

1 **A Machine-Learning Approach to Mitigate Ground Clutter in the GPM
2 Combined Radar-Radiometer Algorithm (CORRA) Precipitation Estimates**

3 Mircea Grecu,^{a,b} Gerald M. Heymsfield,^a Stephen Nicholls,^{a,c} Stephen Lang,^{a,c} William S.
4 Olson,^{a,d}

5 ^a *NASA Goddard Space Flight Center, Greenbelt, Maryland*

6 ^b *Morgan State University, Baltimore, MD*

7 ^c *Science Systems and Applications, Inc., Greenbelt, Maryland*

8 ^d *University of Maryland, Baltimore County, Baltimore, Maryland*

9 Corresponding author: Mircea Grecu, email:mircea.grecu-1@nasa.gov

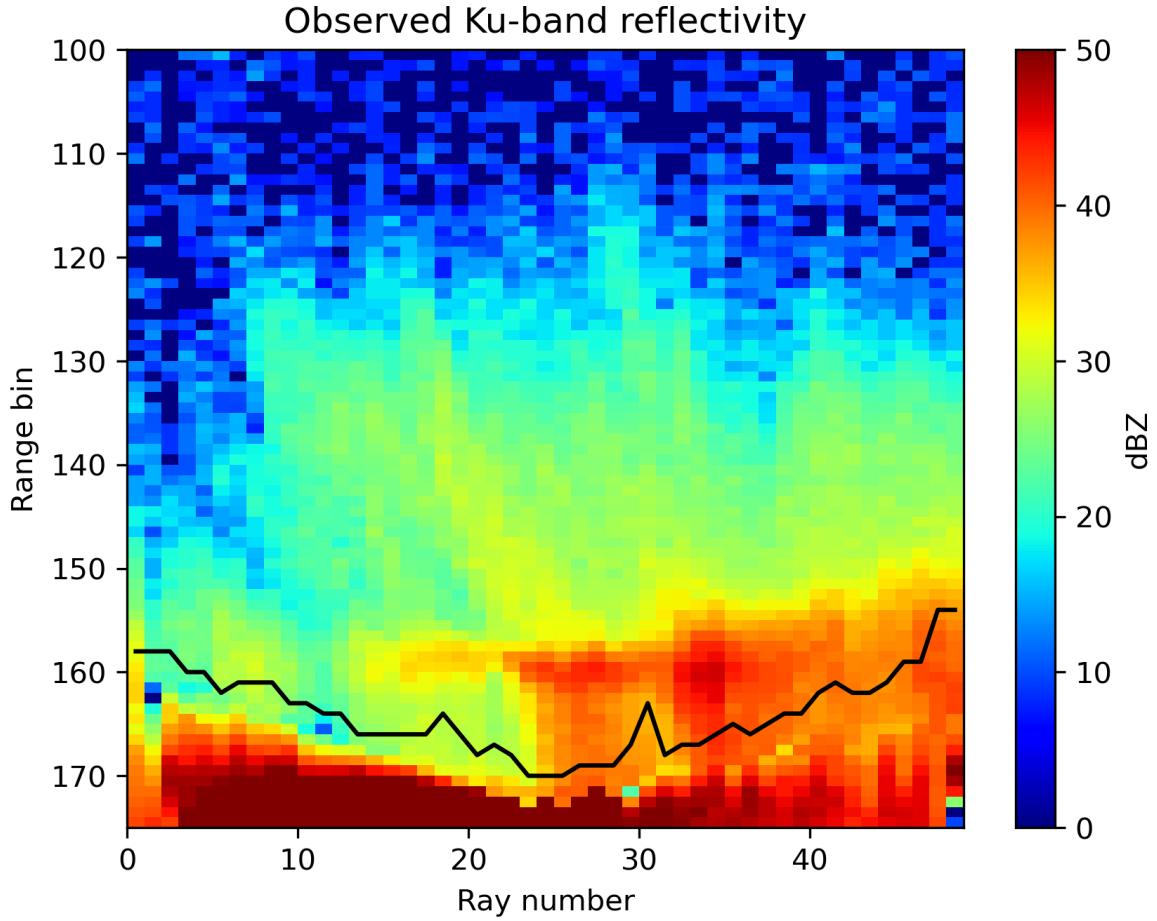
¹⁰ ABSTRACT: In this study, a machine-learning based methodology is developed to mitigate
¹¹ the effects of ground clutter on precipitation estimates from the Global Precipitation Mission
¹² Combined Radar-Radiometer Algorithm. Ground clutter can corrupt and obscure precipitation
¹³ echo in radar observations, leading to inaccuracies in estimates. To improve upon the previous work,
¹⁴ this study introduces a more general ML approach that enables a more systematic investigation
¹⁵ and a better understanding of uncertainties in clutter mitigation. To allow for a less restrictive
¹⁶ exploration of conditional relations between precipitation above the lowest clutter free bin and
¹⁷ surface precipitation, reflectivity observations above the clutter are included in a fixed-size set of
¹⁸ predictors along with the precipitation type, surface type, and freezing level to estimate surface
¹⁹ precipitation rates, and several ML-based estimation methods are investigated. A Gradient Boosting
²⁰ Model (GBM) is ultimately identified as the best candidate for systematic evaluations, as it is
²¹ computationally fast to train and apply while effective in applications. The GBM appears effective
²² in providing unbiased estimates; however, it is not much more effective in reducing random errors
²³ in the estimates relative to a simple bias correction approach. The fact that other ML approaches
²⁴ such as the k nearest neighbor method, the random forest method, and feed forward neural network
²⁵ approach showed similar performance in initial evaluations suggests that the inability of the GBM
²⁶ to achieve much more than bias removal is the result of indeterminacy in the data rather than
²⁷ limitations in the ML approach.

28 SIGNIFICANCE STATEMENT: Ground clutter can obscure and corrupt the precipitation echo
29 in the reflectivity observations by space borne radar, leading to inaccuracies and biases the surface
30 precipitation estimates. In this study, a machine learning approach is developed to mitigate the
31 effects of ground clutter on precipitation estimates from the Global Precipitation Mission (GPM)
32 Combined Radar-Radiometer Algorithm (CORRA). The approach is shown to be effective in
33 removing the biases associated with the simplest ground clutter mitigation approach and reducing
34 the random errors associated with more complex climatologically based bias-removal approaches.

35 1. Introduction

36 In radar meteorology, the echo originating in power emitted by the radar and reflected by the
37 ground is called ground clutter. Ground clutter has as negative impact on observations collected by
38 Dual Frequency Precipitation Radar (DPR) of the NASA Global Precipitation Measurement (GPM)
39 mission (Skofronick-Jackson et al. 2017), as it may obscure or corrupt radar signal associated with
40 precipitation. The extent of ground clutter in space-borne radar observations increases with
41 incidence angle (Kubota et al. 2016). Shown in Fig. 1 is a single scan representation of the Ku-
42 band reflectivity observed by the DPR from GPM orbit 50853 on 9 February 2023. The enhanced
43 reflectivity values at ranges close to (and larger than) 170 are contaminated by ground clutter. The
44 lowest bins that are deemed clutter-free by the DPR algorithm (Iguchi et al. 2021) are indicated
45 by the black line in the plot. As apparent from the figure, the number of bins affected by clutter
46 is quite significant for observations near the edge of the swath. Relative to the sea-level, 20 range
47 bins (bin width of 125 m) affected by clutter at the maximum scan incidence angle of 17° are
48 equivalent to a clutter height of about 2.4 km at the edge of the DPR swath. While the assumption
49 that the precipitation flux does not change significantly with height may be reasonable in some
50 situations, it is likely to result in significant biases in the surface precipitation estimates in weather
51 systems with freezing levels close to the ground. This is because ice processes such as riming and
52 depositional growth can result in significant flux changes.

55 To mitigate such biases, statistical correction methodologies, akin to those used to estimate the
56 surface reflectivity from ground-based precipitation radar observations, may be used. Specifically,
57 in ground-based radar, as the height of horizontally scanning radar beams increase with range, the
58 lowest-elevation reflectivity observations may be significantly elevated above the ground at large



53 FIG. 1. Cross-track section of observed Ku-band reflectivity field. Range bin spacing is 125 m and the black
 54 line indicates the lowest clutter-free bins. Bin 175 corresponds to the Earth ellipsoid.

55 ranges. For example, the beam center height is about 1500 m for an elevation angle of 0.5° and a
 56 range of 100.0 km (NWS 2023). Traditionally, to estimate the surface reflectivity from the lowest
 57 elevation angle reflectivity observations of ground radars, short range reflectivity observations
 58 from multiple elevation angle scans are used to derive statistical relationships between surface
 59 reflectivities and reflectivities aloft (Koistinen 1991). A similar approach can be applied to mitigate
 60 ground clutter in space-borne radar observations, with the difference being that the relationships
 61 between surface precipitation rates and precipitation rates aloft are derived from near-nadir space-
 62 borne radar observations and associated precipitation estimates that are minimally impacted by
 63 ground clutter. This approach has already been applied by Hirose et al. (2021) to refine the GPM
 64 DPR surface precipitation estimates.
 65
 66

In this study, we present a machine learning (ML) based methodology (Géron 2022) to investigate and mitigate ground clutter effects on precipitation estimates from the GPM combined radar-radiometer algorithm (CORRA). While conceptually similar to the approach of Hirose et al. (2021), our methodology is different in several key aspects and provides additional insight into ground-clutter-related uncertainties in the surface precipitation estimates and the best strategies to mitigate them. Unlike Hirose et al. (2021), we use reflectivity profile observations (rather than profiles of estimated precipitation rates) in the derivation of relationships between the precipitation rate in the lowest clutter free bin and the surface precipitation rate. The benefit of using reflectivity rather than precipitation profiles is that it enables the development of more physically consistent estimates. That is, radar profiling algorithms (Iguchi et al. 2021; Grecu et al. 2016) require assumptions regarding precipitation structure in the clutter to accurately incorporate estimates of the path integrated attenuation (PIA) from the Surface Reference Technique (SRT) to correct for attenuation down to the surface. However, if the clutter mitigation technique requires precipitation estimates, it can only be applied after the radar estimation process is complete. This may result in inconsistencies between the assumptions regarding the attenuation due to precipitation in the clutter and the actual precipitation estimates. While such inconsistencies may be addressed through iterative procedures, they result in a more computationally intensive retrieval process. In contrast, a clutter mitigation technique that uses reflectivity observations directly to derive relations between information above the clutter and precipitation in the clutter can be explicitly incorporated into the attenuation correction and precipitation estimation process, and this eliminates the need for iterative procedures to ensure the consistency of results. It should be mentioned, however, that the benefit (if any) of estimating the reflectivity in the clutter is limited in deep convection, because there are large uncertainties in the attenuation correction process both above and in the clutter. In this case, additional uncertainties caused by physical inconsistencies may not matter. Another distinction relative to Hirose et al. (2021), is that our methodology is based on ML, which is beneficial from the feature engineering perspective (Zheng and Casari 2018). Specifically, machine learning models can effectively extract relevant information from the data without having to resort to the explicit identification of features (defined as numerical attributes uniquely derived through a computational procedure applied to input data), thereby reducing the need for manual feature engineering, which can be time-consuming and error-prone for human experts. For example, the precipitation gradient

99 with respect to radar range is an intuitively-derived feature in the surface precipitation estimation
100 approach of Hirose et al. (2021). While features that make intuitive sense are valuable, questions
101 regarding their optimality are difficult to objectively address without tedious investigation. From
102 this perspective, ML procedures that do not require explicit features are worth considering. In
103 addition, the organization of data in a format that facilitates the development of an ML model
104 automatically facilitates the model's evaluation.

