Evaluation of Post Hoc Uncertainty Quantification Approaches for Flood Detection From SAR Imagery

Jakob Ludwig and Ronny Hänsch, *Senior Member, IEEE*

***Abstract*—Deep neural networks are the current state-of-the-art for analysis of remote sensing imagery. While they often provide accurate results, they are usually prone to be overconfident with respect to their predictions. In particular when these predictions are used by human decision makers in high stake scenarios, e.g., during detection and monitoring of natural disasters, trustworthiness is a necessary feature. In the context of flood detection from SAR imagery, this work evaluates a variety of uncertainty quantification methods that are applicable to already trained models (i.e., post hoc approaches) and provides detailed experiments evaluating the quantification quality of the different methods. The results show typical and pronounced uncertainty features, such as class boundaries, inaccurate predictions, and out-of-distribution regions. In summary, the overall best epistemic estimations are (albeit with only a small margin) provided by student ensembles, followed closely by the well established Monte Carlo dropout quantifier. However, surprisingly nearly all evaluated methods (incl. using the original softmax estimate of the trained model) provide rather similar results.**

***Index Terms*—Convolutional neural networks, floods, remote sensing (RS), semantic segmentation, synthetic aperture radar (SAR), uncertainty estimation.**

I. Introduction

THE combination of Earth observation imagery and deep learning allows for efficient detection and monitoring of natural disasters, such as wildfires [1], [2], [3], Earthquakes [4], and floods [5], [6], [7]. However, deep networks are usually treated as black-box systems [8] which give little insights into how decisions are made and how trustworthy they are beyond an average accuracy score estimated over a test set. While Explainable AI [9], [10] addresses the first part, i.e., why a machine learning model made a certain prediction, uncertainty quantification (UQ) aims to answer the question how reliable a prediction is [11] (see Fig. 1 for an illustration).

Approaches for UQ can be roughly grouped into two categories: Methods that inherently provide uncertainty estimates while predicting the target variable and post hoc approaches that can be applied to already trained models. The former have the advantage of usually providing very accurate uncertainty estimates, but often at the cost of higher training and inference times and being limited to certain model architectures. Examples include Bayesian neural networks (BNNs) [12], which model network parameters (e.g., weights) as distributions and deep ensembles [13], which train multiple model instances in parallel to fuse their predictions at run time. Post hoc approaches aim to provide accurate uncertainty values for predictions of an already trained model without changing its architecture or training procedure allowing for a larger applicability to a wide range of model architectures. Examples include data augmentation [14] or dropout [15] during inference to create varying model predictions.

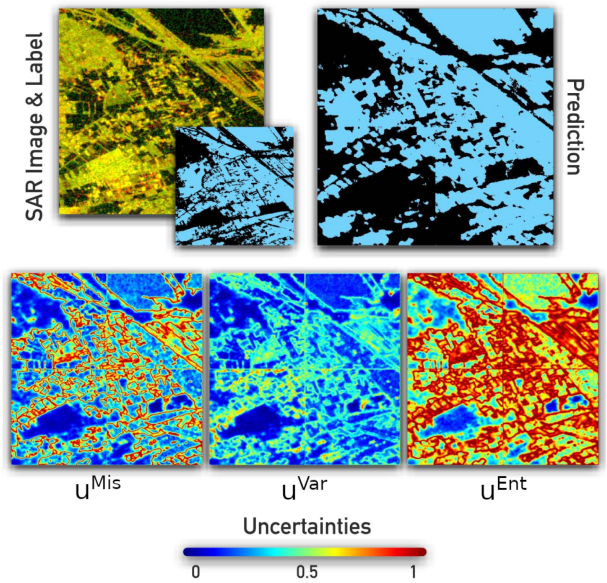


Fig. 1. SAR image, reference label, prediction and multiple UQ maps of a flood event in Spain. Uncertainty values are estimated with student ensembles. Three different quantification metrics are used: the negative maximum class probability u^{Mis}, the standard deviation of the ensemble predictions u^{Var}, and the entropy of the averaged class posterior u^{Ent}.

While UQ remains an active field of research and there are several general review articles, a structured comparison of UQ approaches is missing—in particular in the field of the semantic analysis of remote sensing imagery. For the use case of flood prediction from synthetic aperture radar (SAR) images, we compare several established post hoc UQ methods including the softmax baseline, Monte Carlo (MC) dropout, weight noise, and test-time data augmentation (TTDA), and add student ensembles allowing to use deep ensembles in a post hoc setting. For student ensembles, the already trained network serves as a teacher model in a knowledge distillation framework, i.e., an ensemble of

Received 13 October 2024; revised 27 February 2025 and 19 May 2025; accepted 29 May 2025. Date of publication 9 June 2025; date of current version 30 June 2025. *(Corresponding author: Ronny Hänsch.)*

The authors are with SAR Technology Department, German Aerospace Center (DLR), 82234 Oberpfaffenhofen, Germany (e-mail: ronny.haensch@dlr.de).

