.2 + T12.3	0.28	3.95	-1.80
T11.2 + T6.2 + T12.3 + T7.3	0.30	3.91	-1.80
T11.2 + T6.2 + T12.3 + T7.3 + T8.5	0.31	3.90	-1.69

Table 2 reports the validation scores. It can be observed that the highest accuracy scores are obtained when all BTs from five channels are used. In particular, adding the WV channel at 6.2μm to the legacy 11.2μm channel considerably improves scores compared to other channel additions. For example, CC increases from 0.20 to 0.28 and Rbias reduces from -4.78%

to -2.18%. Similar improvements are also observed for precipitation type classification (Upadhyaya et al., 2021b), with about a 5% accuracy gain for the No-Precipitation and Convective types, and more than 10% gain for all other precipitation types.

The introduction of T12.3 observations in addition to T11.2 and T6.2 does not show improvement in most scores except a small reduction in overall bias. This is consistent with the results in Section 2.1, where the highest contributing satellite predictors do not include T12.3 channel observations. Adding T12.3 observations for detection and classification of precipitation shows modest improvements in the range of 2-4% gain for most types (Upadhyaya et al., 2021b), suggesting the significance of T12.3 observations for classification rather than for quantification.

Adding more channels (T7.3 and T8.5) contributes to an overall improvement in the accuracy scores. Overall, using the 5 multispectral channel BTs improves CC from 0.2 to 0.31, RMSE from 4.01 mm h^{-1} to 3.90 mm h^{-1}, and bias ratio from -4.7% to -1.6%. Note that their impact further improves when using BTDs, D-BTDs and textures (see Section 3.3). Therefore, it is recommended to incorporate at least one additional channel, T6.2, with the legacy channel T11.2 to global precipitation retrievals using GEO sensors, and to revisit climate data records with the two channels.

3.3. Relative contribution of satellite, environmental predictors, and precipitation types

Different experimental set-ups that are listed in Table 3 were conducted to address the following key points:

1. understand the relative impact of combining satellite and environmental predictors.
2. understand the impact of precipitation type probabilities that allow the design of a single quantification model, and compare this design with the common approach that uses separate quantification models trained per precipitation type.

To address point 1, E1 is a model developed with environmental predictors only (19 predictors), E2 is developed with satellite predictors only (241 predictors), and E3 involves both satellite and environmental predictors. To answer point 2, E4 builds different models for each precipitation type deterministically identified by the RF based classification model (Upadhyaya et al., 2021a), while E5 is a single model built with satellite, environmental, and precipitation type probabilities (Upadhyaya et al., 2021b). Other than experiments E1 and E2, all experiments (E3, E4 and E5) use the parsimonious model predictors identified in Section 3.1 (i.e., the RF model with 14 predictors listed in Figure 3).

Table 3. Experimental setup details.

Experiment	*Model Type*	*Predictors*
E1	Single	Environmental Predictors
E2	Single	Satellite Predictors
E3*	Single	Environmental + Satellite Predictors
E4*	Multiple	Environmental + Satellite Predictors
E5*	Single	Environmental + Satellite Predictors + Classification Probabilities

* E3, E4 and E5 use only the predictors included in the subset of 14 parsimonious model predictors

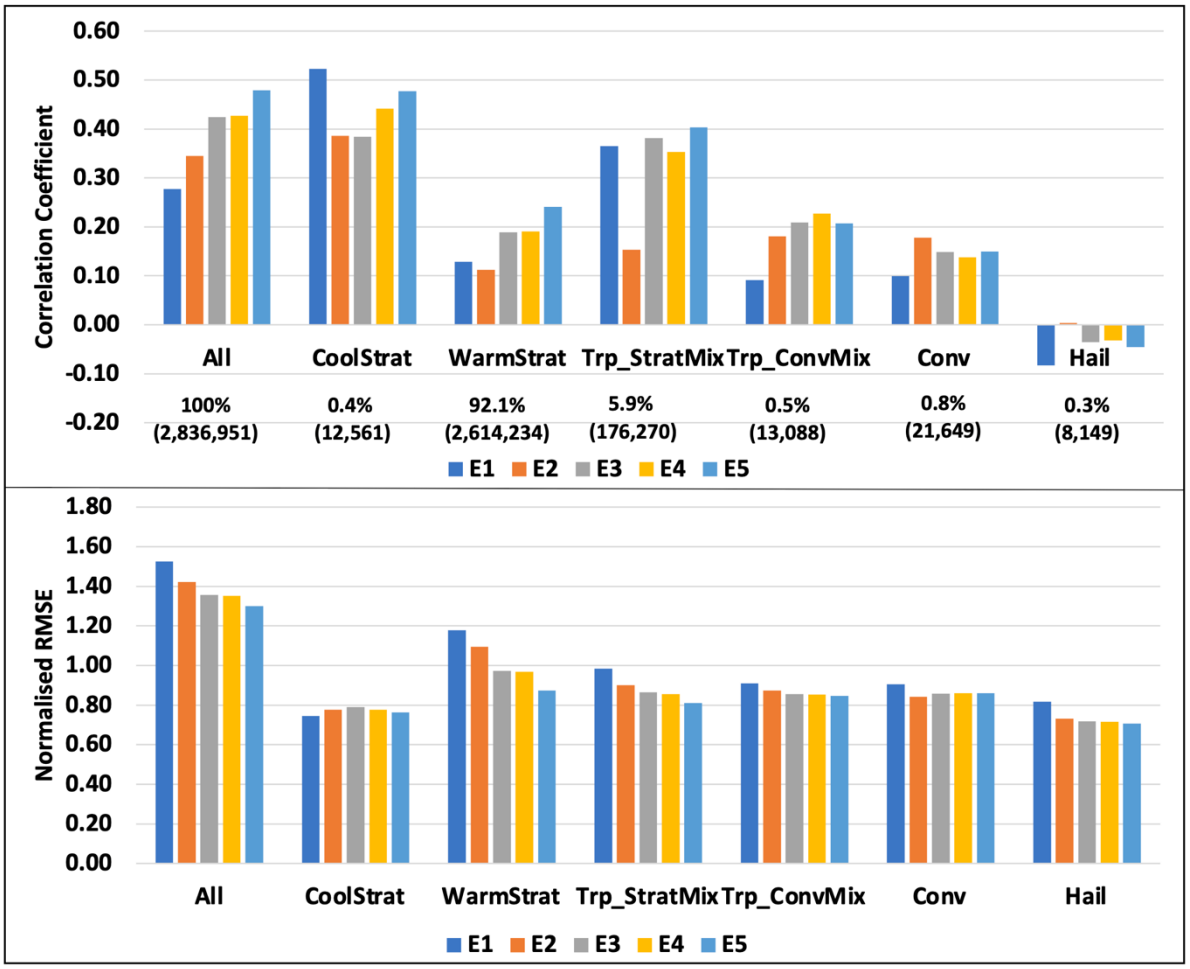


Figure 4. Overall scores and conditioned scores across GV-MRMS precipitation types. (a) Correlation coefficient and (b) normalised RMSE for a set of experiments in Table 3. The

numbers in the bottom of the top panel represent the proportion (sample size) in each precipitation type.

Figure 4 provides the overall evaluation scores and conditioned scores across GV-MRMS precipitation types. From Figure 4, comparing scores of models E1 and E2 shows that the model built with satellite-only predictors (E2) has higher correlation coefficient (0.34) and lower nRMSE (1.42) than the model built with environmental variables only (CC=0.28 and nRMSE=1.52). Combining satellite and environmental predictors further improves scores (E3) with CC=0.42 and nRMSE=1.36, which confirms that the two sets of predictors bring complementary information to quantify precipitation. Focusing on convective precipitation types (Convective and Tropical Convective/Mix types), CC values are generally lower than overall (< 0.2), as expected, because convective precipitation generally displays higher variability. This variability seems better captured with satellite observations than with environmental parameters (E2 displays higher correlation and lower RMSE than E1). The reverse is observed with stratiform types (Warm Stratiform, Cool Stratiform and Tropical Stratiform/Mix). In stratiform precipitation, the influence of processes that significantly affect precipitation fluxes in the vertical and at the surface may be better captured through environmental conditions than with the cloud-top information provided by the ABI. Convective precipitation that is easier to diagnose from Tbs and is more vertically homogeneous may be more appropriate for using cloud-top information to infer surface precipitation. Combining satellite and environmental predictors generates the highest improvement with the Warm Stratiform type with some degradation of Convective types. A possible explanation is that the combined model gives higher importance to environmental predictors so the E3 predictor set is closer to E1 than to E2. Note the high proportion of Warm Stratiform type (92.1%) drives the overall accuracy scores.