105 The paper is organized as follows. In Section 2, we present the ML methodology used to
106 estimate the surface precipitation rate from reflectivity observations not affected by clutter as well
107 as additional information such as the precipitation and surface type and the zero degree isotherm
108 height. In Section 3, we present the results of the application of the ML methodology to the GPM
109 CORRA precipitation estimates. In Section 4, we offer some conclusions from the study.

110 2. Methodology

111 a. General considerations

112 The simplest method to estimate the precipitation rate at a given height above the sea level (and
113 for a given precipitation type PT , surface type ST , and freezing level (FL)) from a precipitation
114 rate at a higher level is to re-scale the higher level value by the ratio of the climatological mean
115 precipitation rates at the two levels. Mathematically, this may be written as

$$P_{rate}(H_1, PT, ST, FL) = P_{rate}(H_2, PT, ST, FL) \frac{< P_{rate}(H_1, PT, ST, FL) >}{< P_{rate}(H_2, PT, ST, FL) >} \quad (1)$$

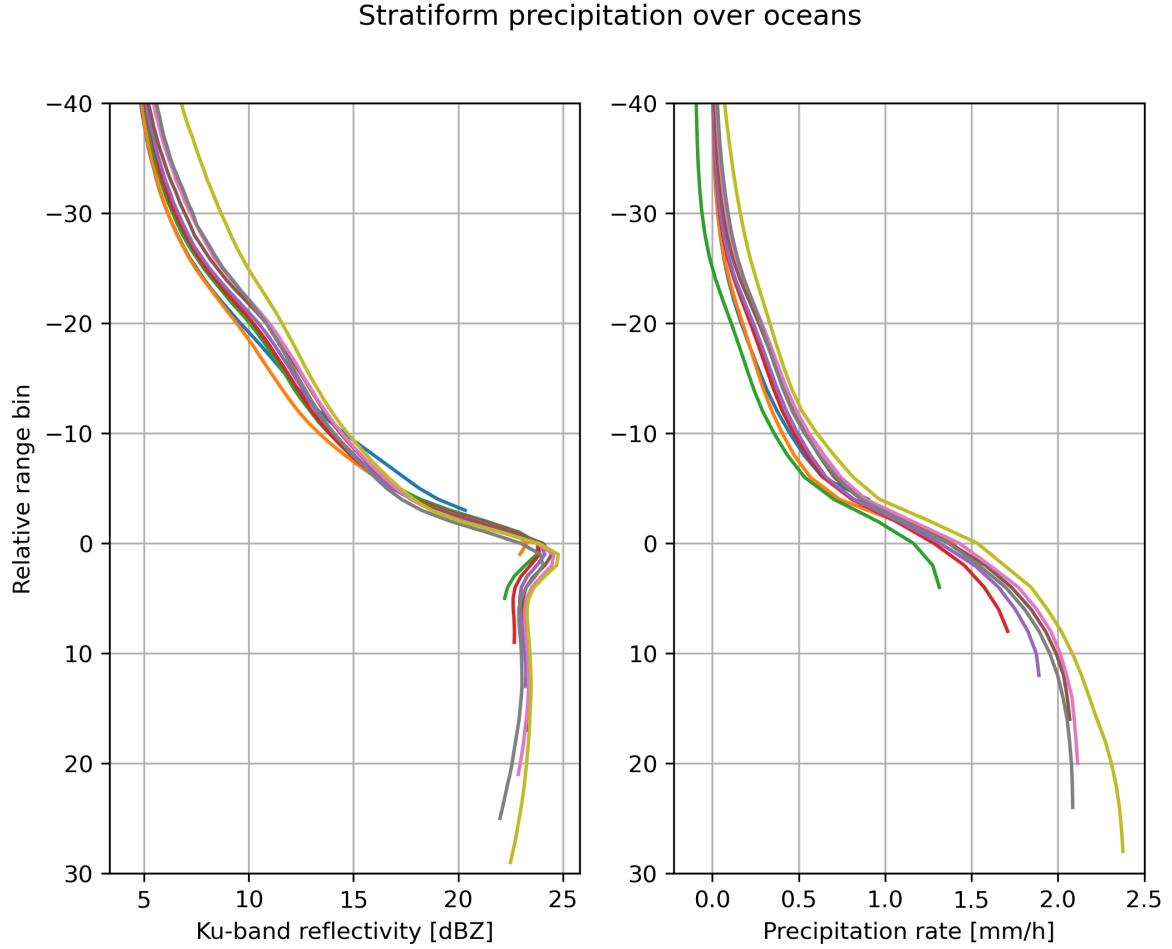
116

117

118 where P_{rate} is the precipitation rate, H_1 is the height where the estimate is needed, but for which
119 no direct measurement is available, H_2 is lowest clutter free height ($H_2 > H_1$) where a radar
120 measurement is available, and operator $< \square >$ denotes the climatological mean over a large dataset
121 characterized by the same freezing level, surface and precipitation type.

124 While simple in form, the challenge in applying a clutter correction methodology based on Eq.
125 (1) is the derivation of the correction factors $\frac{< P_{rate}(H_1) >}{< P_{rate}(H_2) >}$ for all possible (H_1, H_2) pairs. Nevertheless,
126 because the ground clutter depth is a function of the scanning incidence angle, estimates of the

Fig. 2,3: What do different color lines mean, can you please add a legend or colorbar?



122 FIG. 2. Conditional mean reflectivity and precipitation rate profiles over oceans for stratiform precipitation
123 with various freezing level heights.

127 climatological correction factor derived from near-nadir reflectivity observations and precipitation
128 estimates may be used to mitigate the clutter near the edges of the swath. Shown in Fig. 2 are
129 over-oceans conditional mean reflectivity and precipitation rate profiles from the GPM CORRA
130 algorithm (Grecu et al. 2016) for stratiform precipitation with various freezing level heights. The
131 profiles are plotted relative to the 0°C bin to emphasize similarities rather than differences due
132 to temperature-dependent processes. One year's worth (i.e. 2018) of DPR observations and
133 associated GPM CORRA retrievals characterized by fewer than eight bins affected by clutter are
134 selected and used in calculations of the mean profiles. The data are partitioned based on the
135 freezing level height in 12 distinct subsets, with the freezing level heights of each subset within 125

136 m of $1.875 + k * 0.25$ km with k varying from 0 to 11, resulting in 12 conditional mean profiles. As
 137 shown in the figure, the mean reflectivity and the associated precipitation profiles tend to align with
 138 one another. This behaviour may be used to mitigate the impact of clutter, even in near-nadir DPR
 139 observations that are affected by clutter at relatively low altitudes that make direct precipitation
 140 rate estimation at or near the surface impossible. Specifically, the data in Fig. 2 suggests that

$$\frac{\langle P_{rate}(H_1, PT, ST, FL) \rangle}{\langle P_{rate}(H_2, PT, ST, FL) \rangle} \approx \frac{\langle P_{rate}(H_1 + dFL, PT, ST, FL + dFL) \rangle}{\langle P_{rate}(H_2 + dFL, PT, ST, FL + dFL) \rangle} \quad (2)$$

141

142

143 where dFL is the difference between two distinct freezing level heights (FLH). The veracity of
 144 Eq. (2) is supported by the fact that in plots relative to the 0°C isotherm, the conditional mean
 145 precipitation profiles in Fig. 2 look very similar to profiles characterized by higher FLH and
 146 extending to greater depths below the 0°C . Here, the conditional mean precipitation rate refers to
 147 the mean precipitation rate in situations where precipitation is occurring in the LCFB (non-zero).
 148 The selection of profiles (with a maximum of eight radar bins impacted by ground clutter) results
 149 in a minimum value of H_1 of 1,000m (for a climatology derived from profiles with at most six
 150 bins affected by clutter). However, one can use Eq. (2) to approximate $\frac{\langle P_{rate}(0, PT, ST, FLH) \rangle}{\langle P_{rate}(H_2, PT, ST, FLH) \rangle}$ as
 151 $\frac{\langle P_{rate}(1,000m, PT, ST, FLH+1,000m) \rangle}{\langle P_{rate}(H_2+1,000m, PT, ST, FLH+1,000m) \rangle}$.

152 Shown in Fig. 3 are conditional mean reflectivity and precipitation rate profiles from CORRA
 153 for stratiform precipitation with various freezing level heights over land. The conditional mean
 154 precipitation profiles over land exhibit more variability than over oceans. However, this may be
 155 a consequence of precipitation retrieval artifacts rather than differences in temperature-dependent
 156 physical processes. Specifically, given that the SRT PIA estimates are noisier and less reliable over
 157 land, their impact on precipitation estimates may be less systematic, which could result in a larger
 158 spread of conditional mean estimates. Nevertheless, Eq. (2) is still a reasonable assumption.

159 The mean reflectivity profiles shown in Figs. 2 and 3 are stratified by precipitation type (strati-
 160 form), freezing level and surface type only, but it is conceivable that features that further separate
 161 the relationships between the reflectivity observations and the final precipitation estimates exist.
 162 As previously mentioned, Hirose et al. (2021) use the precipitation slope to stratify the database of
 163 near-nadir precipitation supporting their precipitation refinement process. In the current study, we

