UNCERTAINTY AWARE TROPICAL CYCLONE WIND SPEED ESTIMATION FROM SATELLITE DATA

Nils Lehmann
Technical University of Munich
n.lehmann@tum.de

Nina Maria Gottschling
EO Data Science, DLR
nina-maria.gottschling@dlr.de

Stefan Depeweg
Siemens AG
stefan.depeweg@siemens.com

Eric Nalisnick
University of Amsterdam
e.t.nalisnick@uva.nl

ABSTRACT

Deep neural networks (DNNs) have been successfully applied to earth observation (EO) data and opened new research avenues. Despite the theoretical and practical advances of these techniques, DNNs are still considered black box tools and by default are designed to give point predictions. However, the majority of EO applications demand reliable uncertainty estimates that can support practitioners in critical decision making tasks. This work provides a theoretical and quantitative comparison of existing uncertainty quantification methods for DNNs applied to the task of wind speed estimation in satellite imagery of tropical cyclones. We provide a detailed evaluation of predictive uncertainty estimates from state-of-the-art uncertainty quantification (UQ) methods for DNNs. We find that predictive uncertainties can be utilized to further improve accuracy and analyze the predictive uncertainties of different methods across storm categories.

1 INTRODUCTION

The tremendous success of Deep Learning approaches to natural images is increasingly being explored on EO data that is becoming available in ever greater quantities (Tuia et al., 2023). Due to their often vast global coverage, EO data is an indispensable source of information for assessing the state of our planet as well as extreme events that are increasing in frequency and intensity (Kikstra et al., 2022). One category of such extreme events are tropical cyclones. Tropical cyclones - in the US alone - have lead to 6,789 deaths and caused financial damages amounting to a staggering $1,333.6 billion between 1980-2022, with an average instance cost of $22.2 billion and covering 53.9% of all costs caused by US extreme weather disasters (Smith, 2020). Although, satellite data and other in-situ measurements are often available, reliable wind speed estimation remains a challenging task. For example in October 2023, hurricane Otis underwent a rapid intensification of almost 80 kts in 12 hours before causing devastating damage in the city Acapulco.[1] The failure of satellite based wind speed estimation methods (Krämer, 2023) and the need for improving these has been highlighted after this tropical cyclone[2]. Moreover, rapidly intensifying storms near coastlines have shown a trend to become more frequent (Li et al., 2023) and, hence, this demonstrates the need for improved monitoring of wind speeds and better prediction methods to yield improved warning systems. Because data to train such prediction methods can be limited and unevenly distributed, making a perfect prediction is not always possible. However, based on the general viability of DNNs for predicting and estimating wind speeds from satellite data (see e.g. Pradhan et al. (2017)), one possible approach is to equip DNNs with modern uncertainty-quantification (UQ) methods to enhance the quality of predictions and mitigate data imbalances, as well as label and input noise. This uncertainty is important for EO applications, as in practice, a prediction model is only an element of a complex decision making process. For instance, the confidence in the prediction of a

[1] "Hurricane Otis Causes Catastrophic Damage in Acapulco, Mexico", NOAA accessed 31.01.2024.

[2] "Hurricane Otis smashed into Mexico and broke records. Why did no one see it coming?" accessed 31.01.2024.

tropical cyclone category is a key factor for deciding on public safety measures. This paper has the following contribution: Using the dataset proposed in Maskey et al. (2021), we show that equipping DNNs with predictive uncertainty can be utilized to further improve accuracy via selective prediction based on predictive uncertainty. To the best of our knowledge no previous related work (see Section 1.1) considered an evaluation of uncertainty aware regression models in this domain. We compare state-of-the-art UQ methods, (see Section 3), and demonstrate differences across storm categories according to the Saffir-Simpson scale and different dataset splits. We show that UQ can improve real-time wind speed estimation and thus outline the way to apply UQ to DNN forecasting models by a detailed assessment of existing UQ methods.

