Uncertainty Calibration of Passive Microwave Brightness Temperatures Predicted by Bayesian Deep Learning Models

PEDRO ORTIZ,[a] ELEANOR CASAS,[b] MARKO ORESCANIN,[a] SCOTT W. POWELL,[a] VELJKO PETKOVIC,[c] AND MICKY HALL[a]

[a] *Naval Postgraduate School, Monterey, California*
[b] *Department of Earth Sciences, Millersville University, Millersville, Pennsylvania*
[c] *Earth System Science Interdisciplinary Center, Cooperative Institute for Satellite Earth System Studies, University of Maryland, College Park, College Park, Maryland*

(Manuscript received 18 July 2022, in final form 10 July 2023, accepted 12 July 2023)

ABSTRACT: Visible and infrared radiance products of geostationary orbiting platforms provide virtually continuous observations of Earth. In contrast, low-Earth orbiters observe passive microwave (PMW) radiances at any location much less frequently. Prior literature demonstrates the ability of a machine learning (ML) approach to build a link between these two complementary radiance spectra by predicting PMW observations using infrared and visible products collected from geostationary instruments, which could potentially deliver a highly desirable synthetic PMW product with nearly continuous spatiotemporal coverage. However, current ML models lack the ability to provide a measure of uncertainty of such a product, significantly limiting its applications. In this work, Bayesian deep learning is employed to generate synthetic Global Precipitation Measurement (GPM) Microwave Imager (GMI) data from Advanced Baseline Imager (ABI) observations with attached uncertainties over the ocean. The study first uses deterministic residual networks (ResNets) to generate synthetic GMI brightness temperatures with as little mean absolute error as 1.72 K at the ABI spatiotemporal resolution. Then, for the same task, we use three Bayesian ResNet models to produce a comparable amount of error while providing previously unavailable predictive variance (i.e., uncertainty) for each synthetic data point. We find that the Flipout configuration provides the most robust calibration between uncertainty and error across GMI frequencies, and then demonstrate how this additional information is useful for discarding high-error synthetic data points prior to use by downstream applications.

KEYWORDS: Microwave observations; Satellite observations; Bayesian methods; Deep learning

1. Introduction and motivation

Passive microwave (PMW) radiometers have been utilized to collect Earth science observations since the 1970s (Carver et al. 1985; Kummerow et al. 1996) due to their valuable contributions toward numerous applications, such as establishing climatologies of columnar water vapor (e.g., Hilburn and Wentz 2008), clouds (e.g., Greenwald et al. 2018), and precipitation (e.g., Adler et al. 2003); validating numerical weather prediction (NWP) model accuracy (e.g., Wang et al. 2022); enhancing NWP accuracy through data assimilation in both clear-sky (e.g., Derber and Wu 1998; Pu et al. 2019) and cloudy and/or precipitating conditions (e.g., Weng 2017; Migliorini and Candy 2019); and improving real-time estimates of tropical cyclone structure and intensity (e.g., Weng 2017). These and additional applications are made possible through utilizing knowledge of how water in all three phases impacts radiative transfer processes, such as emission and scattering (Kummerow et al. 1996). For example, estimates of humidity are relatively straightforward to retrieve because microwave radiation is generally increasingly sensitive to absorption by water vapor molecules as frequency increases from 10 to 200 GHz, excluding water vapor absorption bands such as 23 and 183 GHz. In contrast, estimates of cloud and precipitation are more challenging to derive from microwave brightness temperatures T_b^{mw} due to greater nonlinear interactions between radiation and both liquid water and ice. For example, 37-GHz brightness temperature is sensitive to both liquid water and ice, such that clouds generally appear warmer than their surrounding environment due to liquid water emission, but very deep convection appears much colder than the surrounding environment due to large amounts of scattering by ice (Guilloteau and Foufoula-Georgiou 2020).

However, there are several drawbacks to using PMW data. First, upwelling microwave radiance from Earth's surface or the atmosphere is much lower than infrared radiance for Earth-based bodies with temperatures roughly between 180 and 330 K. Therefore, PMW sensors must observe radiation across a larger footprint on Earth's surface to gather meaningful data. For example, the nominal at-nadir field of view for the thermal channels of the Advanced Baseline Imager (ABI; Schmit et al. 2005) aboard the current generation of Geostationary Operational Environmental Satellites (GOES) is 2 km, and GOES provides full-disk imagery every 10–15 min depending on the data retrieval strategy (or "mode") used. In contrast, the nominal field of view for the 23-GHz PMW observation band for the Global Precipitation Measurement (GPM) Microwave Imager (GMI; Draper et al. 2015) is only about 16 × 10 km^2. In addition, microwave radiometers must fly in low-Earth orbit to achieve even these spatial resolutions. As a result, PMW data are only available in swaths beneath the satellite; therefore,

Corresponding author: Marko Orescanin, marko.orescanin@nps.edu

DOI: 10.1175/AIES-D-22-0056.1 e220056

many hours may pass between successive PMW observations at any point on Earth. This shortcoming is partially remedied by collecting PMW data using multiple satellites simultaneously. However, at any time, large spatial gaps remain in PMW datasets, which strongly limits the ability to follow the development of specific meteorological features of interest. Despite these limitations, PMW data continue to provide valuable information that improves the quality of numerous downstream applications. In an ideal world, meteorologists could have access to PMW data everywhere as frequently as visible and infrared data are available.

Therefore, in this study, we explore the feasibility of using deep learning to generate synthetic PMW data from infrared data. Numerous infrared wavelengths, such as those observed by the ABI aboard the current generation of GOES platforms, are sensitive to water vapor in varying degrees. Since microwave radiation is also sensitive to absorption by water vapor molecules, we hypothesize that multispectral ABI radiances contain information that may help predict what PMW radiances would be if a PMW radiometer were observing the same location. Deep learning, a type of supervised machine learning that uses deep neural networks to represent complex functions (Goodfellow et al. 2016), is one possible way to emulate PMW data in locations where it is not available but geostationary radiances are observed.

Previous literature suggests that predicting synthetic microwave data from infrared data may be possible, since a variety of information that is implicitly contained within satellite brightness temperatures has successfully been inferred. For example, Chen et al. (2019), Wimmers et al. (2019), Lee et al. (2019), and Maskey et al. (2020) each used a convolutional neural network (CNN) to estimate tropical cyclone intensity using PMW observations and/or infrared radiances. Others (Giffard-Roisin et al. 2020) built a CNN that demonstrated improvements in tropical cyclone track forecasts relative to consensus model forecasts. Hilburn (2020) has suggested that column-integrated atmospheric properties such as convective available potential energy (CAPE) or composite reflectivity can be derived using GOES ABI radiances. Petković et al. (2019) successfully explored the use of a deep learning approach in extracting the information content from PMW observation vectors to help identify precipitation types. Such applications of deep learning demonstrate benefits for maximizing information content extraction. However, they lack the ability to provide the uncertainty of predictions, which limits assessment of confidence in any prediction.

Not only is uncertainty helpful for establishing trust in predictions, but estimates of observational uncertainty are also necessary for many downstream, atmospheric applications of PMW data, such as data assimilation in NWP models (e.g., Weng 2017; Geer et al. 2017; Bonavita et al. 2020) and retrieval methods utilizing optimal estimation theory (e.g., Kummerow et al. 1996; Kulie et al. 2010; Schulte et al. 2022). In both data assimilation and optimal estimation, an a priori distribution (which is typically assumed to be Gaussian) is combined with a distribution of observations that is weighted by its uncertainty to produce a posterior distribution that better reflects the true state of the atmosphere. In optimal estimation, this is achieved through utilizing Bayes's theorem to solve an inverse matrix problem (Rodgers 2000). In data assimilation, the corresponding a priori and observation distributions and uncertainties are combined by utilizing knowledge of dynamical processes in various ways that typically involve minimizing a cost function (e.g., Dee 2004). While assimilating overocean, clear-sky PMW radiances into operational NWP models is now routine because of its lower retrieval uncertainties and clear positive impact on model performance, effective assimilation of non-clear-sky and overland PMW radiances into NWP models remains a challenge (Errico et al. 2007; Geer et al. 2017; Bonavita et al. 2020).

Since many applications of PMW directly incorporate uncertainties via Bayesian methods, Bayesian deep learning (BDL; Kendall and Gal 2017) could provide an alternative to deterministic deep learning methods that also accounts for model uncertainty in the weight space. While BDL is more computationally expensive than deterministic deep learning, it is capable of capturing both epistemic and aleatoric uncertainty in predictions (Kendall and Gal 2017; Ortiz et al. 2022), which is not possible using deterministic models. Aleatoric uncertainty captures noise inherent in the data and is irreducible (Kiureghian and Ditlevsen 2009). Epistemic uncertainty captures uncertainty in the model parameters and can be reduced given enough data (Kiureghian and Ditlevsen 2009; Kendall and Gal 2017). In this work, we focus only on modeling epistemic uncertainty.

Additionally, BDL is more robust against overfitting on training data distributions in comparison with deterministic deep learning, resulting in models that generalize better to unseen data (Neal 2012). Kendall and Gal (2017) demonstrate that Bayesian deep learning methods improve performance of neural networks while providing uncertainty estimation on predictions. Orescanin et al. (2021) and Ortiz et al. (2022) demonstrated both the uncertainty quantification and the utility of expressing uncertainty for precipitation type classification by applying BDL to the GMI dataset. In this work, we explore a more complex regression problem of predicting synthetic microwave brightness temperatures.

Multiple ways to construct a Bayesian model architecture are possible, and whether the choice of Bayesian architecture impacts predictive skill is unclear. Therefore, this study has three main scientific objectives:

1) Quantify errors in predicted synthetic passive microwave brightness temperatures using a deterministic model trained on a dataset of limited size.
2) Ascertain whether predictive skill is sacrificed relative to a deterministic model when using BDL to quantify variance (a metric of uncertainty), and
3) Explore how the choice of Bayesian architecture impacts predictive skill and interpretation by focusing on the calibration of predictive error and uncertainty.

To achieve these objectives, we first extract information from infrared radiances to generate synthetic passive microwave data by using residual network (ResNet) deep learning models (section 3a). We then provide additional evidence

TABLE 1. Infrared ABI bands used in this study. All bands have a 2-km field of view at nadir (NASA 2017).

ABI band	Central wavelength (μm)
7	3.9
8	6.2
9	6.9
10	7.3
11	8.4
12	9.6
13	10.3
14	11.2
15	12.3
16	13.3

that deterministic ResNets can be adapted to Bayesian ResNets without loss of skill and with the added benefit of quantified uncertainty (section 3b). Next, we show to what extent each choice of Bayesian architecture is well calibrated, such that we can infer expected error from predictive variance (section 3c). Finally, we demonstrate how the quantified uncertainty from Bayesian ResNets can be used to select low-error synthetic passive microwave data for downstream applications that depend on PMW data (section 3d).

The current article is primarily intended as an early proof of concept of our method. The results herein are highly encouraging and demonstrate the applicability of BDL to regression tasks applied to multispectral satellite observations. As explained and illustrated in the following sections, however, additional steps can be taken to improve predictions in locations where the model presented lacks skill.

2. Data and methods

a. Data, labeling, and dataset description

The fundamental task presented is prediction of GMI brightness temperatures using multispectral infrared radiances observed by the ABI. The ABI observes upwelling radiation from Earth in 16 different wavelength bands (Schmit et al. 2005). Bands 7–16 were used in this study as our input features (see Table 1); these correspond to central wavelengths between 3.9 and 13.3 μm that are emitted by Earth's surface and its atmosphere and, with the exception of band 7, contain negligible solar radiation reflected off of Earth. Differences in the radiance detected in these bands are strongly impacted by the temperature of Earth's surface, the vertical structure of temperature and humidity in the atmosphere, and the presence and properties of clouds. The same factors also impact passive microwave brightness temperatures; therefore, we expect that ABI data alone contains information that will allow us to make accurate predictions of GMI brightness temperatures, at least in noncloudy regions.

In this study, we first use deterministic ResNets to predict GMI brightness temperatures derived from upwelling radiation detected from low-Earth orbit for each of the 13 frequencies contained in Table 2. The deterministic models provide a baseline for comparison to evaluate the performance of Bayesian models. We then use Bayesian ResNets to predict a subset of GMI brightness temperatures to test whether predictive skill is consistent between deterministic and Bayesian ResNets at the following central frequencies at vertical polarization: 183 ± 3, 166, 36.6, and 23.8 GHz; 166- and 183 ± 3–GHz brightness temperatures are known to be useful for data assimilation into numerical models (Pu et al. 2019), and 36.6 and 23.8 GHz are potentially useful for revealing tropical cyclone core structure (Slocum and Knaff 2020). For completeness, one type of Bayesian implementation was used to train one model for each of the 13 GMI frequencies (see appendix for results).

TABLE 2. GMI frequencies, polarization, and field-of-view size; V and H denote vertical and horizontal polarizations, respectively [from Draper et al. (2015)].

