Submitted to the Annals of Applied Statistics

A GENERALIZED ADDITIVE MODEL FOR CLOUDSAT RADAR SURFACE REFLECTIVITY REGRESSION: CONVOLUTIONAL NEURAL NETWORK AS A BASIS FUNCTION

BY DING LI [1] , ALEX STRINGER [2]

[1] *Computational Mathematics, University of Waterloo,* *d376li@uwaterloo.ca*

[2] *Statistics, University of Waterloo,* *alex.stringer@uwaterloo.ca*

1. Introduction.

1.1. *Background and research question.* Problem-solving in climate change heavily relies on cloud observation, as many meteorological estimates within the water cycle are derived from it. CloudSat, one of the few satellite missions offering a quasi-global view of cloud structures, suffers from ground clutter contamination, which overwhelmes near-surface hydrometeors due to strong surface echoes Battaglia (2019). The community had some attempts to quantify its impact. For example, research quantified the impact of ground clutter on snowfall rate anomalies over ice sheets Palerme (2019). Yet, a more general model for ground clutter's strength and impact based on a wider spectrum of surface conditions is still needed, which is crucial for determining the extent of bias introduced by the ground clutter in our data retrieval under different environments. Building on the previous findings, we emphasize the necessity of establishing a quantitative relationship between geographical conditions and surface reflectivity.

1.2. *Model selection.* Modeling reflectivity is determined by a series of very complex fluids-electromagnetic wave interations within a chaotic climate system, hindering the derivation of a purely deterministic model. Instead, we expect the "true underlying" model to have a very high variability and complex non-linear relationships Skolnik (2001). However, it is crucial to avoid relying solely on empirical models without physical mechanisms, as the interpretation of results is vital for addressing the research question. To bridge this gap, we incorporate theoretical insights to construct a regression model underpinned by a suitable physical mechanism, which ensures that our model remains both empirically robust and theoretically informed.

Among various regression model families, linear models are prized for their intuitiveness and interpretability. However, in highly complex systems such as ours, more powerful and flexible models, like deep learning models, are often necessary. Nonetheless, the opaque nature of these "black box" models makes it challenging to validate and elucidate the underlying physical mechanisms, which is a critical aspect of our research objectives. Additionally, the difficulty in obtaining confidence intervals for predictions from these models cannot be overlooked Hastie, Tibshirani and Friedman (2009). Therefore, we propose a novel approach that harnesses the robust capabilities of deep learning models while incorporating the interpretability of generalized additive models (GAM). This method aims to blend the best of both worlds, offering both deep learning's predictive power and the clarity and insight provided by GAMs.

Keywords and phrases: Generalized Additive Models, Ground Surface Reflectivity, CloudSat Radar Profiles, Convolutional Neural Network.

In this project, we initially pretrain a convolution kernel for multi-channel mosaic data to leverage its advanced capabilities in processing geometric and spatial inputs. This pretrained convolution kernel is then integrated as a fixed basis function within a larger GAM, alongside other significant covariates. The kernel is designed to efficiently extract pivotal features from the mosaic geographical profile data, serving as a basis—such as orographic complexity—in a linear model. This model assumes an additive relationship among covariate bases for predicting responses, which will be satisfied by our careful model setup based on the physical mechanism.

2. Data.

2.1. *CloudSat 2B-GEOPROF data product.* The CloudSat official data processing center (DPC) releases a hierarchy of data products. We download the 2B-GEOPROF data product that includes all the variables of our interests, convert the binary profiles into a tabular format within a database, selecting only high-quality, inland footprints for inclusion CloudSat Data Processing Center (2014). The sampled footprint profiles for the project is demonstrated in the top-right plot in Figure1, which spans over the whole earth inland randomly. Each footprint is characterized by attributes including latitude, longitude, surface reflectivity, and vertical gas attenuation path integral over the footprint's vertical column. The reflectivity is set to be the response variate of our model, the rests form the covariates. Geolocation data from these footprints enable the collocation of additional geographical covariates that we will use in the later parts. Also, due to the local continuity nature of the climate system, the proximity in geolocation among footprints leverage the autocorrelation of other unobserved covariates to enhance model accuracy. The gas attenuation data accounts for the influence of atmospheric conditions on surface reflectivity estimates, which helps refine our regression model by integrating crucial atmospheric effects. After the final data cleaning, the sample size for the project is 2,792.