Overall, models that include precipitation type predictors (E4 and E5) outperform models that do not. For example, the single model with precipitation type probabilities (E5) displays the best scores with CC=0.48 and nRMSE=1.30 and outperforms other models for most precipitation types. Interestingly, it performs better than a combination of models designed for each precipitation type (E4; CC=0.43; nRMSE=1.35). These results highlight the advantage of the proposed precipitation type probabilities that constrain the precipitation retrievals within a single model against the commonly used two-step approach that classifies different precipitation types first before applying quantification models for each type (e.g., Kuligowski et al., 2016; Hirose et al., 2019; Ouallouche et al., 2018). A single parsimonious

model applied on all precipitation types also improves retrieval consistency and latency against using a model for each type.

Figure 5 displays the overall and conditional bias values for different classes of precipitation rates and for all models. The overall bias scores are less than 10%, which illustrates the overall good performance of the RF retrievals. Overall bias follows a trend similar to the CC and nRMSE scores reported in Fig. 4. For example, the smallest bias is displayed by the E5 model, which includes precipitation type probabilities as additional predictors. These bulk metrics are complemented with a conditional bias for an in-depth assessment. The conditional biases show significantly higher values than the overall bias depending on precipitation magnitude. All models overestimate lower rain rates and underestimate high-intensity precipitation. This behavior is expected and consistent with other satellite precipitation products (e.g., Upadhyaya et al., 2018; 2020; Kirstetter et al., 2013) as well as with any method that targets overall performance as a driving metric (Ciach et al., 2000). Combining both satellite and environmental predictors (E3) improves bias at very low rain rates (< 0.25 mm h^{-1}) compared to E1 and E2. Furthermore, including precipitation type predictors improves the conditional bias (e.g., E4 and E5 compared to E3) across low rain rates GV-MRMS precipitation bins. The best conditional bias is reported when including precipitation probabilities (E5).

	Overall	0.01-0.25	0.25-0.5	5.0-7.5	7.5-10.0	10.0-25.0	25.0-50.0	50.0-100.0
E1	6	1396	43	-35	-44	-54	-78	-82
E2	9	1491	50	-40	-55	-63	-72	-80
E3	2	1305	36	-36	-45	-54	-70	-77
E4	1	1279	35	-37	-47	-55	-71	-77
E5	0	1209	32	-37	-46	-55	-70	-76
Sample Size	2836951	22543	2498037	209117	51644	39767	21434	3704

Figure 5. Relative bias (in %) conditioned for GV-MRMS precipitation rate classes for the models listed in Table 3. Red (blue) bars represent underestimation (overestimation). The size of each bar is relative to the maximum relative bias in each GV-MRMS precipitation rate bin. The last row gives the sample size in each bin.

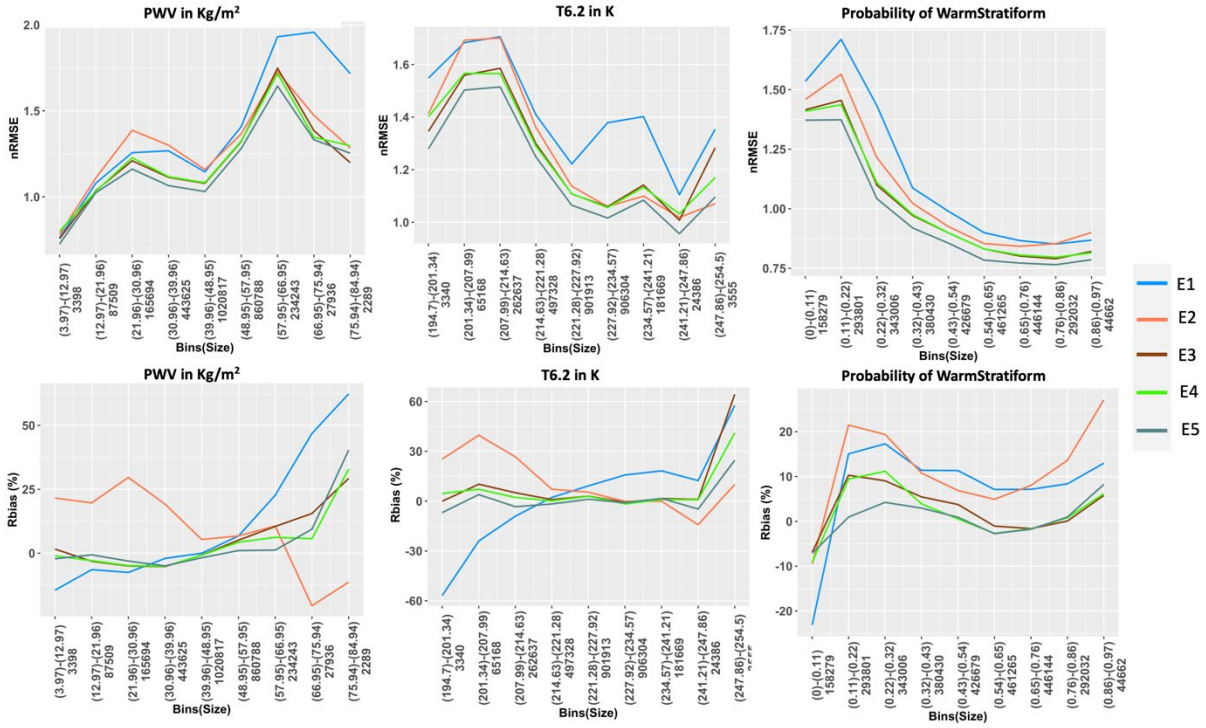


Figure 6. nRMSE and relative bias as a function of the values of selected predictors from environmental, satellite, and precipitation type probabilities for the models listed in Table 3. The x-axis provides the bin-width and the sample size in each bin.

As observed from Fig. 5, bulk metrics average out conditional dependencies. For example, the E5 model shows an overall bias of ~ 0% whereas the conditional analysis displays significantly higher overestimation of low rain rates and underestimation of higher rain rates. To get more insight into how the skill of the model depends on the values of specific predictors, Fig. 6 shows how nRMSE (absolute error magnitude) and Rbias (systematic error) are conditioned on the values of selected predictors. For purposes of illustration, one predictor was selected to represent each of the categories; i.e. Vertically integrated precipitable water (PWV; Environmental Predictor), T6.2 (Satellite Predictor), and Probability of Warm Stratiform (P_{WS}: Classification probabilities). It can be observed that the scores display large conditional dependencies. Regarding nRMSE (upper panels of Fig. 6), the largest dependency is seen with the environmental variable PWV, with nRMSE values in the range [0.75-2] compared to T6.2 (range [1-1.7]) and P_{WS} (range [0.75-1.75]). It can be related to the importance of PWV as a predictor (Section 3.1; Fig. 3). These retrievals and to some extent their biases are expected to be driven more by PWV than by other predictors. For all models, nRMSE shows a positive dependency with PWV, indicating higher errors in environments characterized by high precipitable water. Opposite trends are noted with T6.2 and P_{WS}, i.e. larger errors are noted in

situations associated with cold BTs and low probabilities of warm stratiform precipitation. Collectively, it indicates that larger errors are related to convective situations with higher rain rates and variability compared to lower PWV and more stratiform precipitation (also observed in Figs. 4 and 5).