Frequency (GHz)	Polarization	Field of view (km)
10.65	V	19 × 32
10.65	H	19 × 32
18.7	V	10 × 18
18.7	H	10 × 18
23.8	V	10 × 16
36.6	V	9 × 16
36.6	H	9 × 16
89	V	4 × 7
89	H	4 × 7
166	V	4 × 6
166	H	4 × 6
183 ± 3	V	4 × 6
183 ± 7	V	4 × 6

Data labeling began by collocating GMI and ABI observations. Only 10% of all GMI swaths (every 10th GMI swath starting from the first one observed on 1 January 2020) during the study period were collocated with ABI observations to keep the training dataset size manageable given our computing resources. This training strategy also mitigates concerns of inadvertently introducing spurious inferencing relationships from synoptic decorrelation time-scales, since we only retain approximately 1–2 GMI swaths per day, and the fraction of observations sampling within 0.01° of a given location within three days in the training dataset is approximately 0.3%. The distributions of observed GMI brightness temperatures in the data used in this study are shown in Fig. 1. Once matched, 39 × 39 pixel patches of ABI bands 7–16 were labeled using the corresponding GMI pixel at the center of the patch. These label–feature pairs were accompanied by geolocation, the viewing angle, a 39 × 39 matrix of surface type flags (ocean/land), and a unique integer identifier. Any records located far on the limbs (west of 140°W or otherwise within 20 data pixels from the edge of the available data) of the ABI viewing disk were discarded. Based off of our previous work that identified significantly different data distributions when using GMI data to predict precipitation type over land versus predicting precipitation type over ocean (Ortiz et al. 2022), we also discarded records that had center pixels over land.

For our experiments, we used data from January, February, and May 2020. Three full days from the month of January were randomly selected for a validation dataset, and an additional three days in January were randomly selected for a test

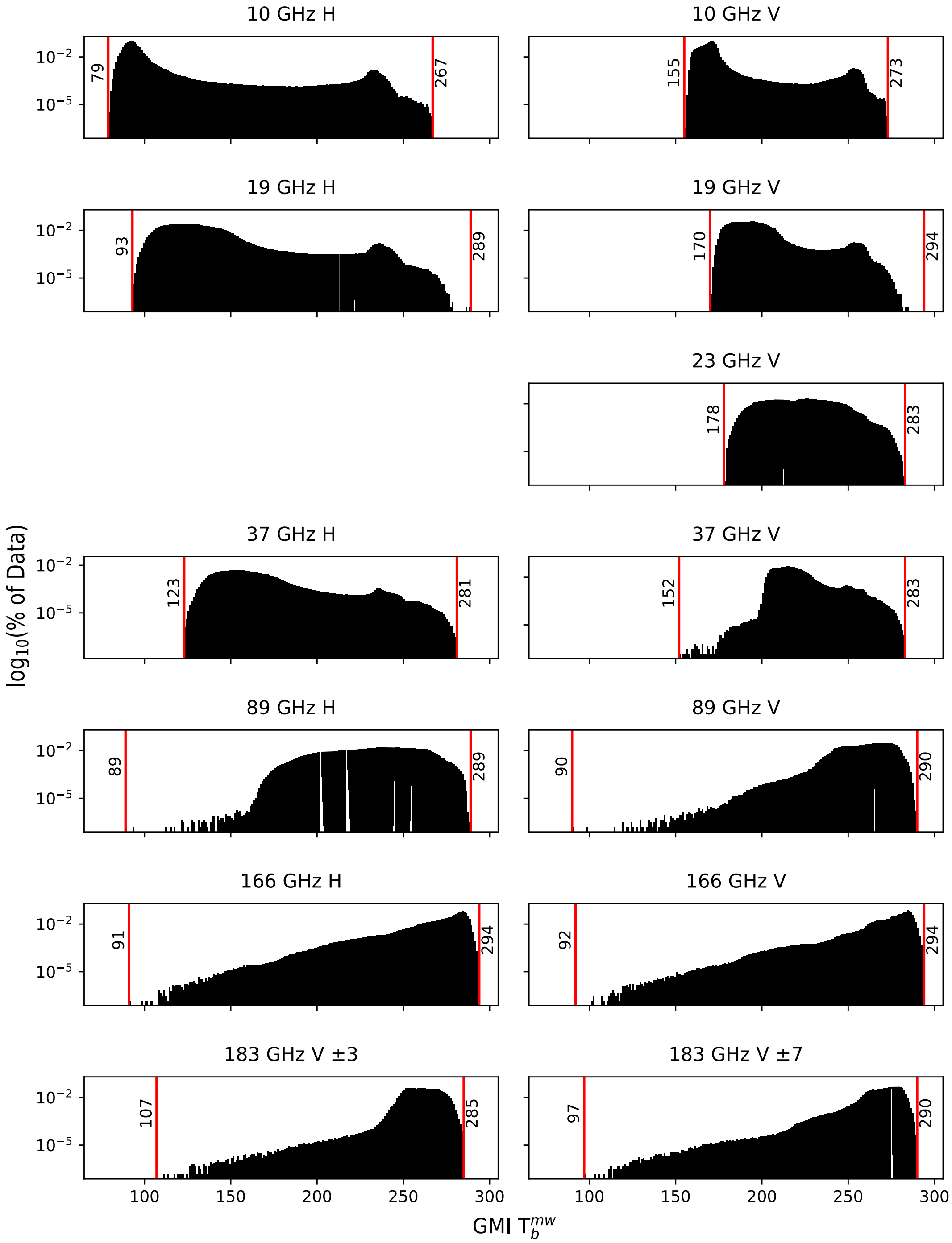

FIG. 1. Histograms of the GMI T_b^{mw} labels from the January training dataset with bins of size 1 K. The y axis is in logarithmic base-10 scale. Red lines mark the highest and lowest T_b^{mw} value contained in the dataset for each GMI frequency and are labeled with the corresponding value.

dataset. The remaining days in January were used to create a training dataset. We created two additional datasets for posttraining evaluation to test the ability of our model to retain accuracy on unseen data from the future: one using data from the first week of February, and one using data from the first week of May 2020. In other words, no data collected after January was used in training and model development.

Given the real-world nature of our data, the resulting training dataset exhibits inherent imbalance in both the label (GMI pixels) and input features (ABI channels). Training on imbalanced datasets can lead to developing overconfident models biased

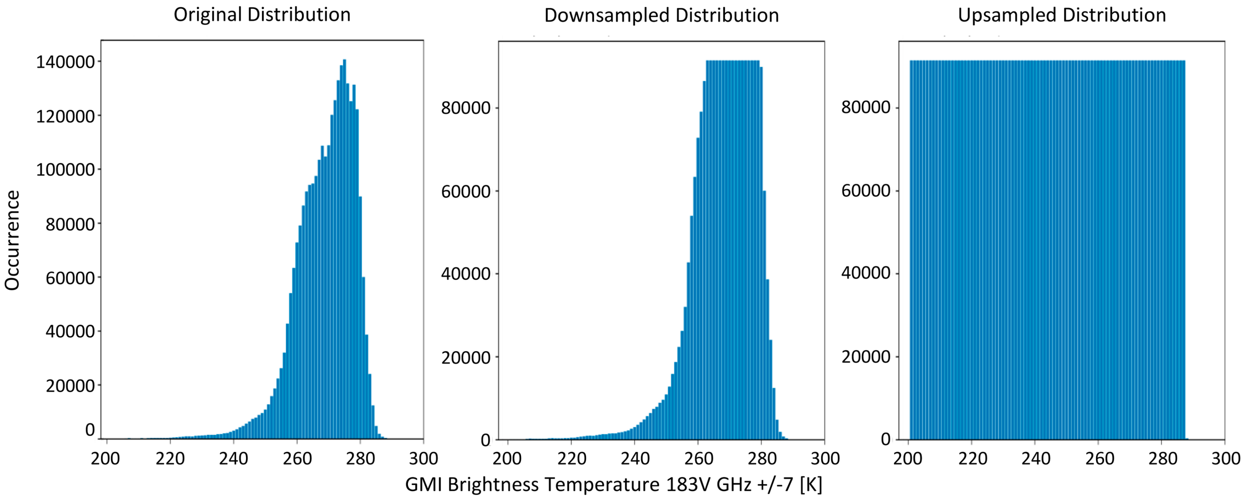

FIG. 2. Label data distributions: (left) The original distribution of 183 ± 7–GHz T_b^{mw}. (center) The resulting label distribution after the downsampling to 65% of the most populated bins. (right) The resulting uniform distribution after upsampling with replacement.

toward the most represented values in the dataset. To manage the imbalanced nature of the data one can include a custom loss function that weighs more learning on underrepresented values as is commonly done with image classification (Filos et al. 2019; Leibig et al. 2017). However, for regression, such an approach would require estimating a weighting curve over the continuous range of labels (Ebert-Uphoff et al. 2021) and would include multiple hyperparameters that would require evaluation.

Rather, in this work we focus on producing a more balanced dataset by sampling data with replacement for GMI temperature labels with low occurrence (see Fig. 2). For each GMI band, we binned the temperature label occurrences, using 1 K increments for bin sizes. We randomly selected 65% of the samples with the most frequently occurring label and made them part of our training set. For all remaining bins, if the number of samples was greater than or equal to 1% of the total number of samples in the dataset, we randomly sampled data with replacement to reach the same threshold. If the number of samples was less than 1% of the total number of samples in the dataset, the samples were included in the dataset but the bin was not upsampled. The end result was a training set with an equal number of samples in all label bins except for the bins in the tails of the original data distribution (not shown). Moreover, we used NumPy to reshape (not transpose) each input from 10 × 39 × 39 to 39 × 39 × 10, preserving the volume-based information.

b. BDL

The recent rise of deep learning applications in remote sensing applications is driven by the nonlinear modeling ability of neural networks that enables recognizing complex patterns and relationships better than the classical parametric modeling approach informed by physics. However, the majority of deep learning models currently used across remote sensing applications lack the ability to provide uncertainty in prediction. BDL combines the nonlinear modeling power of deep learning with Bayesian inference (Blei et al. 2017) enabling machine learning models to provide information about uncertainty in prediction. Assuming a supervised learning setup and a regression task of predicting a continuous value, a training dataset is defined as $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$, where N represents the dataset size, $\mathbf{x}_n$ represents an input feature vector (where $\mathbf{x}_n \in \mathcal{R}^m = [x_{1,n}, x_{2,n}, \ldots, x_{m,n}]$) and y_n represents the corresponding label (where $y_n \in \mathcal{R}$).

We assume that a neural network model with L layers is parameterized by the set of weights $\mathbf{w} = \{W_i\}_{i=1}^{L}$. If one assumes a prior distribution over neural network parameters $p(\mathbf{w})$, then the goal of Bayesian method is to quantify a posterior distribution over the network parameters $p(\mathbf{w}|\mathcal{D})$ conditioned on the distribution of the training data, $p(\mathcal{D})$:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{\int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,d\mathbf{w}}. \qquad (1)$$

The denominator in Eq. (1) often has no closed form solution and is computationally intractable (Blei et al. 2017). As a result, an approximation of the $p(\mathbf{w}|\mathcal{D})$ is computed instead.

Variational inference is one method of posterior approximation that involves an optimization problem to identify the parameters θ of a distribution in a family of distributions $q_\theta(\mathbf{w}) \in Q$ that has the smallest Kullback–Leibler divergence (KL) from the target distribution, $p(\mathbf{w}|\mathcal{D})$:

$$\mathrm{KL}[q_\theta(\mathbf{w})\|\,p(\mathbf{w}|\mathcal{D})] = \int q_\theta(\mathbf{w})\,\log\frac{q_\theta(\mathbf{w})}{p(\mathbf{w}|\mathcal{D})}\,d\mathbf{w}. \qquad (2)$$

However, Eq. (2) still contains $p(\mathbf{w}|\mathcal{D})$, which is intractable. To solve the optimization problem without explicitly calculating $p(\mathbf{w}|\mathcal{D})$, Eq. (2) can be rewritten as (Dürr et al. 2020)

$$\mathrm{KL}[q_\theta(\mathbf{w})\|\,p(\mathbf{w}|\mathcal{D})] = \log\,p(\mathcal{D}) - \underbrace{\int q_\theta(\mathbf{w})\log\frac{p(\mathbf{w})p(\mathcal{D}|\mathbf{w})}{q_\theta(\mathbf{w})}\,d\mathbf{w}}_{\text{Evidence Lower Bound (ELBO)}}. \qquad (3)$$

Since the first term of Eq. (3) does not depend on q_θ, it can be ignored to solve the minimization problem. Instead, the minimization problem is solved by maximizing the second term in Eq. (3), the evidence lower bound (ELBO). The ELBO is a tractable substitute used in practical application. The resulting optimization problem is (Dürr et al. 2020)

$$\theta^* = \arg\max_\theta \int q_\theta(\mathbf{w}) \log \frac{p(\mathbf{w})p(\mathcal{D}|\mathbf{w})}{q_\theta(\mathbf{w})}\, d\mathbf{w}. \tag{4}$$

In practice, rather than maximizing the above expression, we minimize the negative ELBO in the process of optimization.