1.1 Related Work

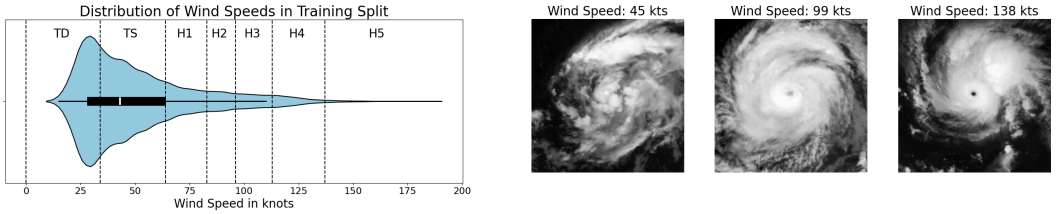
Several works have tackled the task of applying Deep Learning methods to tropical cyclone intensity estimation as a classification (Wimmers et al., 2019) or regression (Chen et al., 2019; Ma et al., 2024; Zhang et al., 2021) task. Based on a dataset of 25k images of infrared satellite imagery matched with storm data from the HURDAT2 database (Landsea & Franklin, 2013), Pradhan et al. (2017) train a CNN architecture for storm-category classification, as well es wind-speed estimation, and demonstrate improvements over previously applied statistical techniques like Advanced Dvorak Technique (ADT) (Piñeros et al., 2011), and Deviation-Angle Variance Technique (DAVT) (Ritchie et al., 2014). Maskey et al. (2020) improve the dataset quality and size by using GEOS Geostationary Operational Environmental Satellite (GEOS) and demonstrate a live production system. Our work is mostly comparable to Maskey et al. (2020) as we use their published dataset that was part of the Driven Data Challenge (Maskey et al., 2021).

2 Tropical Cyclone Dataset

Dataset name	Satellite	Spatial Res	Temporal Res	Train Samples	Val Samples	Test Samples
Tropical Cyclone	GOES	2km	15 min	53k	11k	43k

Table 1: Dataset Overview

The imagery represents single channel long-wave infrared measurements captured every 15 minutes, at 10.3 microns, that can capture the spatial structure of the storm in terms of measurements of the brightness temperature, as seen in Figure 1b. For more details about dataset collection, we refer the reader to the methodology section of (Maskey et al., 2020). We resize the images to 224x224 pixels and employ common image augmentations during training. We follow the datasplits by storm of the challenge and use dataloading available through the TorchGeo library (Stewart et al., 2022), which yields 53k training, 11k validation and 43k test samples. As Figure 1a shows, the distribution of targets is highly skewed with the majority of samples falling beneath hurricane categories defined by the Saffir Simpson Scale, Simpson (1974). We conduct experiments with the full target range but also subsets that only contain hurricane categories.



(a) Label distribution and storm categories.

(b) Dataset samples

Figure 1: Visualization of Tropical Cyclone Dataset.

3 Methods

Given a set of input-target pairs $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$, (x_i, y_i), the task of the neural network is to predict a target $y^\star \in Y$ given an input $x^\star \in X$. The input is a triplet of monochrome satellite

images at time steps $[t-2, t-1, t]$ and the target is the maximum sustained wind speed in knots (kts) at time step t. [3] This is sometimes referred to as "nowcasting". For this task, we compare five classes of UQ methods: deterministic, ensemble, Bayesian, quantile and diffusion based methods. Firstly, deterministic UQ methods use a DNN, $f_\theta : X \to \mathcal{P}(Y)$, that map inputs x to the parameters of a probability distribution $f_\theta(x^\star) = p_\theta(x^\star) \in \mathcal{P}(Y)$. These include Deep Evidential Networks (**DER**) Amini et al. (2020), where we use the correction proposed by Meinert et al. (2023), and Mean Variance Networks (**MVE**) (Nix & Weigend, 1994) which output the mean and standard deviation of a Gaussian distribution $f_\theta^{\text{MVE}}(x^\star) = (\mu_\theta(x^\star), \sigma_\theta(x^\star))$. Secondly, the broadly considered state-of-the-art method Deep Ensembles (**DeepEnsembles**) proposed by Lakshminarayanan et al. (2017) utilizes an ensemble over MVE networks. Thirdly, Bayesian methods aim at modelling a distribution over the network parameters and are commonly used to approximate the first and second moment of a marginalized distribution. These include Bayesian Neural Networks with Variational Inference (**BNN VI ELBO**) Blundell et al. (2015), MC-Dropout (**MCDropout**) Gal & Ghahramani (2016), the Laplace Approximation (**Laplace**) Ritter et al. (2018)Daxberger et al. (2021) and **SWAG** Maddox et al. (2019) with partially stochastic variants presented in Sharma et al. (2023). Gaussian Process based methods model a distribution over functions that also approximate the fist and second moment of the marginalized distribution. These include Deep Kernel Learning (**DKL**) Wilson et al. (2016) and an extension thereof Deterministic Uncertainty Estimation (**DUE**) (van Amersfoort et al., 2021). Fourthly, quantile based models $f_\theta : X \to Y^n$ that map to n quantiles, $f_\theta(x^\star) = (q_1(x^\star), ..., q_n(x^\star)) \in Y^n$, such as Quantile Regression (**Quantile Regression**) and the conformalized version thereof (**ConformalQR**) suggested by Romano et al. (2019). Lastly, we also consider a diffusion model (**CARD**) as introduced by Han et al. (2022). A detailed description of the methods is provided in the supplementary material. Depending on underlying assumptions UQ, methods are regarded to express two different types of uncertainties (Hüllermeier & Waegeman, 2021). **Aleatoric uncertainty** refers to inherent randomness in the data and **epistemic uncertainty** to a lack of knowledge in the modelling process. From a statistical perspective Gruber et al. (2023) allude that such a distinction is often not possible. Thus, we focus solely on predictive uncertainty.