Digital Object Identifier 10.1109/JSTARS.2025.3577690

TABLE I
(NONEXHAUSTIVE) OVERVIEW OF UQ APPROACHES FOR DEEP NETWORKS

	Method	Post-hoc	Aleatoric, Epistemic	Applications
Determ.	**Softmax** [16]	✓	Epistemic	[17], [18], [19], [20]
	Deterministic UQ [21]	✗	Combined	[22], [23]
	Deep deterministic uncertainty [24]	✗	Aleatoric, Epistemic	[17]
Bayesian	**MC Dropout** [15]	✗	Epistemic	[25], [26], [27], [28], [29], [30]
	BNN [12]	✗	Epistemic	[31], [32], [33], [34]
	Bayes by Backpropagation [35]	✗	Epistemic	[25], [20], [29]
	Weight Noise [36]	✓	Epistemic	
Ensemble	Deep Ensembles [13]	✗	Epistemic	[26], [27], [25], [37]
	Deep Ensemble Approximations [42], [43], [44], [45]	✗	Epistemic	[38], [39], [40], [41]
	Student Ensembles	✓	Epistemic	
	TTDA [14]	✓	Aleatoric	[46], [47]

Methods evaluated in this work are marked in bold.

students is trained on the model predictions of the teacher to mimic its behavior.

The main contributions of this work are as follows.

1) Benchmarking several post hoc UQ methods on the task of flood prediction from SAR.
2) Proposing student ensembles as a post hoc version of deep ensembles.
3) Evaluating various UQ metrics.
4) Providing a versatile toolbox for UQ for semantic segmentation.

II. RELATED WORK

There are several general review papers for UQ in deep learning discussing different sources of uncertainty predictions and approaches to quantify them. These include UQ in generic applications establishing a systematic taxonomy of UQ methods [11], [48], [49] as well as being more focused on remote sensing and Earth observation [50].

Here, we provide a comprehensive summary of such methods but focus on methods that are later used in the comparative evaluation (see Table I).

Deep neural networks for classification tasks often model the prediction via a softmax layer that transforms the model logits into a probabilistic estimate of the class posterior. However, this softmax posterior is prone to be biased to overconfident estimates, i.e., having high probabilities for correct as well as wrong predictions [16]. Deterministic UQ models aim to mitigate this issue by leveraging novel loss functions [21] or regularizations [24] with limited success.

A more intuitive approach to model uncertainty is to leverage network architectures that inherently provide well-calibrated probabilistic estimates. The most common approach are BNNs that represent model weights not as trainable scalar values but as learnable distributions tracking the uncertainty over these parameters. Despite their potential, BNNs have issues with high dimensional parameter spaces and large datasets [31], [32]. While variational methods [33], [34] try to mitigate these problems, they often fail to model parameter correlation within the posterior. Furthermore, inference requires extensive sampling from the learned weight distributions.

Sampling free variations usually restrict the estimation of distributions to certain parameters (e.g., only of input and output layers) and propagate input variance throughout the network to model uncertainty [51], [52].

Another approach to circumvent the challenges posed by BNNs and their variants is to rely on fully deterministic networks only and create the necessary variation via combining multiple model instances within an ensemble [13], [42], [43], [44], [45]. Such Deep Ensembles have been shown to provide excellent uncertainty estimates at the cost of multiple training and inference runs. Lightweight versions of deep ensembles aim to create multiple predictions from a single model making multiple training runs obsolete while still requiring multiple inference runs. One approach is to leverage different versions of a model as they occured during the training, i.e., using model weights of different training epochs [53]. This disadvantage of multiple training runs can be mitigated to some extent by knowledge distillation [54], i.e., training a single model (the student) to predict the posterior distribution as estimated by the ensemble (the teacher) [19], [55], [56], [57].

A particular ensemble variant is stacking [58], [59] where the output of several models and their predictive uncertainty are used to fuse them into a single prediction via a weighted estimate [60]. Stacking has been used with Gaussian processes to transform uncertainties of geospatial input data into uncertainty of the predictions [61]. It also has been empirically shown that stacking improves the calibration quality and thus the basis of UQ in the case of probabilistic estimators, such as random forests [62].

Neither BNNs nor deep ensembles allow to analyze the predictions of a given, trained model regarding their uncertainty. The former requires using a specific model architecture while the latter would at least require retraining multiple instances. Gal and Ghahramani [15] introduced the approach of performing dropout at test time while Wang et al. [14] showed the applicability of data augmentation at test time for UQ. Both approaches create multiple estimates by either using slightly different versions of the network (dropout) or input data (data augmentation) and leverage the variation in these model predictions to quantify uncertainty. Kendall et al. [30] introduced one of the first segmentation networks approximating a BNN with dropout. These methods have the advantage that they decouple architectural choices to solve a given task from the additional requirement of modeling the uncertainty of the predictions. Thus, such post hoc approaches can be applied to any given model.