2.2. *Digital Elevation Model.* According to classical radar theory, orography complexity significantly and nonlinearly impacts the surface reflectivity Skolnik (2001). In geosciences, digital elevation models (DEMs) are utilized to depict orography. A DEM is composed of mosaic raster data that subdivides a spatial window into multiple sub-pixels containing the elevation of that specific location. The top layer-2 left plot in Figure1 shows an example structure. This DEM data originates from the Shuttle Radar Topography Mission (SRTM) conducted by NASA, which offers a global perspective on surface elevation with a spatial resolution of 30 meters. Consequently, for each radar footprint, we compile approximately 1000 DEM pixels based on their geolocations, then condense them into a more manageable 5-by-5 mosaic format NASA JPL (2013).

2.3. *Land surface type.* Radar theory shows that different land types have varied reflectivity levels; for example, ice sheets and bare land reflect differently Skolnik (2001). To collocate the global land surface types to each footprint as a covariate, we use the MODIS/006/MCD12Q1 data product from NASA's MODIS instrument on the Terra and Aqua satellites. This dataset offers annual global land cover classifications into 17 categories according to the International Geosphere-Biosphere Programme standards, including sea ice, urban, water, etc. Despite its annual temporal resolution limits the tracking seasonal surface variations, such as sea ice melting, it's already the finest data source we can find to have a global coverage we require NASA LP DAAC (2020).

3. Methods. The general methodology workflow and model structures adopted in this project is depicted as Figure1. Each radar footprint serves as a data unit, with orography information represented by the combination of the mosaic DEM data and the surface type data, which is fed into the Convolutional Neural Network (CNN) basis function. Other covariates are organized in tabular form and input into the other traditionally structured basis functions. These basis functions are meticulously determined to ensure that the post-processed covariates indeed form an uncorrelated additive relationship contributing to the surface reflectivity estimate interpretable by the radar theory.

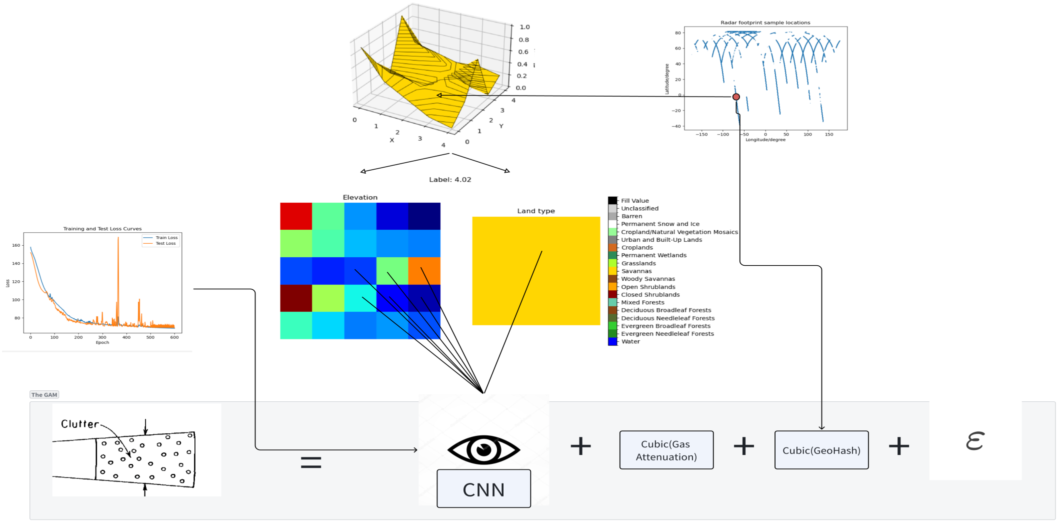

FIGURE 1. *Model Structure Overview. Layer 1: Dataset footprint locations (right) and DEM visualization for a sample footprint (left). Layer 2: CNN pretraining dataset; reflectivity labels (top), color-encoded DEM mosaic (left), surface type (right). Layer 3: GAM structure featuring surface reflectivity, CNN basis (with its pretraining loss curves), cubic spline for gas attenuation, Morton-encoded geolocation basis, and residuals*