Evaluation methodology: In addition to standard metrics for regression, such as root-mean-squared error (RMSE), we utilize proper scoring rules such as the negative log-likelihood (NLL) and continuous ranked probability score (CRPS) (Gneiting & Raftery, 2007) and the mean absolute calibration error (MACE). To evaluate the merit of UQ methods for decision making, we use selective prediction as a downstream task. Here, samples with a predictive uncertainty above a given threshold are omitted from prediction and can be referred to an expert or other estimation methods. Based on the Saffir-Simpson Scale (Simpson, 1974) bin intervals, we chose the threshold such that it would on average shift the category of the regression prediction. Hence, we take the threshold to be the mean over categories of the wind speed interval from categories 1 to 4, which is approximately 9 kts. We experiment with different threshold choices which are reported in the supplementary material and in Fig. 3b. All methods have an ImageNet pretrained ResNet-18 (He et al., 2016) backbone available frome the timm library (Wightman, 2019). Metrics are computed with the UQ-toolbox by Chung et al. (2021).[4]

4 RESULTS

We show fine grained results for storm categories Tropical Depression (TD), and Hurricane categories 1, 3, and 5 for better visualization. Additional results for different dataset splits and thresholds including all categories are included in the supplementary material.

How effective is selective prediction? As Table 2 shows, selective prediction - enabled through uncertainy aware models - can yield significant accuracy improvements for selected methods. The best performing methods obtain an RMSE between $9.27 - 10.95$ kts, yet the accuracy improvement obtained by selective prediction varies significantly. However, the coverage - the remaining samples after selective prediction - also varies considerably. For higher hurricane categories, accuracy and uncertainty metrics worsen substantially as shown in Figure 2 and different ranges of improvement are obtained by selective prediction, as shown in the supplementary material. When averaging

[3] We choose this input image composition, as it was utilized in the winning solution of the challenge (Maskey et al., 2021), which improved reported accuracy significantly compared to (Maskey et al., 2020).

[4] Code available under https://github.com/nilsleh/tropical_cyclone_uq

UQ group	Method	RMSE ↓	RMSE Δ ↑	Coverage ↑	CRPS ↓	NLL ↓	MACE ↓
None	Deterministic	10.50	0.00	1.00	NaN	NaN	NaN
Deterministic	**MVE**	9.95	2.10	**0.62**	**5.31**	**3.64**	0.04
	DER	10.14	NaN	0.00	10.07	4.60	0.35
Quantile	QR	10.95	3.28	0.44	5.82	3.73	**0.01**
	CQR	10.95	**6.18**	0.08	5.98	3.79	0.10
Ensemble	Deep Ensemble	16.19	0.00	0.00	8.83	4.15	0.05
Bayesian	MC Dropout	10.23	6.12	0.00	5.78	3.81	0.16
	SWAG	9.78	5.42	0.11	5.40	3.71	0.13
	Laplace	10.53	0.00	0.00	7.96	4.31	0.28
	BNN VI ELBO	**9.27**	0.00	1.00	6.28	52.60	0.41
	DKL	12.59	0.00	0.00	6.84	3.95	0.06
	DUE	9.95	0.00	0.00	5.43	3.73	0.08
Diffusion	CARD	10.86	1.50	0.60	5.84	3.92	0.05

Table 2: Evaluation Results on test set. RMSE Δ shows the improvement after selective prediction, where 0.00 indicates that all samples were withdrawn, while Coverage denotes the fraction of remaining samples that were not omitted. Threshold 9 kts.