Another example of post hoc approaches are surrogate models that train a UQ model to mimic the behavior of a given trained model (e.g., based on density estimation of the model outputs [63]) or aim to directly estimate the uncertainty based on whether the underlying model made a correct or wrong prediction of the training data [64], [65].

Given the advantages of deep ensembles, we aim to include them into the set of approaches evaluated in this work. To enable their use in a post hoc setting, we combine the ideas of surrogate models, deep ensembles, and knowledge distillation. However,

in contrast to the approaches mentioned above, we do not distill a given ensemble into a single model. Instead, a given trained model serves as teacher and is distilled into a deep ensemble of networks. This allows for accurate post hoc UQ with all the advantages (and limitations) of deep ensembles for any given black-box machine learning model (i.e., not being limited to deep networks).

The main focus of this work is on the evaluation of several standard post hoc UQ approaches. The majority of publications benchmarking different UQ methods are within the biomedical field [27], [37], [66], [67]. Ng et al. [25] test a significant number of UQ approaches including Bayes by backpropagation, MC dropout parameterized multiple times, deep ensembles, and stochastic segmentation networks [68] for the segmentation of MRI data. Another very recent work by Jacobsen et al. [26] compared MC dropout and deep ensembles for the use case of gastrointestinal polyp segmentation. A recent study benchmarks several standard UQ approaches in the context of classification of natural images (i.e., ImageNet) and with a focus on uncertainty disentanglement (i.e., the ability to distinguish different sources of uncertainty) and finds that none of the evaluated methods is truly able to solve this task in practice [69].

UQ of deep learning models applied to SAR imagery is rather limited. Dera et al. [70] applied an adapted BNN with variational inference to airborne polarimetric SAR data for classification. Chen et al. [71], [72] used spaceborne C-band SAR imagery from the RADARSAT-2 satellite and apply a BNN as well. Haas and Rabus [73] discussed the potential of deep ensembles and MC Dropout for road detection in RADARSAT-2 imagery. UQ has also been applied in object detection in SAR imagery to improve the interpretability of the results [74].

We evaluate the abovementioned post hoc UQ approaches in the context of flood prediction from SAR imagery. SAR data are highly relevant for flood prediction [7] as it is independent of daylight and can be acquired during cloud cover allowing for timely predictions. For detecting and monitoring natural disasters, reliable UQ is of utmost importance since model predictions are used by decision makers to organize response efforts. Understanding model outcomes and having a notion which predictions are reliable and can be acted upon is a necessary prerequisite to enhance the trust in the automatic analysis of remote sensing imagery in such applications.

III. Uncertainty Quantification

The uncertainty of predicting the label $\mathbf{y}$ from the input $\mathbf{x}$ based on a model with trainable parameters θ that have been optimized on a dataset D can be modeled as

$$p(\mathbf{y}|\mathbf{x}, D) = \int \underbrace{p(\mathbf{y}|\mathbf{x}, \theta)}_{\text{Aleatoric}} \underbrace{p(\theta|D)}_{\text{Epistemic}} \, d\theta \tag{1}$$

where the aleatoric and epistemic aspects of uncertainty cover uncertainty based on the randomness of the data itself (e.g., due to noise) that cannot be reduced and uncertainty based on the model itself that can be reduced, e.g., with using more data, stronger optimization, or a model with more capacity [11]. Most approaches of UQ address either epistemic uncertainty or the overall predictive uncertainty (see Table I).

In the following we briefly discuss standard approaches to estimate uncertainty, i.e., MC Dropout (see Section III-B), Weight Noise (see Section III-C), and TTDA (see Section III-D), before introducing student ensembles (see Section III-E) as an additional method enabling to use deep ensembles in a post hoc setting. Most of these approaches model uncertainty as variation of the posterior. Different ways to quantify uncertainty based on this variation are discussed in Section III-F.

A. Softmax Baseline

This method simply uses the estimated posterior $p(\mathbf{y}|\mathbf{x})$ based on a softmax transform over the logits of the neural network.

B. MC Dropout

This method falls into the variational inference category. The epistemic uncertainty $p(\theta|D)$ within the overall prediction $p(\mathbf{y}|\mathbf{x}, D)$ [see (1)] is intractable, but can be approximated with a family of variational distributions with support of variational inference [11], [75]. MC dropout achieves the family of distributions by dropout layers within the network that set the output of randomly selected units to zero during training and test time.

C. Weight Noise

Srivastava et al. [36] introduced dropout, but generalize the technique to multiplicative Gaussian noise as well. The weight noise uncertainty estimation method follows this idea but uses Gaussian additive noise instead. For one sampling run we add the noise ϵ to the network parameters θ by

$$\tilde{\theta}_i = \theta_i + \alpha \cdot \epsilon, \epsilon \sim \mathcal{N}(0, 1) \tag{2}$$

where $\alpha \geq 0$ controls the amount of noise ϵ to be added. Weight noise is a light variant of dropout, where just a small amount of noise is perturbing the weights and therefore achieves varying model outputs for multiple inference runs.

