



APPLICATION PAPER

# Informing synthetic passive microwave predictions through Bayesian deep learning with uncertainty decomposition

Pedro Ortiz<sup>1</sup> , Eleanor Casas<sup>2</sup>, Marko Orescanin<sup>1</sup> , Scott W. Powell<sup>3</sup> and Veljko Petkovic<sup>4</sup>

<sup>1</sup>Department of Computer Science, Naval Postgraduate School, Monterey, CA, USA

<sup>2</sup>Department of Earth Sciences, Millersville University, Millersville, PA, USA

<sup>3</sup>Department of Meteorology, Naval Postgraduate School, Monterey, CA, USA

<sup>4</sup>Cooperative Institute for Satellite Earth System Studies/Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA

**Corresponding author:** Marko Orescanin; Email: [marko.orescanin@nps.edu](mailto:marko.orescanin@nps.edu)

**Received:** 28 February 2023; **Revised:** 19 September 2023; **Accepted:** 24 May 2024

**Keywords:** Bayesian deep learning; infrared; microwave; uncertainty quantification

## Abstract

Space-borne passive microwave (PMW) data provide rich information on atmospheric state, including cloud structure and underlying surface properties. However, PMW data are sparse and limited due to low Earth orbit collection, resulting in coarse Earth system sampling. This study demonstrates that Bayesian deep learning (BDL) is a promising technique for predicting synthetic microwave (MW) data and its uncertainties from more ubiquitously available geostationary infrared observations. Our BDL models decompose predicted uncertainty into aleatoric (irreducible) and epistemic (reducible) components, providing insights into uncertainty origin and guiding model improvement. Low and high aleatoric uncertainty values are characteristic of clear sky and cloudy regions, respectively, suggesting that expanding the input feature vector to allow richer information content could improve model performance. The initially high average epistemic uncertainty metrics quantified by most models indicate that the training process would benefit from a greater data volume, leading to improved performance at most studied MW frequencies. Using quantified epistemic uncertainty to select the most useful additional training data (a training dataset size increase of 3.6%), the study reduced the mean absolute error and root mean squared error by 1.74% and 1.38%, respectively. The broader impact of this study is the demonstration of how predicted epistemic uncertainty can be used to select targeted training data. This allows for the curation of smaller, more optimized training datasets and also allows for future active learning studies.

## Impact Statement

In this study, Bayesian deep learning was used to create synthetic microwave (MW) brightness temperatures that deliver both the high spatial and temporal resolutions of infrared data and some information about the internal structure of clouds. Additionally, we provide the estimated variance of each predicted MW brightness temperature to help scientists and downstream users discern whether or not each predicted MW brightness temperatures are reliable. Furthermore, by decomposing variance into its aleatoric and epistemic components, scientists can discern whether predictions can be improved with additional training data (e.g., high epistemic uncertainty) or if there are inherent limitations of the model that warrant further scientific investigation into its utilization and potential reformulation (e.g., high aleatoric uncertainty). This methodology also demonstrates how regression models can utilize epistemic uncertainty predictions to actively select new training data that can lead to more optimized datasets and improved skill in future models.

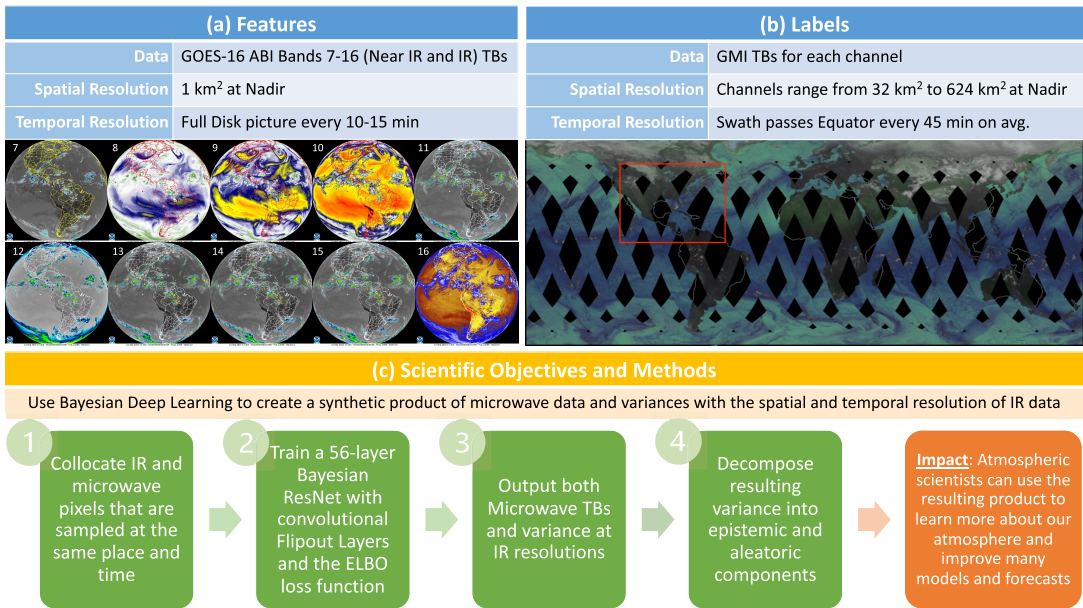
## 1. Introduction

To fully understand and simulate how the Earth's climate is changing, climate scientists need to understand end-to-end how energy is transferred into, within, and out of the Earth's climate system at all wavelengths and time scales (Loeb et al., 2009; Trenberth et al., 2009; Stephens et al., 2012). While much progress has been made toward closing this complete energy transfer “budget” since the development of weather satellites (Vonder Haar and Suomi, 1971; Dines, 2007; Trenberth et al., 2009; Stephens et al., 2012), one of the larger remaining sources of uncertainty is the budget of longwave radiation, which is strongly influenced by net latent heat exchange into or out of the atmosphere (Stephens et al., 2012). Earth's heat budget is partially dependent on precipitation, which is extremely challenging to measure but can be somewhat constrained using passive microwave (PMW) data. Furthermore, the concentrations of ice and liquid water as well as limited information about ice habit in clouds (Sun and Weng, 2012; Kroodsma et al., 2022) can be derived from microwave (MW) observations and also impacts radiative heat budget in the atmosphere.

Constraining information about atmospheric water in various phases is challenging because no single observing instrument can fully observe the distributions of each water phase over a large area at high temporal resolution. For example, infrared (IR) wavelengths measured by the Advanced Baseline Imager (ABI; Schmit et al., 2017) on Geostationary Operational Environmental Satellite (GOES)-16 provide continuous observations with high spatial resolution over a large area (e.g., Figure 1a). In clear air, ABI IR data are well-suited for providing information about water vapor concentration because its wavelengths have various sensitivities to absorption by water vapor and therefore provide information about water vapor concentration in multiple layers of the atmosphere. However, ABI IR observations in regions where thick, precipitating clouds occur are mostly representative of cloud top emissions. Thus, additional information provided by the longer, less scattered MW emissions detected by the Global Precipitation Measurement (GPM) Microwave Imager (GMI; Draper et al., 2015) are needed to provide information about liquid and ice structures in precipitating clouds (Petković and Kummerow, 2017). However, spatial resolutions for MW data are much coarser, and large gaps between the narrow data “swaths” are a consequence of the low-earth orbits of MW-detecting satellite platforms (Figure 1b). The low-earth orbits mean that clouds—despite their important impacts on Earth's energy budget and their prominent role in many extreme weather events—are persistently undersampled. This undersampling therefore imposes limitations on a wide variety of climate and weather applications, including models, forecasts, and scientific understanding (Trenberth et al., 2007; Stephens and Hu, 2010; Stephens et al., 2012; Petković and Kummerow, 2017; Pu et al., 2019).

As computational power has become more readily available over the past several years, deep learning has become increasingly utilized for addressing earth science research problems that have an abundance of data and require nonlinear mapping of inputs to output (e.g., Barnes and Barnes, 2021; Barnes et al., 2021; Foster et al., 2021; Guillaumin and Zanna, 2021). Deep convolutional networks or deep residual networks (He et al., 2016) are often used to facilitate this complex mapping. Moreover, such networks can also quantify uncertainty, either through the modeling of an additional output in the last model layer or through the use of dropout layers throughout the model (e.g., Gal and Ghahramani, 2016). However, when quantified without separating uncertainty into epistemic and aleatoric components, the advantages of having access to the uncertainty of each model prediction are not being fully leveraged. In contrast, Orescanin et al. (2021) and Ortiz et al. (2022) used Bayesian deep learning (BDL) to quantify and decompose the uncertainty in an earth science classification setting and proposed a methodology for applying this additional knowledge to make decisions about collecting additional input data, handling predictions with high uncertainty, choosing between several probabilistic models, and conducting targeted data analysis. Finally, the calibration between error and total uncertainty for different Bayesian architectures was explored in Ortiz et al. (2023).

In this study, we improve the regression model architectures presented in Ortiz et al. (2023) that used BDL to generate synthetic MW brightness temperatures ( $T_b^{mw}$ ) to deliver both the high spatial and temporal resolutions of IR data and information related to water below cloud tops (Figure 1c).



**Figure 1.** Executive summary of this study, where: (a) training dataset features are the brightness temperatures (TBs) measured by GOES-16 ABI Bands 7–16 (near IR and IR wavelengths) and are exemplified by full disk imagery from 17:40 UTC on September 12, 2022 (GOES Image Viewer, 2017); (b) training dataset labels are the microwave TBs from GMI, and are exemplified by the global, full-day GMI coverage on February 01, 2020, where shading denotes data coverage (Negri et al., 1989; EOSDIS Worldview, 2013), black denotes data gaps, and the red box denotes the domain in Figure 10); and (c) a flowchart outlines the methods and scientific impact.

Furthermore, whereas the prior models could only predict total uncertainty, the models presented in this study can have variance decomposed into its aleatoric and epistemic components. Through this uncertainty decomposition, we seek to help scientists discern whether predictive skill from our product can be improved with additional training data (e.g., high epistemic uncertainty) or if there is inherent atmospheric variability that warrants further scientific investigation into the utility of the model (e.g., high aleatoric uncertainty).

Although not yet demonstrated, a product containing both synthetic  $T_b^{mw}$  and decomposed uncertainty metrics in each prediction could then be utilized by atmospheric scientists to generate far-reaching, downstream impacts wherever existing MW data are currently being used in models, forecasts, and scientific analysis. For example, MW data are commonly used to: (1) retrieve precipitation rates (e.g., Adler et al., 1994; Cecil, 2007); (2) investigate moisture and precipitation evolutions (e.g., Wang and Hanks, 2016; Hristova-Veleva et al., 2022); (3) estimate tropical cyclone intensity (e.g., Yang et al., 2014; Olander and Velden, 2019); and (4) improve initial conditions through data assimilation in numerical weather prediction models. By predicting both MW brightness temperatures and uncertainties, future studies could then use Bayesian optimal estimation and assimilation techniques that require uncertainty quantification to investigate whether utilizing ML predictions produces reasonable values in derived products and therefore establish trust in the synthetic MW product. Additionally, directly predicting MW brightness temperatures and uncertainties allows for multiple applications stemming from one set of models, and forecasters could also directly use the MW brightness temperatures and uncertainties to diagnose convective structures in near real-time.

Ultimately, with this new synthetic product, one would be able to investigate weather systems using synthetic MW data and uncertainties that has unprecedented temporal (10–15 min) and spatial resolution

(2–4 km). This would allow for near real-time analysis and characterization of high-impact, fast-evolving weather systems and their convective structures, such as tropical cyclones. For comparison, current GMI MW data swath typically only partially sample a tropical cyclone 3–4 times daily. Additionally, weather events that GOES-16 observed in the past can also be analyzed in future case studies. While maximizing the synthetic MW product's skill will ultimately require more careful spatiotemporal co-location of ABI and GMI data (e.g., correcting the ABI data for parallax and imposing a more stringent time difference requirement between ABI and GMI scans), the current article demonstrates a pathway toward eventually generating a reliable, globally available synthetic PMW product that can enhance current and prior IR observations in locations where no PMW data exist.

Therefore, to work toward these long-term goals, this study has the following three primary objectives:

1. Establish a baseline of the relative differences in skill and decomposed uncertainties for predictions at all GMI MW frequencies using a consistent training methodology (Sections 3.1 and 3.2).
2. Introduce a method for identifying optimal, additional training data by leveraging predicted epistemic uncertainty (Section 3.3).
3. Assess the impacts of the new data selection method for a relatively small addition of training data to better understand the implications of uncertainty decomposition (Section 3.4).