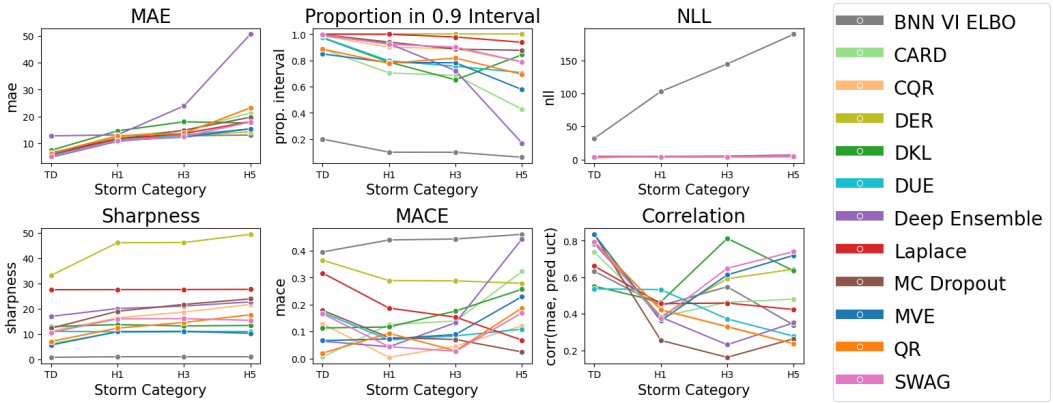


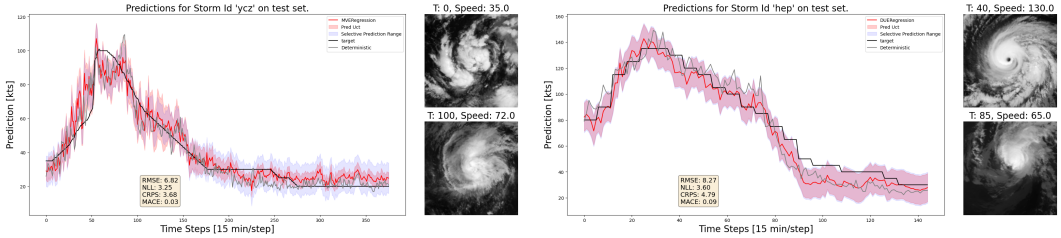
Figure 2: Uncertainty Metrics over different storm categories. We find that VI BNNs under cover (e.g. see proportion in interval), DER tends to over cover (e.g. see sharpness), with many other methods performing in between.

over all categories Table 2 shows that SWAG and CQR obtain relatively low RMSE after seletive prediction, 4.36 and 4.77 kts, while maintaining a coverage of 11% and 8%.

Error and Predictive Uncertainty across Categories: We evaluate the predictive uncertainty across storm categories with three criteria: the correlation between predictive uncertainty and MAE, sharpness, and MACE. Fig. 2, on the bottom right, shows the correlation is best for the TD case and is fairly consistent for most models. On the higher categories H3 and H5 we observe a larger spread between models, with SWAG, MVE, DKL and DER demonstrating higher correlation values. Accurate predictive uncertainties need to be both well calibrated - obtain a low MACE - and be sharp Kuleshov et al. (2018). MVE, SWAG, QR and CQR most closely fulfill this criteria. In contrast, Laplace and MC-Dropout obtain a low MACE on higher categories but are also less sharp and show lower correlation. Fig. 3a gives a "qualitative" example of MVE predictions for a selected storm track which generally follows the trend of the underlying target. Samples with a predictive uncertainty that exceeds the selective prediction threshold could be referred to an expert or postprocessing step.

4.1 Detailed discussion per UQ method group

Deterministic UQ methods: Table 2 shows MVE obtains an RMSE of 7.85 kts after selective prediction while maintaining a coverage of 62% and the lowest scoring rules, NLL and CRPS, which may be correlated to the fact that the loss objective is the NLL. At the same time MVE remains well calibrated compared to all other methods. Table 1 in the Appendix, Section 1, shows that MVE also obtains a comparably low RMSE and NLL per category. DER obtains a higher RMSE and no improvement with selective prediction, Table 2. This may be due to the fact that the predicted standard deviations of DER are relatively high compared to the selective prediction threshold, which is reflected in the sharpness accross storm categories in Figure 2. **Quantile based UQ methods**:



(a) MVE prediction Example with a visualized threshold of 9 kts.