D. Test-Time Data Augmentation

TTDA [14] aims at simulating artificial variations of the signal to model data uncertainty. For each query sample, multiple inference runs are performed where variations are created by deterministically applying a combination of data augmentations to the input image. We use a combination of horizontal and vertical flips, rotation, additive Gaussian noise, and multiplication by a constant [76].

E. Deep Student Ensemble

Ensemble learning aims to increase accuracy while providing improved uncertainty estimates by creating multiple models and fusing their output instead of relying on a single instance [77]. In particular deep ensembles have been shown to be able to quantify model uncertainty with high accuracy [13]. We combine the idea of deep ensembles and knowledge distillation to obtain a post hoc UQ method that can be applied to any given trained model with access to data.

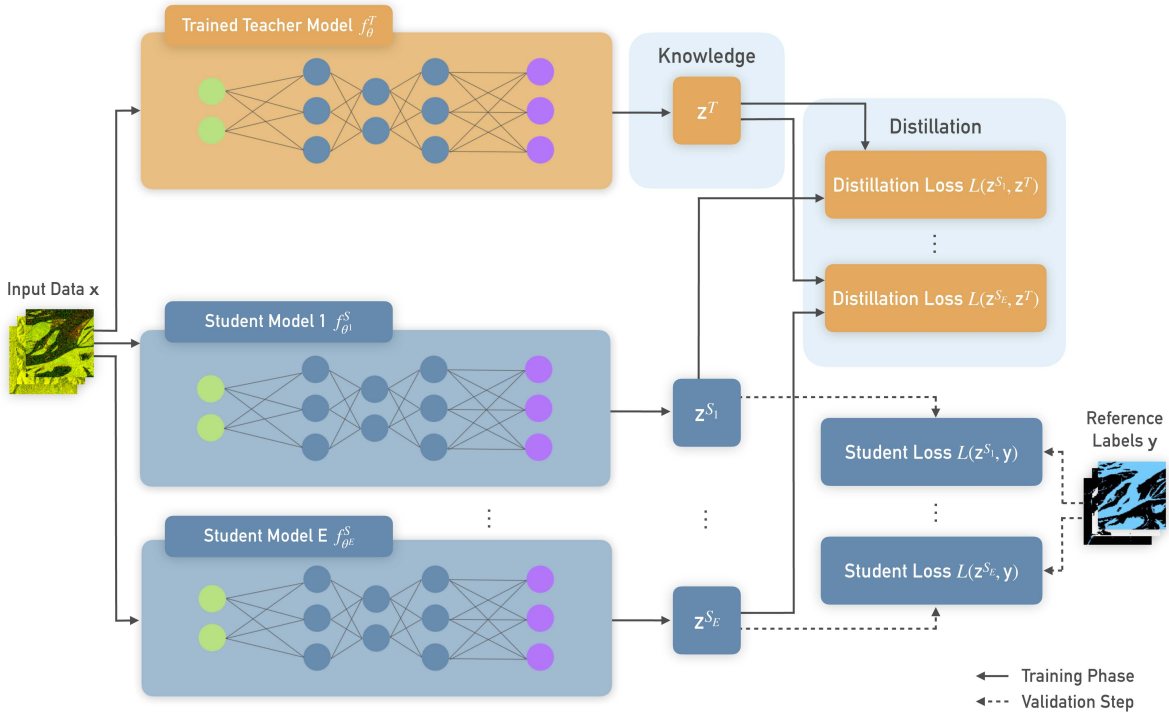


Fig. 2. *Training* process for an ensemble of students. The main training step consists of minimizing the distillation losses $\mathcal{L}(\mathbf{z}^T, \mathbf{z}^{S_e})$ per student model e. For informative purposes, the student models can be validated with reference annotations.

In the following, f_θ^T is such a trained model with parameters θ. Let $\mathbf{z}^T \in \mathbb{R}$ be the logit output of the model for a given sample $\mathbf{x}$, i.e., the model output before applying the final activation function (usually softmax in the case of a categorical target variable).

As illustrated in Fig. 2, we create a set of E new neural networks $f_{\theta_e}^S$ (with $1 \leq e \leq E$) with randomly initialized model weights θ_e. There are different typical architectural choices for student models, e.g., smaller networks that distill knowledge from a large teacher model into a more efficient architecture [54]. For simplicity, we choose to work with the same model architecture of the teacher [78], [79].

In contrast to standard deep ensembles, the goal of these models is not to estimate the target variable $\mathbf{y}$ for which the given model f^T was trained. Instead, they are supposed to match the logits $\mathbf{z}^T$ of f^T for a given input image $\mathbf{x}$ (note that this also means that ensemble training does not require reference data $\mathbf{y}$ but only input images $\mathbf{x}$).