Among models, the intercomparison of nRMSE for E1 (environmental predictors only) and E2 (satellite predictors only) shows larger errors with E2 at low PWV (<50 kg/m2) and with E1 at higher PWV. It clearly shows the complementary advantages of environmental predictors at lower PWV and of satellite predictors when PWV is high in probably convective situations. It is demonstrated with E3 (combining both categories of predictors), where nRMSE is consistently improved across most of PWV, T6.2 and P_{WS} values. Introducing precipitation types, either as probabilities (E5) or as two-stage modelling (E4) improves nRMSE across most cases, with higher improvements with model E5.

Rbias (lower panels in Figure. 6) shows the largest range with predictor PWV (range [-25; 80%]), followed by T6.2 (range [-15; 60%] except E1) and P_{WS} (range [-20; 25%]). PWV is confirmed to be the strongest driver of errors. Again, overestimation and underestimation driven by various conditions average into overall biases of less than 10% (Figure. 5). For all models except E2, there is a general tendency to underestimate precipitation rates (up to -20%) associated with low PWV values (<30 kg/m2) and overestimate (up to 40%) at high PWV values (> 66 kg/m2). This overestimation can be attributed to challenges in detecting tropical precipitation types that are often associated with high PWV values (> 66 kg/m2); it will be investigated in a future study. Note that high PWV values contribute to only 1% of the total validation sample. This trend is the opposite for the E2 model, most likely because it does not include environmental predictors. Rbias conditioned by T6.2 highlights a general trend of overestimation at high BTs whereas for intermediate to low BTs, Rbias is limited and shows a slight decreasing trend (Except for E1 and E2). As expected, the E1 model (without satellite predictors) shows a larger conditional and increasing bias with T6.2 that confirms the significance of satellite observations. Rbias conditioned with P_{ws} displays large underestimation (up to -20% for E1) for low probabilities ($P_{ws} < 0.1$), again indicating challenges in quantifying convective precipitation, whereas overestimation is noted (Rbias up to 25% for E2) for higher P_{ws}. Like nRMSE, Rbias is larger for extreme values for all three predictors. It highlights that extreme retrieval conditions pose quantification challenges and a room for improvement exists.

Collectively, larger Rbias that is noticed with predictors not included in models confirm how models are sensitive to parameters that have not been accounted for (Stephens and

Kummerow, 2007). Overall, the E5 model displays the lowest conditional bias which remains within the ±15 % range in most situations. It supports the benefits of using precipitation type probabilities in a single model to constrain the precipitation retrievals.

3.4. RF and SCaMPR evaluation across different climate regions, land/coastal/ocean areas and complex/flat surfaces:

In this section, the RF retrievals are intercompared with SCaMPR with an objective of performance evaluation across different climate regions which is further sub-categorized in terms of coastal areas and Complex/Flat surfaces.

3.4.1. Precipitation Detection

Table 4. Detection scores for RF and SC

	Threshold	*POD*	*POND*	*FAR*	*HSS*	*CSI*	*VHI*
RF	0.1 mm h^{-1}	0.98	0.94	0.02	0.90	0.96	0.98
SC		0.38	0.99	0.01	0.19	0.38	0.45

Table 4 shows the detection scores (Table A1) for RF and SC computed with a precipitation rate threshold of 0.1 mm h^{-1}. SC shows a detection score of 38% by occurrence (POD) and 45% by volume (VHI), indicating that it misses 55% of the precipitation volume across the period of comparison (summer 2018) and across the U.S. By comparison, RF detects most of the precipitation occurrence (POD = 98%) and volume (VHI = 98%). Both algorithms correctly detect high fractions of non-precipitating events (POND > 0.94). RF detects more non-precipitation pixels as precipitating since POND is 5% lower than SC, yet both SC and RF show similar and very low proportion of falsely detected precipitation (FAR = 0.02 for RF and 0.01 for SC). SC has a comparatively lower POD than RF, hence the proportion of false alarms is the same. As a result, the RF HSS is high (0.90), showing excellent ability to separate rain from no rain, and much better than SC (HSS = 0.19). Note that this analysis has been repeated with higher thresholds of precipitation rate (not shown) and similar trends are observed between RF and SC.

Table 6 breaks down detection scores by precipitation types. Note that only POD is given in Table 6 because SC does not define corresponding precipitation types; in other words, MRMS provides a precipitation type for computing POD but SC does not provide one for false

alarms. For a detailed analysis on the performance of the RF model across different precipitation types, please refer to Upadhyaya et al. (2021a, b). In Table 6 RF displays excellent performance in detecting precipitation occurrence across all precipitation types (POD > 0.98). RF detection scores are also consistently higher than SC. SC detection is higher than 75% for all three convective types, followed by Tropical Stratiform/Mix type (69%), Warm Stratiform (36%) and Cool Stratiform (8%), which clearly indicates that SCaMPR has better ability to detect convective types than stratiform types (as expected from Section 3.3). RF detection of stratiform precipitation occurrence is noteworthy, since it has been a reported challenge with GEO precipitation retrievals (Upadhyaya et al., 2020). This performance is attributed to the use of texture-based satellite indices and environmental predictors (Upadhyaya et al., 2021b).

Table 6. Probability of detection (POD) scores for RF and SC for different GV-MRMS precipitation types.

Precipitation Type	*POD: RF*	*POD: SC*
Cool_Strat	0.99	0.08
WarmStart	0.98	0.36
Trp_StratMix	1.00	0.69
Trp_ConvMix	1.00	0.86
Convec	0.99	0.75
Hail	0.99	0.86

Detection performances display a regional dependence that is illustrated in Fig. 7 and 8, which show the POD and HSS, respectively, for both RF and SC retrievals across different regions. RF shows consistently higher POD (0.69-0.99) and HSS (0.78-0.93) compared to SC POD (0-0.52) and HSS (0-0.48) across all regions. Overall, lower POD detection scores are observed in the western US for both retrievals (around 5-12% reduction in POD for RF and 8-20% for SC compared to the eastern US), where the ABI zenith angle and the associated footprint are larger. Note that the degradation of performance is greater with SC while RF displays more consistent detection performance across regions. This can be attributed to the inclusion of the zenith angle as an additional predictor in the RF model. HSS is slightly lower

across the eastern CONUS for both RF and SC than over the central and western CONUS, which may be due to higher FAR in the east.

As expected, both RF and SC detection is generally lower over coastal regions than over land across all regions except in the southern and southeastern CONUS. The POD across coastal regions falls to a range of 1-21% for RF and 1-29% for RF. These performance differences are larger in the western and northwestern regions of the CONUS, where the degraded ABI spatial resolution may combine with gradients in these transition zones to make precipitation detection more challenging. With regards to oceanic regions, both RF and SC show higher POD over the EastOcean region (0.98 for RF and 0.36 for SC) than over the WestOcean region (0.94 for RF and 0.24 for SC), again, presumably because of the degraded ABI resolution. This trend is reversed with HSS, possibly due to higher false alarms as indicated earlier. The HSS for RF (SC) improves from 0.86 (0.13) to 0.92 (0.23) from the EastOcean region to the WestOcean region. Note also that the sample size across the WestOcean region is considerably lower compared to the EastOcean region.

From Fig. 7, it can be observed that across the eastern regions of the CONUS, RF POD is up to 1% higher in complex terrain than flat terrain. On the contrary, in the western CONUS, POD in complex terrain drops by 2% with respect to flat terrain. This trend is not observed with SC: in all regions except NorthEast and EastCentralNorth, SC POD is lower in complex terrain by 2-18%. HSS shows lower values in complex terrain than flat regions in all regions except NorthEast, EastCentralNorth, WestNorthCentral for both SC and RF. This analysis suggests that the degradation of precipitation delineation across the western CONUS is due both to the degrading ABI resolution as well as complex terrain. This impact is higher on SC than on RF.