By demonstrating the utility of our methodology in the prototype models presented herein, we conclude with discussion on how our results could be extended into an automated active learning framework that would ultimately utilize smaller and optimized training datasets in the future.

## 2. Data and methods

### 2.1. Data and dataset construction

This study used data collected by GPM GMI and GOES-16 ABI sensors during January–June 2020. We used the 1st, 11th, and 21st days in each month for the validation set; we used the remaining days as a training set. We withheld 12 days from July and August as test sets, using days from the beginning (4th–7th), the middle (15th–18th), and the end (26th–29th) of each month. A 4-day separation between the training dataset and the two withheld test datasets ensure that the two datasets are independent and not linked to the same synoptic weather patterns.

To create our dataset, we follow the general collocation methodology of Ortiz et al. (2023) by first selecting 10% of all GMI swaths (every tenth file from an S3 bucket) during the study period, which kept the training dataset manageable given the computing resources. Next, we temporally matched ABI and GMI observations by choosing the ABI disk file with the closest end time to each GMI pixel. Because GOES-16 does not scan the entire disk at once, co-located ABI and GMI data could be offset in time by as much as 15 min, which was 1.5 times the temporal frequency of ABI data collection. Then, ABI and GMI observations were spatially collocated by labeling  $39 \times 39$  pixel ABI patches from Bands 7–16 (near IR and IR wavelengths) with a single GMI  $T_b^{mw}$  corresponding to the center location of the ABI patch, such that independent BDL models with identical ABI patch training datasets are trained for each of the 13 GMI channels. The location of ABI data was not corrected for parallax, meaning that in deep clouds observed not at nadir, there is some offset between the reported latitude/longitude of the data and the coordinates from where the radiation is actually upwelling. This co-location error will generally increase as distance from the GOES-16 nadir point increases. Additionally, any records containing an ABI patch with a pixel over land, containing any missing data, or too close to the limb of the ABI viewing disk (west of  $140^\circ\text{W}$  or otherwise within 20 data pixels from the edge of the available data) were discarded, as in Ortiz et al. (2023). Land pixels were removed from training and predictions to constrain effects from additional sources of uncertainty related to highly variable surface emissivities. Thus, the training dataset is composed of ocean-only ABI patches within the GOES-16 viewing disk that are not at the limbs.

**2.2. Bayesian deep learning (BDL)**

BDL assumes distributions over weights,  $\mathbf{w}$ , of a neural network. If  $p(\mathbf{w})$  is a prior distribution over  $\mathbf{w}$  and  $p(\mathcal{D})$  is a distribution over the training data, then the goal of a BDL is to quantify a posterior distribution over the network weights conditioned on the distribution of the training data,  $p(\mathbf{w}|\mathcal{D})$ . For high dimensional problems, the calculation of  $p(\mathbf{w}|\mathcal{D})$  has no closed formed solution making it intractable (Blei et al., 2017).

We use variational inference to approximate  $p(\mathbf{w}|\mathcal{D})$  via the evidence lower bound (ELBO; Blei et al., 2017) by implementing a 56-layer Bayesian Residual Network (ResNet) version 2 (He et al., 2016) model architecture, using convolutional Flipout layers (Wen et al., 2018). This model architecture performs heteroscedastic regression, where the model outputs the parameters of a Gaussian distribution,  $\mu$  and  $\sigma^2$ , using a pre-built distribution layer from the Tensorflow Probability library version 0.12.1 (Dillon et al., 2017). Both the prior distribution and the posterior distributions are initialized using the Tensorflow Probability default settings of  $\mu = 0.0$  and  $\sigma = 1.0$  for the prior and of  $\mu = 0.0$  and  $\sigma = 0.05$  for the posterior.

Since the trained models have distributions over weights, we use Monte Carlo integration to conduct model inference (Filos et al., 2019; Feng et al., 2021), sampling from the weight distribution  $T$  times and averaging over the output values. Following the approach described in Filos et al. (2019), we chose  $T = 100$  to calculate  $\hat{y}_n$ . As the output of our regression models represents the parameters of a Normal distribution, we calculate  $\hat{y}_n$  as a mixture of Normal distributions (Lakshminarayanan et al., 2017) with uniform mixture weights:

$$\hat{y}_n \sim \frac{1}{T} \sum_{t=1}^T \mathcal{N}(\mu_t(x_n, \mathbf{w}_t), \sigma_t^2(x_n, \mathbf{w}_t)) = (\hat{\mu}_n, \hat{\sigma}_n^2), \tag{1}$$

where  $\frac{1}{T} \sum_{t=1}^T \mu_t(x_n, \mathbf{w}_t) = \hat{\mu}_n$  is the model’s predicted  $T_b^{mw}$  based on the  $T_b^r$  input features. Following the law of total variance, we calculate the variance of the  $\hat{y}_n$  mixture distribution as

$$\hat{\sigma}_n^2 = \frac{1}{T} \sum_{t=1}^T \underbrace{\hat{\sigma}_t^2(x_n, \mathbf{w}_t)}_{\text{Aleatoric Uncertainty}} + \underbrace{\hat{\mu}_t^2(x_n, \mathbf{w}_t) - \hat{\mu}_n^2}_{\text{Epistemic Uncertainty}}, \tag{2}$$

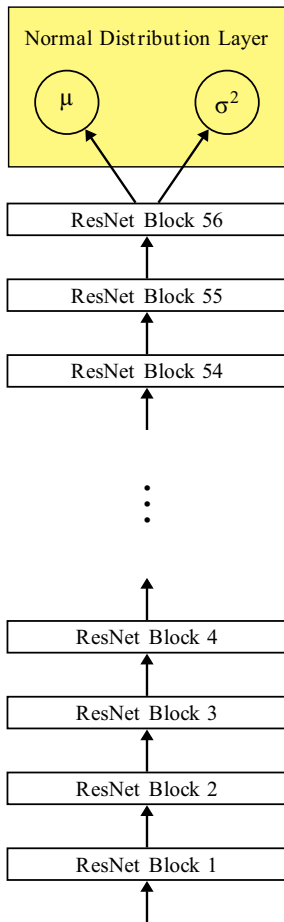
where  $\hat{\sigma}_t^2(x_n, \mathbf{w}_t)$  and  $\hat{\mu}_t(x_n, \mathbf{w}_t)$  are the outputs of the last layer of the model (heteroscedastic regression).

Aleatoric uncertainty is irreducible (Kiureghian and Ditlevsen, 2009), and epistemic uncertainty can be reduced given enough data (Kendall and Gal, 2017).

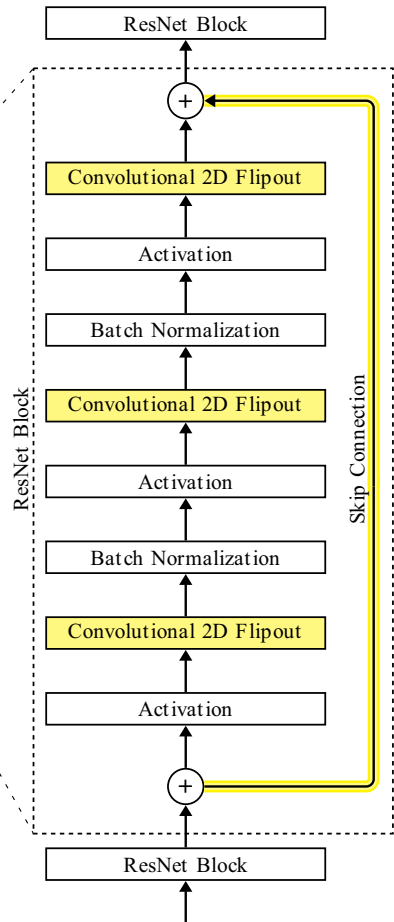
**2.3. Model architecture and training methodology**

Our BDL models build off of the prior architecture reported in Ortiz et al. (2023) to quantify both aleatoric and epistemic uncertainty (see Eq. (2)), with user-specified hyperparameters remaining consistent between the two studies. The model architecture used for this study is shown in Figure 2a. It is a 56-layer ResNet with an output layer that explicitly models a Normal distribution (highlighted in yellow in Figure 2a). The defining feature of ResNets is the skip connection (highlighted in yellow in Figure 2b), which allow for easier training of deep networks that generalize better (He et al., 2016). For each prediction, the distribution variance ( $\sigma^2$ ) is the aleatoric uncertainty of the predicted  $T_b^{mw}$  value ( $\hat{\sigma}_t^2(x_n, \mathbf{w}_t)$  in Eq. (2)). Quantifying the heteroscedastic aleatoric uncertainty in this way using deterministic models was previously demonstrated to lead to better performance on geoscience regression tasks (Barnes et al., 2021; Gordon and Barnes, 2022). The distribution mean ( $\mu$ ) is the predicted  $T_b^{mw}$  value ( $\hat{\mu}_t(x_n, \mathbf{w}_t)$  in Eq. (2)). To implement the BDL probabilistic framework into these models, convolutional two-dimensional Flipout layers are used inside each ResNet block (highlighted in yellow in Figure 2b) instead of deterministic convolutional two-dimensional layers. Since the layer weight distributions are sampled  $T$  times for each input to make  $T$  predictions (see Eq. (1)), the variance of the predicted  $T_b^{mw}$  value of a single input provides quantification of the epistemic uncertainty. The addition of epistemic uncertainty through BDL can help identify out-of-distribution samples (concept drift) (Ortiz et al.,

a) 56-Layer ResNet with Normal Distribution Output



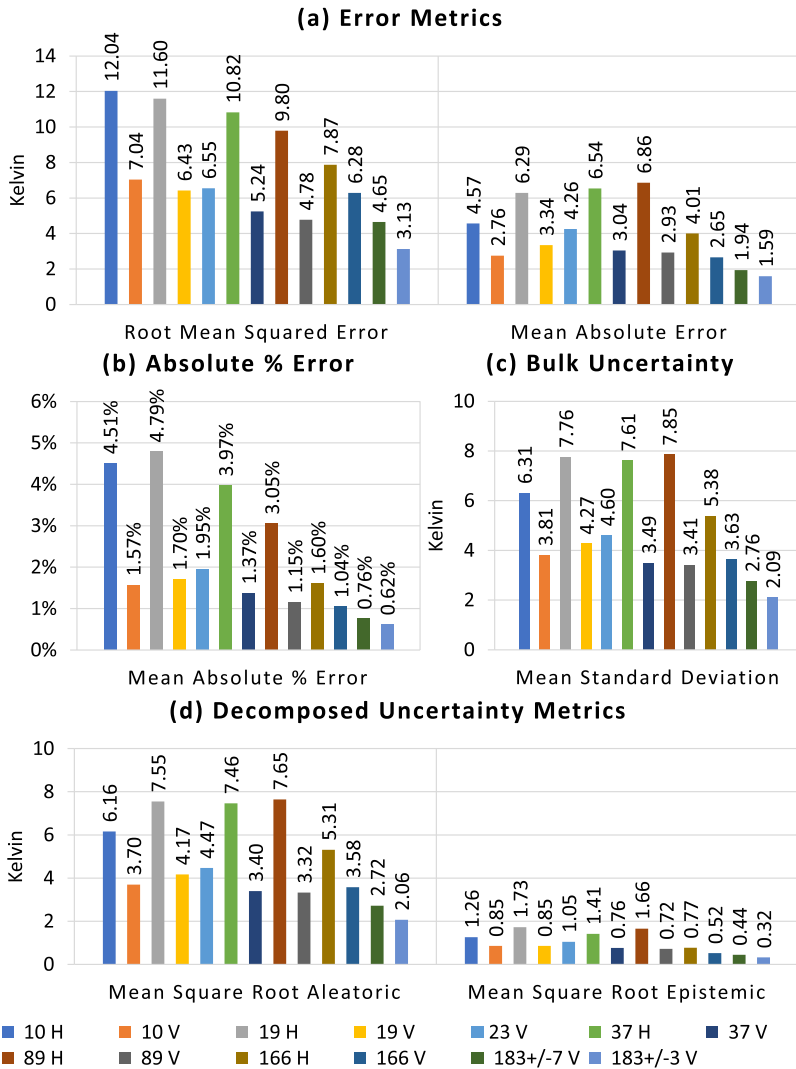
b) Bayesian ResNet Block



**Figure 2.** Network architecture. (a) 56-block Residual Network (ResNet) with output modeled as a Normal distribution (highlighted in yellow). (b) Internal structure of each ResNet block depicted in panel (a). The skip connection (highlighted in yellow) inside each block is the defining characteristic of a ResNet architecture. The convolutional 2D Flipout layers (highlighted in yellow) implement variational inference, making this a Bayesian model architecture.