While there are various types of knowledge distillation schemes [54], we focus on *response-based knowledge* combined with *offline distillation:* During training of the ensemble, the trained model f^T is applied to an input sample $\mathbf{x}$ leading to logits $\mathbf{z}^T$ and eventually the model estimate $\mathbf{y}^T$.

We apply the mean-squared-error between the logits $\mathbf{z}^T$ of the model and the output of the E ensemble members $\mathbf{z}^{S_e}$ over all M samples as distillation loss, i.e.,

$$\mathcal{L} = \frac{1}{M \cdot E} \sum_{m=1}^{M} \sum_{e=1}^{E} ||\mathbf{z}_m^T - \mathbf{z}_m^{S_e}||_2^2. \tag{3}$$

For validation purposes during training, the student model quality can be examined with standard performance metrics (e.g., the cross-entropy loss for categorical variables) with respect to the target variable, which we note as student loss $L(\mathbf{y}^{S_e}, \mathbf{y})$ in Fig. 2. Note, however, that this loss is purely informative and not used during optimization.

Due to random initialization of the ensemble weights θ_e and the independent shuffling of training samples, each ensemble member is slightly different after training despite being optimized on the same task. Thus, when presented with a query sample, each ensemble member will provide a slightly different output. The remaining steps (illustrated in Fig. 3) are virtually identical to any sampling-based UQ approach, such as TTDA and MC dropout: The computed logits $\mathbf{z}^{S_e}$ are converted into probabilities $p(\mathbf{y}|\mathbf{z}^{S_e})$ via softmax and leveraged to estimate the model uncertainty by applying an uncertainty metric (see Section III-F).

Unrelated to the uncertainty results, we are also able to obtain the final class prediction of the whole ensemble. While there are multiple combination approaches [13], we use the ensemble average due to its simplicity and good performance

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{E} \sum_{e=1}^{E} p_{\theta^e}(\mathbf{y}|\mathbf{x}). \tag{4}$$

F. UQ Metrics

Given the model estimate of the class posterior $p(\mathbf{y}|\mathbf{x})$ (either the softmax output of a single model or the averaged softmax output of the ensemble), there are several metrics to obtain a

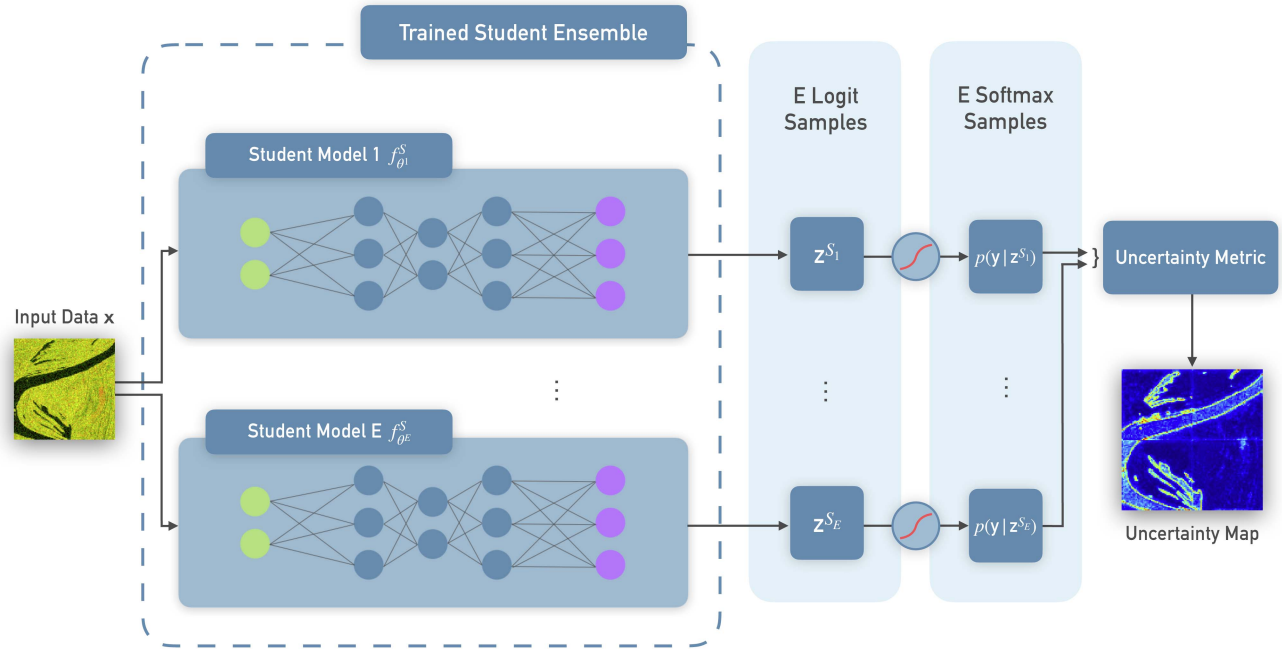


Fig. 3. Schema of the *UQ* process for an ensemble of students. Prerequisites are E trained student models $f^S_{\theta^e}$. After obtaining logit samples, the softmax function $\sigma(\mathbf{z}_e)$ is applied. Per sample, we receive a probabilistic distribution.