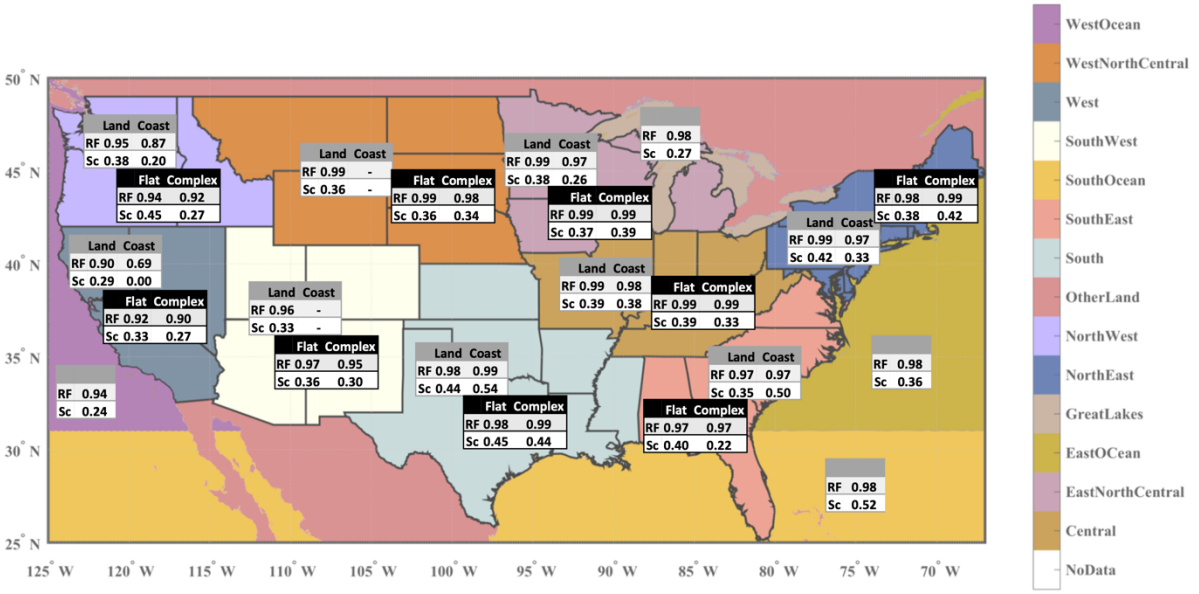


Figure 7. Probability of detection for RF and SC across different climate regions and broken down into land vs. coast areas and flat vs. complex terrain.

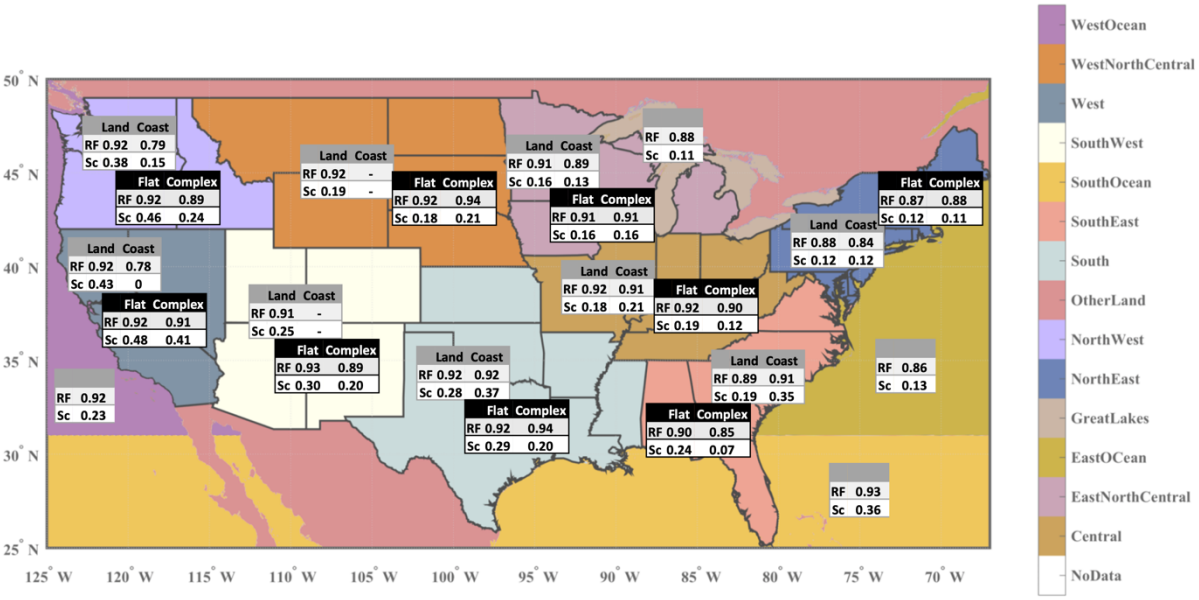


Figure 8. HSS of RF and SC across different climate regions and broken down per Land/Coast areas and Flat/Complex terrain.

To get further insight into the impact of the ABI resolution, the detection scores are broken down per ABI satellite zenith angle bins in Table 7.

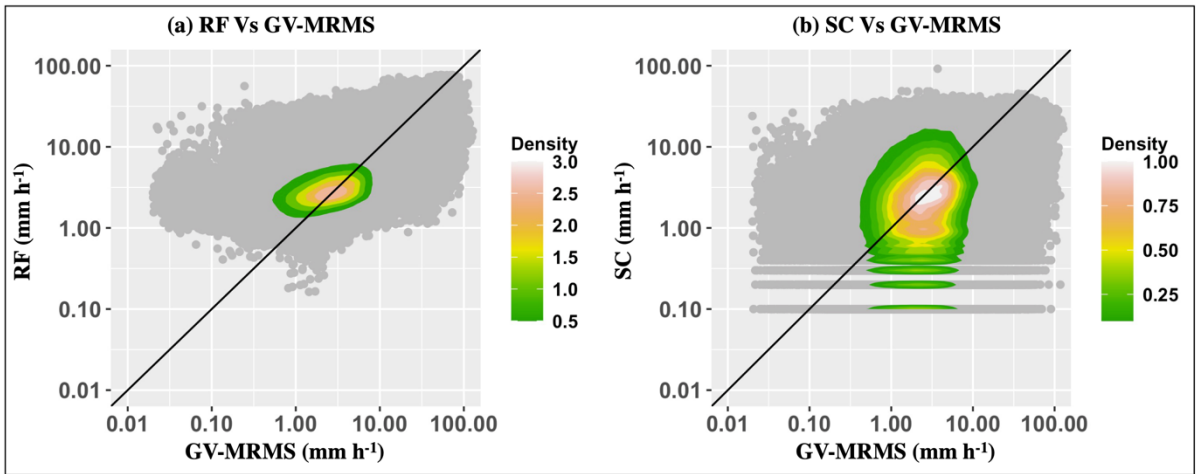
Table 7. Probability of detection (POD) and Heidke Skill Score (HSS) for the developed model (RF) and SCaMPR (SC) across different ABI zenith angle bins

ABI Zenith Angle Bin	*POD: RF*	*POD: SC*	*HSS:RF*	*HSS:SC*
30.0 – 40.0	0.98	0.43	0.92	0.30
40.0 – 50.0	0.98	0.40	0.90	0.19
50.0 – 60.0	0.97	0.35	0.89	0.16
60.0 – 75.0	0.96	0.27	0.93	0.21

From Table 7, it can be observed that as the satellite footprint resolution decreases with zenith angle, the detection scores drop in terms of POD by 2% for RF and 16% for SC. HSS shows a similar trend overall. The impact of zenith angle and resolution is higher on SC than RF.

3.4.2. Quantification

Figure 9 shows the scatter plots (grey points) overlaid by density plot of RF and SC against GV-MRMS along with overall quantification scores. Overall, the RF estimates are more centered around the 1:1 line than SC estimates. It is confirmed that the RF model has better scores than SC. SC underestimates precipitation rates by ~17% overall, while RF slightly underestimates (~2%) with respect to. GV-MRMS. There is considerable difference in correlation, with RF CC = 0.47 and SC CC = 0.26, and 14% better RMSE is observed with RF (5.01 mm h^{-1}) than with SC (5.87 mm h^{-1}).