3.1. *CNN pre-training.* Before using the GAM, we pretrain a CNN to extract complex orographic information from DEM mosaic data, recognizing the significant, non-additive influence of surface type on surface reflectivity. Ideally, adapting the GAM to condition the CNN's output on surface type would refine predictions, but would complicate the model and increase workload, suggesting a direction for future research.

To manage complexity, we integrate surface type as 17 one-hot binary layers in the CNN, ensuring unique representation for each class. This structure aids the CNN in learning orographic details pertinent to each surface class, assisting in accurate reflectivity predictions. The CNN design, featuring two convolutional layers with batch normalization, balances simplicity with the need to avoid underfitting and overfitting, optimized via cross-validation. This configuration uses full gradient descent with the batch size equating to the entire training dataset, avoiding dropout for simplicity.

Layer 2 in Figure1 illustrates the CNN input structure with a plot showing DEM variations and another displaying surface types, each transformed into a one-hot layer for CNN processing. This setup expects the CNN to effectively learn from the DEM to represent orography complexity.

3.2. *GAM construction.* After pretraining, we use the CNN as a fixed basis function $Conv(D_i, L_i)$, where D_i and L_i are the DEM mosaic and categorical land surface type of a footprint, respectively. To account for spatial relationships not captured by simple additive models, we employ Morton's encoding $M(x_i, y_i)$ to map geolocations into a scalar field, preserving geographical proximity, where (x_i, y_i) represents the latitude-longitude pair of a footprint, and M indicates the Morton encoding function. This approach allows us to exploit spatial autocorrelation for the unseen covariates effectively within our model.

Additionally, while the surface type has been encoded as one-hot channels within the CNN to differentiate the impact of orography complexity on surface reflectivity, the surface type itself inherently influences reflectivity. For instance, regardless of orography, open water will reflect differently compared to a snow layer. This effect is distinct from orography complexity, prompting us to also include it as a separate basis function in the GAM, outside of the CNN.

It's important to note that although the surface type appears twice within the GAM framework, it represents different impacts on the response. The nonlinear basis function within the CNN transforms these inputs into different representations, ensuring they do not exhibit collinearity within the model.

For encoding the surface type as a basis function, we use dummy variable encoding, similar to one-hot encoding, to convert it into multiple binary variables. This basis function is expressed as:

$$f(L_i) = \beta_0 + \sum_{c=1}^{17} \beta_c 1[L_i = c]$$

, where f is the dummy encoding function (i.e., the basis function), 1 is a boolean function.

We assume a nonlinear relationship between surface reflectivity and gas attenuation due to the absence of a specific model for electromagnetism-fluid interaction. To address this, we use a cubic spline basis $s(\mu_i)$ for smoothing, where s is the spline basis function, and μ_i is the gas attenuation.

Summarizing, the finalized structure of the GAM is given by:

$$K(D_i, L_i, x_i, y_i, \mu_i) = \alpha_0 + \alpha_1 Conv(D_i, L_i) + \alpha_2 M(x_i, y_i) + \alpha_3 f(L_i) + \alpha_5 s(\mu_i)$$

The model assumes residual normality, expressed as:

$$\sigma_0 \sim N(K(D, L, x, y, \mu), \epsilon^2)$$

and the absence of multicollinearity between bases. These assumptions will be verified through a series of diagnostic analyses. Although the first assumption might be challenged if the smooth penalty is overly stringent, it is generally expected to hold reasonably well, given that most of our basis functions—excluding the cubic spline and the CNN—are not excessively complex (i.e., not highly 'wiggling').