2022) and data distributions that would be beneficial for additional training to improve a model's skill on a given task.

We maintained consistency with our prior work (Ortiz et al., 2023) in our training approach. Model weights followed He initialization (He et al., 2016), utilizing the Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of 0.001. We monitored validation loss for learning rate annealing (Li et al., 2019) and implemented a 4x learning rate reduction if no improvement occurred over five consecutive epochs. Early stopping (Goodfellow et al., 2016) was applied to prevent overfitting, and training could have continued up to 500 epochs if early stopping criteria were not met. Our Bayesian models used a batch size of 2,048 and were trained on 4 NVIDIA RTX 8000 48GB GPUs.



**Figure 3.** Results using models trained on 26,587,000 samples from January to June 2020 to generate synthetic GMI data using the July test dataset of 1.63 million samples. H = horizontal polarization; V = vertical polarization.

### 3. Results and discussion

#### 3.1. Initial model performance on July test set

Together, the BDL models are first used to predict synthetic GMI  $T_b^{mw}$  and uncertainties from the ABI  $T_b^{ir}$  July test set to establish a baseline skill for later comparison. Figure 3 shows the mean error and uncertainty statistics for the GMI channels predicted (see Appendix Table A1). The 0.62% and 0.76% mean absolute percentage error (MAPE) produced at  $183 \pm 3$  GHz and  $183 \pm 7$  GHz are examples of how well neural networks can learn any nonlinear function with a large enough training dataset (26.6 million examples). These models are expected to have the highest skill since, like the ABI Band 8 ( $6.2 \mu\text{m}$ ) in the input features, these GMI channels also primarily sense emissions from the upper troposphere due to strong water vapor absorption at these frequencies and wavelengths. In contrast, the 19 GHz H model has an MAPE of 4.79%. We also expect models for lower frequencies to not perform as well because emissions measured at these MW frequencies are less scattered and originate from lower in the

atmosphere than IR wavelengths. For the July test set, the MAPE and root mean square error (RMSE) trends of higher frequencies having less MAPE and RMSE than lower frequencies corresponds with the information content overlap between the ABI inputs and GMI frequency for which the model is predicting.

When model results for July 2020 are organized by polarization, several items are noted. In general, the MAPE and RMSE mostly decrease as frequency increases for each polarization. However, in the vertical polarization, 23 GHz predictions have the largest MAPE (Figure 3b), and 166 GHz has a larger RMSE than 37 GHz, and 89 GHz (Figure 3a). For horizontally polarized channels, the general error trends are similar, but average error magnitudes are larger because the innermost 95% of brightness temperature distributions at horizontal polarization are wider than those at vertical polarization for the same frequency (Appendix Figure A1). For mean absolute error (MAE), each polarization exhibits different error characteristics. For channels with vertical polarization, the MAE mostly decreases as frequency increases above 10 GHz with the exception of 23 GHz, which has the largest MAE (Figure 3a). For channels with horizontal polarization, the MAE peaks at the moderate frequency of 89 GHz (Figure 3a). Overall, the horizontally polarized error metrics have larger magnitude than the vertically polarized error metrics due to the fact that horizontally polarized channels have wider  $T_b^{mw}$ -value distributions than the vertically polarized channels (see Appendix Figure A1).

The mean standard deviation (MSD) provides a measure of total uncertainty, and unlike MAPE and RMSE, a primary benefit of MSD is that this metric provides additional information in a deployed setting that is not readily accessible with deterministic models because  $T_b^{mw}$ -value labels are not required to calculate the MSD of collected ABI data. Similar to RMSE, Figure 3c also shows that MSD generally decreases as frequency increases. Exceptions to this behavior for the vertical polarization are that predictions at 166 GHz have higher MSD than at 89 GHz and predictions at 23 GHz have higher MSD than at 19 GHz (Figure 3d); for both frequencies, this behavior is the same as with RMSE (Figure 3a). For horizontally polarized predictions, the relative pattern of MSD and RMSE are very similar, but predictions at 89 GHz have a comparatively high MSD that is slightly higher than 19 GHz and 37 GHz. Similar to MAPE and RMSE for the July test set, the MSD magnitudes also correspond with the information content overlap between the ABI inputs and GMI frequency for which the model is predicting.

When the total uncertainty is decomposed into aleatoric and epistemic uncertainty, different patterns emerge. First, aleatoric uncertainty represents the dominant source of uncertainty, such that the relative differences between frequencies for each polarization are the same between the mean square root aleatoric uncertainty (MSRA) and MSD (Figure 3c,d). However, while epistemic uncertainty contributes a small portion of total uncertainty (roughly an order of magnitude smaller), the relative pattern of mean square root epistemic uncertainty (MSRE) is similar to the MAE pattern for each polarization with 19 GHz H being the exception. Moreover, the quantified epistemic uncertainty provides a metric that indicates the relative benefit of providing each model with additional training data (i.e., the 19 GHz H model would benefit the most and the 183 ± 3 GHz V would benefit the least).

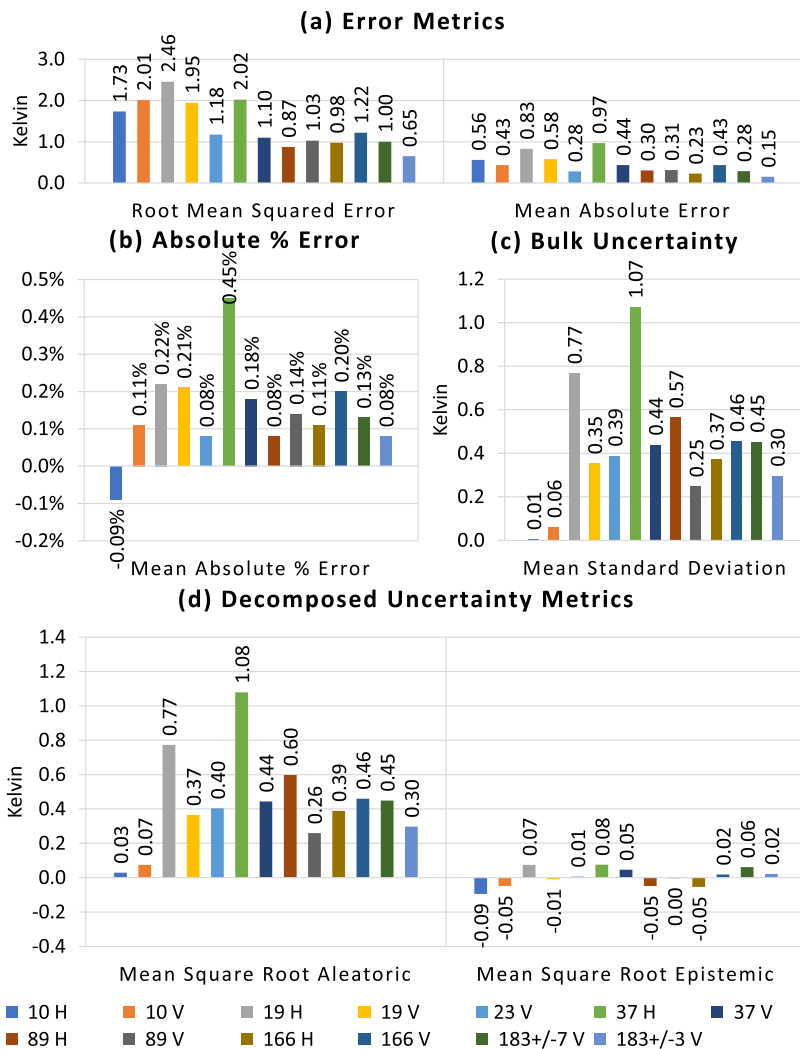
Together, comparison between the error and decomposed uncertainty metric characteristics indicate several relationships of interest. For example, the overall relationship between prediction error and uncertainty corresponds well with the overall atmospheric opacity at each frequency, such that higher frequencies are associated with higher opacity and higher skill (e.g., 166–183 GHz). This relationship is expected, because the weighting functions for brightness temperatures peak at higher altitudes for higher MW frequencies and are therefore more similar to the shape of weighting functions for IR wavelengths. Conversely, the lowest frequencies (19–23 GHz) have some of the highest amounts of error and aleatoric uncertainty, which is also expected because the weighting functions peak at lower altitudes for lower MW frequencies and are therefore less similar to the distributions of emission altitudes observed at IR wavelengths. Additionally, comparison of the middle frequencies (23–89 GHz) indicates that these frequencies have less extreme outlier error (due to atmospheric transparency at 23 GHz and higher-altitude peaks in weighting functions at 37 and 89 GHz), since the difference between MAE and RMSE is smaller. Finally, the middle frequencies appear to have the most potential for skill improvement with



additional training data, since these frequencies have larger MSRE relative to MSRA. However, before testing this hypothesis, predictions will be tested for potential over-fitting by examining model performance on temporally distant days.

### 3.2. Comparative model performance on August test set

One of the main challenges with training machine learning models is generalizing to unseen data because models tend to experience increased error when making model predictions on new data. Therefore, we tested the same models from Section 3.1 by making predictions on data from a different month (August 2020) to allow for the atmospheric state to be sufficiently decorrelated across multiple timescales to be considered “unseen” data. The overall results within this set of results remain consistent with the results described in Section 3.1 (see Appendix Table A2). Thus, we will instead focus on the differences between the results from July and August. Figure 4 reflects the change in error and uncertainty metrics (August



**Figure 4.** Change in metrics for the August test dataset of 1.53 million samples compared to July metrics in Figure 3. H= horizontal polarization; V= vertical polarization. Table of August metrics is in Appendix Table A2.

values from Appendix Table A2 minus July values from Appendix Table A1), where positive numbers reflect an increase for the August test set.

When the same models used in Section 3.1 are used to create synthetic MW data from the August test set, there is a less than 1 K increase in the MAE, ranging from 0.15 K to 0.97 K, and a less than 2.5 K increase in RMSE, ranging from 0.65 K–2.46 K. In relative terms, for the 13 GMI channels in this study, this translates into an increase of less than 0.45% MAPE. Since average model skill for each channel on the August test dataset is minimally different than July predictions, we conclude that these models are not likely to be overfitting to the training dataset and generalize to unseen data.