(c) Overall quantification scores for RF and SC

	CC	RMSE	nRMSE	Abias	Rbias (%)
RF	0.47	5.01	1.35	-0.07	-1.89
SC	0.26	5.87	1.59	-0.64	-17.34

Figure 9. Overall quantification scores of RF and SC. Panels a and b: Scatter plot (grey points) overlaid by density plot of RF and SC against GV-MRMS; panel c is overall quantification scores.

Table 8. Quantification scores of RF and SC for various GV-MRMS precipitation types

	CC		*RMSE (mm h^{-1})*		*nRMSE*		*Abias (mm h^{-1})*		*Rbias (%)*		*Data size*
	RF	SC	RF	SC	RF	SC	RF	SC	RF	SC	
Cool_Strat	0.39	0.44	2.03	2.37	0.73	0.86	-0.78	-0.27	-28.04	-9.87	2114
WarmStart	0.16	0.09	2.48	3.56	0.95	1.37	0.58	0.21	22.36	8.01	1217554
Trp_StratMix	0.42	0.12	5.16	5.80	0.84	0.95	-0.27	-2.21	-4.34	-36.0	130218
Trp_ConvMix	0.21	0.23	31.72	34.1	0.83	0.89	-26.66	-29.8	-69.81	-77.9	12843
Convec	0.17	0.15	17.84	18.6	0.84	0.87	-14.43	-15.1	-67.81	-70.8	18008
Hail	-0.07	0.03	26.93	30.7	0.69	0.79	-23.54	-28.1	-60.69	-72.4	7269

Table 8 breaks down quantification scores by precipitation types. In general, RF overestimates precipitation rates associated with WarmStratiform (by 22%), and it underestimates for all other types (from -4% to -69%). SC also overestimates with WarmStratiform (by 8%) only and underestimates precipitation rates for all other precipitation types (from -9% to -77%). Generally, both products show larger underestimation with convective precipitation types (< -60%) than stratiform types (-36 to 22%). Note that the

proportion of Warm Stratiform is large compared to other precipitation types, thus influencing the overall scores. Except for Cool Stratiform and Tropical Convective/Mix, RF shows higher CC than SC for all precipitation types, with maximum improvement over SC for Tropical Stratiform/Mix precipitation type (RF=0.42 and SC=0.12). As expected, Hail shows the lowest CC, followed by WarmStratiform for both RF and SC. Due to smaller precipitation rate variations within each precipitation type, CC values for individual precipitation types are generally lower than the overall CC. RMSE is better for RF than SC for all precipitation types. Generally, stratiform types show higher nRMSE (0.73-1.37) than convective types (0.69-0.89) for both RF and SC.

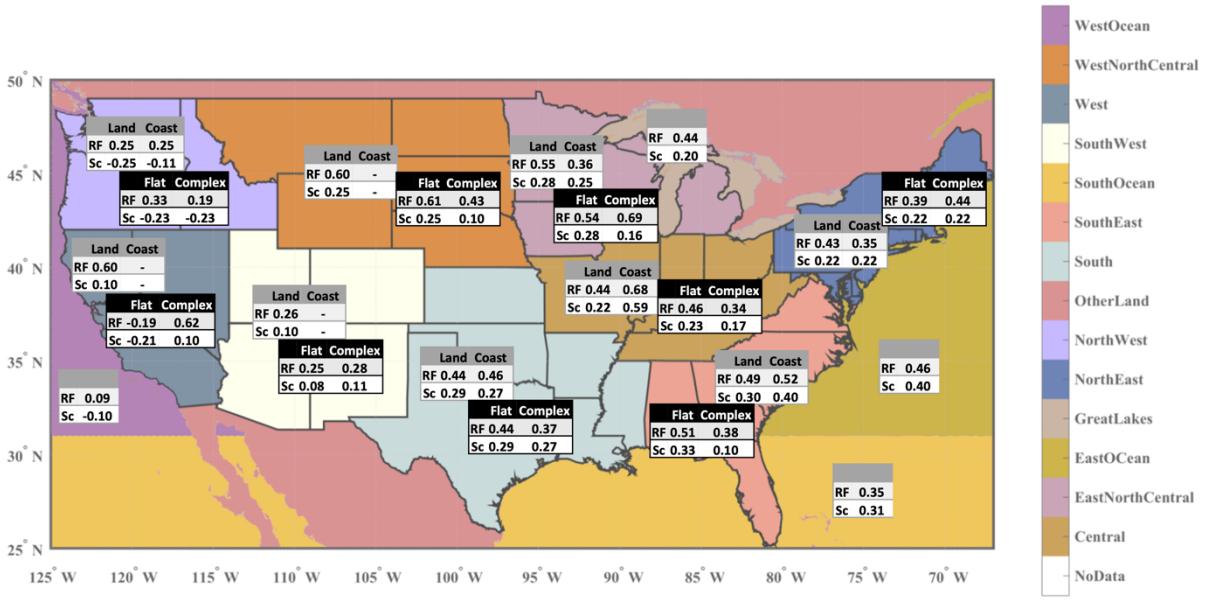


Figure 10. Correlation Coefficient of RF and SC across different climate regions which further segregated across Land/Coast and Flat/Complex terrain regions

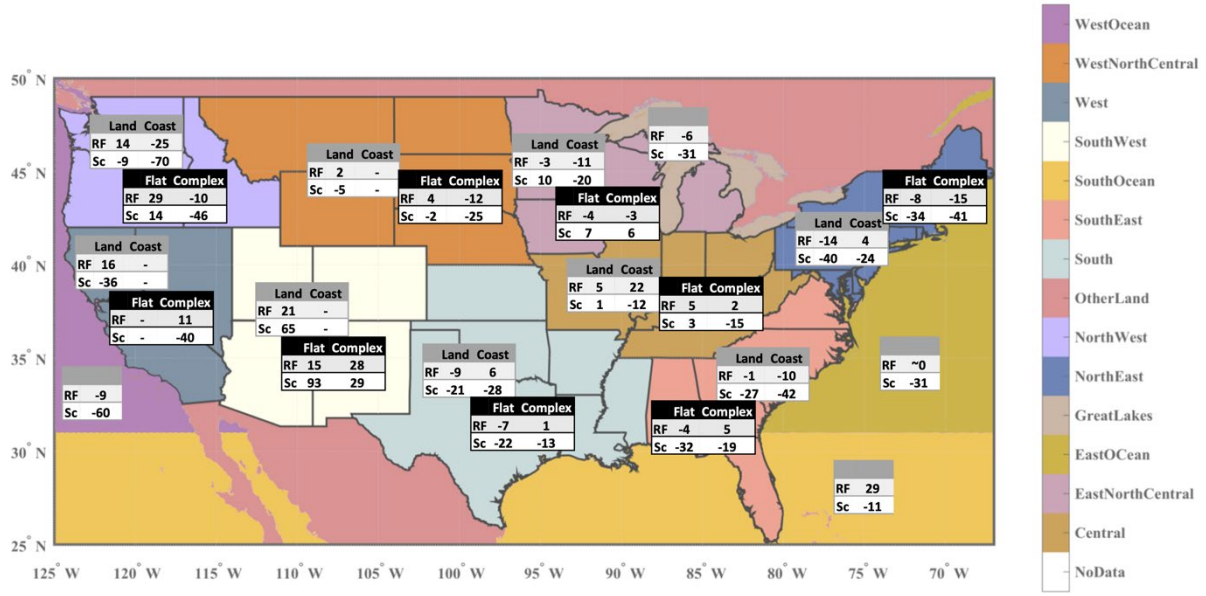


Figure 11. Rbias of RF and SC across different climate regions which further segregated across Land/Coast and Flat/Complex terrain regions

Figures 10 and 11 provide the quantification scores in terms of CC and Rbias across different climate regions and segregated by land vs. coast surfaces and complex vs. flat terrain. As with detection scores, it can be observed that the RF quantification scores are generally better than SC. The improvement with RF over SC is noticeable when comparing CC across land regions (greatest improvement from 0.16 to 0.69 in the EastNorthCentral complex terrain region) and ocean regions (greatest improvement from 0.40 to 0.46 over EastOcean). As seen earlier, SC has a general tendency for underestimation (except SouthWest, flat terrain of NorthWest, and land of EastNorthCentral) whereas no such trends are seen with RF. Overall, the Rbias of RF is within 15% (except for the West and SouthWest CONUS and SouthOcean), while the Rbias for SC varies in the range [-70; +93]% across all regions. Note that RF's Rbias is more degraded across the western CONUS than over other regions (except NorthEast land), as observed with detection scores.