The goal of inference with BDL is to make a prediction, $\hat{y}_n$, for each new input $\mathbf{x}_n$. Each $\hat{y}_n$ is calculated by using Monte Carlo integration with T samples of the weight distribution to make T predictions for each $\mathbf{x}_n$ (Feng et al. 2021; Filos et al. 2019). For this study, we followed the approach described in Filos et al. (2019) and chose $T = 100$. Using our Bayesian models, we made 100 predictions $\hat{y}_t$ for each input and calculated $\hat{y}_n$ as

$$\hat{y}_n = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t(x_n, \mathbf{w}_t). \tag{5}$$

Since the weights of the models are distributions, each $\mathbf{w}_t$ is different, and $\hat{y}_n$ is the average predicted value (referred to as Monte Carlo integration). This results in additional computation cost that is at least linear in the size of T in comparison with deterministic deep learning models that only make a single prediction per input; however, as a by-product of Monte Carlo integration, the uncertainty of each prediction is easily quantified.

Providing a measure of epistemic uncertainty, the same T predictions were also used to estimate the variance of $\hat{y}_n$ (Harris et al. 2020):

$$\mathrm{Var}(\hat{y}_n) = \frac{1}{T}\sum_{t=1}^{T} [\hat{y}_t(x_n,\mathbf{w}_t) - \hat{y}_n]^2. \tag{6}$$

To provide a measure of uncertainty in the same units of the GMI labels and the model predictions (Kelvin), the mean standard deviation (MSD) for a set of predictions, $\hat{y}$, of size N was calculated using the variance from Eq. (6):

$$\mathrm{MSD}(\hat{y}) = \frac{1}{N}\sum_{n=1}^{N} \sqrt{\mathrm{Var}(\hat{y}_n)}. \tag{7}$$

c. Approaches to variational inference

The training of a neural network using Eq. (4) requires the calculation of the derivative of the ELBO with respect to both θ and $\mathbf{w}$. One approach to calculating this derivative is to sample from the distribution $q_\theta(\mathbf{w})$ and then average over the samples, referred to as Monte Carlo estimation (Mohamed et al. 2020). However, Monte Carlo estimation can yield gradients with high variance that inhibit a model from learning (Kingma et al. 2015); Kingma et al. (2015) introduced a computationally efficient method, known as the local reparameterization trick (LRT), that reparameterizes $q_\theta(\mathbf{w})$ such that the variance is reduced. A drawback to the LRT is that the training examples in a minibatch share the same weight distribution parameters that results in correlated gradients, limiting the variance reduction gained through using larger minibatches. To decorrelate these gradients, Wen et al. (2018) introduced the Flipout method. The Flipout method is computationally more expensive than the LRT, but it produces lower variances as the size of the minibatch is increased.

Assuming a Gaussian distribution over $q_\theta(\mathbf{w})$, using the Flipout method effectively doubles the number of parameters that must be learned relative to a deterministic neural network. However, it is possible approximate variational inference using a deterministic network and the dropout regularization technique. Srivastava et al. (2014) introduced a regularization technique for training neural networks whereby each network weight is set to zero with probability p (we set p to 0.2) each time the weights are sampled for training, resulting in a model with fewer connections between neurons (roughly 80% fewer for our models). Gal and Ghahramani (2016) proved that using L2 regularization in conjunction with dropout during inference is equivalent to variational inference. This approach is known as Monte Carlo (MC) Dropout. In contrast to the other two approaches, MC Dropout models do not contain a direct representation of the weight distributions. In this work, we use all three of the above approaches, which are implemented as part of the TensorFlow software library (Abadi et al. 2015), as described in the next section.

d. Model architecture

A residual network (ResNet), version 2 (He et al. 2016), with 58 convolutional layers (see Fig. 3a) was chosen as a representative deterministic architecture to predict the brightness temperatures from GMI input data, based on the precipitation classification results in Orescanin et al. (2021) using similar data. The key feature of ResNets is the skip connection within each ResNet block (see Fig. 3). The skip connections are identity functions that allow information to be propagated directly to a layer from any preceding layer in the neural network. This architecture enabled deeper convolutional neural networks to be trained than was previously possible (He et al. 2016). For our networks, we stacked the ResNet blocks depicted in Fig. 3a until our models had a total of 58 convolutional layers, consisting of 19 ResNet blocks and an additional convolutional layer immediately after the input layer. Each of these layers is built into Tensorflow; for our experiments, we used Tensorflow graphical processing unit (GPU), version 2.4.

By following the approach in Tran et al. (2019), Bayesian ResNet architectures were adopted in an identical configuration as the deterministic architecture. The changes to the deterministic architectures required to develop the Bayesian models are highlighted in yellow in Figs. 3b–d. For the Flipout and Reparameterization models, we conducted a one-for-one replacement of deterministic convolutional layers with either Flipout (Wen et al. 2018) or Reparameterization (Kingma et al. 2015) convolutional layers contained in the Tensorflow Probability library, version 0.12.1 (Dillon et al. 2017). We

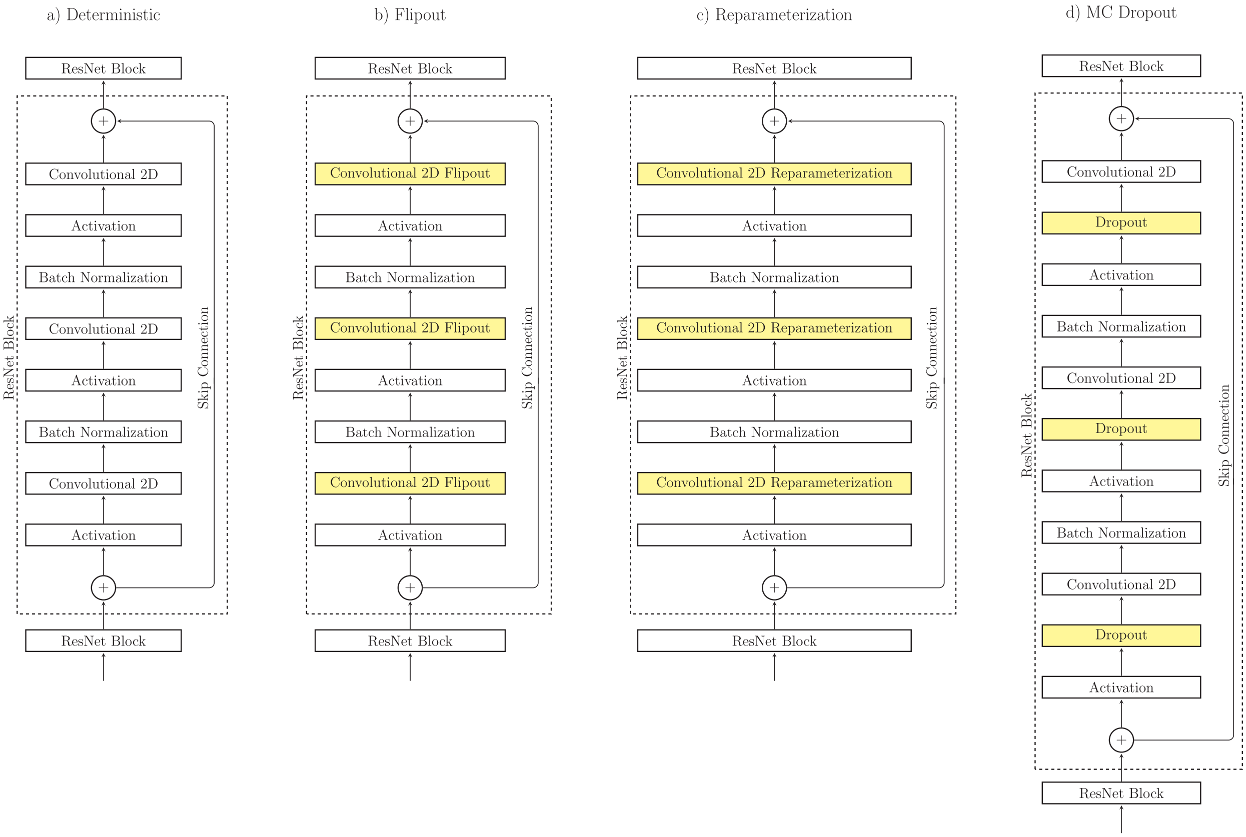

FIG. 3. A depiction of the deterministic and Bayesian network architectures. Inside each dashed rectangle is a single ResNet block. The adaptations from the deterministic model to create our Bayesian models are highlighted in yellow.

constructed our MC Dropout models by placing Dropout layers (built into Tensorflow 2.4) immediately after each activation layer in the original deterministic model. Dropout is applied both during training and during inference (Gal and Ghahramani 2016).

e. *Training methodology*

At the start of the training process, the model weights were initialized using He initialization (He et al. 2016). Each model used the Adam optimizer (Kingma and Ba 2017) with a starting learning rate of 0.001. To conduct learning rate annealing, the validation loss was monitored throughout training (Li et al. 2019). If the validation loss did not improve from the best recorded validation loss after 5 consecutive epochs, the learning rate was reduced by a factor of 4. An early stopping strategy was employed to regularize for overfitting (Goodfellow et al. 2016). If early stopping did not occur, training would have continued for a total of 500 epochs. Our Bayesian models, using a batch size of 2048, required approximately 1.5–3 days to train using 4 NVIDIA RTX 8000 48GB GPUs per model. To support fair benchmarking, both deterministic and Bayesian models were trained with the same strategy. In relative terms, it took roughly twice as long for the Flipout models to train than the models using the LRT, which is consistent with the findings in Wen et al. (2018). Training time for the MC Dropout models fell in between the training times of the other two model types. Unsurprisingly, the deterministic models trained faster than all three Bayesian methods.

f. *Well-calibrated uncertainty*

A model must be well calibrated in order to infer a likely amount of error from a predicted variance. According to Filos et al. (2019), if the performance of a model improves as more high-uncertainty predictions are discarded, then the model has well-calibrated uncertainty. Therefore, for a well-behaved model, as we decrease the standard deviation threshold for discarding data [i.e., we reduce the amount of data used to calculate mean absolute error (MAE) to include only low-standard-deviation predictions], the MAE should also decrease. In other words, a well-calibrated model is one that has a positive, monotonic relationship between the mean absolute error and percent of predictions retained.

Examples of model calibration are shown in Fig. 4, where the lowest 1% percent of predictions retained represent only the top 1% of predictions with the lowest uncertainty and 100% percent of predictions retained represents all predictions. Figure 4a represents an example of an uncalibrated model because it depicts a negative, monotonic relationship between MAE and percent of predictions retained until approximately 90% of predictions are retained with a slight

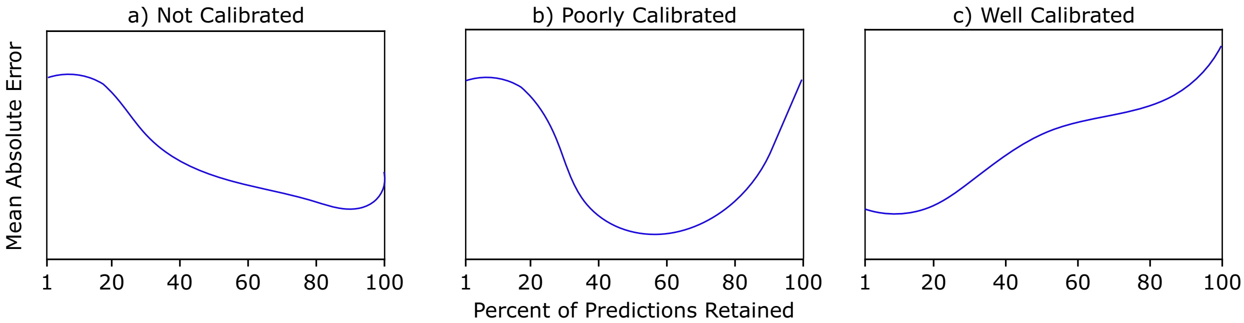

FIG. 4. (a),(b) An example depiction of uncertainty calibration where the change of MAE is shown as a function of the data retained after removing a percentage of the most uncertain samples. The uncertainty in (a) and (b) are both not well calibrated, but for different reasons. (c) The general trend in MAE for a model with well calibrated uncertainty.

increase in MAE thereafter. At best, the relationship between error and uncertainty is weak for this model, which means that predicted variance could not be used at all as proxy for unknown absolute error in downstream applications.

Figure 4b represents an example of a poorly calibrated model, where the MAE decreases until approximately 50% of predictions are retained and then increases as more predictions are retained. This model has a stronger relationship between error and model uncertainty than in Fig. 4a because the most uncertain predictions (far right) correspond to a higher amount of error; however, the decrease at the far left means that the model makes predictions with low uncertainty but high error. This scenario of both very low and very high uncertainty predictions corresponding with high error is still problematic for downstream applications that would use predicted variance as a proxy for unknown absolute error. In both of these two cases, neither model produces quantified uncertainty that meets the definition from Filos et al. (2019).

Figure 4c represents a well-calibrated model, because it depicts a positive, monotonic trend as an increasing number of predictions are retained to calculate the MAE. This model does meet the definition of well-calibrated uncertainty, making it a better candidate for real-world deployment than the other two models. As demonstrated in our previous work (Orescanin et al. 2021; Ortiz et al. 2022), having a deployed model with well-calibrated uncertainty is desirable since the true label (T_b^{mw}) is unknown during live inference and the predictive error cannot be calculated. For deployed models with well-calibrated uncertainty, a prediction with high standard deviation likely has a high amount of error; conversely, a prediction with low standard deviation likely has a low amount of error.