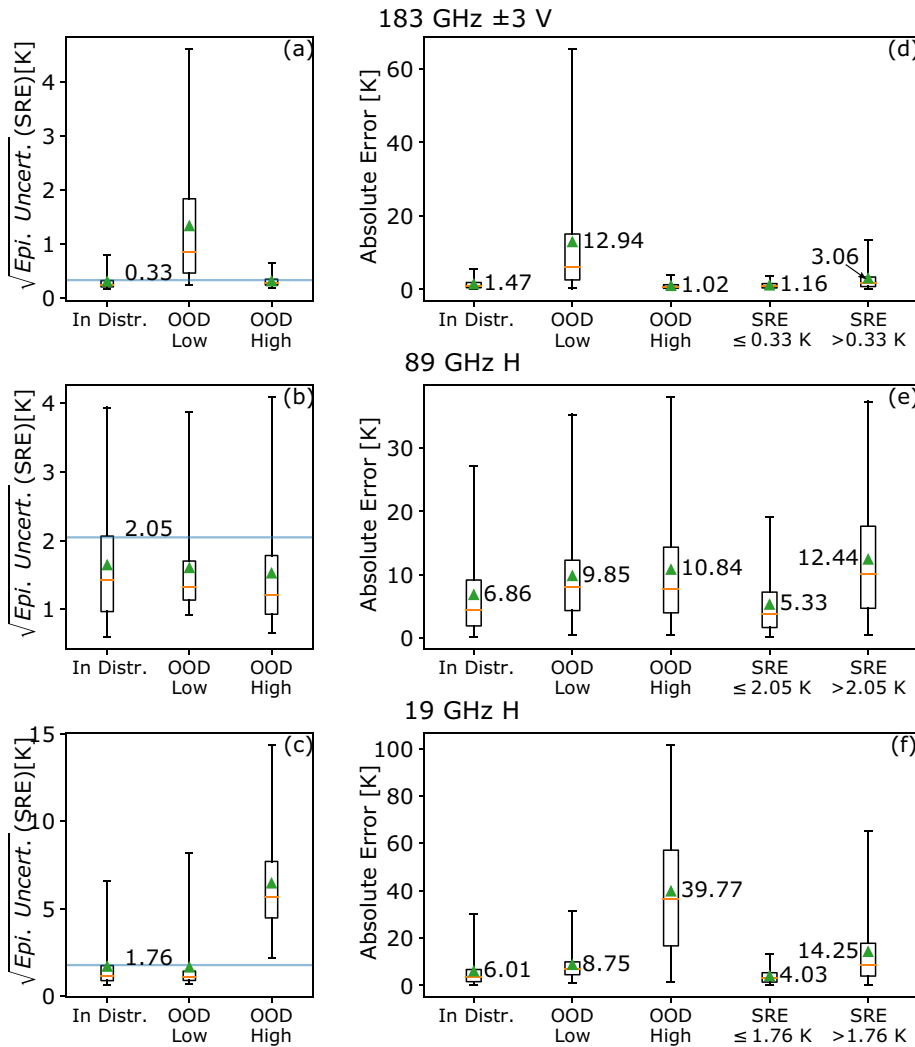
When comparing the uncertainty metrics (Figure 4c,d), there is a trend that the percentage increase in total uncertainty and aleatoric uncertainty is higher for the V channels than for the H channels. There is a corresponding trend between the two polarizations where the percentage increase in error is higher for the vertical polarization than for the horizontal polarization at each frequency. This is likely due to the fact that the H distributions contain a wider range of  $T_b^{mw}$  values than the V distribution for each frequency (see Appendix Figure A1). The width of the horizontal channel  $T_b^{mw}$ -value distributions make it more likely that a model has seen a sample with a similar  $T_b^{mw}$ -value; however, the width of the horizontal distributions also diffuses the  $T_b^{mw}$ -value distribution density such that they inherently have higher aleatoric uncertainty. Conversely, the vertical distributions are narrower, and these models have likely learned the center of the training distribution and are being presented with inputs that were either in the tails of the training distribution or outside of the distribution completely.

The change in the epistemic uncertainty from July to August enables some insights about which month contains data that would be useful for future training. The 10 GHz H, 10 GHz V, 19 GHz V, 89 GHz H, and 166 GHz H models produced predictions with less uncertainty for the August dataset than for the July dataset. For these models, training on additional data from July would be more beneficial than training on data from August. Conversely, the 19 GHz H, 37 GHz H, and the 23 GHz–183 GHz V models except 89 GHz V produced predictions with more uncertainty for the August dataset than for the July dataset (see Appendix Figure A2). For these models, training on additional data from August would be more beneficial than training on data from July. Since the 89 GHz V predictions had the same amount of epistemic uncertainty for both months, training on data from either month would be equally beneficial. These trends demonstrate one way the epistemic uncertainty of model predictions can be used to make decisions about whether or not to select particular data for additional model training.

### 3.3. Relationship between labels, epistemic uncertainty, and error

Since researchers are often limited by the availability of labeled data and of computing resources, it would aid the research process if the epistemic uncertainty could be used to identify the most beneficial samples for future training. In contrast to data with high aleatoric uncertainty, data with epistemic uncertainty can be used to reduce the model uncertainty. Selecting the data with the most epistemic uncertainty helps identify the data that also most likely produces the most error. Furthermore, by only training on the data that produces the highest epistemic uncertainty, any increase in the size of the training dataset can be limited to a percentage of any newly collected data.

The first step in leveraging quantified epistemic uncertainty to reduce model error is to identify the data for which each model produces predictions with the highest epistemic uncertainty. Figure 5 illustrates the relationship between the label distribution, epistemic uncertainty, and error. Figure 5a–c shows the relationship between the label distribution and epistemic uncertainty. In these figures, each sample in the July training dataset was categorized as either in-distribution or out-of-distribution (OOD). In-distribution samples had  $T_b^{mw}$  values that fell in the middle 95% of the data distribution (see Appendix Figure A1). All other samples were categorized as OOD. OOD samples were further sub-categorized as either being below the lowest in-distribution  $T_b^{mw}$  value (Low) or above the highest in-distribution  $T_b^{mw}$  value (High). The light blue horizontal lines denote the 75th percentile epistemic uncertainty value for all samples in the July dataset, which we will use as the threshold to identify high epistemic uncertainty predictions for the remainder of this work.



**Figure 5.** Box plots for square root epistemic uncertainty (SRE) and absolute error in Kelvin (K) for July data, where orange lines denote the median values and green triangles denotes the mean value for a given frequency. The whiskers span from 2.5% to 97.5%. (a–c) Each label ( $T_b^{mw}$  value) is categorized as “in distribution” if the  $T_b^{mw}$  belongs to middle 95% of the label distribution (see Appendix Figure A1). Out of distribution (OOD) labels are further divided into “low” and “high” by  $T_b^{mw}$ , where “low” means lowest 2.5% of  $T_b^{mw}$  values and “highest” means highest 2.5% of  $T_b^{mw}$  values. The blue horizontal line denotes the 75th percentile square root epistemic uncertainty (SRE) value for all labels. (d–f) The three leftmost box plots depict the prediction error using the label categorization scheme from panels (a) to (c). The two rightmost box plots depict the prediction error using the 75th percentile SRE value for categorization. The mean absolute error for each category is labeled in Kelvin.

The  $183 \pm 3$  GHz model produces the least amount of predictive error of all the models; however, calculating the predictive epistemic uncertainty may help to lower the amount of error even further. Selecting July data with epistemic uncertainty values greater than the 75th percentile (above the blue line in Figure 5a) for future training has the potential to help the model learn in two ways. First, this threshold identifies almost all OOD low samples and approximately 25% of the high OOD samples, which

correspond to the lowest density  $T_b^{mw}$  values for this GMI frequency (see Appendix Figure A1). In Figure 5d, the box plot for low OOD samples shows that these samples produce some of the largest errors. In addition to OOD samples, the 75 percentile of epistemic uncertainty aids in identifying the 25% of in-distribution samples for which the model parameters are most uncertain. Figure 5d shows that median, mean, and 97.5 percentile errors are all higher for samples with SRE greater than the 75th percentile. Selecting July data with epistemic uncertainty values greater than the 75th percentile for future training has the potential to help the 183 GHz model reduce errors by allowing it to train on samples that are currently outside of training distribution and that are inside the current training distribution but produce high epistemic uncertainty.

Similarly, for 19 GHz H, the 75th percentile for epistemic uncertainty helps identify almost all of the high OOD samples and approximately 25% of the low OOD samples (see Figure 5c). Again, this is useful as training the model on samples that are dissimilar to the original training dataset distribution may help decrease the model error. However, since all of the OOD samples only comprise 5% of a single month of data and less than 1% the size of the training dataset, any decrease from these samples alone would likely be minimal. If the data are already labeled, these samples could be selected by using the label value; however, there is still value to be gained from having the epistemic uncertainty. Again, for in-distribution samples, the 75th percentile helps identify samples that the model has already seen but for which it still had high model parameter uncertainty. Even with access to labels, identifying these samples as the most beneficial would be difficult and require more time than using a model to produce predictions. Using the 75th percentile epistemic uncertainty value to select new training data would likely benefit the 19 GHz H model in the same way as the  $183 \pm 3$  GHz V model.

In Figure 5f, the 75th percentile epistemic uncertainty value is used to illustrate how uncertainty relates to error. For 19 GHz H, the OOD samples have higher MAE (green triangles) and higher 97.5 percentile absolute error values. When using the 75th percentile epistemic uncertainty value to categorize the samples, the error trend is similar to the trend for in-distribution and OOD samples. For samples with epistemic uncertainty less than 1.76 K (75th percentile), the MAE is lower than for samples with epistemic uncertainty greater than 1.76 K, and the low epistemic uncertainty samples have a lower 97.5 percentile absolute error value than the high uncertainty samples.

Unlike the  $183 \pm 3$  GHz V and 19 GHz V models, the 89 GHz H model uncertainty identifies less than 25% of both types of OOD samples (see Figure 5b); however, the model uncertainty still identifies the 25% of in-distribution samples with the highest epistemic uncertainty. This is an indicator that this model has not learned the underlying training distribution as well as the other two models. This is likely due to the fact that 89 GHz H has the widest in-distribution  $T_b^{mw}$  value range of all GMI frequencies (see Appendix Figure A1). While this model would likely benefit for any additional training data, Figure 5e illustrates that training on sample with epistemic uncertainty greater than 2.05 K would be the most beneficial, as the median, mean, and 97.5 percentile error is higher than for samples with less than or equal to 2.05 K epistemic uncertainty. Using the 75th percentile epistemic uncertainty value to select new training data would likely benefit the 89 GHz model differently than the 183 GHz and 19 GHz models since fewer OOD samples have epistemic uncertainty greater than 2.05 K.

In all three cases, the highest epistemic uncertainty samples correspond to higher error than samples with lower epistemic uncertainty. Epistemic uncertainty is attributable to the uncertainty in the model weight distributions and is reducible with more data that is representative of the underlying training distribution. In theory, selecting additional training data from the July dataset that has epistemic uncertainty greater the 75th percentile for each GMI frequency should reduce the epistemic uncertainty and, consequently, the amount of error. The next section presents the results of applying this methodology.

### 3.4. Selective model updating

Since the results in Figures 3 and 4 show epistemic uncertainty greater than zero, none of the models have learned the entire underlying data distribution. However, as discussed in Section 3.1, higher MW frequencies (166–183  $\pm$  3 GHz) were associated with the least error and epistemic uncertainty, which

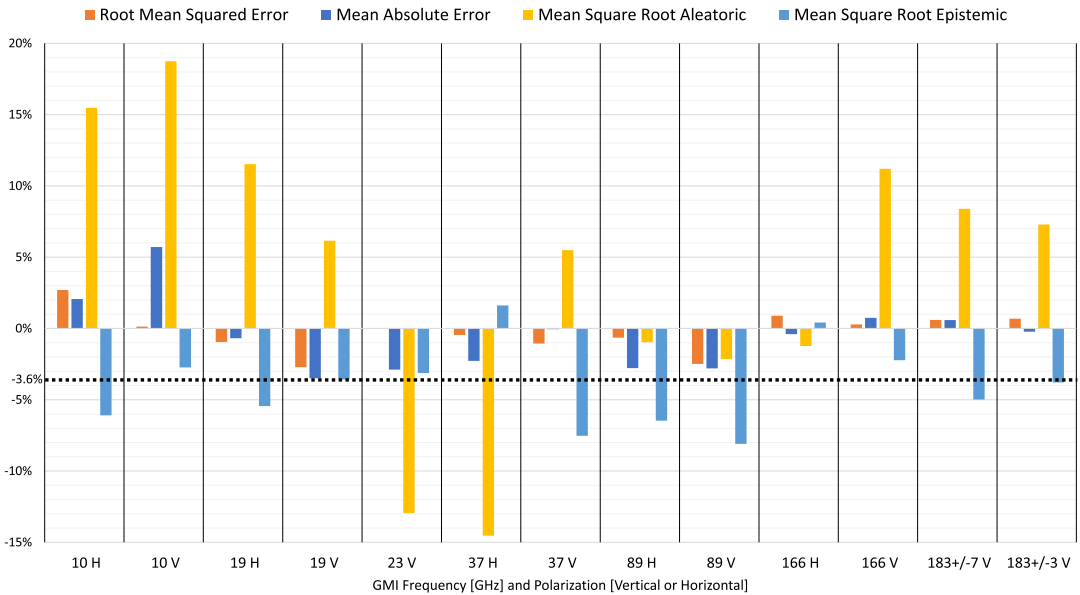
implies that additional training data would not provide a large improvement of skill. In contrast, the lower MW frequencies (19–23 GHz) were associated with a relatively large amount of error and aleatoric uncertainty, which implies that additional training data may not be as helpful at reducing error for the lowest frequencies. Finally, the middle MW frequencies (23–89 GHz) were associated with a moderate level of error and the highest fraction of epistemic uncertainty out of total uncertainty, which implies the middle MW frequencies could benefit the most from additional training data. Quantified epistemic uncertainty serves as a metric that enables insights as to how additional training data might affect the model performance at each frequency. Therefore, this section has two objectives: (1) test how the above noted relationships inferred from uncertainty decomposition are affected with additional training data, and (2) demonstrate how to use the quantified epistemic uncertainty of each model to help select the samples that are most likely to reduce the error without using 100% of the data.