Contrasting land and coastal areas, RF and SC both display degraded Rbias and varying CC over coastal regions. The bias is larger in most coastal regions (e.g., +14% Rbias for RF in NorthWest land versus -25% over coast) except in the East coastal region for RF and NorthEast for SC. These varying performances across coastal regions reflect the challenges in quantifying precipitation from space in transition zones, and the need to analyze and treat such regions separately for algorithm development.

Over ocean, it can be observed that both products (RF/SC) show lower CC (0.09/-0.10) over the WestOcean region than over other water regions (e.g., EastOcean CC=0.46/0.40; SouthOcean=0.35/0.31). Again, this can be attributed to the coarser spatial resolution of ABI that misses more precipitation variability. RF exhibits underestimation over the West Ocean region (-9%) and the Great Lakes region (-6%), close to no bias over the EastOcean region, and overestimation in the SouthOcean region (+29%). SC underestimates across all water regions ranging from -11% in the SouthOcean region to -60% in the WestOcean region.

Contrasting flat and complex terrain, CC tends to be lower over complex terrain than over flat terrain as expected, except for the SouthWest, NorthEast and EastNorthCentral regions. Rbias takes on larger values across complex terrain than flat terrain except for the Central, NorthWest and EastNorthCentral regions for RF. This contrast in Rbias values between terrain types is higher in the west CONUS (e.g., for RF in NorthWest the difference in RBias is 40%) than other regions. Note the sample size in the flat terrain region in the western

CONUS is low (138 samples); therefore, no significant conclusion can be made about the performance difference between complex and flat terrain in this region.

CC also degrades with Ze for both retrievals and the impact is larger on SC (Table 9). The general degradation of performance in the West is explained by a combination of coarser resolution and complex precipitation mechanisms in mountainous terrain.

Table 9. Quantification score CC for developed model (RF) and SCaMPR (SC) across different ABI zenith angle bins

ABI Zenith Angle Bin	*CC: RF*	*CC: SC*
30.0 – 40.0	0.48	0.33
40.0 – 50.0	0.46	0.29
50.0 – 60.0	0.46	0.21
60.0 – 75.0	0.30	0.14

4. Summary and Conclusions

The study focuses on the development and comprehensive evaluation of a Random Forest QPE using ABI observations onboard the GOES-16 satellite across the CONUS. Key innovations concern the introduction of novel predictors and the use of a high-quality and high-accuracy reference:

1. For the first time, 241 different satellite predictors are investigated for QPE that are derived from five ABI infrared channels observations. Categories of predictors are: Brightness Temperatures (BTs), Difference of BTs (DBTs), Difference of DBTs (D-BTD) and their textures which accounts for spatial information.
2. A set of 19 NWP-based predictors are used to complement the satellite observations with mid- and low-level environmental conditions.
3. The high-quality and high-accuracy gauge-radar ground reference GV-MRMS is used to develop and evaluate the QPEs.
4. A novel approach is explored to account for precipitation types and associated processes that drive surface precipitation. By using probabilities of precipitation types as additional predictors, a single model is developed that seamlessly accounts for varying relationships between satellite observations and surface precipitation. It differs

from the commonly used multi-step approach that detects and classifies various precipitation types then applies a quantification model for each type.

The benefit of these novel features is systematically evaluated in a series of experiments. Results are summarized as follows:

1. The number of model predictors can be reduced from 260 to 14 without significantly reducing the quantification accuracy.
2. The satellite predictors that contribute the most information include T8.5 – T11.2, texture features of the brightness temperature at 6.2μm (T6.2) and texture-based index derived from D-BTD: T6.2 – T8.5 and T8.5 – T11.2 mean.
3. Some of the important environmental predictors include the relative humidity, temperature, wind shear and the vertically integrated precipitable water.
4. The relative contribution of satellite and environmental predictors depends on the precipitation type. Satellite predictors contribute more towards high intensity (probably convective) precipitation with overall CC=0.34, whereas environmental predictors contribute more to low-intensity precipitation with overall CC=0.28.
5. Combining both categories of predictors improves overall performances with overall CC=0.42 and especially with the Warm Stratiform precipitation type.
6. Precipitation type predictors further constrain the retrievals and improve performance. The best improvement is obtained with precipitation type probabilities (CC = 0.48). This approach benefits operational implementation with a single model.
7. Conditional analysis showed that the RF retrieval with classification probabilities displays the least conditional bias, which remains within the ±15 % range in most cases. However, room for improvement is noticed for extreme rain rates.
8. The additional channels provided by the ABI have value for QPE. It is confirmed that the highest accuracy is obtained when all five channels are combined. In particular, adding the WV channel at 6.2μm to the 11.2μm legacy channel significantly improves the retrieval performance.

The RF retrieval was compared with the operational Self-Calibrating Multivariate Precipitation Retrieval (SCaMPR; Kuligowski 2002; Kuligowski et al., 2016) across different climate regions, surfaces (ocean, coast, land) and regions (complex/flat terrain). RF displays higher precipitation detection and quantification skills than SC. Both products display similar trends across regions.

- *Detection:* While SC's detection performance varies with precipitation types and is improved for convective precipitation, it misses significant amounts of stratiform precipitation. On the contrary, RF detection is consistent across precipitation types and displays better scores overall than SC. Both retrievals show lower detection across the western CONUS than the eastern CONUS, which is attributed to the degraded ABI spatial resolution and complex topography. This degradation in performance is greater for SC than RF. Detection along the coast (complex terrain) is generally lower than land (flat terrain) for both QPEs.
- *Quantification*: RF displays better quantification scores than SC. SC underestimates by ~17% while RF slightly underestimates (~2%) compared to GV-MRMS precipitation rates. Considerable differences in correlation are noted, with RF CC=0.47 and SC CC=0.26 overall. The improvement in RF over SC is noticeable in terms of CC across land regions, whereas over ocean RF and SC have closer CC. Quantification across coastal regions (complex terrain) is more challenging than over land regions (flat terrain). Again, the degradation of performance in the western CONUS is attributed to the degraded resolution of ABI as well as complex terrain.

Recommendations from the study are:

1. Predictors derived from the five satellite brightness temperatures (e.g., texture, inter-band differences) add information with respect to single-pixel, single-channel values and are recommended to include in addition to brightness temperatures in precipitation rate retrieval algorithms.
2. Environmental predictors from NWP, such as vertically integrated precipitable water and the relative humidity, complement the satellite brightness temperatures and are recommended as predictors in precipitation rate retrieval algorithms.
3. Constraining precipitation quantification with precipitation type is advantageous. It is recommended to incorporate precipitation type probabilities as additional predictors for seamless integration across different precipitation types and also benefit from a single model for operational implementation.
4. When possible, it is recommended to include the heritage channel T6.2 in global precipitation retrieval algorithms and for precipitation reanalyses.

The findings of this study are expected to contribute to improved precipitation estimation and characterization from space, which can complement the incomplete weather radar coverage of the Weather Service Radar-1988 Doppler (WSR-88D) network in the

western United States (Gebregiorgis et al. 2018; Upadhyaya et al., 2020). The approach could be directly adapted to other regions with similar precipitation climatologies, environment, and satellite observations. Note that the QPE is developed and evaluated across the CONUS and during the summer season. Future work will focus on winter precipitation and precipitation outside the CONUS.