3. Results and discussion

a. Deterministic ResNets

We trained one deterministic model for each GMI frequency as outlined in section 2a. The results of these experiments, using the January test dataset, are captured in Table 3. MAE was smallest at the 183 ± 3–GHz bands (1.71 and 2.31 K) and was largest for 89 GHz horizontal (13.70 K). The magnitude of MAE generally corresponded to the frequency of occurrence of observed T_b^{mw} in the tails of distributions (Fig. 1), with wide and more uniform distributions unsurprisingly posing a greater challenge to the model. Because the ocean is a poorer emitter of horizontally polarized microwave radiation, observed horizontally polarized T_b^{mw} over ocean in clear-air is much lower than vertically polarized T_b^{mw} at the same frequency, causing wider distributions of T_b^{mw}. Therefore, for each frequency sampled in both horizontal and vertical polarizations, the horizontal polarization model produced more error than the corresponding vertical polarization model. While the ranges of observed T_b^{mw} were large for the 166- and 183 ± 3–GHz bands, the frequency of T_b^{mw} observed was low in the observed distribution's tails at low brightness temperatures. In other words, for bands in which the observed T_b^{mw} distribution has lower variance, our models produce synthetic T_b^{mw} with less error.

To evaluate the temporal persistence of skill in our models, we generated T_b^{mw} predictions for data collected in the next month (February) and 4 months later (May). Table 4 contains the results of using each model to predict T_b^{mw} using ABI brightness temperatures T_b^{ir} collected daily from 1 to 7 February 2020 (see section 2a). The absolute error amplitude for predictions in this dataset was lower for 18.7 GHz (vertical), 23 GHz, and 89 GHz (vertical) (1.5%, 5.5% and 0.3% decrease, respectively)

TABLE 3. Summary statistics of deterministic model performance in terms of MAE, RMSE, and R^2 score (the coefficient of determination) for each GMI channel for the January test set of 723 000 samples.

	10.6 GHz		18.7 GHz		23 GHz	37 GHz		89 GHz		166 GHz		183 GHz	
Polarization	H	V	H	V	V	H	V	H	V	H	V	±3 V	±7 V
MAE	6.93	4.86	12.1	7.50	10.27	13.45	7.03	13.70	6.32	5.79	3.62	1.71	2.31
RMSE	16.52	11.93	18.48	12.64	14.41	20.18	10.30	18.38	8.70	9.40	6.55	2.86	4.61
R^2	0.32	0.11	0.19	−0.05	0.12	−0.06	−0.10	0.46	0.57	0.66	0.73	0.90	0.79

TABLE 4. As in Table 3, but for the February test set of 857 000 samples.

	10.6 GHz		18.7 GHz		23 GHz	37 GHz		89 GHz		166 GHz		183 GHz	
Polarization	H	V	H	V	V	H	V	H	V	H	V	±3 V	±7 V
MAE	7.04	5.40	12.31	7.39	9.71	13.73	7.15	13.97	6.30	6.30	4.10	2.01	2.72
RMSE	16.67	13.21	19.64	12.92	13.76	20.74	10.76	19.18	8.82	10.86	8.18	3.80	5.91
R^2	0.44	0.11	0.27	0.12	0.29	0.13	0.08	0.36	0.50	0.55	0.61	0.84	0.69

and higher for all other GMI frequencies (8.3% average increase) relative to the January dataset. Relative to the results from the January dataset, the February dataset produced higher root-mean-square error (RMSE) for all frequencies (12.8% average increase) except 23 GHz (4.5% decrease). The models predicted $T_b^{\rm mw}$ with the least amount of error for 183 ± 3–GHz channel, with an RMSE of ~3.80 K and an MAE of ~2.01 K. The 37-GHz (horizontal) model predictions had the highest RMSE of ~20.74 K, and the 89-GHz (horizontal) model predictions had the highest MAE of ~13.97 K.

Table 5 contains the results of using each model to predict $T_b^{\rm mw}$ from $T_b^{\rm ir}$ collected daily at a much later date, from 1 to 7 May 2020. Similar to the results using the February dataset, the error amplitude for predictions in this dataset was higher for all but one GMI frequency relative to the January dataset but up to 50% higher for the 10.6-GHz (horizontal) band. Relative to the results from the February dataset, the models generated lower error for 183 ± 7–GHz vertical (decrease in MAE and RMSE by 9.5% and 5.0%) and higher error for the remaining frequencies (average increase in MAE and RMSE of 15.9% and 22.5%). The models still predicted $T_b^{\rm mw}$ with the least amount of error for 183 ± 3–GHz (vertical) channel, with an RMSE of 3.85 K and an MAE of 2.05 K. The 37-GHz (horizontal) model predictions contained the greatest amount of absolute error with an MAE of 14.46 K; however, it did not have the highest RMSE (22.31 K). The vertical channels at 10.6 and 18.7 GHz had higher RMSE of 27.88 and 25.78 K, which surpasses the horizontal 37-GHz error. These two channels have the least atmospheric opacity in clear-air and are thus most sensitive to the surface; therefore, we speculate that the disproportionate increase in error for these two GMI channels may partially result from a warmer surface and lower-tropospheric temperatures that occur in May and are not represented in the training dataset.

Since there is a large range of MAE (1.7–14.5 K) reported in Tables 3–5, a question remains whether the highest errors found in the 37-GHz channel are due to an ill-fitting model or whether predicting 37-GHz $T_b^{\rm mw}$ from IR data is a more challenging regression task than predicting $T_b^{\rm mw}$ of other PMW frequencies. To assess the extent of physically consistent relationships learned by the deterministic models, Fig. 5 shows examples of $T_b^{\rm mw}$ predictions and absolute error from the models with lowest MAE (183 ± 3 GHz) and highest MAE (37 GHz H) on 1 February 2022. The 183 ± 3–GHz predictions have considerable skill, with many clear-sky predictions having absolute error less than 1 K. While predictions in cloudy/precipitating scenes can have error in excess of 3.4 K (2 times the MAE for 183 ± 3–GHz predictions), the deterministic models are still able to capture that the coldest $T_b^{\rm mw}$ s occur in the convection between Florida and Cuba. This result is expected, because both the 183 ± 3–GHz channel and ABI band 8 (6.2 μm) are highly sensitive to mid- to upper-level water vapor. In contrast, the 37-GHz H model predictions have much higher absolute errors, in excess of 26.8 K (2 times the MAE for horizontal 37-GHz predictions). However, Fig. 5 indicates that the horizontal 37-GHz model is able to consistently capture clear-sky predictions with error less than 4 K, and the cloudy and precipitating scenes correctly have warmer $T_b^{\rm mw}$ than the surrounding environment. Additionally, the 37-GHz H model correctly predicts the approximate locations of where the warmest $T_b^{\rm mw}$ magnitudes occur, such as near the equator, between Florida and Cuba, and north of the Bahamas, although the magnitudes of the predictions are lower than observed values. Together, Fig. 5 suggests that the 37-GHz H model is not entirely ill-fitting since it has learned many physically consistent relationships and model skill appears to be consistently high in clear-sky regions. Instead, the large MAE at 37 GHz is more likely to be due to more complex relationships between how infrared and microwave radiation interact with liquid water and ice in comparison with other PMW frequencies. We speculate this because model performance degrades most in regions where 37-GHz and IR-wavelength observations are expected to be most different, such as within clouds, because in cloudy regions a single infrared brightness temperature can correspond to a wide range of 37-GHz brightness temperatures. Therefore, it is plausible that our model is simply making predictions in the center of observed $T_b^{\rm mw}$ distributions in its best effort to minimize its loss function, which suggests that additional input features (e.g., visible radiances) may be required to drive significant improvement in model performance.

TABLE 5. As in Table 3, but for the May test set of 925 500 samples.

	10.6 GHz		18.7 GHz		23 GHz	37 GHz		89 GHz		166 GHz		183 GHz	
Polarization	H	V	H	V	V	H	V	H	V	H	V	±3 V	±7 V
MAE	10.55	6.59	13.77	9.28	11.52	14.46	7.36	14.07	6.82	6.92	4.38	2.05	2.50
RMSE	27.88	17.89	25.78	17.63	16.46	22.31	11.64	19.01	9.53	11.44	8.45	3.85	5.68
R^2	−0.06	−0.09	0.00	−0.27	0.22	−0.02	−0.08	0.37	0.47	0.50	0.57	0.80	0.66

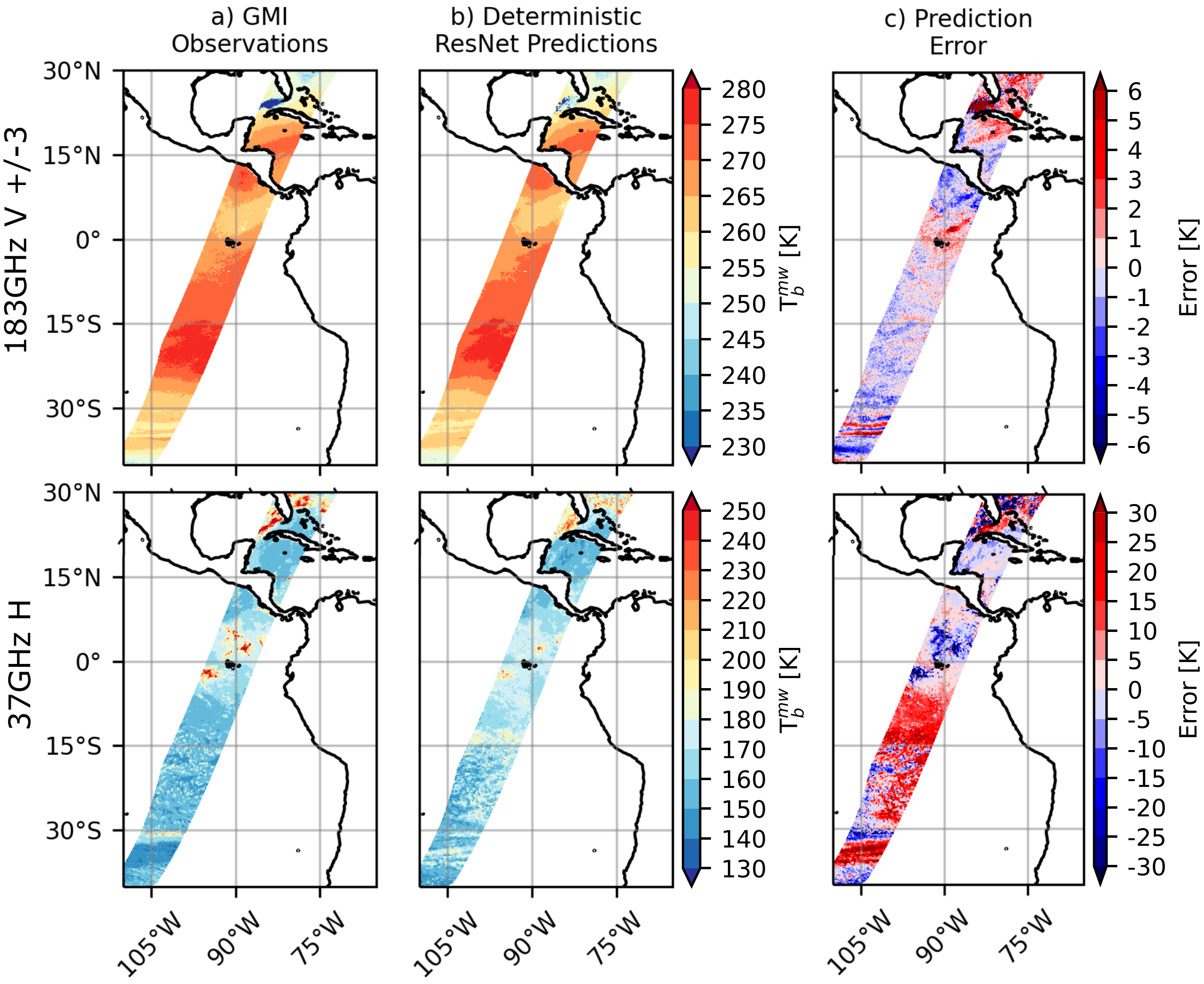

FIG. 5. GMI swath at (top) 183 ± 3 and (bottom) 37 GHz (horizontal) around 1440 UTC 1 Feb 2022 (GPM orbit number 33679): (a) observed T_b^{mw}, (b) predicted T_b^{mw} from the deterministic ResNet model, and (c) prediction error.