Stratiform precipitation over land

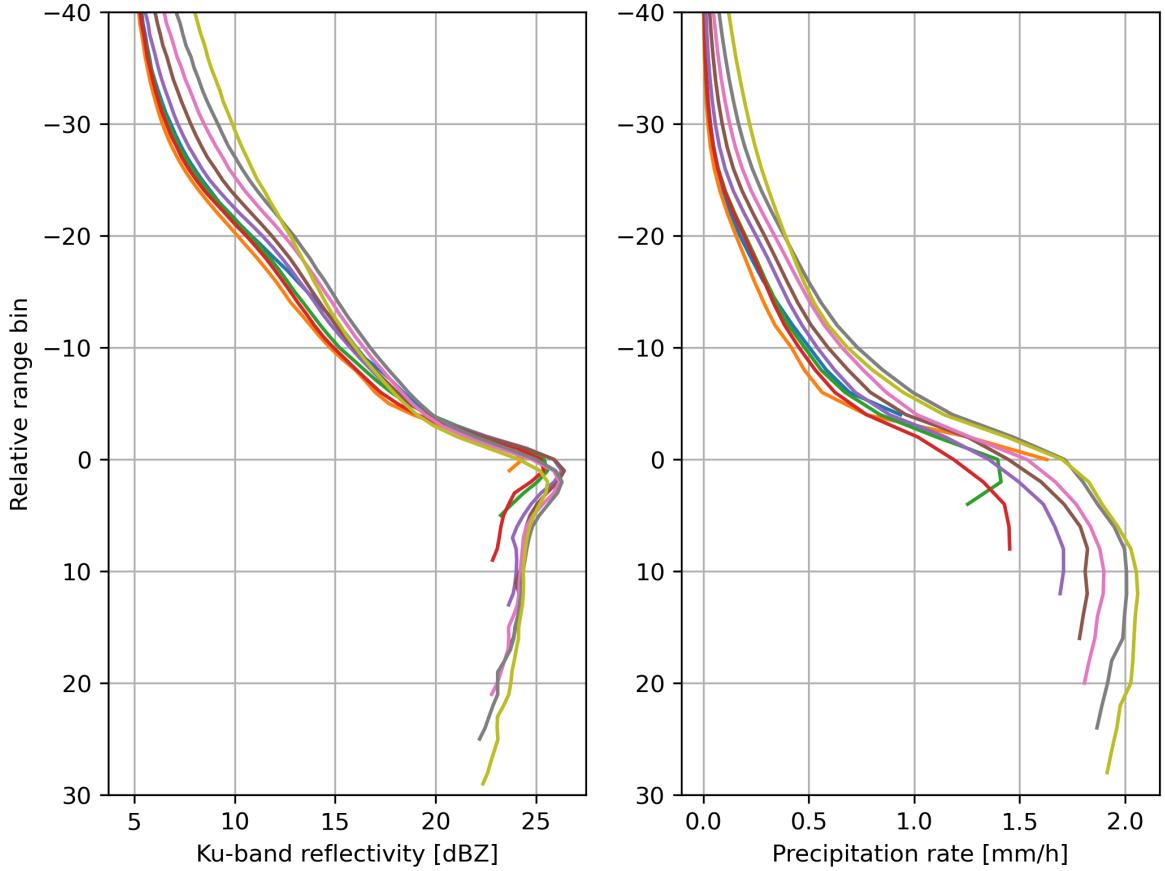


FIG. 3. Same as 2 but over land.

also investigate the slope of the reflectivity profile below the freezing level as a feature potentially useful for predicting the surface precipitation rate from the lowest clutter free precipitation rate. Specifically, the slope of the Ku-band reflectivity observations in the first six radar bins below the bright-band bottom Iguchi et al. (2021) is used to stratify the observations into five categories. The resulting mean reflectivity profiles and the associated mean precipitation profiles are shown in Fig. 4 for three of these categories, the other two (i.e. associated with slopes with absolute values larger than 0.75dB/bin) accounting for less than 10% of the total number of profiles. As seen in the figure, distinct mean reflectivity profiles result in distinct mean precipitation profiles. This behavior may be used to derive more accurate surface precipitation estimates than those derived from Eq. (1). However, to make effective use of the reflectivity slope and other such features, questions regarding the optimal strategy to calculate the slopes and partition them by value, especially when

175 the ground clutter extends close to or above the freezing level, need to be addressed. As there is
 176 no obvious strategy to address such questions, we resort to a ML approach.

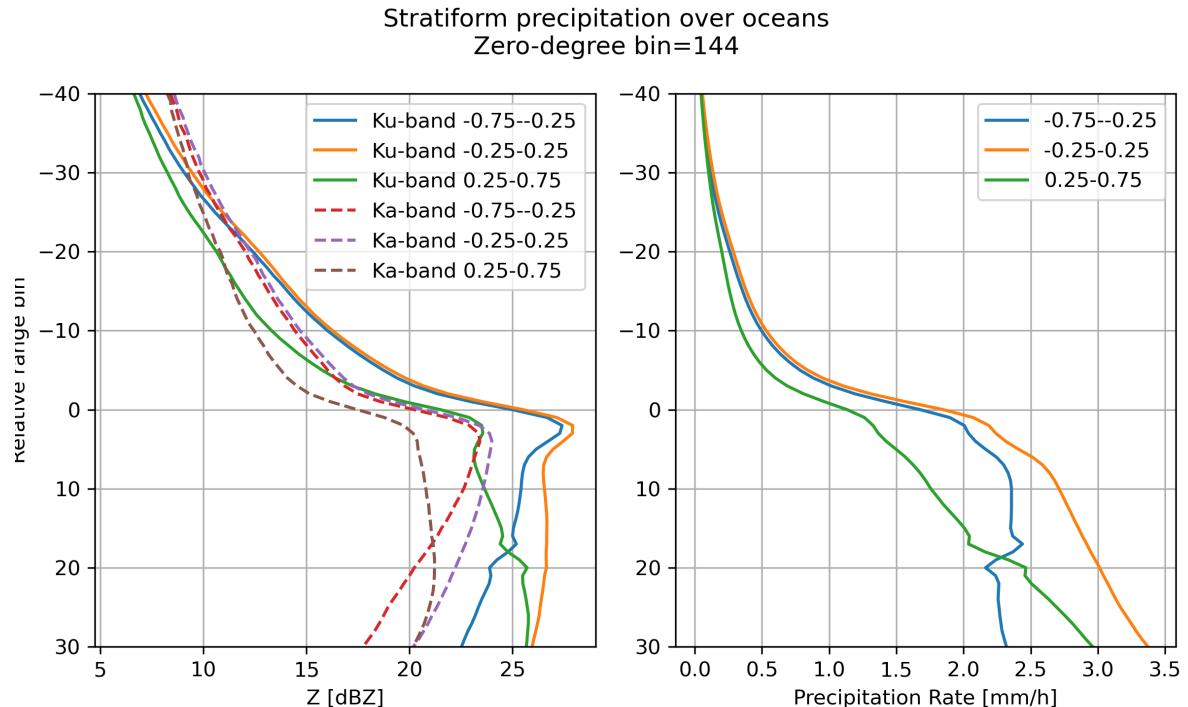


Fig. 4: Left axis label partly cutoff.

177 FIG. 4. Same as 2 but for a zero-degree bin of 144 and stratified by reflectivity slopes. The dashed lines in the
 178 left-hand side panel of the figures indicate the conditional mean reflectivity profiles at Ka-band associated with
 179 the three classes.

180 As previously mentioned, ML approaches do not require the explicit use of Eq. (1) and the
 181 stratification of the dataset by manually designed and optimized features. Instead, they require the
 182 organization of the dataset into a design matrix and a response matrix (Bishop and Nasrabadi 2006).
 183 In machine learning, the concepts of design and response matrices are borrowed from regression
 184 analysis, with the design matrix representing the array of predictor variables, while the response
 185 matrix representing the array of predicted variables. Each row of the design matrix corresponds to
 186 a single observation or data point, while each column represents a different predictor variable or
 187 feature.

188 In our study, the design matrix is an array of reflectivity observations and associated information,
 189 with each row containing the reflectivity values from a fixed-size portion of an observed profile. In
 190 addition to the reflectivity information, the zero degree bin, the position of the lowest clutter-free

Line 188: It would help the reader to provide a table of predictors. I don't see the profile slope mentioned - was 10 predictor?

191 bin (LCFB) bin relative to the zero degree bin, the position of the surface relative to the zero
192 degree bin, and the LCFB precipitation rate are included in the design matrix. To make the ML
193 models computationally efficient, the number of reflectivity observations above the LCFB is set to
194 30. Larger numbers of reflectivity observations above the LCFB were tested, but did not result in
195 improved results.

196 The response matrix is one-dimensional, i.e. a vector, and it contains the associated surface
197 precipitation rates. As explained above, the profiles in the training/evaluation dataset are charac-
198 terized by at most eight bins affected by clutter. Although minimally affected by clutter, there are
199 no surface estimates in the profiles of the training dataset, initially, and the precipitation estimates
200 associated with the clutter-free observations need to be extended to the surface. To achieve this,
201 for every profile with a FL height greater than 1.5 km above sea level, we regress the precipitation
202 rate against range using the estimates associated with the lowest four clutter-free bins and employ
203 the resulting regression to estimate the precipitation down to the sea level (next eight bins). It
204 should be mentioned that this extrapolation does not eliminate the need for more comprehensive
205 methodologies (or make them superfluous), because at higher incidence angles, the ground clutter
206 has a more significant and complex impact on profiles than the eight bins contaminated by ground-
207 clutter in the training dataset. When the FL is below 1.5 km, the precipitation slope may not be
208 reliably derived, as the possible existence of ice-phase and melting precipitation in the four lowest
209 clutter-free bins may significantly affect the vertical distribution of precipitation rates. For such
210 profiles, we use a k-Nearest Neighbor (k-NN) (Friedman 2001) approach to extend the precipitation
211 estimates into the clutter. Specifically, given that bin n_1 in a profile with freezing level height FLH_1
212 is roughly characterized by the same temperature as bin $n_1 - dn$ in a profile with freezing level
213 height FLH_2 , where dn is the integer part of $\frac{FLH_2 - FLH_1}{125m}$, we search for the k nearest neighbors
214 of a profile with $FLH_1 < 1.5 \text{ km}$ among profiles with $FLH_2 = FLH_1 + 1.0 \text{ km}$. The proximity is
215 evaluated using the Euclidean distance in a system of reference relative to the zero degree bin.
216 Then we use the k nearest neighbors to fill information in the clutter region, which is possible
217 because bins affected by clutter for profiles with a given freezing level height FLH_1 are clutter-free
218 in profiles with a freezing level height greater by 1.0 km, i.e $FLH_2 = FLH_1 + 1.0 \text{ km}$. While neither
219 the slope-based extrapolation nor the reflectivity-based extension are error free, they nevertheless
220 provide a reasonable methodology to extrapolate nearest surface precipitation rate rates in the

221 training data down to the sea level. Alternative methodologies based on cloud-resolving models
222 (CRMs) and ground-based radar observations are possible. However, these methodologies are not
223 necessarily bias- or complication-free, because CRMs may exhibit microphysical biases, while the
224 derivation of CRM simulations and the collection of ground observations representative of the
225 global distribution of precipitation events are extremely laborious processes. While CRMs and
226 ground observations may be able to eventually provide better (more complete) datasets to develop
227 methodologies to mitigate ground clutter in space-borne radar observations, the approach in this
228 study is still needed because the number of bins affected by ground-clutter is significantly greater
229 than eight (which is the number of clutter-affected bins in the training dataset in our study).

230 The structured organization of the dataset makes it possible to explore multiple ML models with
231 minimum effort and select the optimal one. While ML models are generally physics-agnostic in
232 the sense that they do not explicitly make use of physical laws, they can exploit physical causality
233 embedded in the dataset. For example, if slopes of the reflectivity profiles above the clutter
234 are reliable predictors of the precipitation rate at the surface relative to the lowest clutter-free
235 precipitation rate(as suggested by Fig.4), then an ML model based on the k-NN (Friedman 2001)
236 will be able to exploit this causality because similar reflectivity profiles in the design matrix result
237 in similar slopes. However, potentially more accurate or computationally more efficient ML models
238 may exist, and so in addition to the k-NN model, we also consider a Gradient Boosting (GB) model
239 (Friedman 2001), and a random forest (RF) model (Ho 1995). Both the GB and RF models are
240 based on decision trees (DTs) that are built through a process that constructs a tree-like structure by
241 recursively splitting the dataset based on feature conditions (Bishop and Nasrabadi 2006). However,
242 while the GB model starts with a weak DT and iteratively adds DTs to minimize residuals, the RF
243 model derives an ensemble of DTs and uses their average for prediction. The k-NN and RF model
244 implementations used in the study are based on the scikit-learn library (Pedregosa et al. 2011),
245 while the GB model is based on the efficient implementation of Ke et al. (2017). In addition to
246 the three scikit-learn based models, we consider a neural network (NN) model (Goodfellow et al.
247 2016; Géron 2022) based on the TensorFlow library (Abadi et al. 2016).