As an illustrative case, the models from Section 3.1 were used to create synthetic GMI data for the entire month of July, and the per-pixel epistemic uncertainty was calculated as discussed in Section 3.3. Each model for each channel was then trained using the data from July that had uncertainty greater than or equal to the 75th percentile value of MSRE. By selecting the 25% of the data for which the model produces predictions with highest epistemic uncertainty, the new training data will contain examples of inputs with in-distribution  $T_b^{mw}$  values that produce predictions with higher uncertainty (and likely produce higher error), as well as examples of inputs that are outside of the previous training distribution (i.e., previously unseen or rarely seen examples). Because each frequency had a different model, the per-sample epistemic uncertainty varied for each GMI frequency. This resulted in 13 unique training sets, but containing the same number of samples (965,000).

Figure 6 shows the overall results for the changes of August predictions in response to updating the model weight distributions by training only on the identified high epistemic uncertainty July pixels (see Appendix Table A3). Training with this additional data resulted in a decrease in MAE for 9 of the 13 models (dark blue). The MAE of the 166 GHz V and  $183 \pm 7$  GHz V increased by 0.03 K (0.075%) and 0.01 K (0.59%), while the 10 GHz H and 10 GHz V models increased by 0.11 K (2.07%) and 0.18 K (5.71%). The models for frequencies from 19 GHz to 89 GHz exhibited a decrease in RMSE (orange), except 23 GHz model (no change). The models for frequencies above 89 GHz experienced less than 1% increases in RMSE, ranging from 0.28% to 0.89%, while the 10 GHz models had increases of 2.69% (horizontal) and 0.13% (vertical). In relative terms, increasing the amount of training data by 3.6% resulted in average decreases of 1.74% in MAE and 1.38% in RMSE for the channels where these metrics decreased. Further examination of the change in error confirms the hypotheses developed in Section 3.1. While the highest MW frequencies had a relatively small change in error with additional training data, the middle MW frequencies (19–89 GHz except 37 V) experienced a considerable skill improvement in both MAE and RMSE from a relatively small amount of additional training data. The lowest frequency (19 GHz) had the largest percentage decrease in RMSE and MAE of any channel for the vertical polarization, while the horizontal polarization had nominal decreases in both RMSE and MAE.

Examination of the decomposed uncertainty changes from training with additional high epistemic uncertainty also reveal interesting characteristics (Figure 6). While the highest MW frequencies had very minimal differences in error, they also had a more than 7% increase in MSRA (Figure 6, yellow bars). Interestingly, out of the highest frequency channels, the 166 GHz H model experienced a decrease in aleatoric uncertainty and an increase in epistemic uncertainty. In contrast, models for the lowest frequency (10 GHz) experienced decreases in epistemic uncertainty similar to the highest frequencies, yet it led to greater increase in aleatoric uncertainty and error. Finally, the middle frequencies (23–89 GHz except 37 GHz H) experienced reductions in epistemic uncertainty; however, the 37 GHz H model was the only model in this range to exhibit an increase in epistemic uncertainty. Four of the middle GMI channel frequencies (23 GHz V, 37 GHz H, 89 GHz H, and 89 GHz V) experienced decreases in aleatoric uncertainty, while 37 GHz V experienced an increase in aleatoric uncertainty. To further investigate, why these changes occurred, spatial plots from three example frequencies are examined.

Figures 7–9 show examples of results both before and after updating the models with additional training data for  $183 \pm 3$  GHz (vertical), 23 GHz (vertical), and 19 GHz (horizontal) in a domain

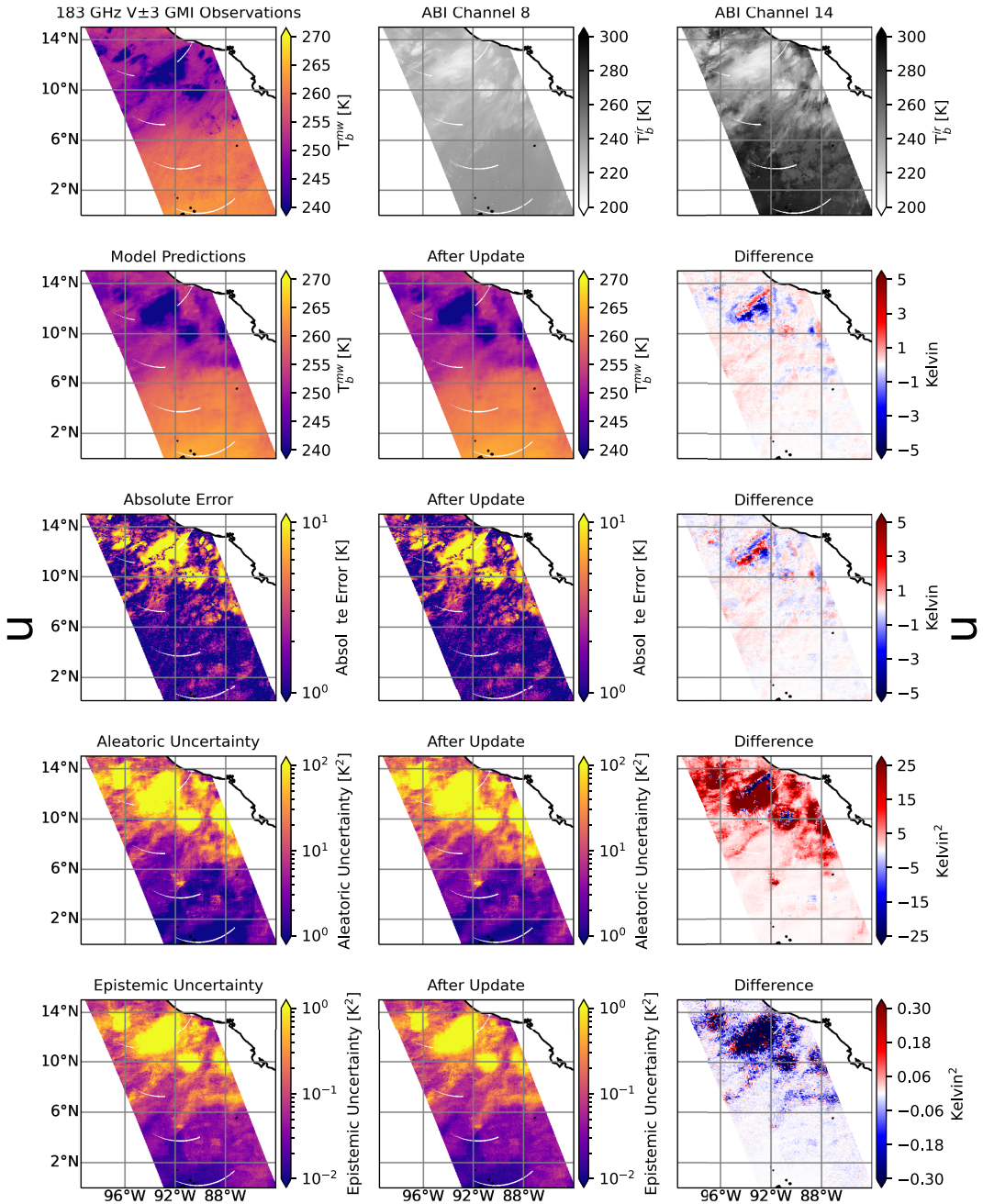


**Figure 6.** Percentage change in error and uncertainty due to updating the model by training on additional data from July that has uncertainty greater than the 75th percentile ([Appendix Table A3 values – Appendix Table A2 values]/Appendix Table A2 values). The dashed line at  $-3.6\%$  indicates a decrease proportional to the size of the growth in the training dataset.

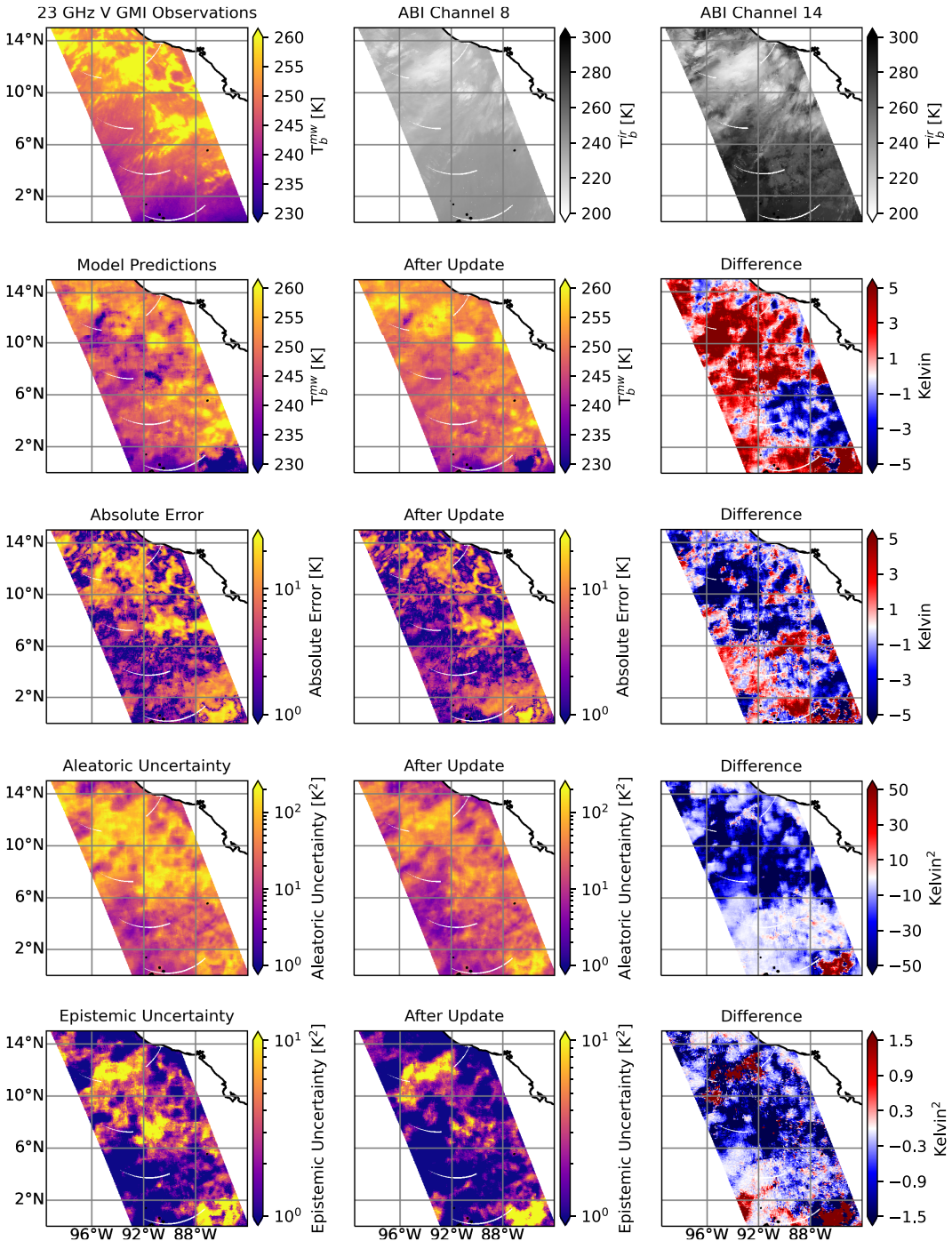
southwest of Mexico that was observed by GMI. The top row of each figure contains the GMI observations and the ABI observations for Bands 8 ( $6.2 \mu\text{m}$ ) and 14 ( $11.2 \mu\text{m}$ ) for a patch of both cloudy and clear conditions at 13:41 UTC on August 26, 2020. The first and second columns of the remaining rows depict the model predictions, the prediction absolute error, the prediction aleatoric uncertainty, and the prediction epistemic uncertainty, where the first column was produced using the model before the update and the second column was produced using the model after the update. The third column contains the difference between the values from column 1 and column 2, where red indicates that a metric increased after the model was updated and blue indicates that a metric decreased after the model was updated.