Together, these results indicate it is possible to take a step toward much higher spatiotemporal resolution than is currently available using real GMI observations by using deep learning models, especially for higher GMI frequencies. However, while deterministic models have the benefit of being relatively computationally inexpensive, they have an intrinsic limitation that there is no way to assess predictive uncertainty or error in the absence of validation data. Therefore, for the remainder of this study, we utilize Bayesian deep learning, which predicts both the magnitudes of GMI brightness temperatures and the variance in each prediction for our regression task.

b. *Comparison of deterministic and Bayesian model errors*

For the remainder of this article, we narrow the focus to only a few GMI channels as representative examples and investigate the performance of three different types of Bayesian models in comparison with the deterministic ResNet results revealed in section 3a. We used a deterministic ResNet as the base model architecture (see Fig. 3) and trained three Bayesian models each (Flipout, MC Dropout, and Reparameterization) for the vertically polarized 23, 37, 166, and 183 ± 3–GHz GMI channels (the vertical polarization is implicit for Bayesian models hereinafter). Table 6 contains the RMSE and MAE results of using the deterministic model and each Bayesian model to predict T_b^{mw} from ABI radiances collected on 8, 21, and 26 January 2020 (the three random days in our test dataset; see section 2a). Both deterministic and Bayesian models predicted 183 ± 3–GHz T_b^{mw} with the least amount of error, with an RMSE of 2.84–2.95 K and an MAE of 1.66–1.84 K. The 166-GHz predictions were more prone to error with an RMSE of 6.25–6.54 K and an MAE of 3.63–3.83 K. The 37-GHz predictions were even more prone to error than 166 GHz with an RMSE of 10.11–10.90 K and an MAE of 7.02–7.67 K. The 23-GHz predictions contained the greatest amount of error with an RMSE of 13.71–17.04 K and an

TABLE 6. Summary statistics of Bayesian model performance in terms of RMSE, MAE, and MSD for the test set of 723 000 samples during January. Models shown predict brightness temperatures from the vertical 183 ± 3–, 166-, 37-, and 23-GHz GMI channels. For each GMI channel shown, there are four model configurations: the deterministic ResNet (Det ResNet) as shown in section 3a and three types of Bayesian models—Flipout, MC Dropout, and Reparameterization (Reparam).

	183 ± 3 GHz V			166 GHz V			37 GHz V			23 GHz V		
Model	RMSE	MAE	MSD	RMSE	MAE	MSD	RMSE	MAE	MSD	RMSE	MAE	MSD
Det ResNet	2.86	1.71	—	6.55	3.62	—	10.30	7.03	—	14.41	10.27	—
Flipout	2.84	1.66	0.48	6.54	3.83	1.30	10.90	7.67	1.15	17.04	12.23	2.08
MC Dropout	2.95	1.84	0.81	6.37	3.63	1.48	10.83	7.36	2.57	15.81	11.13	3.84
Reparam	2.84	1.70	1.56	6.25	3.63	2.49	10.11	7.02	1.71	13.71	10.42	6.02

MAE of 10.42–12.23 K. Together, these results indicate that the Bayesian and deterministic models produce similar errors in January for a given GMI channel. Furthermore, Table 6 also reports the MSD for the Bayesian models, because the model parameter (epistemic) uncertainty is measurable using the 100 predictions made per input [see Eq. (6)]. For all frequencies except 37 GHz, the Reparameterization models are the most uncertain (had largest MSD) in their predictions, followed by the MC Dropout models and the Flipout models (least uncertain and lowest MSD). For 37 GHz, MC Dropout had the highest uncertainty, followed by Reparameterization, and Flipout still being the least uncertain in its predictions.

To evaluate whether the Bayesian models perform better or worse over time since training data, we repeated the T_b^{mw} predictions for data collected in the next month (February) and four months later (May). Table 7 contains the results of using each model to predict T_b^{mw} using ABI brightness temperatures T_b^{ir} collected daily from 1 to 7 February 2020 (see section 2a). The error amplitude for predictions in this dataset was higher for 183 ± 3 and 166 GHz (~16% and ~28.9% increase in MAE and RMSE), comparable for 37 GHz (~0.36% decrease in MAE and ~3.1% increase in RMSE), and lower for 23 GHz (~12.7% and ~11.3% decrease in MAE and RMSE) relative to the January dataset. The models predicted T_b^{mw} with the least amount of error for the 183 ± 3–GHz channel, with an RMSE of 3.61–3.79 K and an MAE of 1.94–2.10 K. The 166-GHz model predictions were more prone to error with an RMSE of 8.06–8.48 K and an MAE of 4.19–4.48 K. The 37-GHz predictions were even more prone to error than 166 GHz with an RMSE of 10.76–11.11 K and an MAE of 7.17–7.36 K. The 23-GHz model predictions contained the greatest amount of error with an RMSE of 12.03–14.40 K and an MAE of 8.96–10.05 K. As compared with the January results, all models make predictions with a similar amount of error for a given GMI frequency. Additionally, for all frequencies except 37 GHz, the Reparameterization models are the most uncertain in their predictions, followed by the MC Dropout models and the Flipout models (least uncertain). For 37 GHz, MC Dropout had the highest uncertainty, followed by Reparameterization, and Flipout still being the least uncertain in its predictions.

Table 8 contains the results of using each model to predict T_b^{mw} from T_b^{ir} collected daily at a much later date, from 1 to 7 May 2020. The error amplitude for predictions in this dataset were higher for all four GMI frequencies relative to the January dataset. Relative to the results from the February dataset, the models generated an additional ~5% MAE and RMSE on average. The models predicted T_b^{mw} with the least amount of error for 183 ± 3–GHz channel, with an RMSE of 3.64–3.77 K and an MAE of 2.00–2.14 K. The 166-GHz model predictions were more prone to error with an RMSE of 7.79–8.24 K and an MAE of 4.11–4.38 K. The 37-GHz predictions were even more prone to error than 166 GHz with an RMSE of 11.34–12.03 K and an MAE of 7.41–7.75 K. The 23-GHz model predictions contained the greatest amount of error with an RMSE of 14.31–16.29 K and an MAE of 10.11–11.46 K. In general, most Bayesian model predictions in May were slightly more accurate than corresponding deterministic model predictions, excluding those at 37 GHz. In terms of uncertainty, Table 8 indicates the same trends seen in January and February; the Flipout model predictions are least uncertain and the Reparameterization model is most uncertain.

Overall, the Bayesian models we trained generated comparable error to their deterministic counterparts. This means that we can use our Bayesian models where deterministic models might normally be used. Additionally, the overall smaller change in Bayesian model error relative to deterministic models from January to May reported in Tables 6–8 is significant because it indicates that the Bayesian models generalize to unseen data better than our deterministic models. Moreover, this reinforces the findings in the existing literature

TABLE 7. As in Table 6, but with data from February with 857 000 samples.

	183 ± 3 GHz V			166 GHz V			37 GHz V			23 GHz V		
Model	RMSE	MAE	MSD	RMSE	MAE	MSD	RMSE	MAE	MSD	RMSE	MAE	MSD
Det ResNet	3.80	2.01	—	8.18	4.10	—	10.76	7.15	—	13.76	9.71	—
Flipout	3.62	1.94	0.52	8.42	4.48	1.36	10.83	7.36	1.21	14.40	10.05	2.07
MC Dropout	3.79	2.10	0.90	8.48	4.19	1.72	11.11	7.30	2.75	13.86	9.75	4.08
Reparam	3.61	1.97	1.56	8.06	4.28	2.66	10.76	7.17	1.84	12.03	8.96	6.12

TABLE 8. As in Table 6, but with data from May with 925 500 samples.

	183 ± 3 GHz V			166 GHz V			37 GHz V			23 GHz V		
Model	RMSE	MAE	MSD	RMSE	MAE	MSD	RMSE	MAE	MSD	RMSE	MAE	MSD
Det ResNet	3.85	2.05	—	8.45	4.38	—	11.64	7.36	—	16.46	11.52	—
Flipout	3.77	2.02	0.52	8.24	4.38	1.31	11.34	7.41	1.23	16.29	11.46	2.10
MC Dropout	3.77	2.14	0.86	8.24	4.29	1.64	12.03	7.75	2.83	15.74	11.17	4.06
Reparam	3.64	2.01	1.62	7.80	4.11	2.67	11.64	7.46	1.90	14.31	10.11	5.85

that Bayesian neural networks are more robust to overfitting on training data (Neal 2012; Filos et al. 2019; Gal and Ghahramani 2016). The Bayesian models also provide the additional benefit of quantifying the uncertainty of each prediction, and results indicate that predicted uncertainty tends to either remain consistent or slightly increase from January to May. Since this presented uncertainty is epistemic, we expect that it is reducible with additional training data. Furthermore, the predicted uncertainty can be useful in selecting synthetic GMI data for ingestion by downstream applications if the uncertainty is well calibrated.

c. *Uncertainty calibration of Bayesian ResNets*

While having an accurate model is desirable, having a high degree of calibration between predicted variance and error is also very important for downstream applications that utilize uncertainty estimates (e.g., section 2f). Figure 6 demonstrates the degree of calibration between error and uncertainty for each Bayesian model and example GMI channel by depicting the MAE calculated using a variable percentage of the January test data based on predicted standard deviation. For example, if only 80% of test data is used, predictions with standard deviation above the 80th percentile among standard deviations at all data points are excluded. The orange, horizontal, dashed line shows the MAE of each corresponding deterministic ResNet model as a benchmark for reference. The other lines depict the MAE for each Bayesian model using the percentage of the data retained, which is denoted by the abscissa value. To calculate the ordinate values, the standard deviation of each prediction is used to determine the standard deviation value for each percentile. Predictions with a standard deviation greater than the allowed threshold are discarded, and the MAE is calculated for the remaining predictions. For example, in Fig. 6b, 80% of the predictions using the MC Dropout model (purple) have a predictive standard deviation less than or equal to 1.65 K. The corresponding MAE is 2.53 K, which is 70% of the 3.63 K MAE that occurred when no standard deviation threshold was used (Table 6).

Using the definition of well-calibrated uncertainty from Filos et al. (2019) as described in section 2f, the curves in Fig. 6a (183 ± 3 GHz) indicate that the Flipout and the MC Dropout models have well-calibrated uncertainty at 183 ± 3 GHz, while the Reparameterization model does not. Both the Flipout (blue) and MC Dropout (purple) MAE values have a monotonic, increasing relationship as the percent of predictions retained increases, whereas the Reparameterization model (green) only has increasing relationships from approximately 1%–5% and 95%–100% of predictions retained, and the remaining range from approximately 5%–95% of predictions retained exhibits a decreasing relationship between MAE and percent of predictions retained. This means that there are many Reparameterization predictions of 183 ± 3–GHz T_b^{mw} with higher error that are associated with lower standard deviations, which is unsuitable for downstream applications that would use standard deviation as proxy for error.

To further demonstrate the differences of model calibration, an example visual comparison between the poorly calibrated Reparameterization and well-calibrated MC Dropout 183 ± 3–GHz predictions is illustrated in Fig. 7.[1] Figure 7a shows the actual GMI observations, and Fig. 7b shows the predictions for the Reparameterization (bottom row) and MC Dropout (top row) models. Comparison between both model predictions and the GMI observational truth shows that both models are capable of producing highly accurate 183 ± 3–GHz predictions, similar to the deterministic results shown in Fig. 5. In addition, both models have similar absolute error characteristics to the deterministic model, such that clear-sky Bayesian predictions are also associated with error less than 1 K, and absolute errors may exceed 3.7 K in clouds (Fig. 7c). Based only on absolute error shown both in Fig. 7c and Tables 6 and 8, it would be difficult to choose between the 183 ± 3–GHz Reparameterization and MC Dropout model architectures for potential real-time deployment. However, a comparison between the absolute error and uncertainty in Figs. 7c and 7d shows that the spatial distribution of MC Dropout model uncertainty strongly resembles the spatial distribution of absolute error, whereas the spatial distribution of Reparameterization uncertainty bears little resemblance to the actual error distribution. Furthermore, Fig. 7 shows that for the poorly calibrated 183 ± 3–GHz Reparameterization model, the absolute highest uncertainty predictions near Florida correspond with high absolute error, and the lowest uncertainty predictions in the North Atlantic correspond with low absolute error, but uncertainty magnitudes between these two extremes do not exhibit a consistent relationship.

However, the degree of calibration for each Bayesian model is not consistent across all GMI frequencies shown (Fig. 6). While the Reparameterization model (green) calibration is poor for 183 ± 3 and 37-GHz predictions (Figs. 6a,c), it

[1] A Jupyter notebook that reproduces the top row of Fig. 7 and a sample calibration curve using the MC Dropout model are available online (https://github.com/marko-orescanin-nps/Uncertainty-Calibration-of-PMW). Data required to run the code are linked to the notebook.

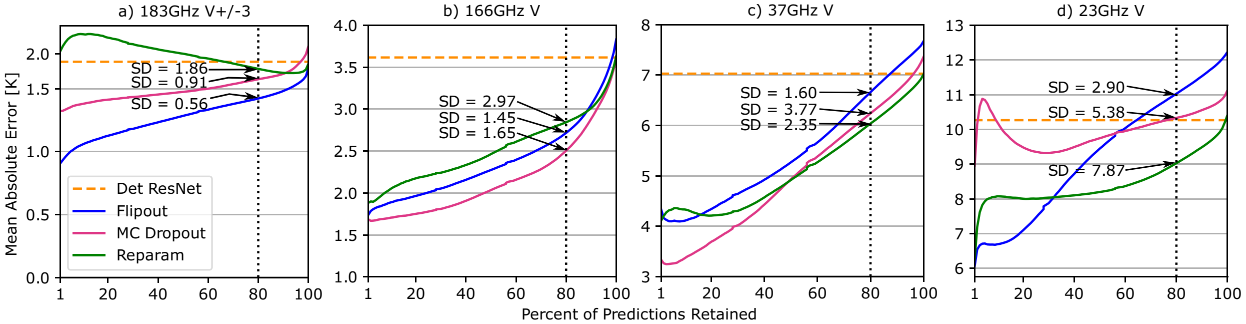

FIG. 6. Change of MAE as a function of retained low-uncertainty data for deterministic ResNet (dashed orange), Flipout (blue), MC Dropout (purple), and Reparameterization (green) models. Standard deviation (SD) values for the 80th percentile of predictive standard deviation of each model are shown in text in each panel; arrows indicate which model each standard deviation describes.

is actually well calibrated at 166 GHz (Fig. 6b). In contrast, the MC Dropout model (purple) is well calibrated for 183 ± 3–, 166-, and 37-GHz predictions (Figs. 6a–c), but not 23 GHz. Instead, the 23-GHz MC Dropout model shows a decreasing relationship from approximately 3%–25% of predictions retained. Finally, the Flipout model (blue) is the most well calibrated across all GMI frequencies shown. Further investigation of Flipout model predictions of $T_b^{\rm mw}$ for the remaining GMI frequencies indicate that the Flipout model is well calibrated across all GMI frequencies (not shown).