(b) DUE Prediction Example with a visualized threshold of 12 kts.

Figure 3: Predictive Uncertainty Examples. Note that models under our setup do not have a concept of time, we merely combine individual nowcasting predictions into a time-series. Red shaded areas exceeding blue areas indicate samples that *would* be omitted during selective prediction. Figure inspired by Zhang et al. (2019).

CQR obtains higher improvements with selective prediction than QR, see Table 2, which is due to conformalization of quantiles and the resulting shift in predictive uncertainty. Yet this comes at the cost of a significantly lower coverage of CQR after selective prediction with only 8% compared to 44% for QR. **Ensemble methods**: Table 2 shows that overall Deep Ensembles obtain a higher RMSE than all other methods and also a significantly higher RMSE on category 5 cyclones, see Figure 2. Although Deep Ensembles are considered state-of-the-art, Seligmann et al. (2024) also find that they do not perform best at every UQ task. As for DER the predictive uncertainty of Deep Ensembles is larger than the selective prediction threshold, resulting in no improvement in RMSE. However, choosing a different threshold may result in accuracy improvements. We hypothesize that the variance of ensemble members might not be large enough and instead have converged to similar solutions, which implies that the ensemble members have similar biases. **Bayesian methods**: MC Dropout obtains an RMSE improvement for selective prediction, resulting in 4.11 kts at the cost of a coverage of approximately 0 %. This means that after selective predictions almost no samples remain, potentially adapting the threshold may result in improvements. SWAG obtains significant improvements with selective prediction at the cost of a low coverage of 11% and obtains relatively low CRPS and NLL as well as MACE averaged over categories, see Table 2, as well as per category, Figure 2. This indicates a good fit, however the coverage after selective prediction may be improved with a different threshold. Laplace obtains no improvement with selective prediction and interestingly also has a constant sharpness across categories as Figure 2 shows. This may be due to the fact that the Laplace approximation uses a second order Taylor expansion with respect to the model parameters of the loss and does not take into account variances in the data to construct a Gaussian approximation to the posterior weight distribution. BNN VI ELBO interestingly obtains the lowest RMSE per category and overall, Table 1 in the Appendix, Section 1, which indicates a good fit of the mean prediction. However, the predictive uncertainties are relatively small as the low sharpness and high negative log likelihood per category suggest, Figure 2. DKL obtains a relatively high RMSE and no improvement with selective prediction, although the correlation between predictive uncertainty and MAE, Figure 2, is also high on higher categories. However this may be due to high errors and high uncertainties. Compared to DKL, DUE obtains a significantly lower RMSE which may be due to the spectral normalization of layers, as this is the only difference between the methods. Otherwise DUE obtains a lower MACE per category than DKL, yet also a lower correlation between predictive uncertainty and MAE. **Diffusion UQ methods**, surprisingly CARD obtains a average RMSE and a significant improvement with selective prediction, while maintaining a coverage of 60 % and a low miscalibration error (MACE) of 0.05.

5 Conclusion

We presented a first analysis of predictive uncertainty for cyclone wind speed estimation. The various methods considered performed quite differently across storm categories and often exhibited a tradeoff between coverage vs accuracy. When predicting the maximum sustained wind speed, MVE demonstrated high coverage and low RMSE. Yet if a lower coverage is tolerable, then SWAG is a more attractive option due to it having a better RMSE than MVE. In future work, we plan to consider autoregressive models for the time series task presented in Figure 3.

6 ACKNOWLEDGEMENTS

This work was supported by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@KIT partition.

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Buo-Fu Chen, Boyo Chen, Hsuan-Tien Lin, and Russell L Elsberry. Estimating tropical cyclone intensity by satellite imagery utilizing convolutional neural networks. *Weather and Forecasting*, 34(2):447–465, 2019.

Youngseog Chung, Ian Char, Han Guo, Jeff Schneider, and Willie Neiswanger. Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*, 2021.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*, 2023.

Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

Jarmo S Kikstra, Zebedee RJ Nicholls, Christopher J Smith, Jared Lewis, Robin D Lamboll, Edward Byers, Marit Sandstad, Malte Meinshausen, Matthew J Gidden, Joeri Rogelj, et al. The ipcc sixth assessment report wgiii climate assessment of mitigation pathways: from emissions to global temperatures. *Geoscientific Model Development*, 15(24):9075–9109, 2022.