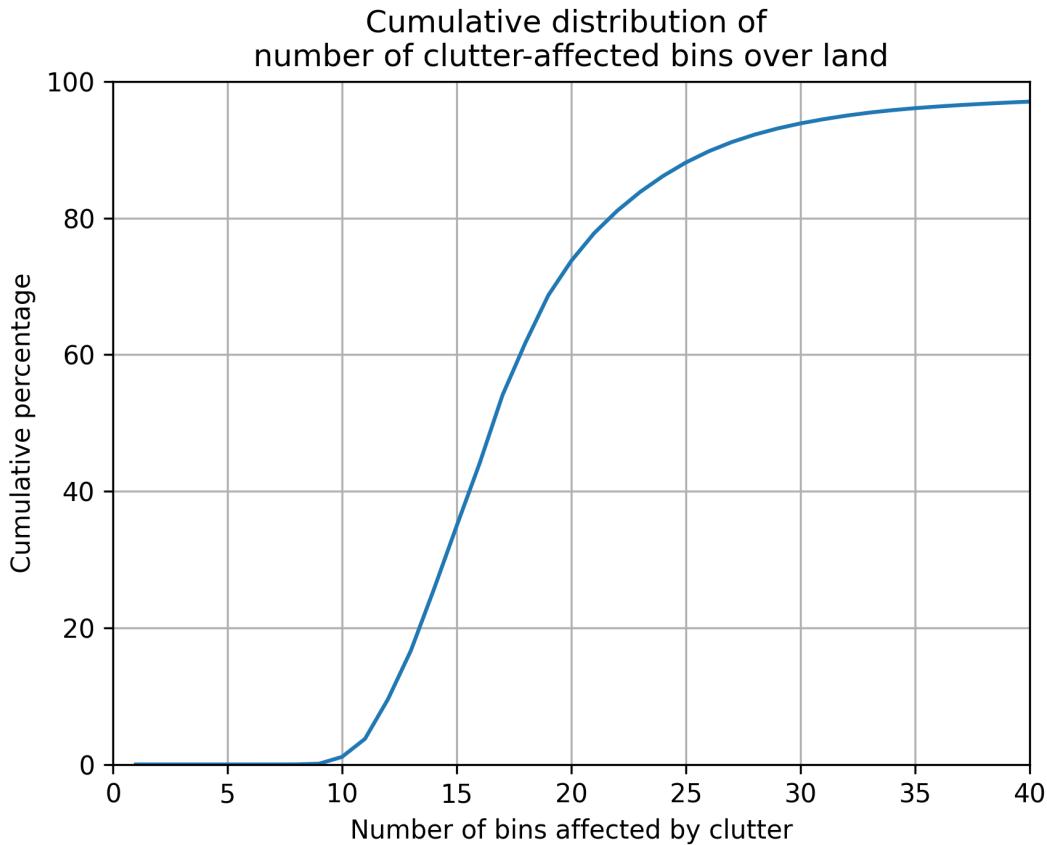
248 The scikit-learn library provides a convenient interface to train and test ML models. As such,
249 the definition of the scikit-learn ML models requires minimal specifications. They include the
250 number of neighbors for the k-NN model and the number of trees for the random forest model,

Line 247: If you want to guarantee that profile slope information was used, did you try a neural network with at least one 1D convolutional layer? If not, is there some reason this is difficult or a bad idea?

and the number of estimators for the gradient boosting model. Similarly, the Light Gradient Boosting Model (LGBM) implementation of Ke et al. (2017) requires minimal specifications. We set the number of neighbors to 20 and use the default values for the RF and LGBM models. The TensorFlow library, on the other hand, requires a more specific definition of the model architecture. We use a simple fully connected feed forward neural network with two hidden layers. The number of neurons is set to 32 in both hidden layers. The activation function is the rectified linear unit (ReLU) (Nair and Hinton 2010) and the output layer is a linear layer. The loss function is the mean squared error (MSE), and the optimizer is the Adam optimizer (Kingma and Ba 2014). The TensorFlow model is included in the study to provide insight into how the complexity of the ML model affects performance. Specifically, while the scikit-learn models and the LGBM are relatively simple, the NN model is more complex (Goodfellow et al. 2016) and may be able to capture more complex relationships between the reflectivities above the clutter and the surface precipitation rate.

Shown in Fig. 5 is the cumulative distribution of number of bins affected by clutter for rays in the DPR's outer swath. As apparent in the figure, more than eight bins are affected by clutter for the vast majority of the DPR outer swath profiles over land, with about half of profiles characterized by more than 15 bins affected by clutter and 10% of profiles characterized by more than 26 clutter-affected bins. The dataset of observations and precipitation profiles minimally affected by clutter and extended to sea-level, as explained above, is used to develop and test the different ML model approaches. To simulate clutter effects, n_c bins are assumed affected by clutter, where n_c is a random integer uniformly distributed between 1 and 26. The upper limit 26 was chosen based on the results shown in Fig. 5. While about 10% of profiles exhibit more than 26 bins affected by clutter, a larger upper limit might result in biases in the estimation because the distribution of the number of pixels affected by clutter is not uniform. A sampling strategy consistent with the cumulative distribution function in Fig. 5 may be used, but it would unnecessarily increase the size of the training dataset, because the sample size necessary to mitigate noise for $n_c > 26$ would result in significant oversampling for $n_c \leq 26$. However, given that the LGBM and the NN models have good extrapolation capabilities, and the number of profiles with $n_c > 26$ is relatively small, deriving ML models for $n_c \leq 26$ and applying them for $n_c > 26$ is not necessarily a poor choice. Moreover, an additional set of models trained exclusively for $n_c > 26$ may be derived. However, it is beneficial to first systematically investigate the performance and behavior of ML for $n_c \leq 26$.

Line 276: "Given that LGBM and NN models have good extrapolation capabilities" My experience has been not to trust NN models for extrapolation... but maybe decision tree-based models are different? Do you have a reference for this statement?



281 FIG. 5. The cumulative distribution of number of bins affected by clutter for rays in the DPR's outer swath
 282 (defined as the portion of the swath within 12 rays from the edges) over land.

Line 284: I don't see any information about the number of samples (training and testing)? It would be most helpful to provide that information as function of stratiform/convective and land/ocean, since those are the categories you are considering - so maybe a table. I don't see information on the data used (e.g., time-period or orbit numbers, geographic restrictions).

284 To evaluate the performance of the ML models, we use a cross-validation approach. Specifically,
 285 the DPR dataset is randomly split into a training and a testing dataset with 70% of profiles in
 286 the training dataset and the remaining 30% in the testing dataset. The training dataset is used to
 287 optimize the ML models, while the testing dataset is used to evaluate them. The evaluation is based
 288 on calculations of the correlation coefficient and bias between the predicted and observed surface
 289 precipitation rates.

290 **3. Results**

291 The reason for considering several ML model architectures is to ensure that there is no latent
292 information in the input data that is not properly captured. The inclusion of multiple ML models
293 reduces the likelihood of such a possibility, as the models are based on different statistical modeling
294 paradigms. However, in our initial model testing, no particular ML model emerged as significantly
295 better than the others. This outcome, which is not totally surprising, may be an indication that the
296 relations between the surface precipitation rate and the precipitation rate at a given height above
297 the surface depend on a multitude of factors that cannot be directly observed or do not have a
298 clear signature in the reflectivity observations. Nevertheless, some models are preferable to others.
299 Specifically, the k-NN model is rather slow in applications, as it requires searches through its
300 supporting database. The NN model on the other hand is slow in training and prone to overfitting
301 (i.e. it tends to produce smaller errors in training, but large biases in the evaluation process).
302 The LGBM and RF models have similar performances, with significantly lower computational
303 costs associated with the LGBM model. Therefore, based on this initial testing, we choose the
304 LGBM as the best option, and instead of exploring additional methodologies or carrying out further
305 tuning, we focus on characterizing its performance, especially in relation to a simple estimation
306 methodology.

7 Line 306: I'm curious how much of the hyper-parameters space was evaluated for these models. Of course, NNs
have tons of hyper-parameters, but knowing how many layers and how many units per layer that were explored
would be helpful. For the tree-based LGBM and RF models, knowing how many estimators and how deep would be
helpful. Also, I don't see where you provide information about the final model that you selected to produce the results

308 Before describing the performance of the different ML estimation methods, we will first examine
309 the persistence solution as a benchmark. In this simple solution, the precipitation rate at the LCFB
310 is assumed to be the same as the surface precipitation rate. Shown in the left-hand side panel of
9. Line 309: I'm curious how your conclusions might change if your benchmark was current Hirose
311 Fig. 6 is the correlation coefficient ~~appearing in the right panel of?~~ between the ~~than~~ surface precipitation rate and the LCFB
312 precipitation rate up to 26 bins above the surface for stratiform precipitation events over land.
313 The vertical axis is the difference between the surface bin and the LCFB, and the horizontal axis
314 represents the zero-degree bin. As seen in the panel, the correlation decreases with the position
315 of the LCFB above the surface. The bins marking a more significant correlation decrease (from
316 above 0.8 to below 0.75) generally occur in the mixed and ice phase. The biases associated with
317 the LCFB-derived precipitation rate relative to the surface precipitation rate are shown in the
318 right-hand side panel of Fig. 6. This type of estimation is referred to as persistence in the figure

8. Line 306: An advantage of tree-based models is that they provide interpretability. I'm curious whether your model
follows your intuition in terms of the predictors it focused on, or where there are any surprises - predictors that have been
previously overlooked that machine learning found useful?

319 and henceforth. Similar to the correlation coefficient, the largest biases occur when the LCFB is
320 in the ice phase. The correlation coefficients exhibit a discontinuous distribution for profiles with
321 a zero degree bin near 160, while the biases exhibit a relatively continuous distribution for profiles
322 with a zero degree bin near 167. This behavior is likely a consequence of the fact that precipitation
323 estimates in the mixed layer may be biased and noisy, and this may impact the procedure used to
324 fill in the precipitation estimates in the eight bins affected by clutter in the database. At the same
325 time, the DPR detection capabilities deteriorate for profiles with only snow above the clutter or if
326 the melting layer is close to the clutter.