When all three implementations display similar uncertainty calibration as if Fig. 6b for 166 GHz, it may be necessary to use other criteria to choose between Bayesian implementations. When averaged across the three datasets, the Reparameterization model produces the lowest average RMSE and average MAE at 166 GHz. Moreover, for a homoscedastic regression problem such as this, the negative log-likelihood is proportional to the mean squared error Dürr et al. (2020). This means that the Reparameterization model also fits more closely to the data for 166 GHz. Additionally, the Reparameterization model has the fastest training time relative to the other two Bayesian implementations. However, MC Dropout is the simplest implementation to code and to train, perhaps making it a more attractive choice. When the uncertainty calibration is similar across Bayesian implementation, researchers should consider the error metrics, the negative log-likelihood (fit to the data), the speed of training, and the ease of implementation/training when choosing what type of Bayesian model to put into production or to use for future experimentation.

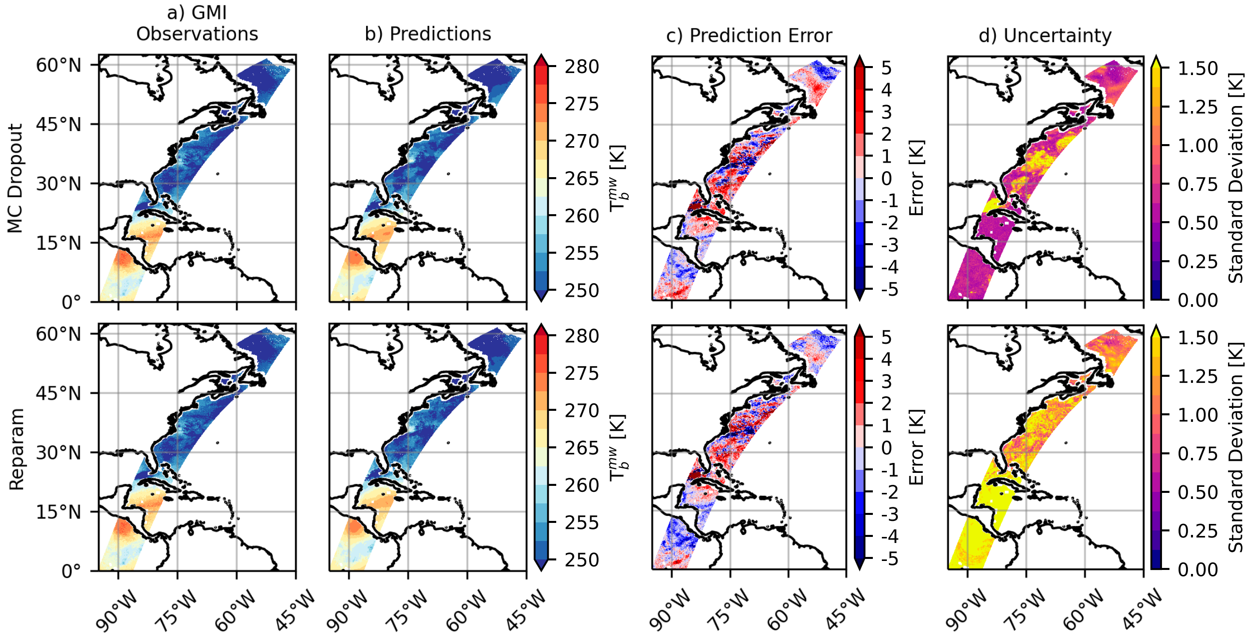

FIG. 7. GMI swath at the 183 ± 3–GHz channel around 1440 UTC 1 Feb 2022 (GPM orbit number 33679): (a) observed $T_b^{\rm mw}$, (b) predicted $T_b^{\rm mw}$ from the (top) MC Dropout model and (bottom) Reparameterization model, (c) prediction absolute error, and (d) predictive standard deviation.

TABLE 9. Error and uncertainty metrics for February test set after removing predictions with variance greater than or equal to the 80th-percentile predictive variance value in January data. The variance values for the 80th percentile of each channel are depicted in Fig. 6. The values for the deterministic ResNet in this table are the same as in Table 7 since deterministic models do not provide predictive variance.

	183 ± 3 GHz V			166 GHz V			37 GHz V			23 GHz V		
Model	RMSE	MAE	MSD	RMSE	MAE	MSD	RMSE	MAE	MSD	RMSE	MAE	MSD
Det ResNet	3.80	2.01	—	8.18	4.10	—	10.76	7.15	—	13.76	9.71	—
Flipout	2.09	1.46	0.41	4.80	2.87	1.01	9.11	6.02	0.85	12.76	8.78	1.49
MC Dropout	2.17	1.58	0.66	3.80	2.40	1.06	9.34	5.84	1.80	12.28	8.45	2.86
Reparam	3.22	1.85	1.40	5.21	3.03	2.04	9.07	5.84	1.27	10.22	7.53	4.88

Together, these results show that while the skill of each Bayesian model per GMI channel was similar to deterministic model skill, differences arise between Bayesian model architecture results when examining the degree of calibration for each model across the example GMI frequencies shown. Neither the Reparameterization nor the MC Dropout model architectures were well calibrated across all GMI frequencies. Therefore, these models can only be reasonably applied toward predicting certain GMI frequencies, and thus have a more limited utility. Only the Flipout model was well calibrated across all GMI frequencies shown when using the January dataset; moreover, we confirmed that the Flipout model had well-calibrated uncertainty for 183, 166 and 37 GHz using the February and May datasets (not depicted). For 23 GHz, the Flipout model had well-calibrated uncertainty using the February dataset but had a loss of calibration for the May dataset for samples with the uncertainty values in the lowest 7%. The Flipout model is therefore the most robust model architecture for our regression task. Thus, the Flipout model architecture is the favored choice of Bayesian model architecture for further model improvement and possible eventual deployment. We use the 183 ± 3–GHz Flipout model results to demonstrate a practical application of using the predictive variance to reduce the amount of error in synthetic GMI data while maintaining high spatiotemporal resolution. The appendix contains Flipout model errors and MSDs of all 13 GMI channels.

d. *Combining high spatiotemporal resolution with quantified uncertainty*

Since many downstream applications of PMW data, such as assimilation of clear-sky PMW brightness temperatures into numerical weather prediction models, are highly sensitive to inaccurate T_b^{mw} magnitudes, having the ability to quickly filter out the least accurate synthetic T_b^{mw} predictions prior to assimilation is very desirable. With deterministic models, this is only possible when there are collocated GMI observations to calculate error, which strongly limits the utility of deterministic models. In contrast, the uncertainty of well-calibrated Bayesian models can be used as a proxy for error even when no corresponding GMI observations exist, and appropriate tolerance thresholds of uncertainty can be tailored to downstream requirements to reduce the total amount of error of synthetic GMI data that would get ingested by these downstream applications while still retaining a vast increase of spatiotemporal resolution relative to existing GMI observations.

As an illustrative case, we use the 80th percentile standard deviation values derived from the January predictions for each model depicted in Fig. 6 to discard elements from the set of February predictions with a standard deviation greater than the January-derived values. In practice, the standard deviation threshold could be selected to suit the needs of downstream applications that ingest this synthetic data. Table 9 contains the recalculated metrics for the February 183 ± 3–, 166-GHz V, 37-GHz V, and 23-GHz V datasets. The values in the first row of Table 9 are unchanged from Table 7 since deterministic models provide no measure of predictive uncertainty; moreover, all three Bayesian models achieve lower error metrics on the February data than the deterministic model does on the January data with the exception of the Reparameterization model for 183 ± 3 GHz, which does not have well-calibrated uncertainty. When compared with the results for the same model in Table 7 (February), all three Bayesian models had a decrease in error for all metrics. This decrease is larger for the Flipout and MC Dropout models than for the Reparameterization models. Furthermore, the difference between the RMSE and the MAE is smaller for all three Bayesian models, meaning that the magnitude of the larger errors has decreased. These results reinforce the choice of the Flipout model for deployment to predict 183 ± 3–GHz T_b^{mw} because it has the lowest error metrics when compared with all other models.

To visually demonstrate the impact of combining high spatiotemporal resolution Bayesian model predictions and quantified uncertainty, we highlight results from the most accurate and well-calibrated model, the 183 ± 3–GHz Flipout model. Figure 8 shows predictions of both T_b^{mw} and MSD over the western Atlantic and eastern Pacific on 1440 UTC 1 February 2020. It illustrates the increase of spatial microwave data coverage for a given time period relative to existing observations, and additional synthetic microwave data and uncertainty can be predicted at the same time interval as ABI full-disk scans are available (10–15 min depending on the scanning strategy). Figures 8b and 8c also shows that the Flipout model predictions tends to have highest uncertainty when T_b^{mw} is coldest, which occurs mostly where scattering by hydrometeors in mid- to upper-tropospheric clouds occurs. Since Fig. 6a previously established that this model is well calibrated, we can infer that the exact representation of T_b^{mw} in clouds likely has

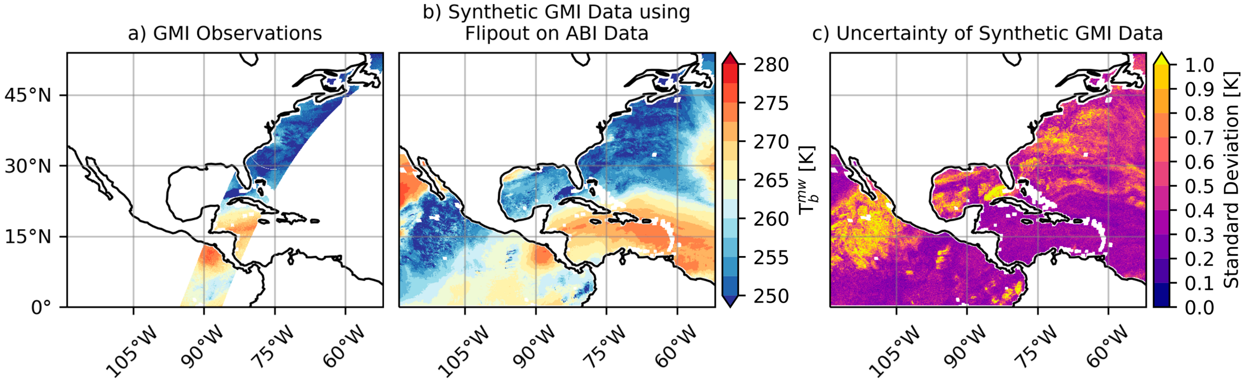

FIG. 8. (a) Observed T_b^{mw} at 1440 UTC 1 Feb 2020 (GPM orbit number 33679), (b) predicted T_b^{mw} from the 183 ± 3–GHz V Flipout model over a wide area, and (c) predictive standard deviation for each synthetic data point in (b).

some error over the entire domain, not just where GMI observations exist.

The same 80th-percentile standard deviation threshold is then applied to the same predictions shown in Fig. 8, and three example zoomed-in features are in shown Fig. 9. Specifically, the three examples of Fig. 9 lie within 15° × 15° areas centered near Central America (left column), western Cuba (middle column), and offshore the northeastern United States (right column). Figure 9a first shows all predictions regardless of predicted variance magnitudes, as in as in Fig. 8b. Figure 9b contains the standard deviation associated with each set of predictions. Using the 80th percentile standard deviation from Fig. 6b of 0.56 K as a threshold, Fig. 9c contains only predictions with standard deviation less than this threshold value; other locations are not filled in and remain white. As expected, Fig. 9c contains about 80% of the predictions from Fig. 9a, but with missing predictions in areas that have high associated standard deviation. Figure 9d contains the ABI data from ABI band 14 (11.2 μm). The areas of high standard deviation appear to correspond closely to areas with thick cloud cover, which is indicated by areas of low T_b^{ir} (brighter grays and white) in Fig. 9d. However, not all predictions in cloudy regions have the same level of uncertainty. Many clouds near the Bay of Campeche, north of 25°N near Florida, and near Massachusetts occur in areas of low predictive uncertainty. These clouds also appear to be lower altitude and are likely to be optically thinner (see Fig. 9d) and possibly contain less ice, which suggests that there may be more skill when there is not significant scattering by hydrometeors.