First, Figure 7 shows an example of how the spatial distributions of  $183 \pm 3 \text{ GHz}$  predictions changed in response to additional high epistemic uncertainty training data. Section 3.1 showed that this channel was already associated with the highest accuracy prior to additional training data, but there was still nonzero epistemic uncertainty. When additional training data are utilized, Figure 7 shows that the largest changes in the predicted  $T_b^{mw}$  and absolute error values occur in and around areas containing clouds, and the rest of the predicted  $T_b^{mw}$  values had changes in  $T_b^{mw}$  and absolute error less than 1 K. However, additional insight is provided through the uncertainty decomposition. With additional training data, epistemic uncertainty decreased the most in the largest area of convection in the same region where aleatoric uncertainty increased (roughly  $12^\circ\text{N}$ ,  $93^\circ\text{W}$ ). Additionally, decreases in aleatoric uncertainty are also associated with decreases in error, which implies that the model is better learning where the edges of the associated convection are located.

Figure 8 depicts the results for the 23 GHz (vertical) model and is an example of the desired result of updating the model using high epistemic uncertainty samples. In the cloudy region located between  $6^\circ\text{N}$  and  $10^\circ\text{N}$  and between  $92^\circ\text{W}$  and  $96^\circ\text{W}$ , the GMI was able to see below cloud top level where the ABI did not for the two channels depicted (8 and 14). In the pre-update model predictions, this area has scattered predicted  $T_b^{mw}$  values below 240 K, which are indicative of clouds for 23 GHz. After the model update, the predicted  $T_b^{mw}$  values below 240 K in this area almost completely disappear. Moreover, north of  $6^\circ\text{N}$  the model predictions increase in value, and south of  $6^\circ\text{N}$  the predicted values decrease, bringing the model



**Figure 7.** GMI  $183 \pm 3$  GHz (vertical) change in metrics due to model update. Row 1: GMI observations and the ABI observations for Bands 8 ( $6.2 \mu\text{m}$ ) and 14 ( $11.2 \mu\text{m}$ ) at 13:41 UTC on August 26, 2020. Rows 2–5: Model predictions, prediction absolute error, prediction aleatoric uncertainty, and prediction epistemic uncertainty. The first column was produced using the model before the update and the second column was produced using the model after the update. The third column contains the difference between the values from column 1 and column 2, where red indicates that a metric increased after the model was updated and blue indicates that a metric decreased after the model was updated.



**Figure 8.** Change in metrics due to model update as in Figure 7, but for GMI 23 GHz (vertical).

predictions in closer alignment with the observed GMI  $T_b^{mw}$  values than before the model update. This is evident when comparing the absolute error before and after the update as there are fewer areas with high error (yellow), particularly between 6°N and 10°N and between 92°W and 96°W. Unsurprisingly, the decrease in this area correlates with the highest decrease in epistemic uncertainty. What is perhaps



unexpected is the changes in aleatoric uncertainty, which is irreducible noise in the data. Before the model update, the bright yellow areas on the aleatoric uncertainty plot indicates that the uncertainty (and associated error) of the model cannot be reduced with more data. However, since these models are Bayesian and not deterministic, the additional information of higher epistemic uncertainty indicates that in fact there may be little aleatoric uncertainty if more training data similar to these inputs can be provided to the model. And, the aleatoric uncertainty for the post-update model has fewer bright yellow pixels. For this reason, an uncertainty analysis without measuring both components, aleatoric and epistemic, may lead to erroneous conclusions.

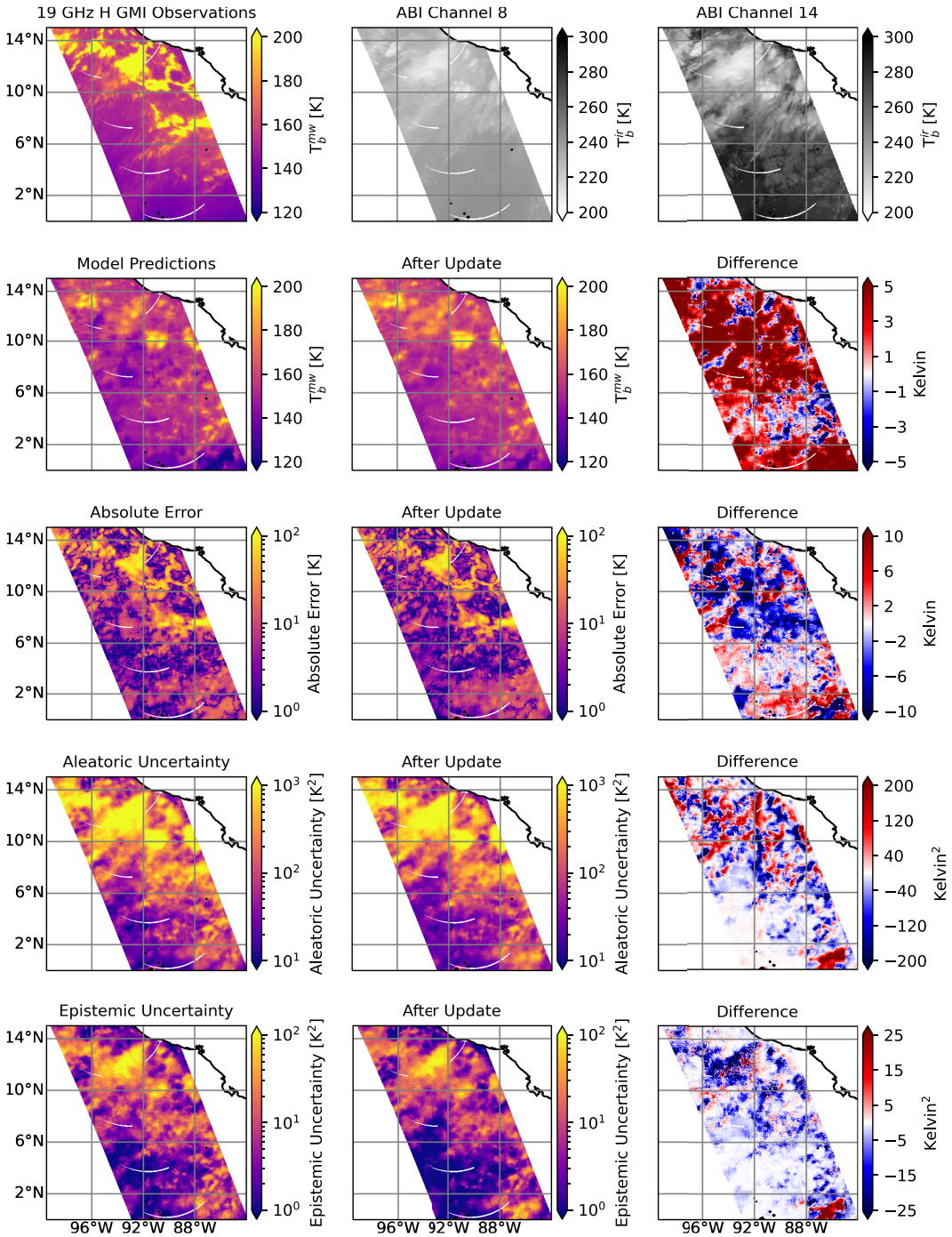
Figure 9 is an example of a low MW frequency (19 GHz horizontal), which was found to be associated with a large fraction of aleatoric uncertainty because this channel is the channel with the least covariance between 19 GHz labels and IR input features in clouds. The greatest decrease in absolute error occurs in the dense cloudy region located between 10°N and 14°N and between 92°W and 96°W, where the model learns that this region should be warmer than previously predicted. However, not all precipitation is predicted by the model, such as the elongated yellow GMI feature from approximately 90°W to 86°W and 7°N to 9°N. Examination of the ABI Bands 8 and 14 indicate that this precipitation is likely shallow and partially obscured by thin cirrus. However, while mean  $T_b^{mw}$  predictions did not capture this feature, aleatoric and epistemic uncertainty is quite high in this region, indicating that the model has identified that precipitation is possible in this area. Finally, epistemic uncertainty changed minimally after the model update, but absolute magnitudes indicate that the model can learn to further warm the warmest  $T_b^{mw}$  associated with heavy precipitation with additional training data.

Together, the example spatial distributions in Figures 7–9 provide context to the overall changes shown in Figure 6. First, the increased mean aleatoric uncertainty was found to be predominately occurring in or around areas of precipitation across frequencies, which suggests that the model is better learning the full potential distribution of  $T_b^{mw}$  distributions. However, decreases in aleatoric uncertainty also occurred, which typically corresponded with noticeable decreases in absolute error. Third, aleatoric and epistemic uncertainty can also be used to identify additional regions of interest that may not be identifiable in mean  $T_b^{mw}$  predictions. Finally, epistemic uncertainty continues to be high in clouds where model skill is currently weakest, which suggests that even more training data can further reduce total uncertainty and error.

#### 4. Conclusion

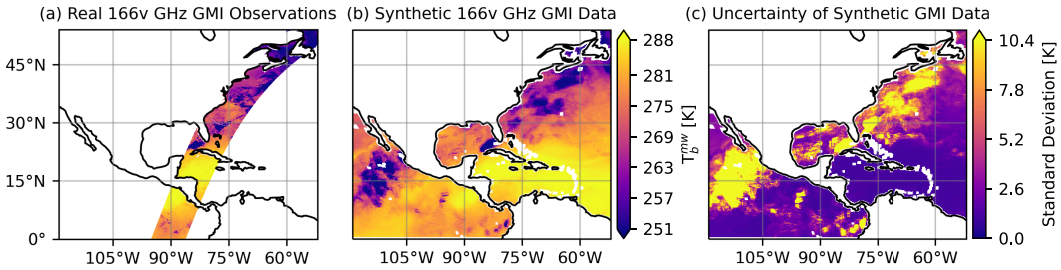
IR, visible, and MW data collected for observing the atmosphere have different strengths and limitations. IR and visible radiances are captured from geostationary orbit, meaning that observations are nearly continuous over a large domain; however, they contain little information about clouds below their tops. MW data, in contrast, can provide more information about internal cloud structure but does not have the continuous, full-disk coverage of geostationary data. We have demonstrated herein the ability to: (1) predict over-ocean synthetic MW brightness temperatures and their uncertainties for all GMI channels using IR radiances obtained from geostationary orbit (Example of 166v GHz in Figure 10); (2) apply a Bayesian technique to provide estimates of both aleatoric and epistemic uncertainty associated with each individual uncertainty prediction; and (3) utilize epistemic uncertainty predictions to identify additional, optimal training data that improved new models in accordance with the expectations inferred from the initial aleatoric and epistemic uncertainty prediction characteristics.

By training models, both before and after the addition of selected training data for the majority of frequencies observed by GMI, we found several results. First, all models performed well in non-cloudy regions, which was expected considering that maximum information overlap exists between IR radiances and MW brightness temperatures given the sensitivity to water vapor shared by both IR and MW radiation. In cloudy areas, error is still large for MW frequencies that are sensitive to scattering by cloud hydrometeors (see, e.g., 37 GHz and 89 GHz in Appendix Table A1). Additionally, lower frequencies that are less sensitive to absorption by water vapor tended to be predicted with larger error as the models



**Figure 9.** Change in metrics due to model update as in *Figure 7*, but for GMI 19 GHz (horizontal).

struggled to make predictions at the tails of observed brightness temperature distributions (i.e., cloudy areas). Together, these results are generally consistent with Ortiz et al. (2023), but the average errors associated with the models in this study are lower due to utilizing a larger training dataset.



**Figure 10.** (a)  $\sim 15$  min of GMI observations. (b) Synthetic  $T_b^{mw}$  generated from 15 min of ABI data (GPM orbit number 33679 at 14:40 UTC on February 1, 2020) corresponding to the box in Figure 1. (c) Standard deviation of each  $T_b^{mw}$  in panel (b).

Nevertheless, the key finding of this study is that further scientific understanding can be derived from uncertainty decomposition. Without the context of the decomposed uncertainty metrics, the channels associated with the highest error metrics, such as 89 GHz H, could be interpreted to mean that our network architecture is not suitable for this regression task. However, the initially high, model-predicted mean square root of epistemic uncertainty (MSRE) for these channels indicated that the error metrics for these frequencies could be further reduced by training on additional data because the MSRE for frequencies below 183 GHz ranges from 2 to 6 times higher than the model with the lowest error. Furthermore, we also demonstrate that the presented square root epistemic uncertainty data selection method presented herein is robust for differently skewed label distributions, such that the selected data consistently indicates that the incorporation of more cloud data relative to the properties of each channel would provide the most benefit in future models. Finally, by training new models on both the original plus new selected data, we found that frequencies associated with higher initial epistemic uncertainty improved more than those with lower initial epistemic uncertainty. Thus, we find that BDL with uncertainty decomposition is a promising technique for increasing understanding of model predictions for our regression task.