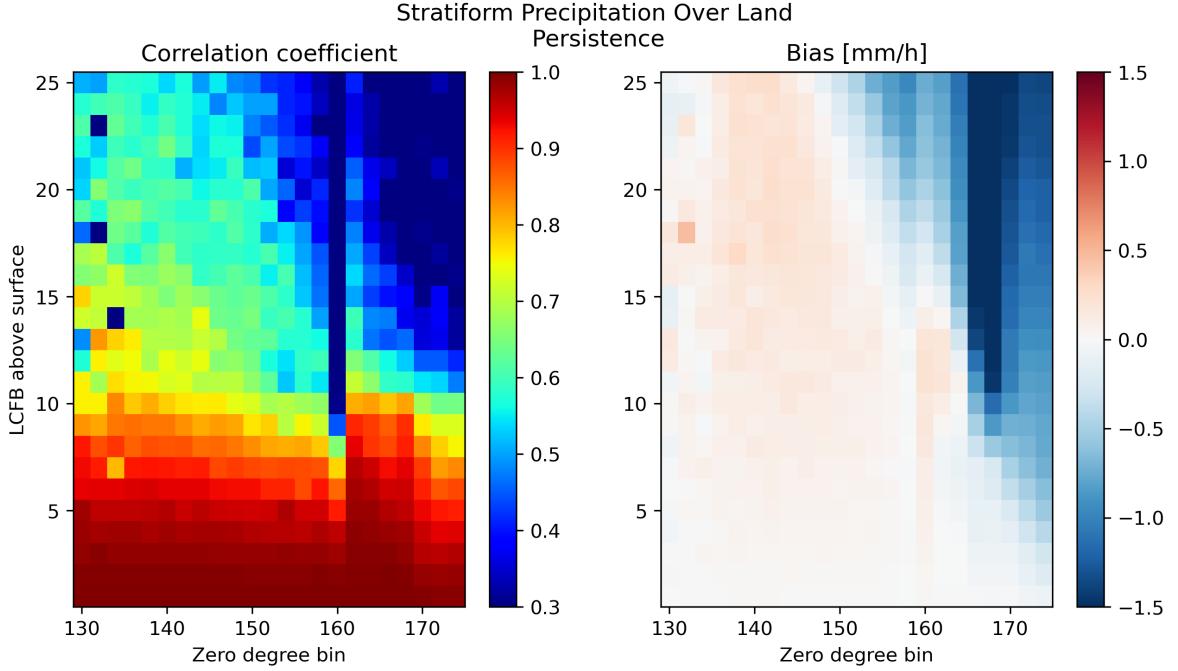
327 Unlike persistence-based estimates, surface precipitation estimates based on Eq. (1) would be
328 bias-free (assuming that the precipitation climatology is bias-free in the training dataset). However,
329 the distribution of correlation coefficients between the estimates and the true surface values would
330 not be different from that shown in Fig. 6. In other words, systematic errors are zero in estimates
331 based on Eq. (1), but the random differences remain largely the same. An estimation superior to
332 bias removal would also show an improvement in the distribution of the correlation coefficients
333 and an overall reduction in the root mean squared error (RMSE). Shown in Fig. 7 are results for
334 the LGBM method. As seen in the figure, the correlation coefficients increase slightly relative
335 to those shown in Fig. 6, while biases are almost zero. In particular, the biases in the ice phase
336 associated with the persistence-based estimates are largely removed. However, the marginal (at
337 best) improvement in the correlations between the estimated surface precipitation rates and those
338 in the databases suggests that there is significant variability of precipitation profiles in the clutter
339 that cannot be reliably predicted from observations in the clutter-free portion of reflectivity profile.

346

349

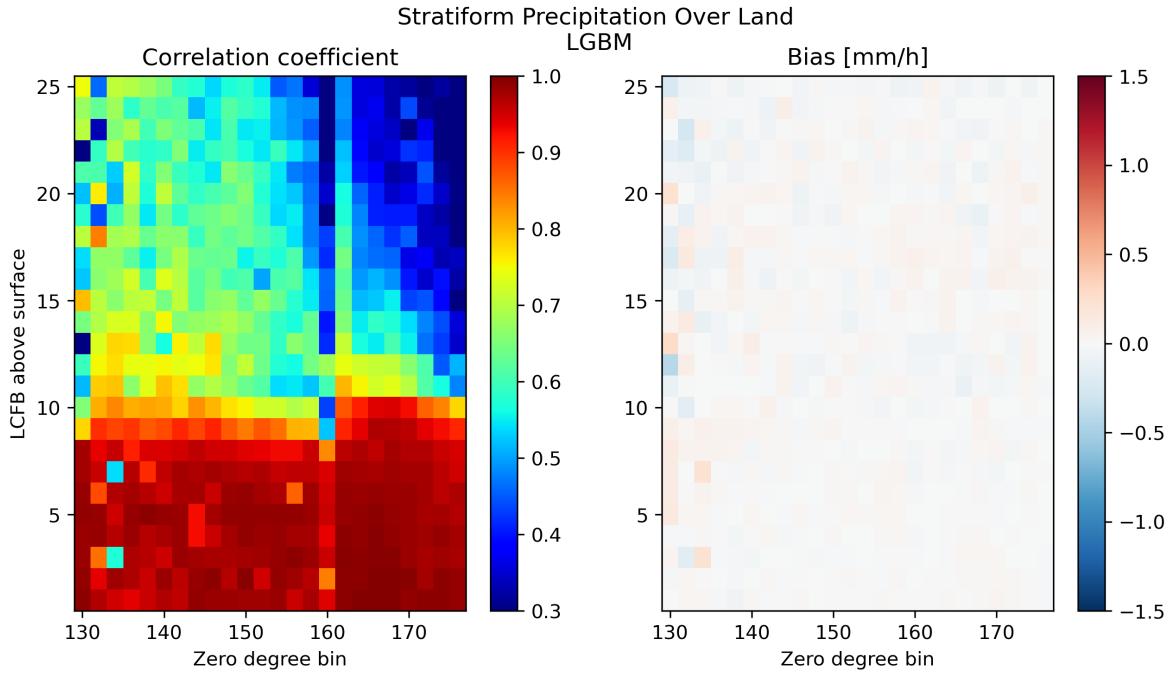
350 *b. Convective precipitation over land*

351 Shown in Fig. 8 are the distributions of correlation coefficients and biases of the persistence-
352 based estimator of surface convective precipitation over land. Results are qualitatively similar to
353 those obtained for stratiform precipitation over land, but with larger biases when the LCFB is in
354 the ice phase. Some positive biases for bins in the mixed phase are also obvious. These biases
355 are most likely the consequence of artifacts in the precipitation estimates across the melting layer



340 FIG. 6. Performance of a persistence-based clutter mitigation method for stratiform precipitation over land. The
 341 left-hand-side panel shows the correlation coefficient of the surface precipitation rates with the precipitation rates
 342 in the LCFB (which serves as the surface estimate in the persistence-based scheme), while the right-hand-side
 343 panel shows mean differences between the surface precipitate rates and the LCFB precipitation rates. Values are
 344 plotted for different surface vs. LCFB bin differences (vertical axis) and for different zero-degree bins (horizontal
 345 axis).

346 due to use of different reflectivity/precipitation lookup tables. The distributions of correlation
 347 coefficients and biases associated with the LGBM for convective precipitation over land are shown
 348 in Fig. 9. As seen in the figure, both the correlation coefficient and the bias improve relatively to
 349 results in Fig. 8. However, the bias distribution exhibits more variability around zero than the bias
 350 associated with stratiform precipitation over land. This is most likely a consequence of convective
 351 precipitation exhibiting more vertical variability while being about five times less frequent than
 352 stratiform precipitation. This makes the statistics of convective precipitation profiles in the training
 353 dataset noisier than those of stratiform precipitation. Noise can be mitigated by extending the
 354 dataset through inclusion of DPR observations and CORRA estimates from other periods, but it is
 355 likely that a portion of the noise is caused by artifacts due to multiple scattering and non-uniform



347 FIG. 7. Performance of the LGBM clutter mitigation method for stratiform precipitation over land. That is,
 348 same as Fig. 6 but for LGBM surface precipitation rate estimates instead of the persistence-based estimates.

366 beam filling in the precipitation estimation procedure. From this perspective, it is beneficial that
 367 convective dataset extension be considered at the same time with or after a refinement of the
 368 convective precipitation estimation methods in CORRA.

369

370

371 *c. Precipitation over oceans*

372 The statistics for precipitation over oceans are qualitatively similar to those over land; see Figs. 10
 373 and 11. The most significant difference is that, as suggested by Figs. 3 and 4, the mean precipitation
 374 profiles have different shapes, with the oceanic precipitation generally exhibiting more systematic
 375 increases with range below the freezing level than precipitation over land. However, the LGBM
 376 clutter correction schemes exhibit behaviors similar to those over land for both stratiform and
 377 convective precipitation types. This is shown in Fig. 10 for stratiform precipitation and in Fig. 11
 378 for convective precipitation. The LGBM model for stratiform precipitation appears rather noisy

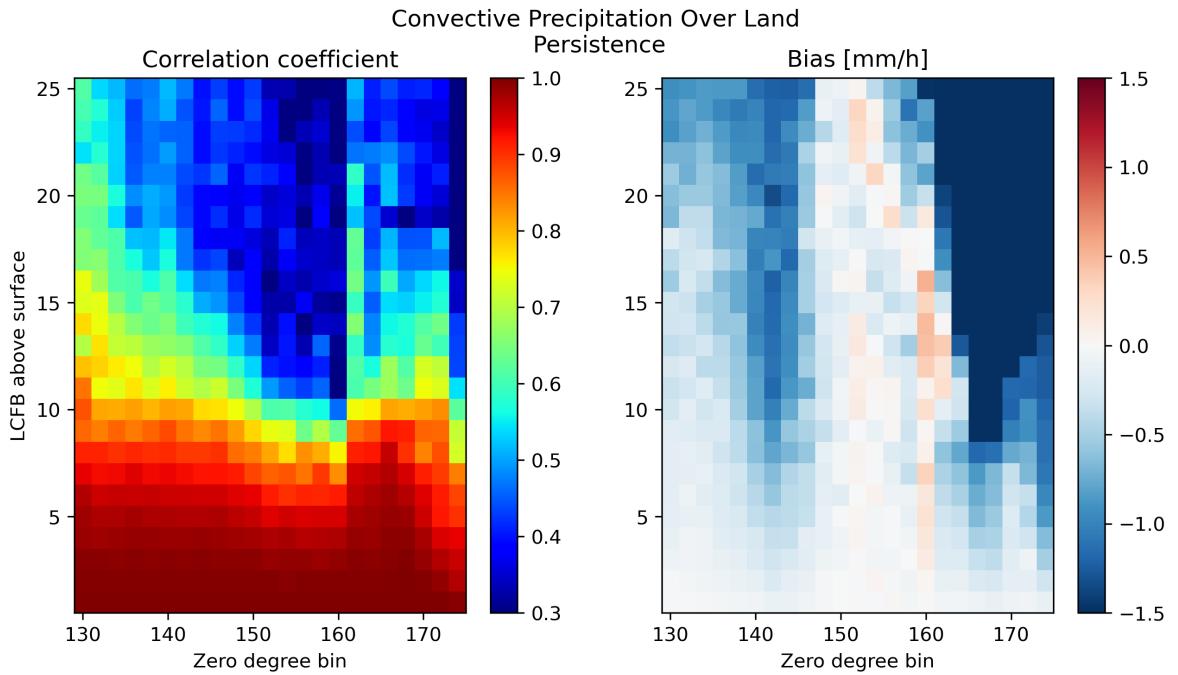


FIG. 8. Same as Fig. 6 but for convective precipitation over land.