Overall, we show that implementing an 80th percentile threshold leads to a reduction in mean absolute error. If error and predicted uncertainty are well calibrated, stricter thresholds below the 80th percentile will yield even lower error predictions. However, there is a trade-off between amount of data retained versus skill, and exact thresholds can vary on downstream application tolerances. We also show that an 80th percentile threshold of uncertainty yields primarily filters out the clouds that are deepest and have the coldest T_b^{ir} magnitudes, but lower clouds with warmer T_b^{ir} magnitudes are retained.

4. Summary and conclusions

In this study, we developed a total of 34 deterministic and Bayesian residual network (ResNet) deep learning models for the regression task of predicting GMI passive microwave (PMW) brightness temperatures over ocean from *GOES-16* ABI infrared brightness temperatures. Deterministic models were developed for each GMI band, and resulting synthetic GMI data produced by a model trained on just data from a single month (January 2020) has a mean absolute error (MAE) as low as 1.72 K for a GMI band centered at 183 ± 3 GHz. Errors for GMI channels generally increased as the observed range and/or variability of the distributions of their brightness temperatures increased (Fig. 1 and Table 3), which resulted in errors that were lower for vertically polarized channels than horizontally polarized ones. GMI channels associated with lower atmospheric opacity (e.g., 10.6–18.7 GHz) also had the greatest increase in error over time. Together, these deterministic models also demonstrate how to generate synthetic GMI data with the same spatiotemporal resolution as ABI, and they establish a baseline skill for later comparison. However, the deterministic models lack estimates of uncertainty that are used as weights in downstream meteorological applications, such as clear- or all-sky retrievals that utilize optimal estimation or numerical weather prediction models utilizing various data assimilation methods.

To address the existing need for quantitative uncertainty estimates, we then adapted three types of Bayesian models [i.e., Flipout, Monte Carlo (MC) Dropout, and Reparameterization] from our deterministic ResNet architecture to produce synthetic GMI data at 23, 37, 166, and 183 ± 3 GHz, which are sensitive in different ways to water vapor, liquid water, and ice. Comparison of the MAE and root-mean-square error between corresponding deterministic and Bayesian models reveal that model skill is not sacrificed to produce quantitative estimates of predictive uncertainty when the same training strategy is utilized. Additionally, our results indicate that the Bayesian models had a smaller decrease in skill from January to May and are therefore more robust to overfitting, consistent with the findings of Neal (2012). In addition to error, we also examine the uncertainty calibration of each of the three

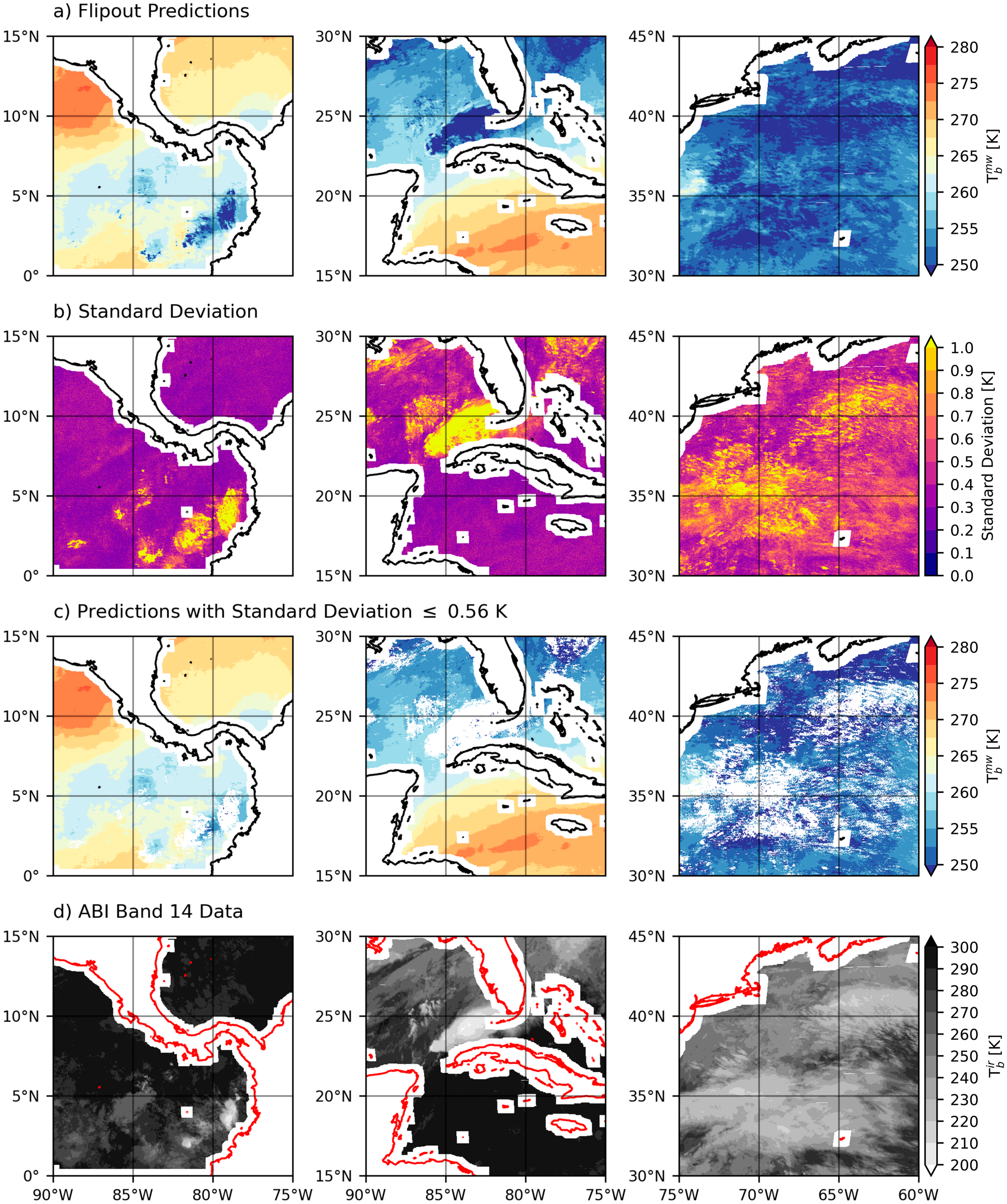

FIG. 9. (a) All model predictions for 183 ± 3 GHz using Flipout at 1440 UTC 1 Feb 2022, (b) standard deviation associated with each prediction, (c) predictions with standard deviation less than the 80th-percentile standard deviation value when predicting on the January training data (see Fig. 6), and (d) the corresponding ABI data from 11.2-μm band (band 14).

Bayesian ResNet models; uncertainty must be well calibrated in order to infer a likely amount of error from predictive variance. We find that the 183 ± 3–GHz Reparameterization model and the 23-GHz MC Dropout model had a higher number of predictions with low predictive standard deviation but high absolute error, and are therefore not well calibrated. For this reason, we generally favor the Flipout configuration since it is the most consistently well calibrated across all GMI channels for which a model was produced.

We used our most accurate Flipout model, for 183 ± 3 GHz, to demonstrate the full benefit of our Bayesian architecture in a meteorological context. Like the deterministic models, the Flipout model produced synthetic PMW brightness temperatures at ABI spatiotemporal resolutions over the entire ocean-only portion of the *GOES-16* domain, which may allow for new pathways of investigation into the evolution of individual meteorological features of interest with passive microwave data in the future. However, the primary benefit came from the predictive uncertainty quantification, which generally showed that the presence of clouds increased the predictive uncertainty in our models. Comparison of variance across different cloud types in Fig. 9 suggested that lower, relatively warmer clouds, such as those shown off the coast of Massachusetts, were associated with lower variance than deeper, relatively colder clouds, such as those shown near Florida and South America. Without knowledge of the predictive variance, we would not have known whether predictions without corresponding labels likely had skill.

This decreased skill in clouds may be due to two reasons. First, the information provided to the model as input features, infrared brightness temperatures, may be insufficient to fully inform a neural network of cloud properties that might contribute to PMW brightness temperatures in cloudy regions. In particular, both the fact that the highest predictive variance is associated with colder, thicker clouds (such as those associated with deep convection) and that the various 37- and 89-GHz models had the largest absolute errors suggests that additional information is likely needed to gain skill when significant scattering from ice is present. We speculate that including multispectral visible radiances/reflectance or products related to microphysical composition of clouds in our training data would increase predictive skill. Second, the model may simply have insufficient data to train on. The general decrease in model skill from February to May predictions suggests that just including additional infrared training data over a longer period of time will likely also help. We also suspect that simply training a model on more clouds instead of relying on upsampling to produce unbiased predictions will help reduce error in cloudy regions. Indeed, ongoing work by the authors indicates that both are likely true. Because the lowest frequency GMI predictions (10.6–18.7 GHz) were associated with the highest increase in error from January to May in deterministic models, we speculate that incorporation of surface-based or low-level input features (such as sea surface temperature, in optically thick regions) may also help to further reduce error for these channels. Together, we expect that making such improvements could yield a Bayesian model based on our initial proof-of-concept methodology that produces higher skill in cloudy and precipitating regimes as well as clear-air regions.

To ascertain whether just additional infrared training data is sufficient to improve future model skill or whether additional input features are needed (such as adding visible data), the predictive variance must be decomposed into its aleatoric and epistemic components. In this work, we specifically focused on modeling epistemic uncertainty, which contains information about the uncertainty in the model and can be reduced by training on additional data of the same set of features. In contrast, aleatoric uncertainty cannot be reduced by additional data of the same set of features, and a combination of high aleatoric uncertainty and near-zero epistemic uncertainty implies that additional input features are needed to further increase model skill. However, to predict aleatoric uncertainty modifications to our model architecture are necessary. Ongoing and future work will add this functionality to our Bayesian model architecture. Additionally, future work will also explore predicting more realistic, nonparametric error distributions, since our current models assume that predictive error distributions are Gaussian.

The results presented provide additional evidence that deep learning in combination with Bayesian models have potential to provide additional high-value information content in meteorological applications (Orescanin et al. 2021; Ortiz et al. 2022). We expect that the ability to produce highly accurate emulations of PMW data at virtually continuous spatiotemporal resolution will open new frontiers in modeling, retrieving, and analyzing atmospheric properties.

Acknowledgments. This work has been supported by the Office of Naval Research Grants N0001421WX00575 and N0001422WX01251 and National Oceanic and Atmospheric Administration Grant NA19NES4320002 to the Cooperative Institute for Satellite Earth System Studies (CISESS) at the University of Maryland.

Data availability statement. The *GOES-16* dataset is available online on a public S3 bucket (s3://noaa-goes16/ABI-L2-MCMIPF/2020/). The GMI brightness temperatures are publicly available online from NASA (https://gpm.nasa.gov/data/directory). This study used the Level-1B product for GMI brightness temperatures.

APPENDIX

Additional Bayesian Model Results

Tables A1–A3 show the MAE, RMSE, and MSD of Bayesian Flipout model predictions for all 13 GMI channels in January, February, and May, respectively. Comparison with Tables 3–5 shows that predictive skill is similar between deterministic and Bayesian models for each channel. For example, both the deterministic and Flipout model architectures have the lowest error at 183 ± 3 GHz, and the horizontal 37- and 89-GHz predictions have the largest error in all three months. Additionally, differences between the skill of each model architecture per channel show that the horizontal 10.6-GHz channel has the largest difference in MAE at 2.01 K in January. Further comparison between the deterministic and Bayesian models shows that in general, the Bayesian models perform marginally better in the scattering-dominant channels of 89-GHz-and-higher frequencies. For channels that are more sensitive to emission, the deterministic models perform slightly better. However, differences in skill are small.

TABLE A1. Summary of Bayesian Flipout model performance for the test set of 723 000 samples during January.

	10.6 GHz		18.7 GHz		23 GHz	37 GHz		89 GHz		166 GHz		183 GHz	
Polarization	H	V	H	V	V	H	V	H	V	H	V	±3 V	±7 V
MAE	8.94	6.21	14.10	8.66	12.23	14.55	7.67	13.44	5.87	5.53	3.83	1.66	2.51
RMSE	17.52	13.08	21.95	14.53	17.04	21.00	10.90	18.00	7.86	8.72	6.54	2.84	4.59
MSD	1.81	1.33	2.94	0.83	2.08	2.18	1.15	3.59	1.89	0.89	1.30	0.48	0.82

TABLE A2. Summary of Flipout model performance on data from February with 857 000 samples.

	10.6 GHz		18.7 GHz		23 GHz	37 GHz		89 GHz		166 GHz		183 GHz	
Polarization	H	V	H	V	V	H	V	H	V	H	V	±3 V	±7 V
MAE	9.16	6.07	12.84	8.13	10.05	14.76	7.36	13.79	6.03	6.26	4.48	1.94	3.06
RMSE	20.40	13.37	20.84	14.01	14.40	21.87	10.83	19.02	8.02	11.08	8.42	3.62	6.21
MSD	1.68	1.33	3.07	0.88	2.07	2.34	1.21	3.64	1.80	0.95	1.36	0.52	0.88

TABLE A3. Summary of Flipout model performance on data from May with 925 500 samples.