Acknowledgements

We are very much indebted to the teams responsible for the GOES-R, MRMS and SCaMPR products. Datasets details for this research are available in these in-text data citation references: Kuligowski et al. (2016), Kirstetter et al. (2012; 2014), Benjamin et al. (2016). The authors acknowledge Y. Derin for helping with downloading and extracting environmental predictors. Funding for this research was provided by the GOES-R Series Risk Reduction program, which provided support to the Cooperative Institute for Mesoscale Meteorological Studies at the University of Oklahoma and by NASA earth science division earth science research from operational geostationary satellite systems (ESROGSS) under Grant NA16OAR4320115. P. Kirstetter acknowledges support from NASA Global Precipitation Measurement Ground Validation program under Grant NNX16AL23G and Precipitation Measurement Missions program under Grant 80NSSC19K0681.

References:

Adler, R. F., and Negri, A. J. (1988). A satellite infrared technique to estimate tropical convective and stratiform rainfall. *Journal of Applied Meteorology and Climatology*, *27*(1), 30-51.

Aonashi, K. and co-authors. (2009). GSMaP passive microwave precipitation retrieval algorithm: Algorithm description and validation. *Journal of the Meteorological Society of Japan. Ser. II*, *87*, 119-136.

ASTER GDEM V3; NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team (2019). *ASTER Global Digital Elevation Model V003* [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2020-12-08 from https://doi.org/10.5067/ASTER/ASTGTM.003.

Ba, M. B., and Gruber, A. (2001). GOES multispectral rainfall algorithm (GMSRA). *Journal of Applied Meteorology*, *40*(8), 1500-1514.

Benjamin, S. G. and Co-authors (2016). A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Monthly Weather Review*, *144*(4), 1669-1694.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Ciach, G. J., Morrissey, M. L., & Krajewski, W. F. (2000). Conditional bias in radar rainfall estimation. *Journal of Applied Meteorology and Climatology*, *39*(11), 1941-1946.

Elmore, K., and Grams, H. (2016). Using mPING data to generate random forests for precipitation type forecasts. In *14th Conf. on Artificial and Computational Intelligence and its Applications to the Environmental Sciences*.

Fowler, H. J. and co-authors (2021). Anthropogenic intensification of short-duration rainfall extremes. *Nature Reviews Earth* and *Environment*, 1-16.

Gagne, D. J., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., and Xue, M. (2017). Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and forecasting*, *32*(5), 1819-1840.

Genuer, R., Poggi, J. M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters*, *31*(14), 2225-2236.

Giannakos, A., and Feidas, H. (2013). Classification of convective and stratiform rain based on the spectral and textural features of Meteosat Second Generation infrared data. *Theoretical and applied climatology*, *113*(3-4), 495-510.

Graham, D., Sault, M., and Bailey, C. J. (2003). National ocean service shoreline—Past, present, and future. *Journal of Coastal Research*, 14-32.

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), 610-621.

Hirose, H., Shige, S., Yamamoto, M. K., and Higuchi, A. (2019). High temporal rainfall estimations from Himawari-8 multiband observations using the random-forest machine-learning method. *Journal of the Meteorological Society of Japan. Ser. II*.

Hong, Y., Hsu, K. L., Sorooshian, S., and Gao, X. (2004). Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *Journal of Applied Meteorology*, *43*(12), 1834-1853.

Hsu, K. L., Gao, X., Sorooshian, S., and Gupta, H. V. (1997). Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology*, *36*(9), 1176-1190.

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., and Yoo, S. H. (2015). NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). *Algorithm Theoretical Basis Document (ATBD) Version*, *4*, 26.

Masson-Delmotte, V. and co-authors (2021). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. . Cambridge University Press. In press.

Janowiak, J. E., Joyce, R. J., and Yarosh, Y. (2001). A real-time global half-hourly pixel-resolution infrared dataset and its applications. *Bulletin of the American Meteorological Society*, *82*(2), 205-218.

Joyce, R. J., Janowiak, J. E., Arkin, P. A., and Xie, P. (2004). CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of hydrometeorology*, *5*(3), 487-503.

Karl, T., and Koss, W. J. (1984). Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983.

Kirstetter, P. E., Hong, Y., Gourley, J. J., Cao, Q., Schwaller, M., and 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. *Remote sensing of the terrestrial water cycle*, 61-79.

Kirstetter, P. E. and Co-authors. (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.

Kirstetter, P. E., Hong, Y., Gourley, J. J., Schwaller, M., Petersen, W., and Zhang, J. (2013). Comparison of TRMM 2A25 products, version 6 and version 7, with NOAA/NSSL ground radar–based National Mosaic QPE. *Journal of Hydrometeorology*, *14*(2), 661-669.

Kirstetter, P. E., Karbalaee, N., Hsu, K., and Hong, Y. (2018). Probabilistic precipitation rate estimates with space-based infrared sensors. *Quarterly Journal of the Royal Meteorological Society*, *144*, 191-205.

Kirstetter, P. E., Karbalaee, N., Hsu, K., and Hong, Y. (2018). Probabilistic precipitation rate estimates with space-based infrared sensors. *Quarterly Journal of the Royal Meteorological Society*, *144*, 191-205.

Kühnlein, M., Appelhans, T., Thies, B., and Nauß, T. (2014). Precipitation estimates from MSG SEVIRI daytime, nighttime, and twilight data with random forests. *Journal of Applied Meteorology and Climatology*, *53*(11), 2457-2480.

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. (2011). Satellite rainfall information for flood preparedness and response. In *Use of Satellite and In-Situ Data to Improve Sustainability* (pp. 31-39). Springer, Dordrecht.

Kuligowski, R. J., Li, Y., Hao, Y., and Zhang, Y. (2016). Improvements to the GOES-R rainfall rate algorithm. *Journal of Hydrometeorology*, *17*(6), 1693-1704.

Lazri, M., and Ameur, S. (2018). Combination of support vector machine, artificial neural network and random forest for improving the classification of convective and stratiform rain using spectral features of SEVIRI data. *Atmospheric research*, *203*, 118-129.

Lazri, M., Labadi, K., Brucker, J. M., and Ameur, S. (2020). Improving satellite rainfall estimation from MSG data in Northern Algeria by using a multi-classifier model based on machine learning. *Journal of Hydrology*, *584*, 124705.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems 26*.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175-2199.

Meyer, H., Kühnlein, M., Appelhans, T., and Nauss, T. (2016). Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. *Atmospheric Research*, *169*, 424-433.

Meyer, H., Kühnlein, M., Reudenbach, C., and Nauss, T. (2017). Revealing the potential of spectral and textural predictor variables in a neural network-based rainfall retrieval technique. *Remote Sensing Letters*, *8*(7), 647-656.

Min, M. and co-authors(2018). Estimating summertime precipitation from Himawari-8 and global forecast system based on machine learning. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(5), 2557-2570.

Neiman, P. J., Martin Ralph, F., Persson, P. O. G., White, A. B., Jorgensen, D. P., and Kingsmill, D. E. (2004). Modification of fronts and precipitation by coastal blocking during an intense landfalling winter storm in southern California: Observations during CALJET. *Monthly weather review*, *132*(1), 242-273.

Ouallouche, F., Lazri, M., and Ameur, S. (2018). Improvement of rainfall estimation from MSG data using Random Forests classification and regression. *Atmospheric Research*, *211*, 62-72.

Pedregosa, F. and Co-authors. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Sadeghi, M., Nguyen, P., Hsu, K., and Sorooshian, S. (2020). Improving near real-time precipitation estimation using a U-Net convolutional neural network and geographical information. *Environmental Modelling* and *Software*, *134*, 104856.

Schmit, T. J., Gunshor, M. M., Menzel, W. P., Gurka, J. J., Li, J., and Bachmeier, A. S. (2005). Introducing the next-generation Advanced Baseline Imager on GOES-R. *Bulletin of the American Meteorological Society*, *86*(8), 1079-1096.