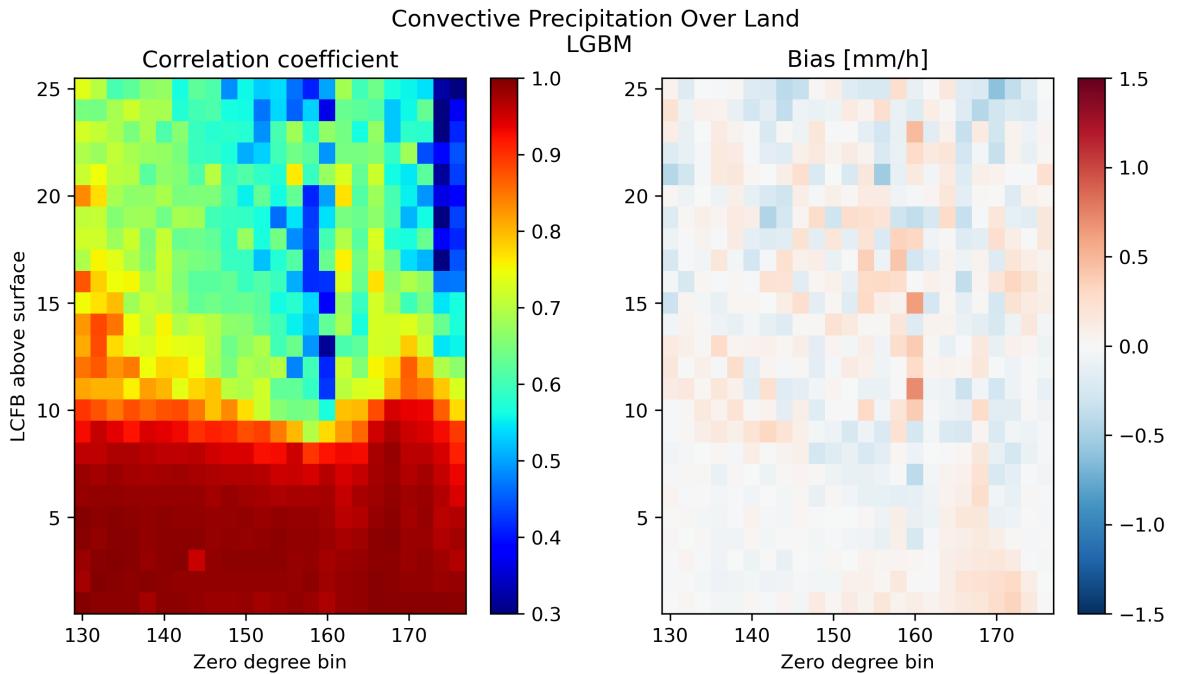


FIG. 9. Same as Fig. 7 but for convective precipitation over land.

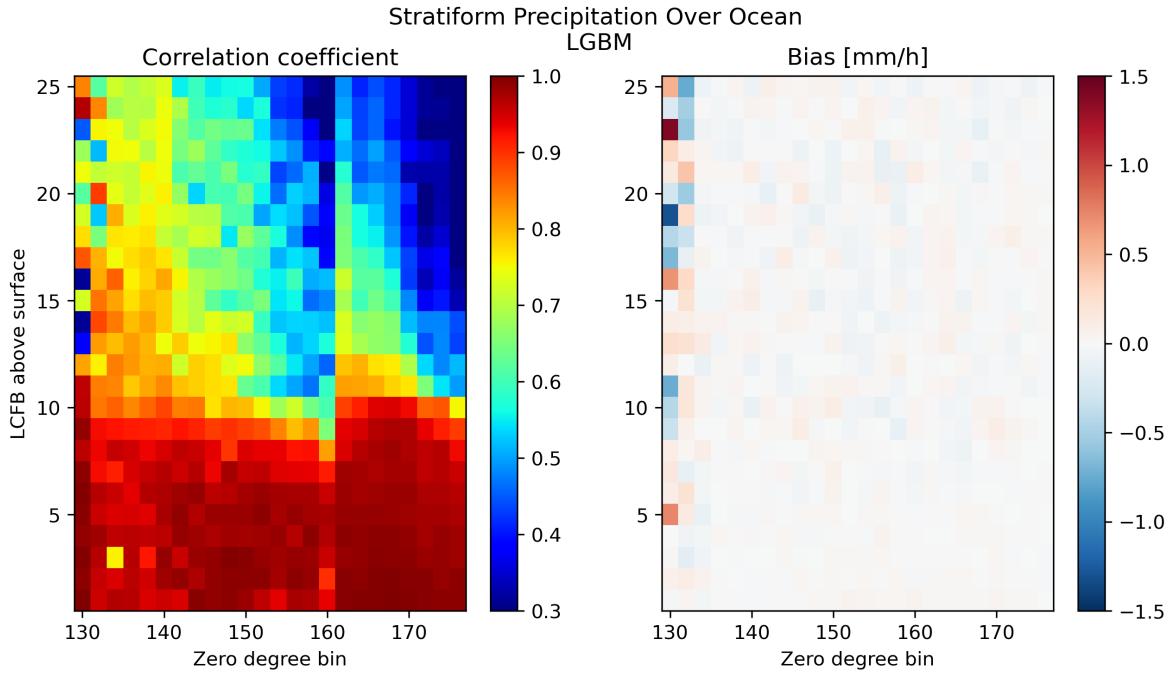


FIG. 10. Performance of the LGBM clutter mitigation method for stratiform precipitation over oceans.

379 for profiles with a zero degree bin less than 133 and an LCBF more than 4 bins above the surface,
 380 which is most likely a consequence of the relatively small number of such profiles in the training
 381 dataset. On the other hand, the LGBM model appears noisy in general for convective precipitation,
 382 which is consistent with its behavior over land.

383

384

385 *d. Evaluation of the random errors in the correction*

386 In the previous section, the LGBM method was shown to be effective in removing the biases
 387 associated with the persistence-based estimates. However, the random errors in the estimates
 388 were not necessarily reduced. Specifically, the correlation coefficients between LGBM surface
 389 precipitation estimates and actual surface precipitation estimates did not appear to be improved
 390 relative to those associated with the persistence-based estimates.

391 To investigate this quantitatively, we calculate the relative RMSE associated with both a climato-
 392 logical scaling correction based on Eq. (1) and the LGBM estimates. The relative RMSE involves

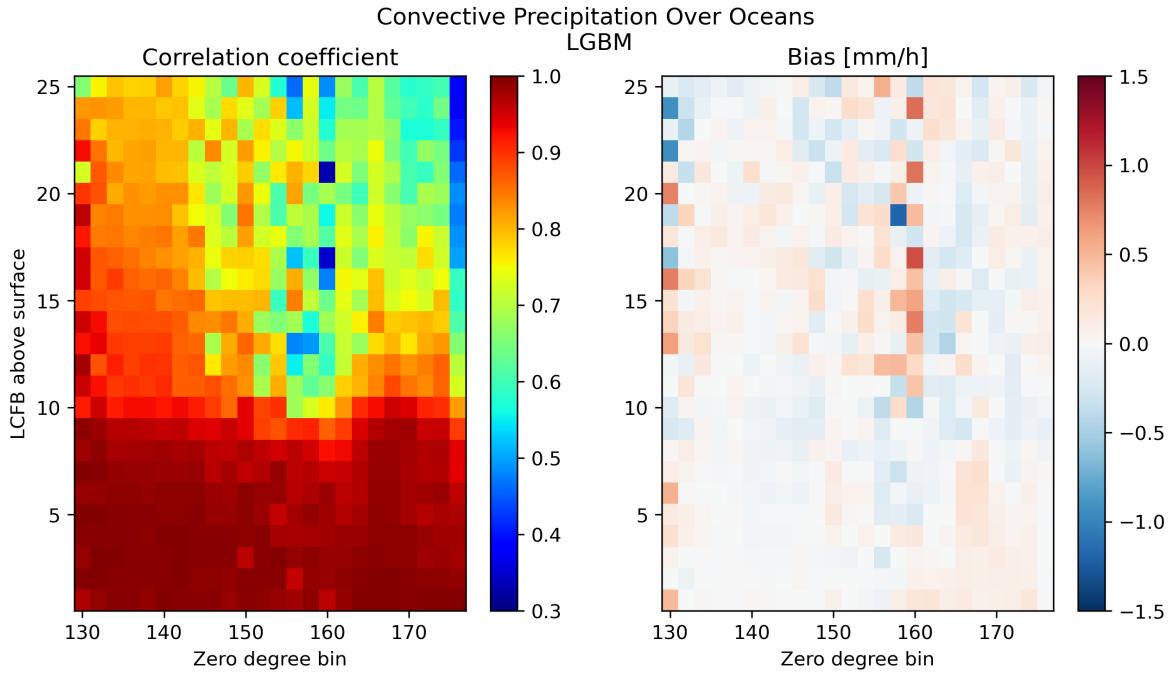
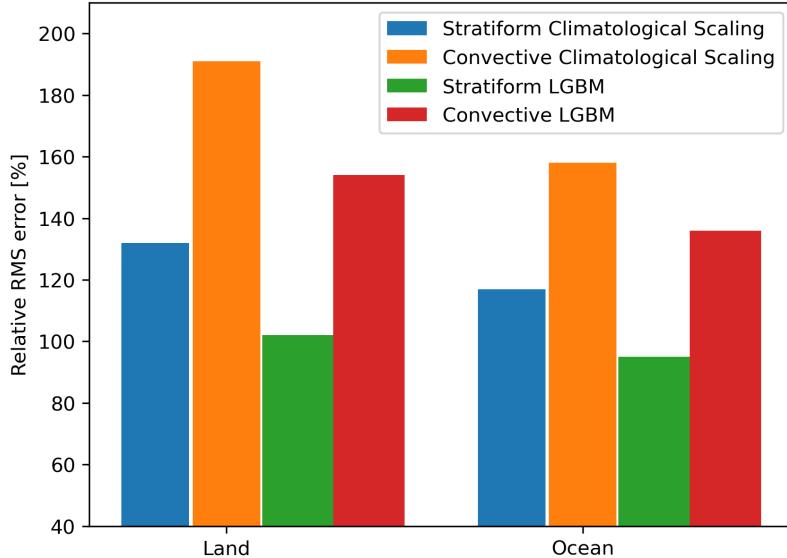


FIG. 11. Same as in Fig. 10 but for convective precipitation.

normalization by the standard deviation of the conditional surface precipitation rates. Results are shown in Fig. 12. Here, the LGBM estimates do not exhibit RMSEs that are much smaller than those of the climatological scaling estimates. This suggests that a simple bias-removal methodology based on Eq. (1) in section a is likely to be satisfactory in many respects. Nevertheless, the application of the LGBM method results in some RMSE reduction. As expected, the relative RMSEs are greater in convective than in stratiform precipitation and greater over land than over oceans. The fact that the LGBM method (which is representative of a broader class of one-dimensional clutter mitigation ML techniques) does not result in significant improvements relative to the simple bias correction provided by Eq. (1) is not necessarily an indication that ML techniques offer no benefit to the clutter mitigation problem. One potential advantage of the ML techniques is that they can incorporate radiometer observations, which may yield a significant benefit in the estimation of light precipitation over oceans. Also, the correction methods explored in this study as well as in the previous work of Hirose et al. (2021) make exclusive use of profile-level information. However, modern deep learning architectures such as U-Nets (Siddique et al. 2021) can readily process 3D information that may be useful for identifying the impacts of phenomena such as the wind shear on



410 FIG. 12. Relative RMSE for both the persistence and LGBM methods as a function of precipitation and surface
 411 type. 11.