3.3. *GAM training.* To train the GAM model outlined in the previous section, we begin by constructing the design matrix using the values of the basis functions for the covariates:

$$B(X) = \begin{bmatrix} 1 & Conv(D_1, L_1) & M(x_1, y_1) & f(L_1) & s(\mu_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Conv(D_N, L_N) & M(x_N, y_N) & f(L_N) & s(\mu_N) \end{bmatrix}$$

, where $X \in R^{N \times 5}$ is the covariate matrix of the training set. Then, based on our assumption, the residual estimator $\sigma_0 - B(X)$ should follow a 0-centered normal distribution, where

$\sigma_0 \in R^N$ is the actual reflectivity vector. Hence, we train the GAM by minimizing such a loss function:

$$L(\alpha, \beta_f, \beta_s) = \|\sigma_0 - B(X)\alpha\|_2^2 + \lambda \alpha^T S \alpha$$

, where $\alpha \in R^6$ is the GAM basis coefficient vector, β_f, β_s are parameters for the cubic spline smoother and dummy encoding function, S is the smooth panelty matrix that avoids overfitting due to too high model variance.

The optimal hyperparameters are determined through a cross-validation process using a 10:1 training-to-validation ratio. Predictions and 95% confidence intervals are generated for the test set, based on the inference on the GAM coefficients and residual variance. To validate model assumptions, we conduct residual analysis using histograms and Q-Q plots, ensuring the assumptions regarding residual distribution are met, thereby confirming the model's validity and accuracy.

4. Results. CNN pretraining loss plot in Figure1 layer 3 shows the training and test loss curves for the CNN during pretraining, with both curves rapidly declining at the start and stabilizing after about 500 epochs, indicating near convergence and suggesting no overfitting as the performance on the unseen test set closely mirrors that on the training set. The CNN has effectively learned to represent orography complexity from DEM data for surface reflectivity modeling. Despite the CNN's R-square value of 20% seeming low, it's adequate considering it includes only two covariates and forms just a part of the broader GAM framework, which will further enhance modeling by incorporating additional information.

The GAM training result shows that the p-values for all basis coefficients were significantly less than 0.05, with the highest being 0.0113, confirming the statistical significance of all covariates as anticipated. The Pseudo R-squared for the GAM stands at 57.2%, which is considered acceptable accuracy in the context of satellite remote sensing regressions. Indeed, the smoothed regression plot on the test set, as shown in Figure2, confirms that the model successfully captures the overall linear trend. The prediction-real value scatterplot also illustrates a good consistency between the predictions and the actual values though a noticeable concentration of points around the true label line, evidenced by a correlation of 0.61. To showcase the model's capability to identify extremely low surface reflectivity nonlinearly, the bottom-middle plot in Figure2 reveals that sample footprints with very low reflectivity also tend to have low predicted values on a log scale. Conversely, the bottom-left exponential-scale plot demonstrates the model's ability to identify very high values nonlinearly, with the evidence that footprints with high reflectivity also tend to yield high predictions. Thus, the model captures both the correct linear trend and the nonlinear variability. An alation experiment is also conducted to quantify each covariate's influence. The strongest basis is the CNN, which drops the R-Squared by 44% by removing it and retrain the GAM, while removing other covariate combinations only drops the R-Squared by up to 12%. The residual histogram and QQ-plot in Figure2 depicts a normal distribution, defends the residual normality assumption of the model, which means the model doesn't suffer from significant systematic bias.

However, according to the smoothed fitted curve in Figure2, 95% confidence interval for the predictions, i.e., the green area, is too conservative to be informative. The predictive curve exhibits a flatter slope compared to the true value curve, indicating that while the model can distinguish between low-reflectivity and high-reflectivity surfaces, it tends to be conservative in predictions at the more extreme ends of the spectrum, which means that the prediction lacks enough variance that it should have had, either due to the powerlessness of the model itself, or the insufficient information exposed.