Katrina Krämer. Daily briefing: Why forecasters failed to predict hurricane otis. *Nature*, 2023.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Christopher W Landsea and James L Franklin. Atlantic hurricane database uncertainty and presentation of a new database format. *Monthly Weather Review*, 141(10):3576–3592, 2013.

Yi Li, Youmin Tang, Shuai Wang, Ralf Toumi, Xiangzhou Song, and Qiang Wang. Recent increases in tropical cyclone rapid intensification events in global offshore regions. *Nature Communications*, 14(1):5167, 2023.

Zhaoyang Ma, Yunfeng Yan, Jianmin Lin, and Dongfang Ma. A multi-scale and multi-layer feature extraction network with dual attention for tropical cyclone intensity estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

M. Maskey, R. Ramachandran, I. Gurung, B. Freitag, M. Ramasubramanian, and J. Miller. Tropical Cyclone Wind Estimation Competition Dataset. `https://doi.org/10.34911/rdnt.xs53up`, 2021.

Manil Maskey, Rahul Ramachandran, Muthukumaran Ramasubramanian, Iksha Gurung, Brian Freitag, Aaron Kaulfus, Drew Bollinger, Daniel J Cecil, and Jeffrey Miller. Deepti: Deep-learning-based tropical cyclone intensity estimation system. *IEEE journal of selected topics in applied Earth observations and remote sensing*, 13:4271–4281, 2020.

Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The unreasonable effectiveness of deep evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9134–9142, 2023.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Miguel F Piñeros, Elizabeth A Ritchie, and J Scott Tyo. Estimating tropical cyclone intensity from infrared image data. *Weather and forecasting*, 26(5):690–698, 2011.

Ritesh Pradhan, Ramazan S Aygun, Manil Maskey, Rahul Ramachandran, and Daniel J Cecil. Tropical cyclone intensity estimation using a deep convolutional neural network. *IEEE Transactions on Image Processing*, 27(2):692–702, 2017.

Elizabeth A Ritchie, Kimberly M Wood, Oscar G Rodríguez-Herrera, Miguel F Piñeros, and J Scott Tyo. Satellite-derived tropical cyclone intensity in the north pacific ocean using the deviation-angle variance technique. *Weather and forecasting*, 29(3):505–516, 2014.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning, 2018.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Florian Seligmann, Philipp Becker, Michael Volpp, and Gerhard Neumann. Beyond deep ensembles: A large-scale evaluation of bayesian deep learning under distribution shift. *Advances in Neural Information Processing Systems*, 36, 2024.

Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do bayesian neural networks need to be fully stochastic? In *International Conference on Artificial Intelligence and Statistics*, pp. 7694–7722. PMLR, 2023.

Robert H Simpson. The hurricane disaster—potential scale. *Weatherwise*, 27(4):169–186, 1974.

Adam B. Smith. U.s. billion-dollar weather and climate disasters, 1980 - present (ncei accession 0209268), 2020. URL `https://www.ncei.noaa.gov/archive/accession/0209268`.

Adam J Stewart, Caleb Robinson, Isaac A Corley, Anthony Ortiz, Juan M Lavista Ferres, and Arindam Banerjee. Torchgeo: deep learning with geospatial data. In *Proceedings of the 30th international conference on advances in geographic information systems*, pp. 1–12, 2022.

Devis Tuia, Konrad Schindler, Begüm Demir, Gustau Camps-Valls, Xiao Xiang Zhu, Mrinalini Kochupillai, Sašo Džeroski, Jan N van Rijn, Holger H Hoos, Fabio Del Frate, et al. Artificial intelligence to advance earth observation: a perspective. *arXiv preprint arXiv:2305.08413*, 2023.

Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv preprint arXiv:2102.11409*, 2021.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pp. 370–378. PMLR, 2016.

Anthony Wimmers, Christopher Velden, and Joshua H Cossuth. Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Monthly Weather Review*, 147(6):2261–2282, 2019.

Chang-Jiang Zhang, Xiao-Jie Wang, Lei-Ming Ma, and Xiao-Qin Lu. Tropical cyclone intensity classification and estimation using infrared satellite images with deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2070–2086, 2021.

Rui Zhang, Qingshan Liu, and Renlong Hang. Tropical cyclone intensity estimation using two-branch convolutional neural network from infrared and water vapor images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):586–597, 2019.