**Figure 12: To me this looks like LGBM is providing value over the climatological scaling approach in terms of I
 reduction, but it sounds like you are underwhelmed? You conclude that simple bias-removal methods are satisf
 408 reflectivity observations and use this kind of information to more accurately predict the distribution
 409 of precipitation in the clutter. These topics will be explored in future studies.**

412 **4. Application to GPM CORRA precipitation estimates over the Continental US in the cold
 413 season**

414 To investigate the impacts of clutter mitigation on the estimation of precipitation over the Conti-
 415 nental US (CONUS) in the cold season, we apply the LGBM method to all GPM CORRA retrievals
 416 over CONUS from 1 December 2021 to 28 February 2022. While the same type of analysis can
 417 be applied to the entire GPM domain over all seasons, given that the focus of this study is on
 418 fundamental benefits and limitations of profile level corrections rather than their climatological
 419 impact, we limit our focus to a single region and season and defer more extensive analyses to future
 420 studies. Only profiles with freezing levels below 1250 m are considered in the analysis because
 421 they are given to the largest corrections (and errors in the absence of any correction), as the LCFB
 422 may be associated with temperatures below freezing, while the surface precipitation may be rain.

423 Shown in Fig 13 is the mean Ku-band reflectivity conditioned on the observed profiles being
 424 classified as precipitating. The means are conditioned on the associated profiles being classified

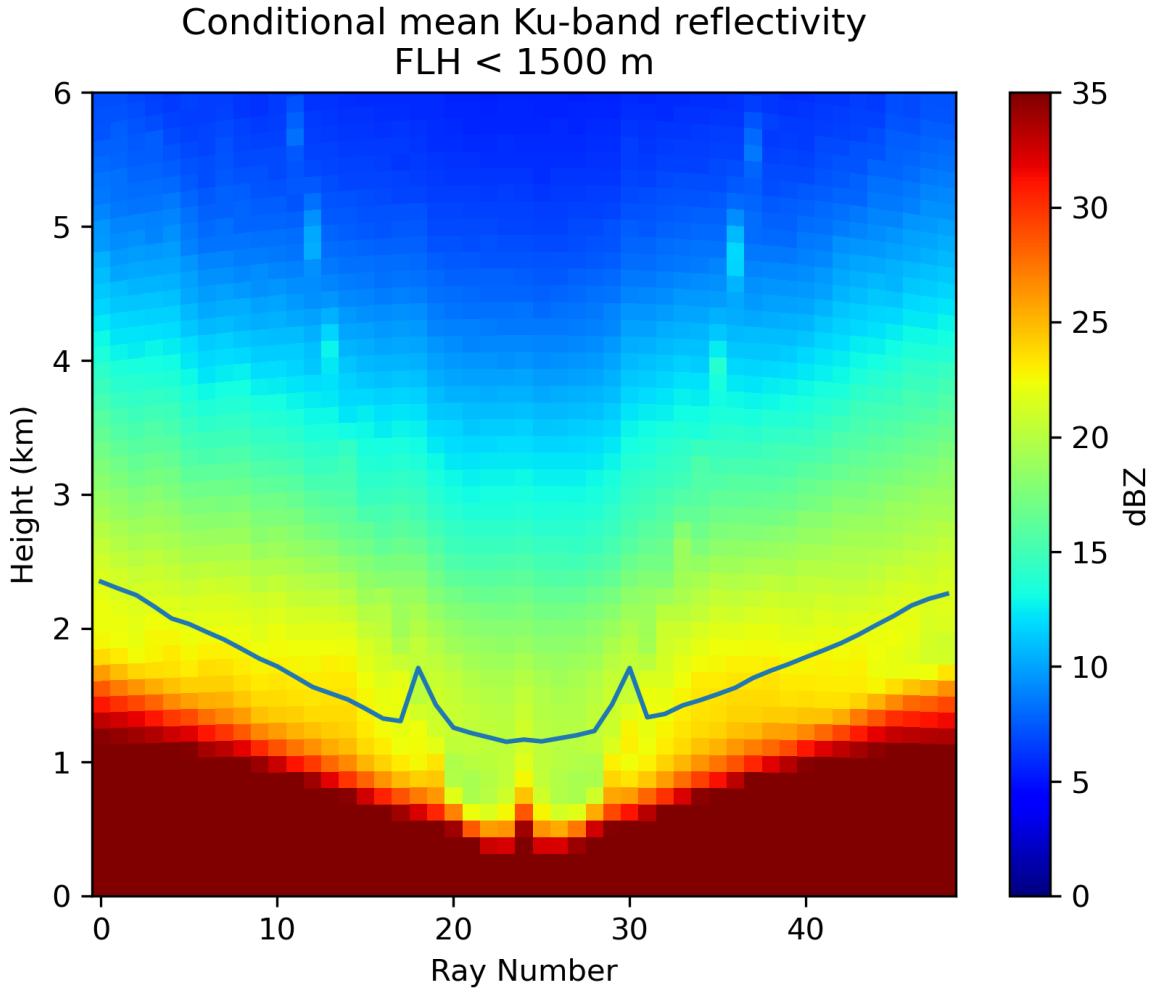
425 as precipitating. As seen in the figure, the region contaminated by clutter (characterized by large
426 reflectivity values) increases in height with the incidence angle. Some artifacts related to the
427 processing of the received power to mitigate sidelobe clutter (Kubota et al. 2016) are also apparent
428 in the figure. Specifically, while some enhanced echo is visible above 4.0 km, a slight reduction
429 in the reflectivities is apparent near the center of the swath (roughly from ray 20 to ray 30). The
430 reduction is more significant below the average height of the LCFB (blue line in the figure), but
431 that reduction does not directly impact the precipitation estimation, as the pixels associated with it
432 are classified as clutter.

433 Shown in the top panel of Fig. 14 are the conditional near-surface precipitation estimated by
434 CORRA and the surface precipitation predicted by the LGBM method. As seen in this figure,
435 the clutter mitigation methodology has a significant impact on the precipitation estimates, with
436 the impact increasing from center towards the edges of the swath. This behavior is, most likely, a
437 consequence of the fact that the DPR’s detection capabilities deteriorate near the edges of the swath
438 for precipitation systems with low FLH. The detected profiles are fewer but are characterized by
439 more intense (and deeper) precipitation that results in reflectivity observations that can be reliably
440 distinguished from clutter. This hypothesis is consistent with the distribution of the number of
441 precipitation profiles as a function of ray, shown in the bottom panel of Fig. 14. However, the
442 clutter correction technique does not result in artificial increases of intensity with distance from the
443 center of the swath in the overall (unconditional) precipitation rate. This point is illustrated in Fig.
444 15. Instead, the opposite effect, i.e. a reduction of the unconditional precipitation rate estimates
445 with distance from the swath center (consistent with the DPR precipitation detection capabilities
446 near the edges of the swath), is apparent in the figure. The overall impact of the LGBM surface
447 precipitation rate estimation procedure is significant for precipitation systems with freezing level
448 heights below 1250 m over CONUS. This is an indication that significant precipitation growth
449 processes such as water vapor deposition and riming occur in the clutter region.

452

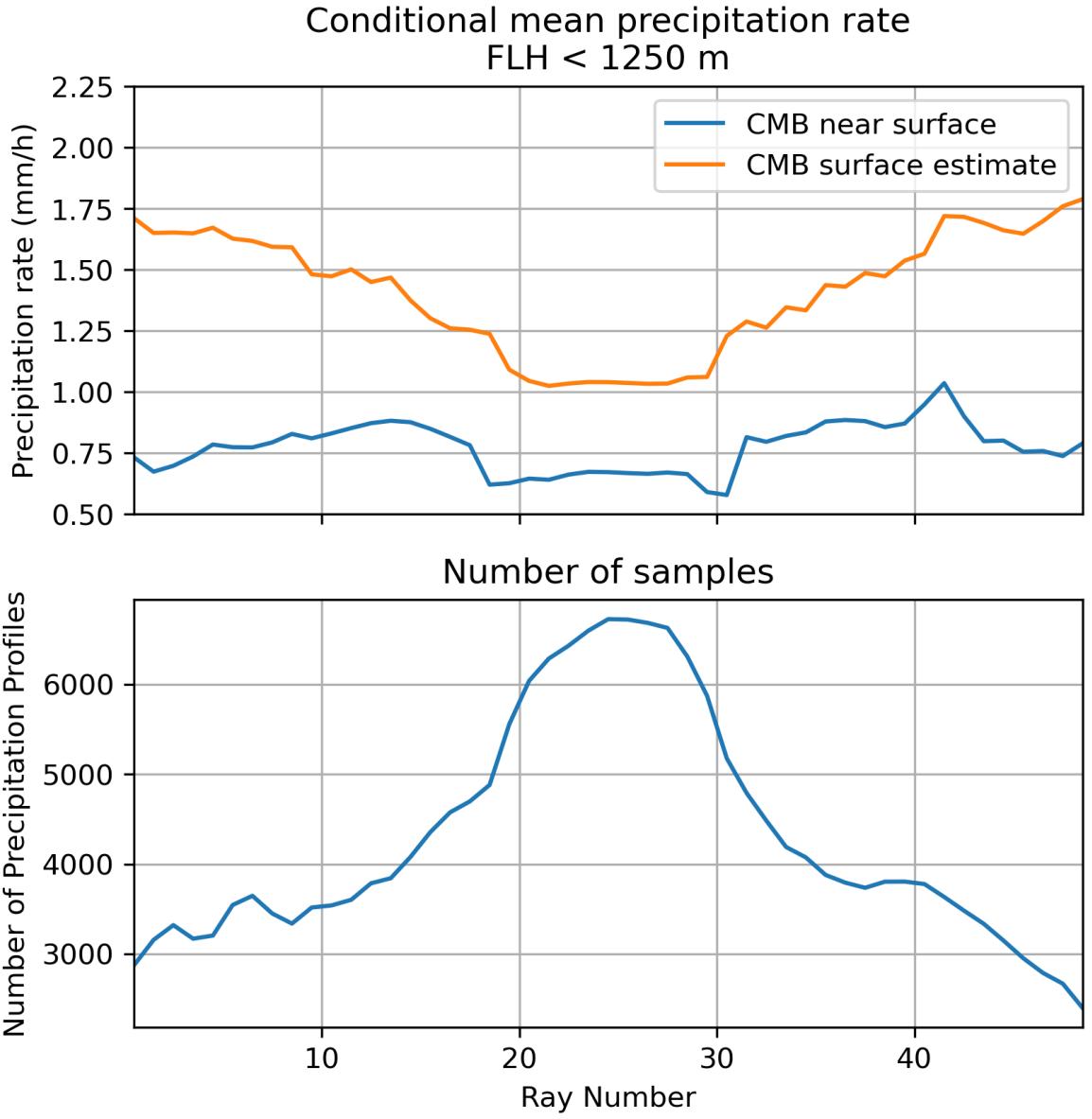
458 5. Summary and Conclusions

459 In this study, a new method for mitigating ground clutter effects in precipitation estimates derived
460 from the GPM mission’s CORRA algorithm is developed. CORRA combines data from the DPR