measure of uncertainty

$$\text{Misclass.: } u^{\text{Mis}} = 1 - \max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) \tag{5}$$

$$\text{Gap: } u^{\text{Gap}} = 1 - \left(\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) - \max^2_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) \right) \tag{6}$$

$$\text{Entropy: } u^{\text{Ent}} = H(p(\mathbf{y}|\mathbf{x})) \tag{7}$$

$$\text{Margin: } u^{\text{Mar}} = \max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) - p(\mathbf{y}_R|\mathbf{x}) \tag{8}$$

where $\max^2$ represents the second largest value. In addition, $p(\mathbf{y}_R|\mathbf{x})$ is the probability of the true class according to the reference data, and H is the entropy of the class posterior $p(\mathbf{y}|\mathbf{x})$ estimated for sample $\mathbf{x}$.

The first metric, i.e., u^{Mis} in (5) uses the probability of misclassification, i.e., if the largest entry in the posterior is 1 (and the others accordingly 0), the uncertainty is measured as 0. Note, that the smallest maximal value in the posterior is $1/C$ where C is the number of classes, i.e., $u^{\text{Mis}} \in [0, \frac{C-1}{C}]$. This metric does not account for the distribution of the remaining probability mass that is not contained in the maximal value. This is taken into account by the second metric, i.e., u^{Gap} in (6), which uses the difference between the first and second maximal values in the posterior. Thus, if two (or more) classes are estimated with the same, maximal probability the uncertainty is the largest. However, this metric does not distinguish between cases where the confusion is between different numbers of classes, i.e., posterior vectors of $[1/3, 1/3, 1/3]$ and $[0, 1/2, 1/2]$ would lead to the same uncertainty value (i.e., 1). This is accounted for in the third metric, i.e., u^{Ent} in (7), which computes the entropy of the class posterior, which is 0 if one class is estimated with 100% and 1 only if all classes obtained the same probability.

The last metric, i.e., u^{Mar} in (8), uses the margin of the posterior as defined by ensemble learning [80], i.e., the difference of the probability of the class with the maximum probability and the probability of the true class. This metric is strongly related to accuracy (i.e., it is 0 if the correct decision is made with 100% and 1 if the wrong decision is made with 100%) and can only be evaluated if reference data are available for the query sample. As such, it is of little use during the actual deployment of a method and can only be applied during evaluation, model selection, etc.

If sampling-based methods, such as the student ensemble are used, another alternative is to measure the variation in the set of estimates, e.g., via their standard variation

$$u^{Var} = \frac{1}{C} \sum_{k=1}^{C} \sqrt{\frac{1}{E-1} \sum_{e=1}^{E} \left(\mathbf{z}_k^{S_e} - \underbrace{\frac{1}{E} \sum_{e=1}^{E} \mathbf{z}_k^{S_e}}_{\text{Mean}} \right)^2} \tag{9}$$

for output logits $\mathbf{z}^s_k$ of a class k and E number of estimates. In contrast to the measures above [see (5)–(8)], this measure is not necessarily in [0,1] but lies within the range of the logit values of the model.

As a final step, we linearly scale all uncertainty measures to [0,1] for better comparability.

IV. Experiments

In the experiments, we focus on semantic segmentation as a frequent application in remote sensing and in particular flood detection from SAR images, i.e., a use case where knowledge about the reliability of a prediction is of particular importance.

TABLE II
ATTRIBUTES OF THE SEN1FLOODS11 DATASET [81]

Satellite	Sentinel-1	**Regions**	11
Frequency Band	C	**Data Points**	432
Polarizations	VV, VH	**Train / Val. / Test**	252 / 90 / 90
Feature Dimension	512 x 512	**Classes**	*flood*, *non-flood*
Cropped Dimension	256 x 256		

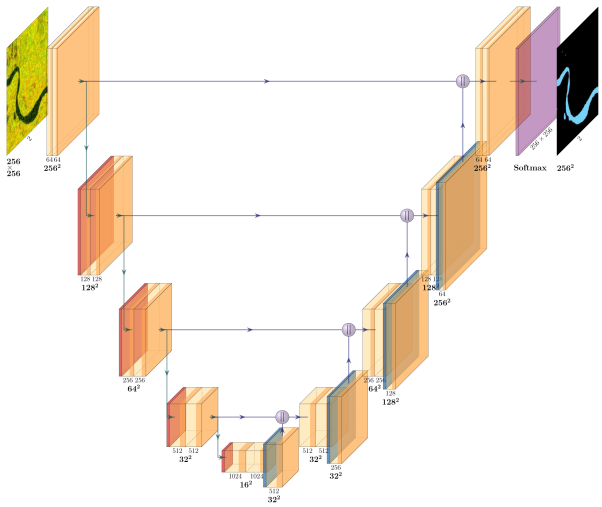


Fig. 4. U-Net architecture used during evaluation. *Bold* numbers under blocks represent the dimensions of the resulting tensor after a convolution or pooling operation. Other numbers in block captions depict the number of channels. Green arrows depict the max-pooling operation. Blue and up-facing arrows are bilinear filters, side-facing arrows skip-connections. The figure was created with the *"PlotNeuralNet"* library [82].