The results from executing a targeted model update informed by the uncertainty decomposition (e.g., Figure 6) suggests that it may be possible to reduce the size of the training dataset by using the model uncertainty as a criteria to select new training data. When this process is conducted in an iterative manner, it is referred to as active learning (Settles, 2010), and Bayesian deep active learning has proven effective for both classification and regression computer vision tasks (Kendall and Gal, 2017). However, active learning has yet to be demonstrated on a large-scale regression task in the Earth Sciences. Applying an active learning framework to this regression problem will likely involve ablation studies to determine the optimal percentage of newly acquired data to use for the next round of training, the optimal interval for acquiring new training data (e.g., daily, weekly, monthly, etc.), and a method for incorporating the existing model weight distributions as a prior distribution. If active learning can successfully be applied, then it would reduce the data storage requirements and the model training time; moreover, it would automate the training data selection process, allowing researchers to focus on other tasks.

While epistemic uncertainty remains the primary focus of this study due to its utility for future active learning applications, value can also be gained from aleatoric uncertainty predictions. First, we find that predicted aleatoric uncertainty tends to be higher in regions associated with extreme brightness temperatures in true GMI data, even if model predictions of mean brightness temperatures are not extreme. Thus, we find that aleatoric uncertainty shows promise for identifying features like precipitation in our synthetic MW product. Furthermore, comparison of aleatoric uncertainty predictions before and after data addition can highlight regions where a model's epistemic uncertainty was too high to produce trustworthy predictions. Since aleatoric uncertainty does not decrease with additional training data when epistemic uncertainty is sufficiently low, the presence of substantial decreases in predicted aleatoric uncertainty with additional training data indicates that the original model had not sufficiently learned enough of the underlying label distribution to be considered trustworthy. An example of this scenario is clearly seen in

the 23 GHz V predictions in Figure 8. In contrast, the minimal changes in aleatoric uncertainty for  $183 \pm 3$  GHz in Figure 7 indicate that epistemic uncertainty is sufficiently low and that predictions from this model are relatively more trustworthy.

To further reduce aleatoric uncertainty after epistemic uncertainty is sufficiently low, changes must be made to the model experimental design, such as including more input features, changing model architectures, or simply better co-locating ABI and GMI data. Specifically, there are several sources of uncertainty that we hope to reduce in future iterations of this synthetic MW product. First, training dataset collocations between the center pixels of IR patch input features and MW pixel labels are associated with very different viewing angles and are currently not parallax-corrected, which likely would result in better spatial co-locations, reduced label variance, and predicted uncertainties for low-latitude and low-altitude cloud tops compared to high-latitude and high-altitude cloud tops. In addition to parallax corrections, future work will also investigate whether adding additional input features (such as visible radiances and ABI viewing angles per pixel), refining the temporal collocation technique, or switching to more efficient model architectures (such as U-Nets) improves model performance.

Looking forward, as we continue to utilize our uncertainty decomposition methodology to improve the predictive skill of  $T_b^{mw}$  from  $T_b^r$ , we hope to provide a highly useful synthetic product for atmospheric scientists. The extreme boost in both spatial and temporal resolution provided by our synthetic product may allow atmospheric scientists to perform previously impossible analyses in the MW spectrum, such as how individual oceanic storms contribute toward the energy budget. Furthermore, the large, continuous spatial coverage could allow scientists to investigate interactions across many temporal and spatial scales, which could lead to increased scientific understanding of weather and climate interactions and more accurate model physics. Additionally, MW data accompanied by uncertainties could be assimilated into models to improve simulated structures, which may lead to more accurate weather predictions. Furthermore, since MW imagery is also used in real-time forecasts of hurricane intensity and other extreme weather phenomena, forecast lead times for high-impact events could also be increased, which helps mitigate loss of life and property. Finally, the uncertainty metrics provided by our product will allow atmospheric scientists to assess which predictions are most trustworthy, which can further improve interpretation of the product and facilitate widespread adoption.

**Author contribution.** Conceptualization: M.O., S.W.P., V.P.; Data curation: E.C., S.W.P.; Data visualization: P.O.; Methodology: P.O.; Writing original draft: P.O., E.C.; Writing—review and editing: M.O., S.W.P., V.P. All authors approved the final submitted draft.

**Competing interest.** The authors declare none.

**Data availability statement.** The GOES-16 dataset is available on a public S3 bucket at [s3://noaagoes16/ABI-L2--MCMIPF/2020/](https://s3://noaagoes16/ABI-L2--MCMIPF/2020/). The GMI brightness temperatures are publicly available from NASA at <https://gpm.nasa.gov/data/directory>. This study used the Level 1B product for GMI brightness temperatures.

**Ethics statement.** The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

**Funding statement.** This research was supported by grants from the Office of Naval Research (N0001421WX00575 and N0001422WX01251) and the Cooperative Institute for Satellite Earth System Studies at the University of Maryland (NA19NES4320002).

## References

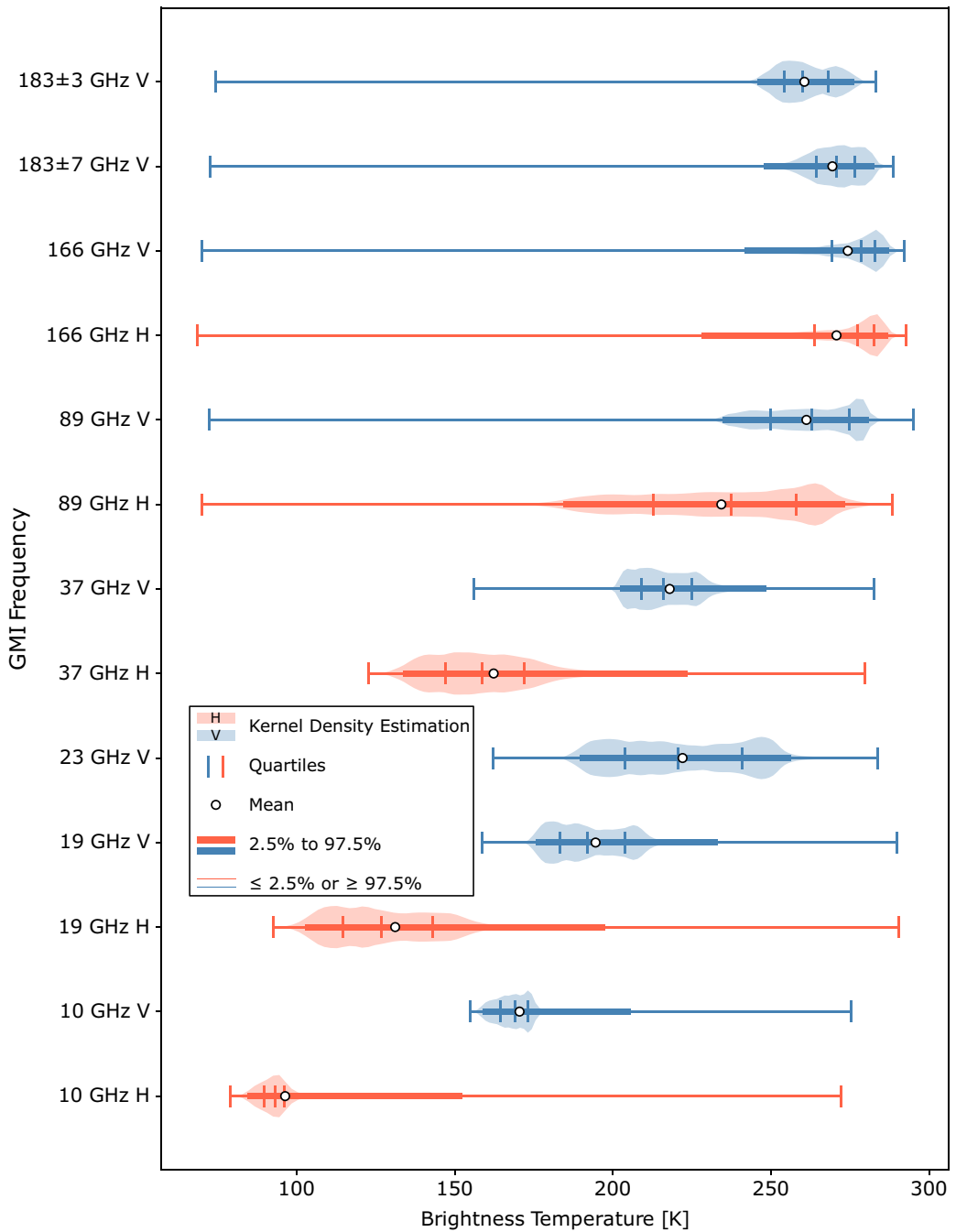
- EOSDIS Worldview.** (2013) Available at [https://worldview.earthdata.nasa.gov/?v=-245.09403723915682,-135.87380640621802,238.6758339845142,113.3700624319035&I=GMI\\_Brightness\\_Temp\\_Dsc,GMI\\_Brightness\\_Temp\\_Asc,Coastlines\\_15m,&lg=true&t=2020-02-01-T21%3A22%3A50Z](https://worldview.earthdata.nasa.gov/?v=-245.09403723915682,-135.87380640621802,238.6758339845142,113.3700624319035&I=GMI_Brightness_Temp_Dsc,GMI_Brightness_Temp_Asc,Coastlines_15m,&lg=true&t=2020-02-01-T21%3A22%3A50Z) (accessed 12 September 2022).
- GOES Image Viewer.** (2017) Available at <https://www.star.nesdis.noaa.gov/GOES/fulldisk.php?sat=G17> (accessed 12 September 2022).
- Adler RF, Huffman GJ and Keehn PR** (1994) Global tropical rain estimates from microwave-adjusted geosynchronous IR data. *Remote Sensing Reviews* 11(1–4), 125–152.
- Barnes EA and Barnes RJ** (2021) Controlled abstention neural networks for identifying skillful predictions for regression problems. *Journal of Advances in Modeling Earth Systems* 13(12), e2021MS002575.