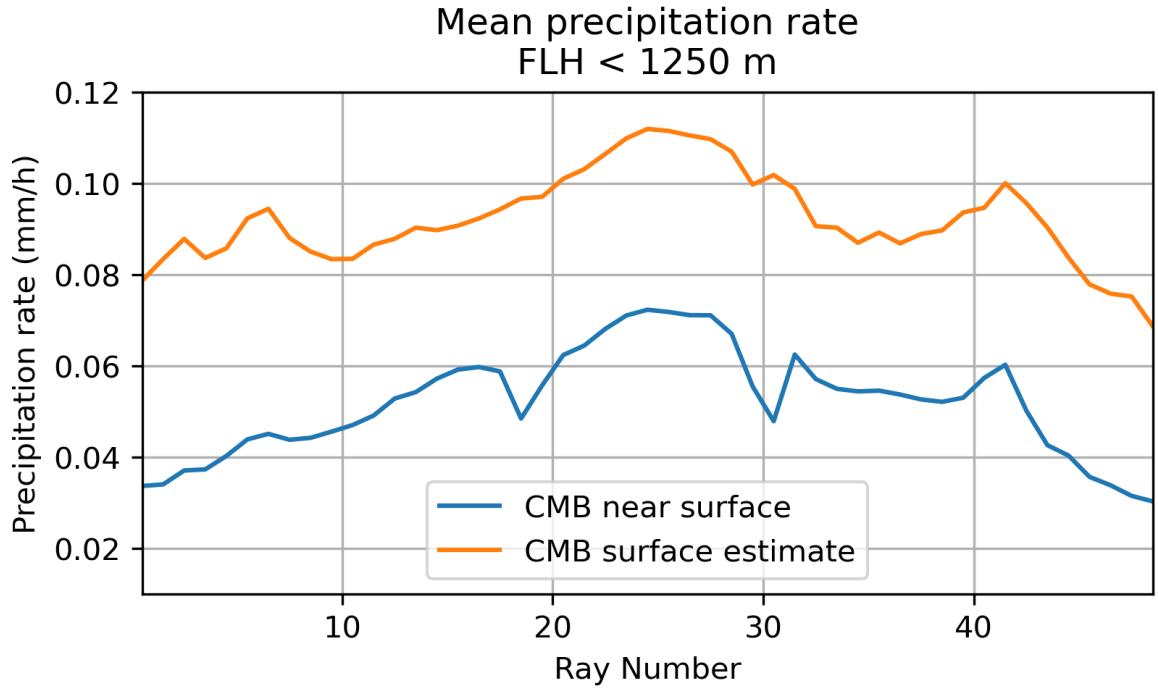
450 FIG. 13. Mean Ku-band reflectivity conditioned on the observed profile being classified as precipitating. The
451 blue line indicates the average height of the LCFB.

461 and GMI on the GPM core satellite to estimate precipitation rate, and ground clutter is a significant
462 problem for spaceborne radar observations, as it can obscure or corrupt the signal associated
463 with precipitation. An approach to mitigate ground clutter using statistical relationships based on
464 precipitation estimates from near-nadir scans has already been developed (Hirose et al. 2021) and
465 applied to precipitation estimates from the DPR algorithm (Iguchi et al. 2021). However, the study
466 of Hirose et al. (2021) did not fully explore the benefits and limitations of statistical methods to
467 mitigate clutter in the DPR reflectivity observations.



453 FIG. 14. Top panel: Conditional near-surface mean precipitation rate from the CORRA and the surface mean
 454 precipitation rate predicted by the LGBM method. Bottom panel: Number of detected precipitation profiles as a
 455 function of ray index.

468 To build upon the previous work, ML approaches are investigated to gain further insight into the
 469 uncertainties of surface precipitate rates derived from information in the portion of the reflectivity
 470 profile not affected by clutter. The ML model uses reflectivity observations, along with additional
 471 information such as precipitation type, surface type, and freezing level, to estimate the surface



456 FIG. 15. Top panel: Near-surface mean precipitation rate from CORRA and the surface mean precipitation
457 rate predicted by the LGBM method.

12. Fig. 15: "top panel" in caption - there is only one panel

472 precipitation rate. The benefits of this approach include the use of ML models efficient at leveraging
473 existing features and capturing complex relationships within the data without relying on explicit
474 feature engineering, and systematic evaluation of estimates is also facilitated. Specifically, various
475 machine learning architectures are investigated to automatically extract information from the data
476 without resorting to subjective efforts. A preliminary evaluation suggests that no architecture offers
477 significantly better performance than the others, and so we select the Light Gradient Boosting Model
478 (LGBM) of Ke et al. (2017) as the best candidate for further systematic evaluations, since it is
479 computationally fast to train and deploy while effective in application.

480 The database used in the training and evaluation of the ML models in the study is derived from one
481 year of DPR near-nadir observed reflectivity profiles that are minimally affected by clutter. A minor
482 deficiency of this database is that while the number of bins affected by clutter is small, one still needs
483 to resort to statistical models and assumptions to derive the surface precipitation estimates. For
484 profiles with a FLH greater than 1.5 km, characterized by a sufficiently large number of observations
485 not affected by clutter in the liquid phase, we simply use linear extrapolation to estimate the surface

486 precipitation rate. For all the other profiles, a k-NN method is used. Specifically, the k-nearest
487 neighbors of a profile are sought among profiles characterized by the same displacement of the
488 LCFB relative to the zero degree bin as that of the surface bin relative to the zero degree bin
489 for the profile in question. The proximity is evaluated using the Euclidean norm in a system of
490 reference relative to the zero degree bin. While the estimation of surface precipitation in the
491 database construction may introduce non-negligible random errors, and even biases, it is most
492 likely preferable to limiting the clutter mitigation to only clutter that extends greater than 1.0 km
493 above the surface (the extent of clutter at nadir view). Nevertheless, this postulate needs to be
494 evaluated in further studies. Over land, high-quality ground radar precipitation estimates adjusted
495 by rain-gauges such as those provided by the MRMS product (Zhang et al. 2016) may be used for
496 this purpose. The evaluation is likely to be more challenging over oceans, as data useful for direct
497 validation of estimates is very limited.

498 Estimates that use the LCFB as a proxy for surface precipitation rate are systematically different
499 from the surface precipitation rates in the training data, prompting an assessment of the LGBM
500 model within this context. Specifically, using the LCFB's precipitation rate as an estimate of surface
501 precipitation rate yields biased results, and so the LGBM model's capacity to produce unbiased
502 estimates is scrutinized. The LGBM model demonstrates effectiveness in providing unbiased
503 estimates, yet does not appear more capable of mitigating random errors than a basic climatological
504 scaling method. Furthermore, the performances of other machine learning techniques like k-nearest
505 neighbor, random forest, and feedforward neural networks mirror that of the LGBM model in initial
506 assessments, implying that the LGBM model's inability to improve upon bias removal stems from
507 the nature of the problem rather than inherent limitations of the approach.

508 6. Acknowledgments.

509 This work was supported by the NASA Global Precipitation Measurement Mission (PMM)
510 project. The authors thank Drs. Tsengdar Lee and Will McCarty (NASA Headquarters) for their
511 support of this effort.

512 **7. Data availability statement.**

513 The version 7 of GPM DPR and CORRA data can be accessed online
514 (<https://arthurhouhttps.pps.eosdis.nasa.gov/gpmdata/>).

515 **References**

516 Abadi, M., and Coauthors, 2016: Tensorflow: a system for large-scale machine learning. *Osdi*,
517 Savannah, GA, USA, Vol. 16, 265–283.

518 Bishop, C. M., and N. M. Nasrabadi, 2006: *Pattern recognition and machine learning*, Vol. 4.
519 Springer.

520 Friedman, J. H., 2001: Greedy function approximation: a gradient boosting machine. *Annals of*
521 *statistics*, 1189–1232.

522 Géron, A., 2022: *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* "O'Reilly
523 Media, Inc.".

524 Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep learning*. MIT press.

525 Grecu, M., W. S. Olson, S. J. Munchak, S. Ringerud, L. Liao, Z. Haddad, B. L. Kelley, and
526 S. F. McLaughlin, 2016: The gpm combined algorithm. *Journal of Atmospheric and Oceanic*
527 *Technology*, **33 (10)**, 2225–2245.

528 Hirose, M., S. Shige, T. Kubota, F. A. Furuzawa, H. Minda, and H. Masunaga, 2021: Refinement
529 of surface precipitation estimates for the dual-frequency precipitation radar on the gpm core
530 observatory using near-nadir measurements. *Journal of the Meteorological Society of Japan.*
531 *Ser. II*, **99 (5)**, 1231–1252, <https://doi.org/10.2151/jmsj.2021-060>.

532 Ho, T. K., 1995: Random decision forests. *Proceedings of 3rd international conference on docu-*
533 *ment analysis and recognition*, IEEE, Vol. 1, 278–282.

534 Iguchi, T., S. Seto, R. Meneghini, N. Yoshida, J. Awaka, M. Le, V. Chandrasekar, and T. Kubota,
535 2021: Gpm/dpr level-2 algorithm theoretical basis document. *NASA Goddard Space Flight*
536 *Center*.

- 537 Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, 2017: Lightgbm:
538 A highly efficient gradient boosting decision tree. *Advances in neural information processing*
539 *systems*, **30**.
- 540 Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. *arXiv preprint*
541 *arXiv:1412.6980*.
- 542 Koistinen, J., 1991: Operational correction of radar rainfall errors due to the vertical reflectivity
543 profile. *Preprints, 25th Int. Conf. on Radar Meteorology, Paris, France, Amer. Meteor. Soc,*
544 Vol. 91, 94.
- 545 Kubota, T., T. Iguchi, M. Kojima, L. Liao, T. Masaki, H. Hanado, R. Meneghini, and R. Oki, 2016:
546 A statistical method for reducing sidelobe clutter for the ku-band precipitation radar on board
547 the gpm core observatory. *Journal of Atmospheric and Oceanic Technology*, **33** (7), 1413–1428.
- 548 Nair, V., and G. E. Hinton, 2010: Rectified linear units improve restricted boltzmann machines.
549 *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807–814.
- 550 NWS, 2023: National weather service beam property calculator. URL [https://training.weather.gov/](https://training.weather.gov/wdtd/tools/beamwidth/)
551 wdtd/tools/beamwidth/.
- 552 Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in python. *the Journal of*
553 *machine Learning research*, **12**, 2825–2830.
- 554 Siddique, N., S. Paheding, C. P. Elkin, and V. Devabhaktuni, 2021: U-net and its variants for
555 medical image segmentation: A review of theory and applications. *Ieee Access*, **9**, 82 031–
556 82 057.
- 557 Skofronick-Jackson, G., and Coauthors, 2017: The global precipitation measurement (gpm) mis-
558 sion for science and society. *Bulletin of the American Meteorological Society*, **98** (8), 1679–1695.
- 559 Zhang, J., and Coauthors, 2016: Multi-radar multi-sensor (mrms) quantitative precipitation esti-
560 mation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, **97** (4),
561 621–638.
- 562 Zheng, A., and A. Casari, 2018: *Feature engineering for machine learning: principles and*
563 *techniques for data scientists.* ” O'Reilly Media, Inc.”.