	10.6 GHz		18.7 GHz		23 GHz	37 GHz		89 GHz		166 GHz		183 GHz	
Polarization	H	V	H	V	V	H	V	H	V	H	V	±3 V	±7 V
MAE	11.78	7.85	14.44	9.32	11.46	14.62	7.41	14.67	6.33	6.79	4.38	2.02	2.61
RMSE	27.83	19.63	25.56	17.18	16.29	22.19	11.34	19.98	9.01	11.45	8.24	3.77	5.38
MSD	1.76	1.31	3.05	0.83	2.10	2.34	1.23	3.80	1.93	0.94	1.31	0.52	0.81

REFERENCES

Abadi, M., and Coauthors, 2015: TensorFlow: Large-scale machine learning on heterogeneous systems. TensorFlow, https://www.tensorflow.org/.

Adler, R. F., and Coauthors, 2003: The Version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.*, **4**, 1147–1167, https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe, 2017: Variational inference: A review for statisticians. *J. Amer. Stat. Assoc.*, **112**, 859–877, https://doi.org/10.1080/01621459.2017.1285773.

Bonavita, M., A. J. Geer, and M. Hamrud, 2020: All-sky microwave radiances assimilated with an ensemble Kalman filter. *Mon. Wea. Rev.*, **148**, 2737–2760, https://doi.org/10.1175/MWR-D-19-0413.1.

Carver, K., C. Elachi, and F. Ulaby, 1985: Microwave remote sensing from space. *Proc. IEEE*, **73**, 970–996, https://doi.org/10.1109/PROC.1985.13230.

Chen, B.-F., B. Chen, H.-T. Lin, and R. L. Elsberry, 2019: Estimating tropical cyclone intensity by satellite imagery utilizing convolutional neural networks. *Wea. Forecasting*, **34**, 447–465, https://doi.org/10.1175/WAF-D-18-0136.1.

Dee, D. P., 2004: Variational bias correction of radiance data in the ECMWF system. *Workshop on Assimilation of High Spectral Resolution Sounders in NWP*, Reading, United Kingdom, ECMWF, 97–112, https://www.ecmwf.int/node/8930.

Derber, J. C., and W.-S. Wu, 1998: The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. *Mon. Wea. Rev.*, **126**, 2287–2299, https://doi.org/10.1175/1520-0493(1998)126⟨2287:TUOTCC⟩2.0.CO;2.

Dillon, J. V., and Coauthors, 2017: TensorFlow distributions. arXiv, 1711.10604v1, https://doi.org/10.48550/arXiv.1711.10604.

Draper, D. W., D. A. Newell, F. J. Wentz, S. Krimchansky, and G. M. Skofronick-Jackson, 2015: The Global Precipitation Measurement (GPM) Microwave Imager (GMI): Instrument overview and early on-orbit performance. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **8**, 3452–3462, https://doi.org/10.1109/JSTARS.2015.2403303.

Dürr, O., B. Sick, and E. Murina, 2020: *Probabilistic Deep Learning: With Python, Keras, and TensorFlow Probability*. Manning, 296 pp.

Ebert-Uphoff, I., R. Lagerquist, K. Hilburn, Y. Lee, K. Haynes, J. Stock, C. Kumler, and J. Q. Stewart, 2021: CIRA guide to custom loss functions for neural networks in environmental sciences—Version 1. arXiv, 2106.09757v1, https://doi.org/10.48550/arXiv.2106.09757.

Errico, R. M., P. Bauer, and J.-F. Mahfouf, 2007: Issues regarding the assimilation of cloud and precipitation data. *J. Atmos. Sci.*, **64**, 3785–3798, https://doi.org/10.1175/2006JAS2044.1.

Feng, R., N. Balling, D. Grana, J. S. Dramsch, and T. M. Hansen, 2021: Bayesian convolutional neural networks for seismic facies classification. *IEEE Trans. Geosci. Remote Sens.*, **59**, 8933–8940, https://doi.org/10.1109/TGRS.2020.3049012.

Filos, A., and Coauthors, 2019: A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks. arXiv, 1912.10481v1, https://doi.org/10.48550/arXiv.1912.10481.

Gal, Y., and Z. Ghahramani, 2016: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proc. 33rd Int. Conf. on Machine Learning*, New York, NY, JMLR, 1050–1059, https://dl.acm.org/doi/10.5555/3045390.3045502.

Geer, A. J., and Coauthors, 2017: The growing impact of satellite observations sensitive to humidity, cloud and precipitation: Impact of satellite humidity, cloud and precipitation observations. *Quart. J. Roy. Meteor. Soc.*, **143**, 3189–3206, https://doi.org/10.1002/qj.3172.

Giffard-Roisin, S., M. Yang, G. Charpiat, C. Kumler Bonfanti, B. Kégl, and C. Monteleoni, 2020: Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data. *Front. Big Data*, **3**, 1, https://doi.org/10.3389/fdata.2020.00001.

Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. Adaptive Computation and Machine Learning Series, MIT Press, 800 pp.

Greenwald, T. J., R. Bennartz, M. Lebsock, and J. Teixeira, 2018: An uncertainty data set for passive microwave satellite observations of warm cloud liquid water path. *J. Geophys. Res. Atmos.*, **123**, 3668–3687, https://doi.org/10.1002/2017JD027638.

Guilloteau, C., and E. Foufoula-Georgiou, 2020: Beyond the pixel: Using patterns and multiscale spatial information to improve the retrieval of precipitation from spaceborne passive microwave imagers. *J. Atmos. Oceanic Technol.*, **37**, 1571–1591, https://doi.org/10.1175/JTECH-D-19-0067.1.

Harris, C. R., and Coauthors, 2020: Array programming with NumPy. *Nature*, **585**, 357–362, https://doi.org/10.1038/s41586-020-2649-2.

He, K., X. Zhang, S. Ren, and J. Sun, 2016: Identity mappings in deep residual networks. *Computer Vision—ECCV 2016*, B. Leibe et al., Eds., Lecture Notes in Computer Science, Vol. 9908, Springer International Publishing, 630–645, https://doi.org/10.1007/978-3-319-46493-0_38.

Hilburn, K., 2020: Inferring airmass properties from GOES-R ABI observations. *2020 Fall Meeting*, Online, Amer. Geophys. Union, Abstract A008–0009, https://ui.adsabs.harvard.edu/abs/2020AGUFMA008.0009H/abstract.

Hilburn, K. A., and F. J. Wentz, 2008: Intercalibrated passive microwave rain products from the Unified Microwave Ocean Retrieval Algorithm (UMORA). *J. Appl. Meteor. Climatol.*, **47**, 778–794, https://doi.org/10.1175/2007JAMC1635.1.

Kendall, A., and Y. Gal, 2017: What uncertainties do we need in Bayesian deep learning for computer vision? *Proc. 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, ACM, 5580–5590, https://dl.acm.org/doi/10.5555/3295222.3295309.

Kingma, D. P., and J. Ba, 2017: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

——, T. Salimans, and M. Welling, 2015: Variational dropout and the local reparameterization trick. *Proc. 28th Int. Conf. on Neural Information Processing Systems*, Montreal, QC, Canada, ACM, 2575–2583, https://dl.acm.org/doi/10.5555/2969442.2969527.

Kiureghian, A. D., and O. Ditlevsen, 2009: Aleatory or epistemic? Does it matter? *Struct. Saf.*, **31**, 105–112, https://doi.org/10.1016/j.strusafe.2008.06.020.

Kulie, M. S., R. Bennartz, T. J. Greenwald, Y. Chen, and F. Weng, 2010: Uncertainties in microwave properties of frozen precipitation: Implications for remote sensing and data assimilation. *J. Atmos. Sci.*, **67**, 3471–3487, https://doi.org/10.1175/2010JAS3520.1.

Kummerow, C., W. Olson, and L. Giglio, 1996: A simplified scheme for obtaining precipitation and vertical hydrometeor profiles from passive microwave sensors. *IEEE Trans. Geosci. Remote Sens.*, **34**, 1213–1232, https://doi.org/10.1109/36.536538.

Lee, J., J. Im, D.-H. Cha, H. Park, and S. Sim, 2019: Tropical cyclone intensity estimation using multi-dimensional convolutional neural networks from geostationary satellite data. *Remote Sens.*, **12**, 108, https://doi.org/10.3390/rs12010108.

Leibig, C., V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, 2017: Leveraging uncertainty information from deep neural networks for disease detection. *Sci. Rep.*, **7**, 17816, https://doi.org/10.1038/s41598-017-17876-z.

Li, Y., C. Wei, and T. Ma, 2019: Towards explaining the regularization effect of initial large learning rate in training neural networks. *Proc. 33rd Conf. on Neural Information Processing Systems*, Vancouver, BC, Canada, NeurIPS, 11 674–11 685, https://proceedings.neurips.cc/paper/2019/file/bce9abf229ffd7e570818476ee5d7dde-Paper.pdf.

Maskey, M., and Coauthors, 2020: Deepti: Deep-learning-based tropical cyclone intensity estimation system. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **13**, 4271–4281, https://doi.org/10.1109/JSTARS.2020.3011907.

Migliorini, S., and B. Candy, 2019: All-sky satellite data assimilation of microwave temperature sounding channels at the Met Office. *Quart. J. Roy. Meteor. Soc.*, **145**, 867–883, https://doi.org/10.1002/qj.3470.

Mohamed, S., M. Rosca, M. Figurnov, and A. Mnih, 2020: Monte Carlo gradient estimation in machine learning. arXiv, 1906.10652v2, https://doi.org/10.48550/arXiv.1906.10652.

NASA, 2017: ABI bands quick information guides. Accessed 10 February 2022, https://www.goes-r.gov/mission/ABI-bands-quick-info.html.

Neal, R. M., 2012: *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics, Vol. 118, Springer Science & Business Media, 204 pp.

Orescanin, M., V. Petković, S. W. Powell, B. R. Marsh, and S. C. Heslin, 2021: Bayesian deep learning for passive microwave precipitation type detection. *IEEE Geosci. Remote Sens. Lett.*, **19**, 1–5, https://doi.org/10.1109/LGRS.2021.3090743.

Ortiz, P., M. Orescanin, V. Petkovic, S. W. Powell, and B. Marsh, 2022: Decomposing satellite-based classification uncertainties in large Earth science datasets. *IEEE Trans. Geosci. Remote Sens.*, **60**, 1–11, https://doi.org/10.1109/TGRS.2022.3152516.

Petković, V., M. Orescanin, P. Kirstetter, C. Kummerow, and R. Ferraro, 2019: Enhancing PMW satellite precipitation estimation: Detecting convective class. *J. Atmos. Oceanic Technol.*, **36**, 2349–2363, https://doi.org/10.1175/JTECH-D-19-0008.1.

Pu, Z., C. Yu, V. Tallapragada, J. Jin, and W. McCarty, 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. *Mon. Wea. Rev.*, **147**, 175–198, https://doi.org/10.1175/MWR-D-17-0200.1.

Rodgers, C. D., 2000: *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific Publishing, 256 pp.

Schmit, T. J., M. M. Gunshor, W. P. Menzel, and J. James, 2005: Introducing the next-generation Advanced Baseline Imager on GOES-R. *Bull. Amer. Meteor. Soc.*, **86**, 1079–1096, https://doi.org/10.1175/BAMS-86-8-1079.

Schulte, R. M., C. D. Kummerow, C. Klepp, and G. G. Mace, 2022: How accurately can warm rain realistically be retrieved with satellite sensors? Part I: DSD uncertainties. *J. Appl. Meteor. Climatol.*, **61**, 1087–1105, https://doi.org/10.1175/JAMC-D-21-0158.1.

Slocum, C., and J. Knaff, 2020: Using geostationary imagery to peer through the clouds revealing hurricane structure. *19th Conf. on Artificial Intelligence for Environmental Science*, Boston, MA, Amer. Meteor. Soc., J43.1, https://ams.confex.com/ams/2020Annual/webprogram/Paper369772.html.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014: Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Tran, D., M. Dusenberry, M. van der Wilk, and D. Hafner, 2019: Bayesian layers: A module for neural network uncertainty. *Proc. 33rd Conf. on Neural Information Processing Systems*, Vancouver, BC, Canada, NeurIPS, https://proceedings.neurips.cc/paper/2019/file/154ff8944e6eac05d0675c95b5b8889d-Paper.pdf.

Wang, Y., and Coauthors, 2022: Verification of operational numerical weather prediction model forecasts of precipitation using satellite rainfall estimates over Africa. arXiv, 2201.02296v1, https://doi.org/10.48550/arXiv.2201.02296.

Wen, Y., P. Vicol, J. Ba, D. Tran, and R. Grosse, 2018: Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *Sixth Int. Conf. on Learning Representations*, Vancouver, BC, Canada, ICLR, https://openreview.net/pdf?id=rJNpifWAb.

Weng, F., 2017: Assimilation of microwave data in regional NWP models. *Passive Microwave Remote Sensing of the Earth*, John Wiley and Sons, 259–297, https://doi.org/10.1002/9783527336289.ch10.

Wimmers, A., C. Velden, and J. H. Cossuth, 2019: Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Mon. Wea. Rev.*, **147**, 2261–2282, https://doi.org/10.1175/MWR-D-18-0391.1.