Network training and UQ utilize an Nvidia A100 SXM4 GPU with 40 GB of GPU memory and an AMD EPYC 7662 CPU with 64 cores, 128 threads, and 1008 GB of RAM. The code is implemented in *Python* 3.10 and uses *PyTorch* for the deep learning parts. All software incl. demo scripts and sample data is available at the project's webpage.[1]

A. Dataset, Network Architecture, and Baseline Results

We use SAR imagery from the *Sen1Floods11* dataset [81] that contains **Sen**tinel-**1** amplitude images of **11** flood events (see Table II for more detailed information).

Fig. 4 shows the architecture of the employed U-Net. The main difference compared to the vanilla architecture [83] are the 3×3 convolution layers with a padding of one between upsampling and downsampling layers. Downsampling max-pooling layers have a kernel size of 2×2. Additional dropout layers are added after each double block of standard convolutions. A group normalization layer [81] follows each convolution within the double block. Table III states the hyperparameters for the (teacher / reference) model training on the flood event dataset. The optimized loss function is a standard pixel-wise cross-entropy loss while prediction performance is measured as

[1][Online]. Available: https://github.com/jakob0246/w2-implementation

TABLE III
HYPERPARAMETERS FOR THE MODEL TRAINING PROCESS (TEACHER / REFERENCE MODEL)

Epochs	100	**Normalization Groups**	32
Batch Size	24	**Dropout Probability**	0.25
Learning Rate	1^{-4}	**Augmentations**	Normalization Random cropping Random flip
Optimizer	AdamW		
Loss Function	$\mathcal{L}_{CE}$		
Class Weights	*flood*: 2, *non-flood*: 1		

intersection of union (IoU) and balanced accuracy (BA, i.e., the average class-wise accuracy).

While the Sen1Floods11 dataset provides SAR imagery of 512×512 pixels, we follow the original setup [81] and use random cropping to 256×256 pixels as data augmentation during training. During inference, the input image is thus divided into four quadrants that are processed individually. The predictions are subsequently patched to the full 512×512 map.

The above described model and training procedure achieved an IoU of 42.30 / 51.60 / 45.07 and a BA of 77.20 / 81.92 / 80.17 on the training / validation / test sets.

The student ensemble's hyperparameters are identical to the model described above,which serves as the teacher model but with disabled dropout as it acts as regularization decreasing variance between models. We suggest to use this heuristic when training a student ensemble with the same model architecture. We use $E = 50$ models in the ensemble. The hyperparameter was chosen based on the grid-search approach when models were validated on the training and validation split. Generally, smaller datasets or inaccurate teacher models require a higher E. Experiments based on more accurate models only needed a value as low as $E = 3$ students.

B. Qualitative UQ Results

For the qualitative evaluation we concentrate on five typical sources of uncertainty as follows:

1) Borders of homogeneous segmentation areas for specific classes.
2) Image boundaries.
3) Regions with image artifacts.
4) Out-of-distribution areas or regions with unusual appearance.
5) Inaccurate or imbalanced class labels.

Fig. 5 shows estimated uncertainty heatmaps for three different SAR images. Boundaries of semantic segments as well as patch boundaries are clearly visible in every map. At segment boundaries the receptive field of a neural network covers areas of different classes, leading to increased uncertainty. Patch boundaries are problematic due to the loss of information caused by downscaling and upscaling within a U-Net architecture (see Fig. 4) and handling cases when the convolution kernel goes beyond the image boundary (i.e., padding). While there are strategies to mitigate these effects, we deliberately selected a model and processing pipeline that is affected by it. On the one