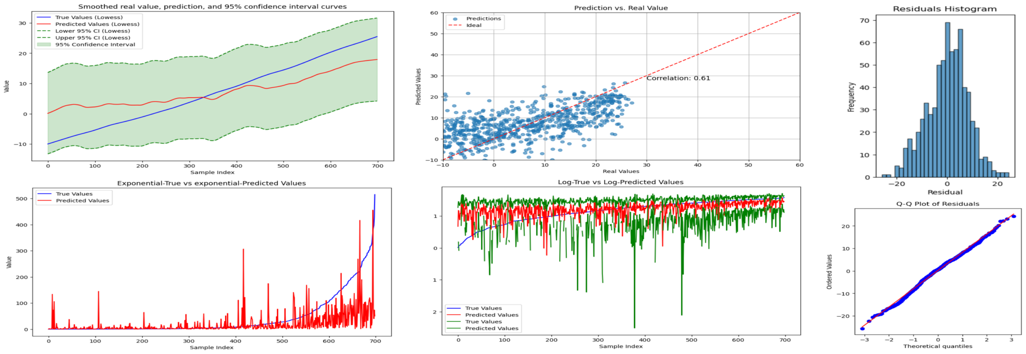

FIGURE 2. *GAM training results. Row 1: smoothed prediction/confidence interval curves, prediction-real value scatterplot, residual histogram (from left to right); Row 2: prediction/real value curves on exponential scale; prediction/real value curves on log scale; residual QQ-plot (from left to right)*

5. Conclusions. The results of the project provides another evidence that the orography complexity is the most deterministic geographical condition for the surface reflectivity. CNN can indeed represent orography complexity by learning more complicated features in DEM. The good performance of the model indeed validates our proposed idea of combining powerful but not-interpretable deep learning models with linear models based on physical mechanism to acquire model capability without losing interpretability. This approach also overcomes the computational inefficiencies and often intractable challenges associated with obtaining confidence intervals in traditional deep learning models, which typically involve a large number of parameters. This improvement is a significant highlight of our project, showcasing the potential of our methodology for application in satellite onboard algorithms. These algorithms demand not only accurate predictions but also reliable, real-time computation of confidence, making our method particularly valuable in practical, operational settings.

However, the model prediction shows some underfitting and too conservative confidence intervals. Since the model is flexible enough and has trained to reach the overfitting threshold, we expect that the lack of prediction variance is not due to the model complexity itself, but the limited covariate availability, which suggests that significant information crucial for modeling surface reflectivity may be missing. Indeed, due to the significant operational differences between various satellite sensors, e.g., the gaps in their spatial/temporal windows, highly limits the data availability for more complete set of covariates. In the future, we should work on yielding a more reliable model, and thus confidence intervals, by overcoming the underfitting problem through collocating a larger set of significant covariates. Additionally, differentiating models based on surface types is expected to satisfy the physical mechanism better and more interpretable than encoding it as a covariate, which is worthy of doing in the future.

REFERENCES

BATTAGLIA, D. KOLLIAS (2019). Spaceborne Cloud and Precipitation Radars: Status, Challenges, and Ways Forward. *Reviews of Geophysics* **58**.

CLOUDSAT DATA PROCESSING CENTER (2014). CloudSat 2B-GEOPROF Data Product. Accessed: 2024-08-02.

NASA LP DAAC (2020). MODIS/006/MCD12Q1 Land Surface Type Data Product. Accessed: 2024-08-02.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second ed. Springer.

NASA JPL (2013). Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM). Accessed: 2024-08-02.

PALERME, W. CLAUD (2019). How Does Ground Clutter Affect CloudSat Snowfall Retrievals Over Ice Sheets? *IEEE GEOSCIENCE AND REMOTE SENSING LETTERS* **16** 342–346.

SKOLNIK, M. I. (2001). *Introduction to Radar Systems*, Second ed. McGraw-Hill, New York.