- Barnes EA, Barnes RJ and Gordillo N** (2021) Adding uncertainty to neural network regression tasks in the geosciences. Preprint, arXiv:2109.07250.
- Blei DM, Kucukelbir A and McAuliffe JD** (2017) Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112(518), 859–877.
- Cecil DJ** (2007) Satellite-derived rain rates in vertically sheared tropical cyclones. *Geophysical Research Letters* 34(2), L02811.
- Dillon JV, Langmore I, Tran D, Brevdo E, Vasudevan S, Moore D, Patton B, Alemi A, Hoffman M and Saurous RA** (2017) TensorFlow distributions. Preprint, arXiv:1711.10604.
- Dines WH** (2007) The heat balance of the atmosphere. *Quarterly Journal of the Royal Meteorological Society* 43(182), 151–158.
- Draper DW, Newell DA, Wentz FJ, Krimchansky S and Skofronick-Jackson GM** (2015) The global precipitation measurement (GPM) microwave imager (GMI): Instrument overview and early on-orbit performance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8(7), 3452–3462.
- Feng R, Balling N, Grana D, Dramsch JS and Hansen T M** (2021) Bayesian convolutional neural networks for seismic facies classification. *IEEE Transactions on Geoscience and Remote Sensing* 59, 8933–8940.
- Filos A, Farquhar S, Gomez AN, Rudner TGJ, Kenton Z, Smith L, Alizadeh M, de Kroon A and Gal Y** (2019) A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks. Preprint, arXiv:1912.10481.
- Foster D, Gagne DJ, II and Whitt DB** (2021) Probabilistic machine learning estimation of ocean mixed layer depth from dense satellite and sparse in situ observations. *Journal of Advances in Modeling Earth Systems* 13(12), e2021MS002474.
- Gal Y and Ghahramani Z** (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR.
- Goodfellow I, Bengio Y and Courville A** (2016) *Deep Learning. Adaptive Computation and Machine Learning*. Cambridge, MA: The MIT Press.
- Gordon EM and Barnes EA** (2022) Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in CESM2. *Geophysical Research Letters* 49(15), e2022GL098635.
- Guillaumin AP and Zanna L** (2021) Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems* 13(9), e2021MS002534.
- He K, Zhang X, Ren S and Sun J** (2016) Identity mappings in deep residual networks. In Leibe B, Matas J, Sebe N and Welling M (eds), *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*. Cham: Springer International Publishing, pp. 630–645.
- Hristova-Veleva S, Haddad Z, Sawaya RC, Zuzow AJ, Vukicevic T, Li PP, Knosp B, Vu Q and Turk FJ** (2022) Understanding and predicting tropical cyclone rapid intensity changes using passive microwave observations from GPM and TRMM. In *IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4457–4460, Kuala Lumpur. IEEE.
- Kendall A and Gal Y** (2017) What uncertainties do we need in Bayesian deep learning for computer vision? In Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.), *Advances in Neural Information Processing Systems*, Vol. 30. Red Hook, NY: Curran Associates, Inc. Available at [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf).
- Kingma DP and Ba J** (2017) Adam: A method for stochastic optimization. Preprint, arXiv:1412.6980.
- Kiureghian AD and Ditlevsen O** (2009) Aleatory or epistemic? Does it matter? *Structural Safety* 31(2), 105–112.
- Kroodsma RA, Berg W and Wilheit TT** (2022) Special sensor microwave imager/sounder updates for the global precipitation measurement v07 data suite. *IEEE Transactions on Geoscience and Remote Sensing* 60, 1–11.
- Lakshminarayanan B, Pritzel A and Blundell C** (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.), *Advances in Neural Information Processing Systems*, Vol. 30. Red Hook, NY: Curran Associates, Inc. Available at [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf).
- Li Y, Wei C and Ma T** (2019) Towards explaining the regularization effect of initial large learning rate in training neural networks. In Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E and Garnett R (eds.), *Advances in Neural Information Processing Systems*, Vol. 32. Red Hook, NY: Curran Associates, Inc. Available at [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/bce9abf229ffd7e570818476ec5d7dde-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/bce9abf229ffd7e570818476ec5d7dde-Paper.pdf).
- Loeb NG, Wielicki BA, Doelling DR, Smith GL, Keyes DF, Kato S, Manalo-Smith N and Wong T** (2009) Toward optimal closure of the earth's top-of-atmosphere radiation budget. *Journal of Climate* 22(3), 748–766.
- Negri AJ, Adler RF and Kummerow CD** (1989) False-color display of special sensor microwave/imager (SSM/I) data. *Bulletin of the American Meteorological Society* 70(2), 146–151.
- Olander TL and Velden CS** (2019) The advanced Dvorak technique (ADT) for estimating tropical cyclone intensity: Update and new capabilities. *Weather and Forecasting* 34(4), 905–922.
- Orescanin M, Petković B, Powell SW, Marsh BR and Heslin SC** (2021) Bayesian deep learning for passive microwave precipitation type detection. *IEEE Geoscience and Remote Sensing Letters* 19, 1–5.
- Ortiz P, Casas E, Orescanin M, Powell SW, Petkovic V and Hall M** (2023) Uncertainty calibration of passive microwave brightness temperatures predicted by Bayesian deep learning models. *Artificial Intelligence for the Earth Systems* 2, 1–42.
- Ortiz P, Orescanin M, Petkovic V, Powell SW and Marsh B** (2022) Decomposing satellite-based classification uncertainties in large earth science datasets. *IEEE Transactions on Geoscience and Remote Sensing* 60, 1–11.
- Petković V and Kummerow CD** (2017) Understanding the sources of satellite passive microwave rainfall retrieval systematic errors over land. *Journal of Applied Meteorology and Climatology* 56(3), 597–614.

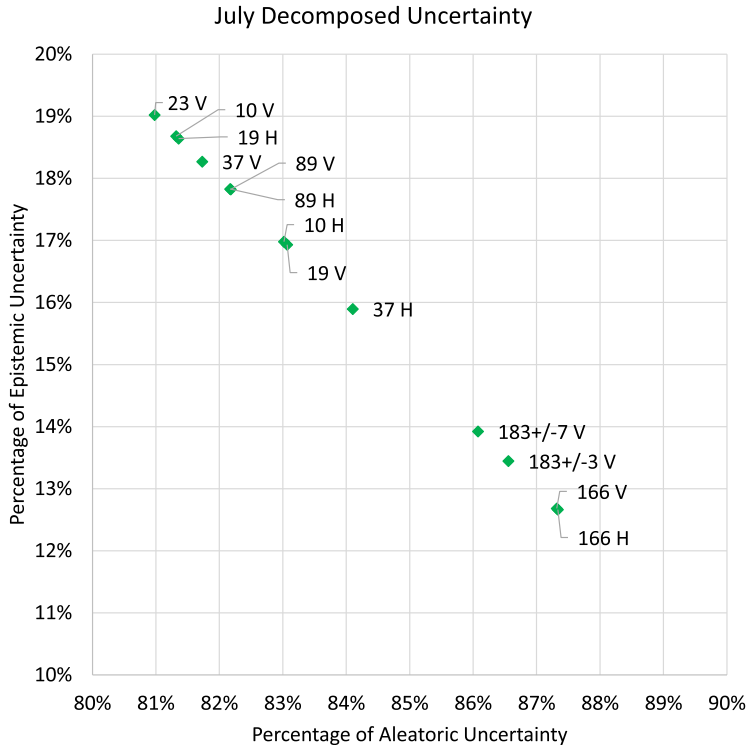
- Pu Z, Yu C, Tallapragada V, Jin J and McCarty W** (2019) The impact of assimilation of GPM microwave imager clear-sky radiance on numerical simulations of hurricanes Joaquin (2015) and Matthew (2016) with the HWRf model. *Monthly Weather Review* 147(1), 175–198.
- Schmit TJ, Griffith P, Gunshor MM, Daniels JM, Goodman SJ and Lehair WJ** (2017) A closer look at the ABI on the GOES-R series. *Bulletin of the American Meteorological Society* 98(4), 681–698.
- Settles B** (2010) *Active learning literature survey*. Technical report, University of Wisconsin–Madison Department of Computer Sciences.
- Stephens GL and Hu Y** (2010) Are climate-related changes to the character of global-mean precipitation predictable? *Environmental Research Letters* 5(2), 025209.
- Stephens GL, Li J, Wild M, Clayson CA, Loeb N, Kato S, L’Ecuyer T, Stackhouse PW, Lebsock M and Andrews T** (2012) An update on Earth’s energy balance in light of the latest global observations. *Nature Geoscience* 5(10), 691–696.
- Sun N and Weng F** (2012) Retrieval of cloud ice water path from special sensor microwave imager/sounder (SSMIS). *Journal of Applied Meteorology and Climatology* 51(2), 366–379.
- Trenberth KE, Fasullo JT and Kiehl J** (2009) Earth’s global energy budget. *Bulletin of the American Meteorological Society* 90(3), 311–324.
- Trenberth KE, Smith L, Qian T, Dai A and Fasullo J** (2007) Estimates of the global water budget and its annual cycle using observational and model data. *Journal of Hydrometeorology* 8(4), 758–769.
- Vonder Haar TH and Suomi VE** (1971) Measurements of the Earth’s radiation budget from satellites during a five-year period. Part I: Extended time and space means. *Journal of the Atmospheric Sciences* 28(3), 305–314.
- Wang Z and Hanks I** (2016) Moisture and precipitation evolution during tropical cyclone formation as revealed by the SSM/I–SSMIS retrievals. *Journal of the Atmospheric Sciences* 73(7), 2773–2781.
- Wen Y, Vicol P, Ba J, Tran D and Grosse R** (2018) Flipout: Efficient pseudo-independent weight perturbations on mini-batches. In *International Conference on Learning Representations*.
- Yang S, Hawkins J and Richardson K** (2014) The improved NRL tropical cyclone monitoring system with a unified microwave brightness temperature calibration scheme. *Remote Sensing* 6(5), 4563–4581.

**A. Appendix. Supporting tables and figures**

The following tables were used to construct Figures 3, 4, and 6.



**Figure A1.** GMI observed microwave temperature brightness ( $T_{bmw}$ ) value distributions for the August dataset. Orange indicates horizontal polarization; blue indicates vertical polarization. Orange and blue shaded areas indicate the  $T_b^{mw}$  value density for a given GMI frequency. Vertical lines mark the 0th, 25th, 50th, 75th, and 100th percentile. White dots mark the mean  $T_b^{mw}$  values. Bold lines indicate  $T_b^{mw}$  values in the middle 95% of the distribution; thin lines indicate  $T_b^{mw}$  values below or above the middle 95% of the distribution.



**Figure A2.** Percentage of uncertainty by component for each GMI channel and polarization using the predictions from the July dataset.



**Table A1.** Results using models trained on 26,587,000 sample from January to June to generate synthetic GMI data using the July test dataset of 1.63 million samples

Polar.	10 GHz		19 GHz		23 GHz	37 GHz		89 GHz		166 GHz		183 GHz	
	H	V	H	V	V	H	V	H	V	H	V	V ± 7	V ± 3
RMSE	12.04	7.04	11.60	6.43	6.55	10.82	5.24	9.80	4.78	7.87	6.28	4.65	3.13
MAE	4.57	2.76	6.29	3.34	4.26	6.54	3.04	6.86	2.93	4.01	2.65	1.94	1.59
MAPE	4.51%	1.57%	4.79%	1.70%	1.95%	3.97%	1.37%	3.05%	1.15%	1.60%	1.04%	0.76%	0.62%
MSD	6.31	3.80	7.76	4.27	4.60	7.61	3.49	7.85	3.41	5.38	3.62	2.76	2.09
MSRA	6.16	3.70	7.55	4.17	4.47	7.46	3.40	7.65	3.32	5.31	3.57	2.72	2.06
MSRE	1.26	0.85	1.72	0.85	1.05	1.41	0.76	1.66	0.72	0.77	0.52	0.44	0.32

Abbreviations: K = Kelvin; MAE = mean absolute error; MAPE = mean absolute percentage error; MSD = mean standard deviation; MSRA = mean square root aleatoric uncertainty; MSRE = mean square root epistemic uncertainty; RMSE = root mean squared error.

**Table A2.** Results as in Appendix Table A1, but for the August test dataset of 1.53 million samples

Polar.	10 GHz		19 GHz		23 GHz	37 GHz		89 GHz		166 GHz		183 GHz	
	H	V	H	V	V	H	V	H	V	H	V	V±7	V±3
RMSE	13.77	9.05	14.05	8.37	7.72	12.84	6.34	10.67	5.81	8.85	7.50	5.65	3.78
MAE	5.13	3.19	7.12	3.92	4.54	7.50	3.48	7.16	3.25	4.23	3.08	2.22	1.74
MAPE	4.42%	1.68%	5.01%	1.91%	2.03%	4.42%	1.55%	3.13%	1.29%	1.71%	1.24%	0.89%	0.70%
MSD	6.31	3.87	8.53	4.62	4.99	8.68	3.93	8.41	3.66	5.75	4.08	3.22	2.39
MSRA	6.19	3.77	8.32	4.54	4.87	8.54	3.84	8.24	3.58	5.70	4.04	3.17	2.36
MSRE	1.16	0.81	1.80	0.84	1.05	1.48	0.81	1.61	0.72	0.72	0.54	0.50	0.34

**Table A3.** Results as in Appendix Table A2, but using the updated models

Polar.	10 GHz		19 GHz		23 GHz	37 GHz		89 GHz		166 GHz		183 GHz	
	H	V	H	V	V	H	V	H	V	H	V	V±7	V±3
RMSE	14.14	9.06	13.92	8.14	7.72	12.79	6.28	10.60	5.66	8.93	7.52	5.69	3.81
MAE	5.24	3.37	7.07	3.79	4.41	7.33	3.48	6.96	3.15	4.21	3.11	2.23	1.74
MAPE	4.41%	1.78%	4.98%	1.85%	1.98%	4.26%	1.55%	3.03%	1.25%	1.71%	1.25%	0.90%	0.70%
MSD	7.24	4.56	9.45	4.89	4.37	7.47	4.12	8.31	3.57	5.69	4.53	3.47	2.56
MSRA	7.15	4.48	9.28	4.82	4.24	7.30	4.05	8.16	3.51	5.63	4.49	3.44	2.53
MSRE	1.09	0.78	1.70	0.81	1.02	1.51	0.75	1.51	0.66	0.72	0.53	0.48	0.33