So, D., and Shin, D. B. (2018). Classification of precipitating clouds using satellite infrared observations and its implications for rainfall estimation. *Quarterly Journal of the Royal Meteorological Society*, *144*, 133-144.

Stephens, G. L., and Kummerow, C. D. (2007). The remote sensing of clouds and precipitation from space: A review. *Journal of Atmospheric Sciences*, *64*(11), 3742-3765.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, *9*(1), 1-11.

Tahir, M. A., Bouridane, A., Kurugollu, F., & Amira, A. (2004, October). Accelerating the computation of GLCM and Haralick texture features on reconfigurable hardware. In *2004 International Conference on Image Processing, 2004. ICIP'04.*(Vol. 5, pp. 2857-2860). IEEE.

Tao, Y., Hsu, K., Ihler, A., Gao, X., and Sorooshian, S. (2018). A two-stage deep neural network framework for precipitation estimation from bispectral satellite information. *Journal of Hydrometeorology*, *19*(2), 393-408.

Tjemkes, S. A., Van de Berg, L., and Schmetz, J. (1997). Warm water vapour pixels over high clouds as observed by Meteosat. *Contributions to atmospheric physics*, *70*.

Upadhyaya, S. A., Kirstetter, P. E., Gourley, J. J., and Kuligowski, R. J. (2020). On the propagation of satellite precipitation estimation errors: From passive microwave to infrared estimates. *Journal of Hydrometeorology*, *21*(6), 1367-1381.

Upadhyaya, S., and Ramsankaran, RAAJ. (2016). Modified-INSAT Multi-Spectral Rainfall Algorithm (M-IMSRA) at climate region scale: Development and validation. *Remote Sensing of Environment*, *187*, 186-201.

Upadhyaya, S., and Ramsankaran, RAAJ. (2014). Multi-index rain detection: a new approach for regional rain area detection from remotely sensed data. *Journal of Hydrometeorology*, *15*(6), 2314-2330.

Upadhyaya, S., and Ramsankaran, RAAJ. (2018). Comprehensive inter-comparison of INSAT multispectral rainfall algorithm estimates and TMPA 3B42-RT V7 estimates across different climate regions of India during southwest monsoon period. *Environmental monitoring and assessment*, *190*(1), 45.

Upadhyaya, S., Kirstetter, P. E., Kuligowski, R. J., and Searls, M. (2021b). Classifying precipitation from GEO Satellite Observations: Diagnostic Model. *Quarterly Journal of the Royal Meteorological Society*, *147*(739), 3394-3409.

Upadhyaya, S., Kirstetter, P. E., Kuligowski, R. J., Gourley, J. J., and Grams, H., (2021a). Classifying precipitation from GEO Satellite Observations: Prognostic Model. *Quarterly Journal of the Royal Meteorological Society*, *147*(739), 3318-3334.

van Emmerik, T., Mulder, G., Eilander, D., Piet, M., and Savenije, H. (2015). Predicting the ungauged basin: model validation and realism assessment. *Frontiers in Earth Science*, *3*, 62.

Yu, Y., Tarpley, D., Privette, J. L., Goldberg, M. D., Raja, M. R. V., Vinnikov, K. Y., and Xu, H. (2008). Developing algorithm for operational GOES-R land surface temperature product. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(3), 936-951.

Zhang, J.and co-authors. (2016). Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621-638.

Zou, X., Qin, Z., and Weng, F. (2011). Improved coastal precipitation forecasts with direct assimilation of GOES-11/12 imager radiances. *Monthly weather review*, *139*(12), 3711-3729.

Appendix:

Table A1. Performance assessment scores used for precipitation quantification and detection

	Quantification	Description
1	$Correlation\ Coefficient\ (CC) = \frac{\sum_{i=1}^{n}(REF_i - \underline{REF})(SRE_i - \underline{SRE})}{\sqrt{\sum_{i=1}^{n}(REF_i - \overline{REF})^2} \times \sqrt{\sum_{i=1}^{n}(SRE_i - \overline{SRE})^2}}$	Linear association between between estimated precipitation and observed precipitation
2	$Root\ Mean\ Square\ Error\ (RMSE;\ \text{mm}\ h^{-1}) = \sqrt{\frac{1}{n}(SRE_i - REF_i)^2}$	Average magnitude of estimation error. RMSE puts greater influence on large errors than smaller errors.
3	$Normalised\ Root\ Mean\ Square\ Error\ (nRMSE) = \frac{\sqrt{\frac{1}{n}(SRE_i - REF_i)^2}}{\overline{REF}}$	nRMSE facilitates comparison across different regions/ precipitation types with varying characteristics/magnitude/scale of precipitation
4	$Additive\ Bias\ (Abias\ ;\ \text{mm}\ h^{-1}) = \frac{1}{n}\sum_{i=1}^{n} SRE_i - REF_i$	Average estimation bias
5	$Relative\ Bias\ (Rbias;\%) = \frac{\sum_{i=1}^{n} SRE_i - REF_i}{\sum_{i=1}^{n} REF_i} \times 100$	Ratio of averaged estimation magnitude and averaged observed magnitude
	Detection	
6	$Probability\ of\ Detection: Rain\ (POD:R) = \frac{h}{h+m}$	Fraction of precipitating events correctly detected by RF
7	$Probability\ of\ Detection: No-Rain\ (POD:NR) = \frac{c}{f+c}$	Fraction of non-precipitating events correctly detected by RF
8	$False\ Alarm\ Ratio\ (FAR) = \frac{f}{h+f}$	Fraction of falsely detected precipitating events among all detected precipitation events by RF
9	$Heidke\ Skill\ Score\ (HSS) = \frac{2(hc-fm)}{(h+f)(f+c)+(h+m)(m+c)}$	Accuracy of precipitation detection relative to that of a random chance

10	$Volumetric\ Hit\ Index\ (VHI) = \frac{\sum_{i=1}^{n}\left((SRE_i > t\ \&\ REF_i > t)\right)}{\sum_{i=1}^{n}\left((SRE_i > t\ \&\ REF_i > t)\right) + \sum_{i=1}^{n}\left((SRE_i \leq t\ \&\ REF_i > t)\right)}$	Volume of precipitation correctly detected by RF

SRE is the satellite-based precipitation product (i.e. RF or SCaMPR) and *REF* is the GV-MRMS reference precipitation, *n* is the validation sample size and *t* is the precipitation threshold above which VHI is computed. h, m, f, c stands for hits, misses, false detections and correct rejections by occurrence, respectively.

Towards Improved Precipitation Estimation with the GOES-16 Advanced Baseline Imager: Algorithm and Evaluation

Shruti A. Upadhyaya[1,*], Pierre-Emmanuel Kirstetter[1,2,3,4,*], Robert J. Kuligowski[5], Maresa Searls[2]

[1] Advanced Radar Research Center, University of Oklahoma, Norman, Oklahoma

[2]School of Meteorology, University of Oklahoma, Norman, Oklahoma

[3]School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, Oklahoma

[4]NOAA/National Severe Storms Laboratory, Norman, Oklahoma

[5]NOAA/NESDIS/Center for Satellite Applications and Research, College Park, Maryland

[*]*Corresponding authors:* Pierre-Emmanuel Kirstetter (pierre.kirstetter@noaa.gov);

Shruti A. Upadhyaya (shruti.a.upadhyaya-1@ou.edu)

Caption:

The study introduces a new quantitative precipitation estimation (QPE) algorithm from Advanced Baseline Imager (ABI) observations from GOES-16 across the CONUS. It is developed and comprehensively evaluated using the Ground Validation Multi-Radar/Multi-Sensor (GV-MRMS) system as a benchmark, and features Random Forest (RF) machine learning-based QPE. The key innovations of the algorithm include a comprehensive set of satellite predictors derived from five infrared ABI channels, complemented by low-level environmental conditions from RAP Numerical Weather Prediction (NWP) model, and

outputs of probability of precipitation type for seamless integration of varying precipitation rates across types.

Graphical Representation: