Decomposing Satellite-Based Classification Uncertainties in Large Earth Science Datasets

Pedro Ortiz, Marko Orescanin, *Member, IEEE,* Veljko Petković, Scott W. Powell, and Benjamin Marsh

***Abstract*—Collection of increasingly voluminous multi-spectral data from multiple instruments with high spatial resolution has posed both an opportunity and a challenge for maximizing their utilization, analysis, and impact. Obtaining accurate estimates of precipitation globally with high temporal resolution is crucial for assessing multi-scale hydrologic impacts and providing a constraint for development of numerical models of the atmosphere that provide weather and climate predictions. Precipitation type classification plays an important role in constraining both the inverse problem in satellite precipitation retrievals and latent heat transfer within weather prediction simulations. Precipitation type, however, is often reported deterministically, without uncertainty attached to an estimate. Machine learning techniques are capable of extracting content of interest from large datasets and accurately retrieving discrete and continuous properties of physical systems, but with limited insights to the retrieval components–such as errors and the physical relationship between the observed and retrieved properties. To address this shortcoming, we perform precipitation type classification to introduce a novel tool for decomposing errors of satellite-retrieved products. We use Bayesian neural networks to map Global Precipitation Measurement mission Microwave Imager observations to Dual-frequency Precipitation Radar-derived precipitation type, which perform comparably to deterministic models, but with the added benefit of providing well calibrated uncertainties. Through uncertainty decomposition, we demonstrate well calibrated uncertainties as useful for making decisions concerning high uncertainty predictions, model selection, targeted data analysis, and data collection and processing. Additionally, our Bayesian models enable mathematical confirmation of a data distribution change as the cause for an unacceptable decline in model accuracy.**

I. Introduction

THE Global Precipitation Measurement (GPM) mission [1] uses a constellation of passive microwave radiometers to offer a nearly global sampling of rain and snowfall rate estimates. The GPM core-observatory carries a passive Microwave Imager (GMI) [2] and an advanced Dual-frequency Precipitation Radar (DPR) system [3]. The two instruments are used to build a link between passive microwave (PMW) brightness temperatures and radar-derived precipitation rates. This link is then employed by an enterprise precipitation retrieval [4] to provide global estimates of precipitation rates. Driven by the globally-observed link, the retrieval delivers global precipitation estimates but suffers from region-specific biases [5] induced by unaddressed variability in precipitation system morphology. If provided, the information about the precipitation type significantly mitigates this problem, as shown in [5] where a simple machine learning model was employed to predict precipitation class (i.e., convective vs. stratiform type). Convective rainfall is usually associated with stronger vertical motions and heavier rainfall than stratiform precipitation [6].

While demonstration studies confirm the great potential of machine learning (ML) methods in solving this particular problem, in order to operationally apply ML, a model must prove not only to be accurate but also to be capable of quantifying how predictive uncertainty varies when a model is applied to different types of precipitation systems (e.g., tropical cyclones, mesoscale convective systems, disorganized precipitation) across various regions on Earth. This becomes especially important when an enterprise retrieval is used, such as the Goddard Profiling Algorithm (GPROF) [4], which must operate over the entire globe as observed data distributions (i.e., brightness temperatures from different PMW radiometric bands) may change over time. The main drivers of variability in the information content are commonly seen in technical characteristics and age of the sensors used for the enterprise products, such as those from the GPM satellite constellation. Although the properties of each sensor are well understood, the effect of their variability on the retrieval performance with deep learning is not. Providing such information remains a challenge; however, this study offers one possible solution to the problem.

Recently, data from the GPM mission was used to apply novel Bayesian deep learning (BDL) models to improve precipitation type classification of multi-spectral PMW observations of precipitation events [7]. Orescanin et al. demonstrated that BDL models can establish a stronger link between raw GMI data and precipitation system morphology over ocean-based precipitation events than deterministic deep learning (DDL) models. Furthermore, these BDL models outperformed the GPROF precipitation type product, part of the standard output of the currently operational precipitation retrieval for the GPM mission [4], [8]. The models successfully combined deep learning with Bayesian statistics to provide accurate precipitation type predictions while simultaneously providing useful measures of uncertainty [7]. Previously, BDL models have been shown to provide uncertainties in decision making, to learn useful information about small datasets, and to be more robust to overfitting to the training data than their deterministic counterparts [9]–[12]. Recent applications of BDL to remote sensing tasks include Active Learning tasks on Synthetic Aperture Radar (SAR) data [13] and hyperspectral imagery [11] using MC Dropout as a variational

Pedro Ortiz, Marko Orescanin, Scott Powell and Benjamin Marsh are with the Naval Postgraduate School, Monterey, CA, 93943 USA (e-mail: pedro.ortiz@nps.edu; marko.orescanin@nps.edu).

Veljko Petković is with ESSIC, CISESS, University of Maryland, College Park, MD 20740 USA.

Manuscript received MONTH DAY, 2021; revised MONTH DAY, 2021.

Bayesian approximation [10] with rudimentary convolutional deep learning models. Additional recent work focuses on seismic facies classification [14] using BDL with a simple convolutional model containing several hidden layers.

The results from Orescanin et al. [7] provide a realistic real-world benchmark using a large Earth Science dataset that provides useful measures of the per-pixel uncertainties, quantified by predictive entropy, which has been a known gap in existing literature [12]. Further, those results demonstrated that high uncertainty was correlated with misclassified pixels [7]. The models had well calibrated uncertainties demonstrated by rejecting data points with high entropy values, which caused model performance on the remaining data points to increase. However, bulk or total uncertainty information, such as predictive entropy or variance [7], [12], fails to identify sources of uncertainty in the developed model. Der Kiureghian and Ditlevsen [15] characterize uncertainty as epistemic if it can be reduced or as aleatoric uncertainty if it can not be further reduced (e.g., caused by noise from the sensor).

Our key contributions in this article are

- Combining BDL with a meaningful real-world remote sensing application to create models with well calibrated uncertainties.
- Decomposing uncertainty into its aleatoric and epistemic components [16] to make decisions about high uncertainty predictions, model selection, targeted data analysis, data collection/processing.
- Providing a method to detect virtual concept drift using the components of the decomposed uncertainty.

Additionally, we systematically benchmark several BDL methods, analyze the quality and consistency of aleatoric and epistemic uncertainty representations, and provide a visual example of handling high uncertainty predictions. We accomplished this by training a deterministic model and five different types of Bayesian models and measuring their accuracy on two temporally distinct case study datasets and one case study dataset with a distinctly different distribution. If shown as robust, this Bayesian approach to error decomposition will provide additional, much needed, information to allow for easier implementation of ML-models into satellite-derived multi-platform retrievals of atmospheric, oceanic, or terrestrial properties.

II. Methodology

A. Bayesian Deep Learning

Many recent machine learning advances can be attributed to deep learning, using artificial neural networks with multiple hidden layers. However, these models are deterministic and do not provide information about the uncertainty of their outputs. By incorporating a Bayesian approach, it is possible to create models that provide information about uncertainty in prediction. This is achieved by replacing the weights of a neural network, θ, with a distribution that is updated as the model is developed on training data, D. Mathematically, the model weights are treated as a prior distribution, $p(\theta)$, and conditioned on the evidence, the distribution of the training data, $p(D)$. When Bayes' Theorem is applied, the posterior distribution is:

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)} = \frac{p(D|\theta) \cdot p(\theta)}{\int p(D|\theta) \cdot p(\theta)\ d\theta} \tag{1}$$

One of the main difficulties with applying a Bayesian approach is that the denominator in Eq. 1 often has no closed form solution and is computationally intractable [9]. As a result, an approximation of the $p(\theta|D)$ is computed instead.

Variational inference is one method to approximate this posterior. The goal of variational inference is to create an optimization problem that identifies the distribution in a family of distributions, $q^*(\theta) \in Q$, that is least distant from the target distribution, $p(\theta|D)$. The measure of distance used is the Kullback-Leibler divergence (KL). The optimization problem is characterized by these two equations [17]:

$$q^*(\theta) = \arg\min_{q \in Q}\ KL(q(\theta)\ ||\ p(\theta|D)) \tag{2}$$

$$KL(q(\theta)\ ||\ p(\theta|D)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta|D)} d\theta \tag{3}$$

However, Eq. 3 still contains $p(\theta|D)$, which is intractable. To solve the optimization problem without explicitly calculating $p(\theta|D)$, Eq. 3 can be re-written as [17]:

$$KL(q(\theta)\ ||\ p(\theta|D)) = \log\ p(D) - \underbrace{\int q(\theta) \log \frac{p(\theta)p(D|\theta)}{q(\theta)} d\theta}_{\text{Evidence Lower Bound (ELBO)}} \tag{4}$$

Since the first term of Eq. 4 does not depend on q, it can be ignored to solve the minimization problem. Instead, the minimization problem is solved by maximizing the second term in Eq. 4, the evidence lower bound (ELBO). The optimization problem then becomes [17]:

$$q^*(\theta) = \arg\max_{q \in Q}\ ELBO(q(\theta)) \tag{5}$$

$$ELBO(q(\theta)) = \int q(\theta) \log \frac{p(\theta)p(D|\theta)}{q(\theta)} d\theta \tag{6}$$

In this study, this problem is further simplified by restricting Q to the fully factorized Gaussian distributions as described in [18]. This simplification allows for the application of the Flipout and Reparameterization methods of variational inference over the weights. Flipout prioritizes a more exact gradient computation over computational efficiency in comparison to other variational inference implementations [19], while Reparameterization prioritizes ease of computation when computing gradients [20]. In this article, these methods are compared to Monte Carlo (MC) dropout and a deterministic model implementation. MC dropout is equivalent to defining Q as Bernoulli distributions but without explicit KL divergence calculation [21].

During training, it is possible for the KL divergence to grow rapidly and prevent model convergence. Since the KL divergence is a penalty on the expected log-likelihood (see Eq 6). One way to address this problem is to decrease the penalty by placing a weight less than one on this term. In [22], this term is given a decreasing weight over the course of M mini-

batches on a schedule of $\pi_i = \frac{2^{M-i}}{2^M-1}$, where $\sum_{i=1}^{M} \pi_i = 1$, and

$$ELBO(q(\theta)) = \sum_{i=1}^{M}(E[\log p(D|\theta)] - \pi_i KL(q(\theta)||p(\theta))) \tag{7}$$

This KL reweighting scheme allows the prior to have a greater effect at the beginning of each epoch and the data to have a greater effect at the end of each epoch [22].

The goal of inference with BDL is to make a prediction, y, from new data, x. For a classification problem with c classes, the model provides the probability that y is a given class, $p(y = c|x, \theta)$. Since the weights of the models are distributions, the average probability is calculated by using Monte Carlo integration with N samples [12], [14]. Using our Bayesian models, we made 25 predictions (samples) for each input. The average probability per class ($\overline{p}_c$) is calculated as:

$$\overline{p}_c = \frac{1}{N}\sum_{n=1}^{N} p(y = c|x, \theta) \tag{8}$$

The class that yields the highest $\overline{p}_c$ is chosen as the predicted class label. The same N predictions are also used to calculate the variance of $\overline{p}_c$, providing a measure of uncertainty.

While having the total uncertainty is useful, knowing the source of the uncertainty is even more helpful. The total uncertainty can be expressed as the sum of the aleatoric uncertainty and the epistemic uncertainty [10], [15], [16].

- **Aleatoric uncertainty** is inherent in the data and cannot be reduced by providing the model more training data.
- **Epistemic uncertainty** is attributed to the uncertainty in the model and can be reduced by increasing the amount of training data available in regions of greater epistemic uncertainty.

Both [23] and [16] propose methods for estimating these individual uncertainties. However, the method described in [23] requires the use of extra variables to explicitly model the mean and the variance on the architecture output, which we call architectural decomposition, to calculate epistemic and aleatoric components. In contrast, [16] proposes a method to calculate epistemic and aleatoric components of the uncertainty without explicit architectural changes. Kwon et al. [16] compared both approaches for uncertainty decomposition on the task of ischemic stroke lesion segmentation. In their analysis of variance decomposition using the method in [16], when the prediction disagreed with the truth, high per-pixel uncertainty correctly identified regions that were misclassified for both false negatives and false positives. On the other hand, their analysis using architectural decomposition [23] did not yield useful information for the same task. In this work, we adopt the uncertainty decomposition approach of [16] and using the following formulation of aleatoric and epistemic components:

$$Var(\hat{p}) = \underbrace{\frac{1}{N}\sum_{n=1}^{N} diag(\hat{p}_n) - \hat{p}_n^{\otimes 2}}_{aleatoric} + \underbrace{\frac{1}{N}\sum_{n=1}^{N}(\hat{p}_n - \overline{p})^{\otimes 2}}_{epistemic} \tag{9}$$

where $\hat{p} = p(y = c|x, \theta)$, $diag(\hat{p}_n)$ is a diagonal matrix, $\hat{p}_n^{\otimes 2} = \hat{p}_n\hat{p}_n^T$, and $(\hat{p}_n - \overline{p})^{\otimes 2} = (\hat{p}_n - \overline{p})(\hat{p}_n - \overline{p})^T$.

B. Dataset Description

This study uses the well established 12-month dataset collected over the oceans in 2017 and approach as in [7]. The standard GMI output provides brightness temperatures observations at 13 different channels, including both vertical (v) and horizontal (h) polarization, with varying FOV size. The available GMI frequencies and corresponding field of views appear in Table I. The GMI product was chosen to define the observation vector over a 125 km $\times$ 125 km area centered on the observing Field of View (FOV), corresponding to a patch of 25$\times$9 individual GMI pixels. Brightness temperatures were collected at these pixels at all of the 13 GMI channels and stored into 9$\times$25$\times$13 arrays. The arrays were then normalized using z-score scaling [24].

Frequency [GHz]	Field of View
10.65v/h	19km x 32km
18.7v/h	10km x 18km
23.8v	10km x 16km
36.6v/h	9km x 16km
89v/h	4km x 7km
166v/h	4km x 6km
183+3v/7v	4km x 6km

TABLE I
GMI FREQUENCIES AND FIELDS OF VIEW

To ensure accurate matching between DPR- and GMI-viewing geometries, each individual GMI pixel was labeled (convective or stratiform) by applying Gaussian weighting to DPR-observed precipitation rates [5] and calculating a convective fraction of precipitation volume within the GMI FOV. Pixels with a fraction of 50% or more were assigned a convective flag; the remaining pixels were labeled as stratiform. Noise in the dataset was minimized by removing observations containing any missing or non-classified data, comprising less than 5% of total data. The remaining $\sim$14 million samples were further split into training/validation/test data with an 80/10/10 ratio respectively, preserving roughly equal representation of both classes (i.e., forming balanced data subsets) [7].

Due to the balanced composition of these data subsets, we only present the accuracy in remainder of this article as a metric of model performance. Accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{10}$$

We treated the DPR-derived classification as the true label for determining whether or not a prediction was correct.

Separately from the traditional training/validation/test datasets, with a goal to demonstrate trained models ability to generalize on unseen data, experiments were also conducted using two temporally independent, single swath case study datasets (Case 1 and Case 2), and one year of data collected over land (Case 3). Case 1 and Case 2 represent instantaneous

observations of two separate precipitation events over ocean, captured by DPR and GMI sensors. Case 1 is a subtropical marine mesoscale convective system (MCS) located near shallow convection on 11 August 2018 over the North Atlantic. Case 2 is a section of Hurricane Lane observed southeast of Hawaii by the GMI on 19 August 2018. The scenes observed by GMI in the 18.7 GHz horizontally polarized band are depicted in Figure 1 with the smaller DPR swaths enclosed by the black and white lines. To test the models on input features with a different distribution, one year of global observations collected over land was used to form the Case 3 dataset. All three case study datasets were collected in 2018 and are temporally independent from the training/validation/test datasets.

C. Model Architecture and Training

Based on the results in [7], a residual network (ResNet) V2 [25] with 38 layers was chosen as a representative deterministic architecture to classify precipitation type as either convective or stratiform. Bayesian ResNet architectures were adopted in identical configuration as the deterministic architecture by following the approach in [26]. Bayesian ResNet model architectures with Flipout layers [19] and Reparameterization layers [20] were implemented utilizing the Tensorflow Probability library [27].

Model weights were initialized for training following He et al. [25]. The Adam optimizer was used with a starting learning rate of 0.001. The validation loss was monitored in order to conduct learning rate annealing [28]. The learning rate was reduced by a factor of 10 if there was no reduction in validation loss after 10 consecutive epochs. To regularize for overfitting, an early stopping strategy was employed [24]. If early stopping did not occur, training was terminated at 600 epochs. Our Bayesian models using a batch size of 128 required approximately 3 weeks to train on a single NVIDIA RTX 8000 48GB GPU. Both deterministic and Bayesian models were trained with the same strategy for the fairness of benchmarking.

III. Results and Discussion

A. Effects of KL Reweighting on Optimization

The results of early experiments indicated that the weight of the KL divergence term (see Eq. 6) for the Flipout and Reparameterization models needed to be reduced. Similarly to [22] and [29], we observed that the KL divergence term rapidly increased during the early stages of the training for the Flipout and Reparameterization models, preventing these models from converging. The results in this section show the accuracy of these models when the KL term is set to zero and when the KL term weight is reweighted according to Eq. 7. The model accuracy achieved when using the KL reweighting scheme in Eq. 7 is comparable to the model accuracy of MC Dropout, a Bayesian model where the KL divergence term is not calculated.

The precipitation classification type derived from GPROF, the NASA operational passive microwave precipitation retrieval for the GPM mission, serves as the benchmark for experimentation. Table II lists the accuracy of all deep learning models, which achieve higher classification accuracy than the GPROF benchmark. For Bayesian models, the predicted class was determined using the mean probability of 25 predictions ($N = 25$ in Eq. 8). Table II lists the model type with KL weighting scheme followed by the classification accuracy of each model on the test set, a swath of the North Atlantic Ocean (Case 1), and a swath of the Pacific Ocean southeast of Hawaii (Case 2).

On the test set, all of the Bayesian deep learning models performed comparably to or better than the deterministic ResNet38 V2 model (0.868 accuracy). For this dataset, setting the KL term to zero produced higher accuracy for the Flipout and Reparameterization models (0.927 and 0.920) than reweighting the KL term (0.866 and 0.864). Since the test and training dataset are subsets of the same dataset, these results indicate that setting the KL term to zero improved classification accuracy when the test data distributions are similar to the distributions within training data. However, while it is possible to control data distributions and splits during model development, such control is not feasible during live inference; there is no way to control the observed data distributions of a live sensor.

All models were also applied to two case study regions of interest (accuracy listed in Table II) where we emulated live inference setting by choosing data significantly temporally separated from the test dataset split. On these two case studies, KL reweighting produced higher accuracy for the Flipout and Reparameterization models (0.794 and 0.834; 0.802 and 0.832) compared to a KL term of zero (0.756 and 0.814; 0.778 and 0.817). Since the case studies were temporally separate from the training data, this higher accuracy indicates that the reweighted KL term helps produce models that generalize better than when the KL term is zero. Furthermore, for the case study datasets, the models with KL reweighting performed comparably to MC Dropout (0.784 and 0.824), a Bayesian model where the KL divergence term is not calculated.

TABLE II
Model Accuracy By dataset for GPROF (benchmark), ResNet38 V2 (deterministic), and remaining Bayesian models. The KL term is either set to zero or reweighted (RW) using Eq. 7 for both Flipout and Reparameterization model.

Model	Test Set	Case 1	Case 2
GPROF	0.743	0.558	0.695
ResNet38 V2	0.868	0.808	0.832
Flipout, KL = 0	0.927	0.756	0.814
Reparam., KL = 0	0.920	0.778	0.817
Flipout, KL RW	0.866	0.794	0.834
Reparam., KL RW	0.864	0.802	0.832
MC Dropout	0.860	0.784	0.824

B. Well-Calibrated Uncertainties

One of the primary reasons to use Bayesian models is to make use of the uncertainty measures that accompany a prediction. According to [12], a model has well-calibrated uncertainty if its performance improves as more high-uncertainty

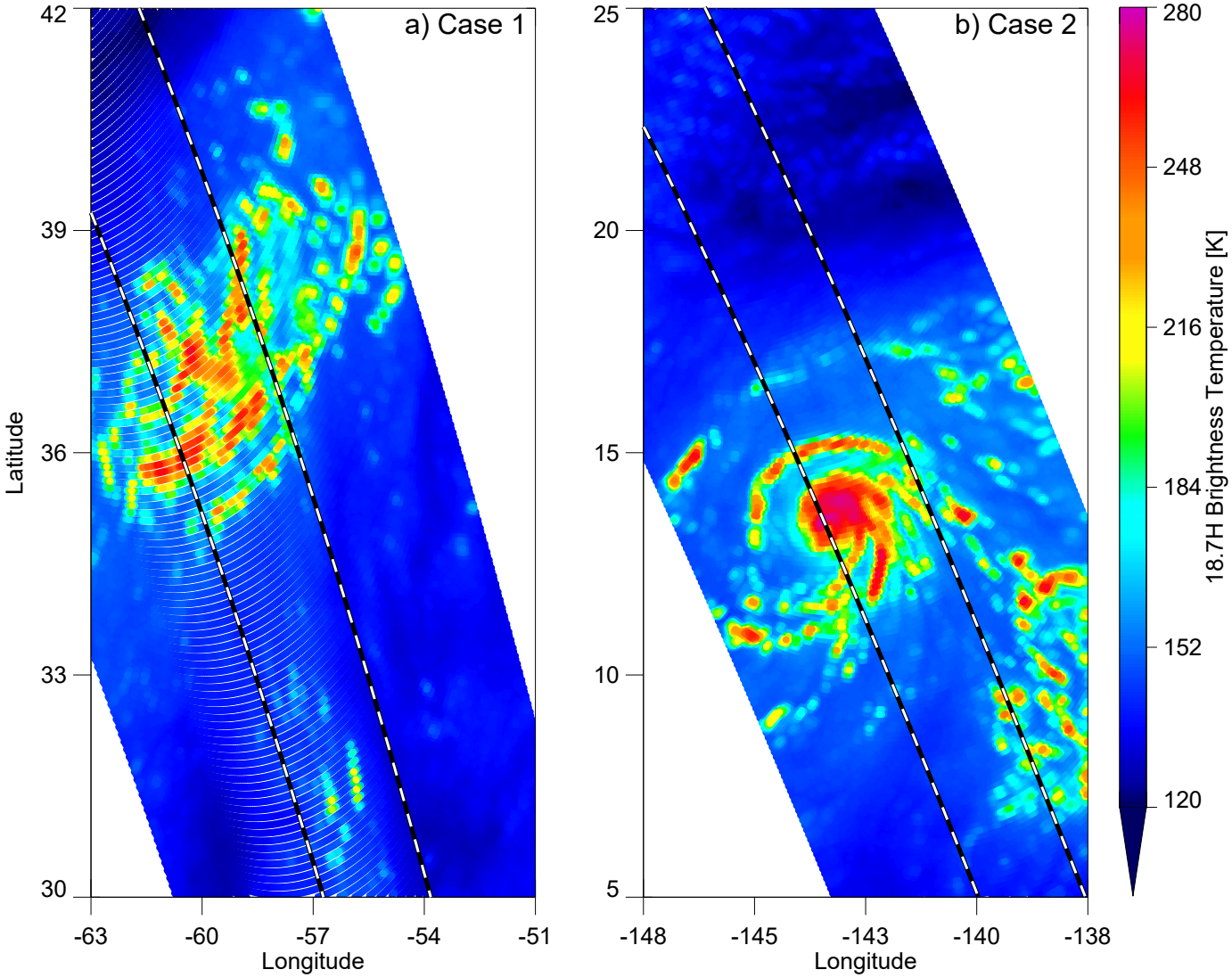

Fig. 1. 18.7 GHz (horizontal polarization) brightness temperatures from GMI depicting the scenes used for analysis in a) Case 1 and b) Case 2. The area enclosed in the black and white lines denotes the smaller DPR swaths where the analysis was conducted.

predictions are discarded. As suggested by the results in Table II, reweighting the KL term plays an important role in producing models with well-calibrated uncertainties that generalize to unseen data because the accuracy on the Case 1 and Case 2 datasets is higher than when the KL term is set to zero. Well-calibrated uncertainties are useful for making decisions about predictions.

The test set was served to each of the models for prediction. Next, the aleatoric and epistemic uncertainties of each test set prediction were calculated using Eq. 9. Table III contains the model type followed by the test set prediction uncertainty value for the 80th percentile of each type of uncertainty. There is relatively little change in the threshold value for the aleatoric uncertainty across all models (approximately 0.4), but the value for epistemic uncertainty thresholds is three orders of magnitude smaller for models with a KL term of zero (10^{-6} and 10^{-5}), meaning these models do not capture as much of the epistemic uncertainty as when the KL term is reweighted (10^{-3} and 10^{-2}). The epistemic uncertainty threshold values for the reweighted KL models (1.544e-03 and 1.033e-02) are comparable to the epistemic uncertainty threshold for MC Dropout (1.438e-02), a Bayesian model where the KL divergence term is not calculated.

TABLE III
THRESHOLD UNCERTAINTY VALUE (80TH PERCENTILE) FOR EACH BAYESIAN MODEL AND EACH UNCERTAINTY TYPE.

Model	Epistemic	Aleatoric
Flipout, KL = 0	4.210e-06	3.831e-01
Reparam., KL = 0	1.328e-05	4.059e-01
Flipout, KL RW	1.544e-03	4.718e-01
Reparam., KL RW	1.033e-02	4.638e-01
MC Dropout	1.438e-02	4.620e-01

The values in Table III were used as thresholds to discard predictions made on the test set and the two case studies. The accuracy values reported in Tables IV-VI were calculated using only the predictions that had uncertainty values less than

or equal to the values in Table III. For example, in Table IV, the Flipout model with a KL term of zero had an accuracy of 0.927 that was calculated using all test set predictions; an accuracy of 0.972 that was calculated using only predictions with epistemic uncertainty less than or equal to 4.210e-06; and an accuracy of 0.975 that was calculated using only predictions with aleatoric uncertainty less than or equal to 3.831e-01.

After removing predictions with uncertainty values above the threshold values, the same accuracy trend appears between the test set and the two case studies that appeared when the accuracy was calculated using all predictions. A KL term of zero led to higher accuracy on the test set. However, this accuracy did not generalize as well to the case studies. A reweighted KL term produced higher accuracy on the case studies. The reappearance of this trend reinforces the conclusion that the reweighted KL term helps produce models that generalize better that when the KL term is zero.

With the exception of the epistemic uncertainty associated with Case 1 predictions made by the Reparameterization model with a KL term of zero, all other uncertainties are well calibrated since model accuracy improved after removing high uncertainty predictions. This demonstrates the additional utility that Bayesian models provide over deterministic models. A deterministic ResNet cannot provide a measure of uncertainty about its predictions, let alone provide a measure of whether or not the uncertainty is well-calibrated. Not only can Bayesian models provide a prediction with uncertainty metrics, it is possible to determine if these uncertainties are well calibrated. Model predictions accompanied by well calibrated uncertainties allow for a decision to be made about whether to keep the prediction, to discard the prediction, or to pass the prediction to another system or a human being, whichever is more useful for the task at hand.

TABLE IV
ACCURACY ON TEST SET USING ALL PREDICTIONS, PREDICTIONS WITH EPISTEMIC UNCERTAINTY LESS THAN OR EQUAL TO THE THRESHOLDS IN TABLE III, AND PREDICTIONS WITH ALEATORIC UNCERTAINTY LESS THAN OR EQUAL TO THE THRESHOLDS IN TABLE III.

Model	All	Epistemic	Aleatoric
Flipout, KL = 0	0.927	0.972	0.975
Reparam., KL = 0	0.920	0.967	0.972
Flipout, KL RW	0.866	0.914	0.930
Reparam., KL RW	0.864	0.913	0.927
MC Dropout	0.860	0.909	0.923

TABLE V
AS IN TABLE IV BUT FOR CASE 1.

Model	All	Epistemic	Aleatoric
Flipout, KL = 0	0.756	0.820	0.817
Reparam., KL = 0	0.778	0.778	0.799
Flipout, KL RW	0.794	0.862	0.871
Reparam., KL RW	0.802	0.849	0.879
MC Dropout	0.784	0.849	0.847

TABLE VI
AS IN TABLE IV BUT FOR CASE 2.

Model	All	Epistemic	Aleatoric
Flipout, KL = 0	0.814	0.859	0.847
Reparam., KL = 0	0.817	0.817	0.842
Flipout, KL RW	0.834	0.904	0.881
Reparam., KL RW	0.832	0.885	0.876
MC Dropout	0.824	0.899	0.876

C. Predictive Implications of Well-Calibrated Uncertainties

Having established that the uncertainties of the models are well-calibrated, these uncertainties can now be used for the task at hand, precipitation type classification. Figure 2a depicts the variance (sum of aleatoric and epistemic uncertainty) associated with each prediction for Case 2. Figures 2b and 2c show the separated uncertainty types as defined by Eq. 9. Brighter colors represent higher levels of variance and uncertainty. Figure 2d and e display the predictions from the GPROF algorithm and the DPR-derived labels (used as true label for training). Figure 2f details the predictions for the Flipout KL Reweighting model on Case 2; lighter shades of blue and red represent classifications with epistemic uncertainty above the threshold of 1.544e-03 from Table III.

When compared side-by side, the aleatoric and epistemic uncertainty maps show exactly how much of the predictive uncertainty is caused by the dataset (aleatoric) and how much is caused by the model (epistemic). The scales of the uncertainty maps show that the majority of the uncertainty is a result of the aleatoric component; even the brightest portions of Fig. 2c are an order of magnitude smaller than the darker portions of Fig. 2b.

When viewed in conjunction with the prediction map (Fig. 2f), the aleatoric and epistemic uncertainty maps in Fig. 2b and c provide information beyond what is available when viewing the prediction map in isolation. In a similar fashion to [16], when the spatial predictions (Fig. 2f) disagree with the DPR-derived labels (Fig. 2e), the epistemic uncertainty (Fig. 2c) identifies the incorrect classifications because of the larger magnitude of these uncertainties compared to the rest of the epistemic uncertainty map. From 12° to 14° latitude and from -143° to -142.5° longitude, the Bayesian model over-predicts convective precipitation, but the epistemic uncertainty map identifies many of these predictions as high uncertainty (depicted in pink on the prediction map). This same type of prediction error occurs along 15° latitude, where the Bayesian model over-predicts convective precipitation, and again, the epistemic uncertainty map identifies these predictions as high uncertainty (depicted in pink in Fig. 2c). If a downstream application requires high accuracy predictions, a decision could be made to discard these types of predictions since they have both high epistemic uncertainty and high aleatoric uncertainty. However, in some instances, it may be beneficial to keep predictions with low epistemic uncertainty, but higher aleatoric uncertainty (see Fig. 2b and c at 13° latitude, -141.5° longitude). For these predictions, the model has low

uncertainty (epistemic) about its prediction despite noise that is inherent in the data (high aleatoric uncertainty). When making these types of decisions, the uncertainty-source component is of particular interest. Compared to deterministic models, this new information about model predictions provides the ability to make informed decisions about how to handle high uncertainty predictions based on the level and the source of the uncertainty. Furthermore, such a decision cannot be made when the variance alone is considered.

D. Utility of Uncertainty Decomposition Beyond Prediction

The results of these experiments offer more than establishing whether or not model uncertainties are well calibrated and making decisions about predictions using well-calibrated uncertainties. Figure 3 shows the mean aleatoric uncertainty (Fig. 3a) and epistemic uncertainty (Fig. 3b) for the predictions made by each model on each dataset. All models have similar levels of aleatoric uncertainty across each dataset ($\sim$0.19 for the test set, $\sim$0.22 for Case 1, $\sim$0.13 for Case 2). However, the models differ with respect to amount of epistemic uncertainty for each dataset ($\sim$0.0005–0.005 for the test set, $\sim$0.001–0.0085 for Case 1, $\sim$0.00075–0.0055 for Case 2). Across datasets, there is a general trend in epistemic uncertainty. MC Dropout has the highest epistemic uncertainty; Reparameterization with KL reweighting has the second highest; and Flipout with KL reweighting has the lowest. By decomposing the variance into aleatoric and epistemic uncertainties, these measures enable informed decision-making about model selection, data collection/processing, and targeted data analysis.

Visualizing the uncertainty decomposition in this way can be useful when making decisions about model selection and data collection. Since all models in Fig. 3 equally represent the aleatoric uncertainty and have comparable accuracy, it is prudent to select the Flipout KL Reweighting model for deployment because it has lower epistemic uncertainty values and a smaller range of epistemic uncertainty ($\sim$ 0.0005-0.001) across the datasets compared to MC Dropout ($\sim$ 0.005-0.0085) and Reparameterization with KL reweighting ($\sim$ 0.003-0.006). From this visualization, it is also possible to gain insight about what type of data to collect. If it is possible to collect more data, the epistemic (model) uncertainty values in Figure 3 indicate that data similar to Case 1 (cyan) would be more beneficial than data similar to Case 2 (green) because the epistemic uncertainty values for Case 1 are higher and epistemic uncertainty can be reduced with more training data. This observation is strengthened when taking the dataset sizes into account. Case 2 epistemic values are close to the Test Set values, but the Test Set has close to 1000 times more samples. Case 1 is only about 200 samples smaller than Case 2. New data similar to Case 1 or augmentation of existing Case 1 data will lower the epistemic uncertainty for each model, which could lead to models that generalize better in live inference than the current models. These types of conclusions cannot be drawn when using a deterministic model since it provides no measure of uncertainty about its predictions. Additionally, these conclusions cannot be made without decomposing the variance into aleatoric and epistemic uncertainty, particularly when the epistemic uncertainty values are much smaller than the aleatoric values (see Table III and the scales of Fig. 3a and b). This is also true when analyzing Bayesian model performance with predictive entropy, a bulk uncertainty metric, such as in Orescanin et al. [7].

This same type of analysis can also be useful to identify when targeted data analysis may be useful. All considered models were less accurate when run on the case study datasets than when run on the much larger training set (see Table II). For Case 1, this difference in accuracy can be explained by the models not seeing enough data during training that is similar in underlying distribution to the data in Case 1 since the mean epistemic uncertainty values in Fig. 3 are higher for Case 1 than for the Test Set. The same holds true when running the MC Dropout model and the Flipout KL Reweighting model on the Case 2 dataset for which the mean epistemic uncertainty values in Fig. 3 are also higher for Case 2 than for the Test Set. However, the Reparameterization KL Reweighting model has lower epistemic uncertainty for Case 2 than for the Test Set. This means the model is less uncertain about its predictions when run on the Case 2 than when run on the Test Set, but it is getting the prediction wrong more often for Case 2 (accuracy 0.832) than for the Test Set (accuracy 0.864). Seeing this curious trend, the false positives and the false negatives for the Reparameterization KL Reweighting model can be aggregated and further analyzed. In this case, it is possible that the false positives and false negatives arise because the model is applied to the inner core of a tropical cyclone, where the precipitation is dynamically neither completely convective nor stratiform. Regardless of the reason, this is yet again another type of observation that cannot be made when using deterministic models or without decomposing the uncertainty.

By exploring the relationship between aleatoric and epistemic uncertainty in another way, it is possible to have still more insight into model selection. In Fig. 4, each line indicates the expected epistemic uncertainty for a prediction given that the aleatoric uncertainty for the prediction is less that or equal to the value on the abscissa. This demonstrates that as aleatoric uncertainty (inherent in the data) increases epistemic (model) uncertainty increases. However, the magenta (Reparameterization KL = 0) and cyan (Flipout KL = 0, hidden by magenta) lines indicate that no change in epistemic uncertainty occurs as aleatoric uncertainty increases when the KL term equals zero. In other words, when the KL term is zero, the model fails to represent the epistemic uncertainty in a meaningful way. The lines for the other models have greater slope that corresponds with an increasing effect on model uncertainty by noise inherent in the data. This trend reinforces the choice of the Flipout KL Reweighting model (red lines) for deployment since it has low epistemic uncertainty even as aleatoric uncertainty increases and has accuracy that generalizes to Cases 1 and 2. The models with KL equal to zero, on the other hand, have low epistemic uncertainty, but their accuracy does not generalize to Cases 1 and 2 (see Table II).

Decomposing the variance into uncertainty types provides the opportunity for yet one more insight, identifying challenges due to data collection and processing. When the observed trend in Fig. 4 is combined with two orders of magnitude

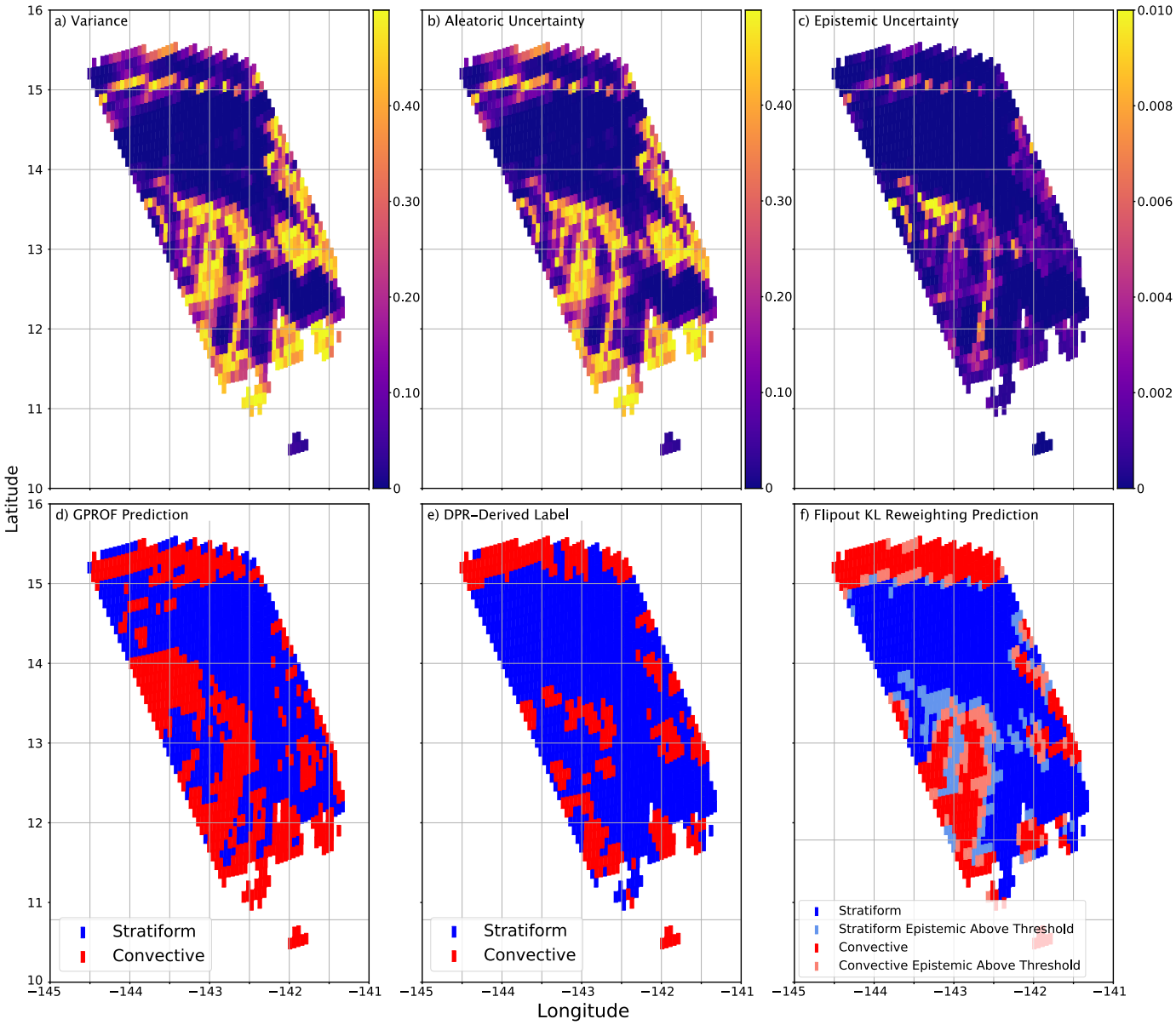

Fig. 2. Plots of the northeast section of Hurricane Lane observed by the GMI instrument at 1645 UTC 19 Aug 2018 (Case 2). The top row contains spatial plots of uncertainty metrics; specifically, a) variance, b) aleatoric uncertainty, and c) epistemic uncertainty. The bottom row contains d) GPROF predictions of precipitation type, e) DPR-derived precipitation type, and f) Flipout KL Reweighting predictions with uncertainty above a threshold of 1.544e-03 (from Table III) shown in lighter shades. Note that the range of magnitudes for variance and aleatoric uncertainty are similar (0.0-0.5) and much larger than the values for epistemic uncertainty (0 to 0.01).

difference in uncertainty values seen in Fig. 3, it can reasonably be concluded in this example that reducing the aleatoric uncertainty would likely yield higher accuracy since the models that have higher aleatoric uncertainty (MC Dropout and Reparameterization with KL reweighting) also have accuracy comparable to the Flipout with KL reweighting model. Having made this conclusion, attempts could now be made to reduce the aleatoric uncertainty, such as by re-calibrating collection sensors, augmenting the existing data with new features, or preprocessing the data differently to increase the signal to noise ratio. Without knowing that the aleatoric uncertainty far outweighs the epistemic uncertainty, it would be impossible to understand that reducing the noise inherent in the data provides more opportunity to improve model accuracy than providing a model more training data.

E. Virtual Concept Drift Detection

Virtual concept drift occurs when the data distribution changes so much that model error is no longer acceptable [30]. Case 3 consists of one year of observations collected over land surface across the entire globe. We expect that the distribution of GMI brightness temperatures over land is completely different than that over ocean because the emissivity of land is significantly different from that of water. Table VII shows the accuracy of the models, which were trained on data collected over oceans, on this third case study

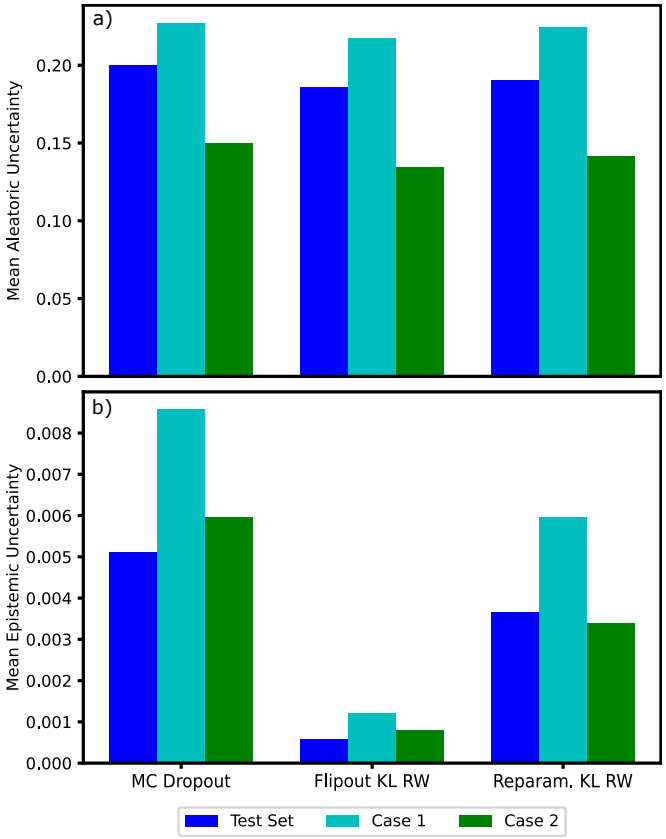

Fig. 3. a) Mean aleatoric uncertainty and b) epistemic uncertainty for each dataset by model (listed on abscissa) for the test dataset (dark blue), Case 1 (cyan), and Case 2 (green).

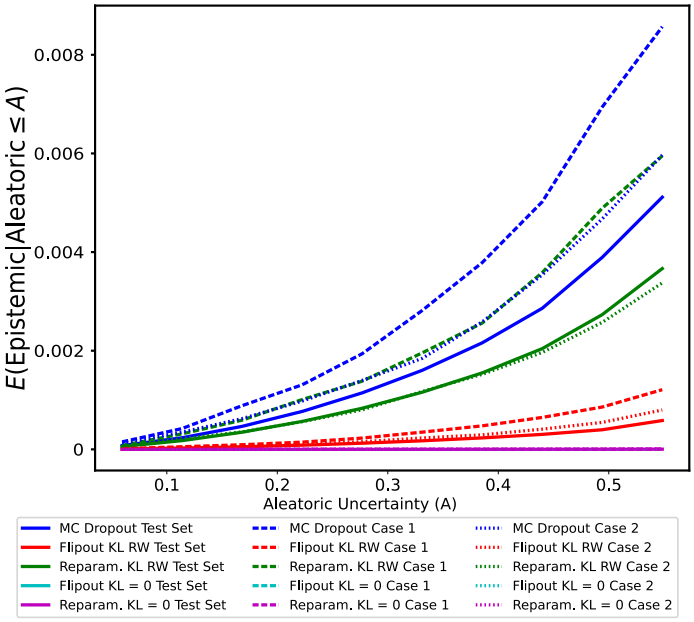

Fig. 4. Expected epistemic uncertainty (E) given that a prediction has aleatoric uncertainty (A) less than or equal to the values on the abscissa for the test dataset (solid lines), Case 1 (dashed lines), and Case 2 (dotted lines). Different models are denoted by different colors as follows: MC Dropout (dark blue), Flipout with KL Reweighting (red), Reparameterization with KL Reweighting (green), Flipout with KL = 0 (cyan), and Reparameterization with KL = 0 (magneta). The models with the KL term of zero (magenta and cyan) are both essentially 0 at all values of A.

data. Because this dataset was balanced, these accuracy results ($\sim 50\%$) are akin to guessing the correct class. Seeing this dramatic decrease in accuracy may be an indicator that virtual concept drift has occurred. The decomposed variance can help confirm this intuition.

TABLE VII
ACCURACY USING MODELS TRAINED ON DATA COLLECTED OVER THE OCEAN TO PREDICT ON DATA COLLECTED OVER LAND IS AKIN TO GUESSING THE CORRECT CLASS ($\sim 50\%$). THE MEAN EPISTEMIC AND ALEATORIC UNCERTAINTIES ARE CLOSER IN MAGNITUDE COMPARED TO THE VALUES IN TABLE III

Model	Accuracy	Epistemic	Aleatoric
ResNet38 V2	0.526	N/A	N/A
Flipout, KL = 0	0.500	5.565e-02	4.557e-02
Reparam., KL = 0	0.500	1.024e-01	7.083e-02
Flipout, KL RW	0.500	3.308e-02	1.426e-01
Reparam., KL RW	0.501	5.828e-02	2.760e-01
MC Dropout	0.501	6.856e-02	2.030e-01

Unlike the previous case studies and the test set where the epistemic uncertainty was much smaller than the aleatoric uncertainty ($\sim$ 0.4 difference), the aleatoric and epistemic uncertainties for Case 3 are much closer in magnitude ($<$ 0.23 difference); the models with the KL term set to zero even have epistemic uncertainty that exceeds the aleatoric. These higher values of epistemic uncertainty suggest that the accuracy is suffering because the model did not see enough similar data in training. In Fig. 5a, the aleatoric values of Case 3 (yellow) are similar to the test set and the other case studies. This makes it unlikely that the decrease in accuracy is due to sensor degradation or some other source of noise inherent in the data. Furthermore, in Fig. 5b, the epistemic values are much higher for this dataset. These high values confirm that the Case 3 data is indeed not part of the same distribution as the model development datasets. This is to be expected given the contrast in brightness temperature (i.e., input features) distributions originating over radiometricaly cold ocean- and warm land-surfaces; however, during live inference, this difference would not be known ahead of time. Separating the uncertainties mathematically confirms that this is indeed virtual concept drift. This confirmation allows model developers to focus decisions on how to handle this new distribution, instead of blindly adding more data to development data splits or conducting time consuming hyperparameter tuning (three weeks for these models).

IV. SUMMARY AND CONCLUSION

Machine learning techniques can efficiently extract discrete and continuous properties of physical systems. However, existing techniques have limited ability to provide insights to errors and the physical relationship between the observed and retrieved properties, a major downside of their application. In the present study, we use a problem of detecting precipitation type from satellite observations to introduce a novel tool for

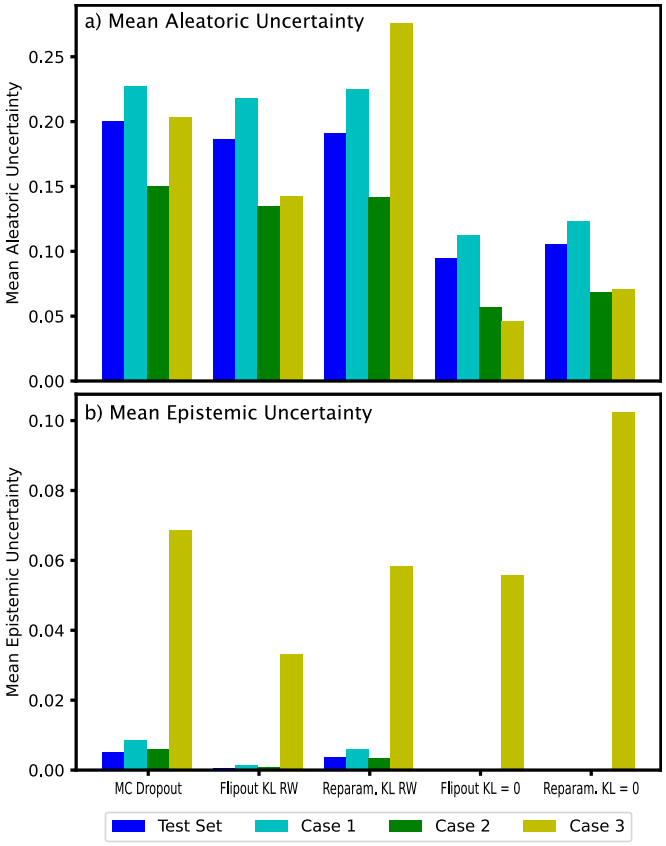

Fig. 5. As in Fig. 3 but for all models and including the mean a) aleatoric and b) epistemic uncertainties for land data in Case 3 (yellow).

decomposing errors of satellite-retrieved products and allow for better understanding of the links between observed and retrieved features. Typically, precipitation type is reported without quantitative uncertainty attached to an estimate. We use Bayesian models to classify precipitation type by mapping Global Precipitation Measurement mission Microwave Imager observations to Dual-frequency Precipitation Radar-derived precipitation type. These Bayesian models perform comparably to deterministic models, but with the added benefit of well calibrated uncertainties. Well calibrated uncertainties are useful for making decisions concerning high uncertainty predictions, model selection, targeted data analysis, and data collection and processing. Additionally, our Bayesian models enable mathematical detection of virtual concept drift, which occurs when the data distribution changes so much that model error is no longer acceptable [30].

From a pool of ~ 14 million samples collected in 2017, we created a development dataset that we used to create traditional training/validation/test datasets with an equal representation of each type of precipitation. To simulate live-inference, we used two temporally independent, single-overpass case study datasets (Case 1 and Case 2) from 2018. Case 1 is a subtropical marine mesoscale convective system (MCS) located near shallow convection. Case 2 is a section of Hurricane Lane observed southeast of Hawaii. To observe model behavior on data with a different distribution, we used a third case study dataset (Case 3) comprised of one year of global observations collected over land in 2018.

In our experiments, we developed Bayesian models using the evidence lower bound (ELBO in Eq. 6) as our loss function, which is dependent on the KL divergence term between the prior and posterior distributions. The KL divergence can grow rapidly and can prevent a model from converging during training. We adopted a KL reweighting scheme (Eq. 7) to control the KL divergence term during optimization. Our results (Table II) indicate that a reweighted KL term helps models achieve accuracy that generalizes better to live inference than setting the KL term to zero. The models with a reweighted KL term had comparable accuracy to the MC Dropout model, a Bayesian model where the KL term is not explicitly computed. All our Bayesian models also had well calibrated uncertainties that proved useful for making decisions about high uncertainty predictions. This information can be used to decide to keep a prediction, to discard a prediction, or to pass a prediction to another system or a human being, whichever is more useful for the task at hand. By decomposing the uncertainty into aleatoric and epistemic components, decisions can be made about how to handle high uncertainty predictions. Figure 2 provides a visualization of different uncertainty metrics and high uncertainty predictions. By separating out the epistemic uncertainty, we identified the Flipout with KL reweighting model as the model most ready for deployment because it had comparable accuracy to all other models but with a lower and smaller range of epistemic uncertainty across the test set and the two case studies (Case 1 and Case 2). Using this same analysis of epistemic uncertainty, we also concluded that false positives and false negatives from the Reparameterization with KL reweighting model were candidates for targeted analysis. Furthermore, we were able to conclude that samples similar to Case 1 would be more beneficial for future training. Analysing the aleatoric uncertainty of predictions made using data collected over the ocean showed that the aleatoric uncertainty far outweighed the epistemic uncertainty. Knowing this fact, attempts can be made to reduce the aleatoric uncertainty, such as by re-calibrating collection sensors, augmenting the existing data with new features, or preprocessing the data differently to increase the signal to noise ratio. The aleatoric uncertainty of predictions made using data collected over land surface was similar to the other datasets collected over the ocean. This similarity coupled with much higher epistemic uncertainty that the other datasets provided a mathematical means to verify that concept drift (a change in distribution) caused the accuracy to plummet.

For precipitation type classification, Bayesian deep learning models perform comparably to their deterministic counterparts. Decomposing the uncertainty available from these Bayesian deep learning models allows users to make informed decisions concerning high uncertainty predictions, model selection, targeted data analysis, data collection/processing, and virtual concept drift. The ramifications of these capabilities for just atmospheric science applications are potentially wide-ranging. For example, if Bayesian neural networks are applied to regression tasks (i.e., predicting microwave brightness temperature using infrared radiances), the uncertainty included

may inform proper weighting of insufficiently certain predictions of synthetic values of commonly assimilated fields into global numerical models of the atmosphere. Features associated with large epistemic uncertainties highlight areas for which additional observations could be beneficial. For example, in this article, the model could improve by training on additional observations of deep convection in tropical cyclones. If predictions using the same model are applied to the same instrument over time (i.e., several years), increasing aleatoric uncertainty could be an early indicator that various issues (e.g., sensor malfunctions, orbital drift) are causing degradation to predictions. None of these are possible using traditional deterministic models.

ACKNOWLEDGMENT

This work has been supported by: ONR grant N0001421WX00575 and NOAA grant NA19NES4320002 (Cooperative Institute for Satellite Earth System Studies - CISESS) at the University of Maryland.

REFERENCES

[1] G. Skofronick-Jackson, W. A. Petersen, W. Berg, C. Kidd, E. F. Stocker, D. B. Kirschbaum, R. Kakar, S. A. Braun, G. J. Huffman, T. Iguchi, P. E. Kirstetter, C. Kummerow, R. Meneghini, R. Oki, W. S. Olson, Y. N. Takayabu, K. Furukawa, and T. Wilheit, "The Global Precipitation Measurement (GPM) Mission for Science and Society," *Bulletin of the American Meteorological Society*, vol. 98, no. 8, pp. 1679–1695, Aug. 2017.

[2] D. W. Draper, D. A. Newell, F. J. Wentz, S. Krimchansky, and G. M. Skofronick-Jackson, "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*, vol. 8, no. 7, pp. 3452–3462, Jul. 2015.

[3] Iguchi, Toshio, S. Shinta, Meneghini, Robert, Yoshida, Naofumi, Awaka, Jun, Le, Minda, Chandrasekar, V, and Kubota, Takuji, "GPM/DPR Level-2 Algorithm Theoretical Basis Document," *NASA Goddard Space Flight Center*, 2010.

[4] C. D. Kummerow, D. L. Randel, M. Kulie, N.-Y. Wang, R. Ferraro, S. J. Munchak, and V. Petkovic, "The Evolution of the Goddard Profiling Algorithm to a Fully Parametric Scheme," *Journal of Atmospheric and Oceanic Technology*, vol. 32, no. 12, pp. 2265–2280, Dec. 2015.

[5] V. Petković, M. Orescanin, P. Kirstetter, C. Kummerow, and R. Ferraro, "Enhancing PMW Satellite Precipitation Estimation: Detecting Convective Class," *Journal of Atmospheric and Oceanic Technology*, vol. 36, no. 12, pp. 2349–2363, Dec. 2019.

[6] S. W. Powell, Robert A. Houze, and S. R. Brodzik, "Rainfall-Type Categorization of Radar Echoes Using Polar Coordinate Reflectivity Data," *Journal of Atmospheric and Oceanic Technology*, vol. 33, no. 3, pp. 523–538, Mar. 2016.

[7] M. Orescanin, V. Petković, S. W. Powell, B. R. Marsh, and S. C. Heslin, "Bayesian Deep Learning for Passive Microwave Precipitation Type Detection," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[8] C. Kidd, J. Tan, P.-E. Kirstetter, and W. A. Petersen, "Validation of the Version 05 Level 2 Precipitation Products from the GPM Core Observatory and Constellation Satellite Sensors," *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. S1, pp. 313–328, 2018.

[9] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, Apr. 2017.

[10] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian Active Learning with Image Data," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1183–1192.

[11] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active Learning With Convolutional Neural Networks for Hyperspectral Image Classification Using a New Bayesian Approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6440–6461, Nov. 2018.

[12] A. Filos, S. Farquhar, A. N. Gomez, T. G. J. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, and Y. Gal, "A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks," *arXiv:1912.10481 [cs, eess, stat]*, Dec. 2019.

[13] H. Ren, X. Yu, L. Bruzzone, Y. Zhang, L. Zou, and X. Wang, "A Bayesian Approach to Active Self-Paced Deep Learning for SAR Automatic Target Recognition," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[14] R. Feng, N. Balling, D. Grana, J. S. Dramsch, and T. M. Hansen, "Bayesian Convolutional Neural Networks for Seismic Facies Classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–8, 2021.

[15] A. D. Kiureghian and O. Ditlevsen, "Aleatory or Epistemic? Does it Matter?" *Structural Safety*, vol. 31, no. 2, pp. 105–112, Mar. 2009.

[16] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, "Uncertainty quantification using Bayesian neural networks in classification: Application to ischemic stroke lesion segmentation," in *1st Conference on Medical Imaging with Deep Learning*, Apr. 2018.

[17] O. Durr, B. Sick, and E. Murina, *Probabilistic Deep Learning: with Python, Keras, and TensorFlow Probability*. Shelter Island, NY: Manning, 2020.

[18] A. Graves, "Practical Variational Inference for Neural Networks," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[19] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse, "Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches," in *International Conference on Learning Representations*, Apr. 2018.

[20] D. P. Kingma, T. Salimans, and M. Welling, "Variational Dropout and the Local Reparameterization Trick," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[21] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *International Conference on Machine Learning*. PMLR, Jun. 2016, pp. 1050–1059.

[22] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight Uncertainty in Neural Networks," *arXiv:1505.05424 [cs, stat]*, May 2015.

[23] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" in *Neural Information Processing Systems*, Oct. 2017.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, ser. Adaptive computation and machine learning. Cambridge, Massachusetts London, England: The MIT Press, 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 630–645.

[26] D. Tran, M. Dusenberry, M. van der Wilk, and D. Hafner, "Bayesian Layers: A Module for Neural Network Uncertainty," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[27] J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. Hoffman, and R. A. Saurous, "TensorFlow Distributions," *arXiv:1711.10604 [cs, stat]*, Nov. 2017.

[28] Y. Li, C. Wei, and T. Ma, "Towards Explaining the Regularization Effect of Initial Large Learning Rate in Training Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[29] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating Sentences from a Continuous Space," *arXiv:1511.06349 [cs]*, May 2016.

[30] A. Tsymbal, "The Problem of Concept Drift: Definitions and Related Work," Trinity College Dublin, Tech. Rep., 2004.