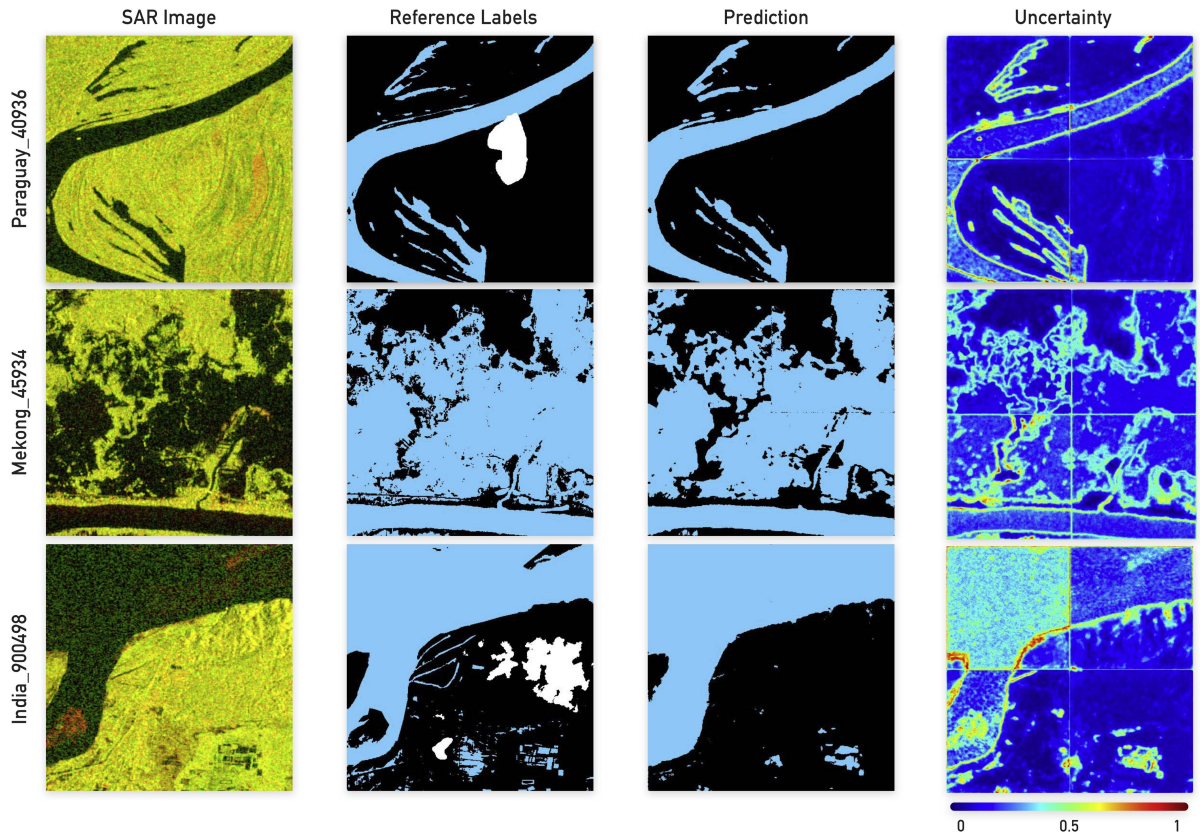


Fig. 5. Example SAR images, reference labels, predictions, and their uncertainty heatmaps for the student ensemble method. The uncertainty metric is u^{Var}.

hand, this paper evaluates post hoc approaches, i.e., methods that leave the underlying model unchanged and need to deal with its peculiarities. On the other hand, it is informative to observe that the predicted semantic maps barely show patch boundaries while they are very dominantly present in the uncertainty maps. The third image shows an example, where a misclassified area, i.e., the small island within the river at the bottom left corner, is marked by high uncertainty values. Interesting is also the role of spatial context: The top-left quadrant of the third image shows mainly water with very little land. While the network makes a correct prediction, the uncertainty is rather high, illustrating that the network struggles to analyze such a large homogeneous area. The top-right quadrant shows basically the same but together with a large portion of land which appears to help the model to put the homogeneous water region into context. The uncertainty values over the water region are visibly reduced, leading to a clear edge between both patches.

Fig. 6 compares the different UQ metrics (all based on the outcome of the student ensemble approach).

Segmentation boundaries between classes are visible for all metrics but the least clearly for u^{Mar} and u^{Ent}. While the former generally results in sparse uncertainty estimates since all pixels with correct predictions are assigned an uncertainty value of zero, the latter shows very broad segment borders and generally leads to rather uniform distributed uncertainty values. On the other hand, u^{Mis}, u^{Gap}, and u^{Var} mark segment boundaries with very distinct and fine lines. Patch boundaries are the most prominent within these three metrics.

The fourth row SAR image shows a line artifact at the bottom of the image which is best highlighted by u^{Var}. While it is also visible in u^{Mis} and u^{Gap}, it is overshadowed by the generally higher uncertainty values of these metrics within the water region. Areas with increased VV response (shown as red area in the pseudo-RGB) have a tendency to be falsely predicted. The associated uncertainty is most prominently shown by the u^{Ent}, u^{Mis}, and u^{Gap}. Other patterns that are difficult to segment or inaccurate flood segmentation results are best highlighted by u^{Var}.

In summary, u^{Ent} provides generally rather uniformly distributed uncertainty estimates, which make it hard to distinguish finer details. The u^{Mar} metric behaves more as a measure of accuracy than uncertainty. Furthermore, it relies on the actual class label rendering it only applicable during evaluation of a method but not during deployment. For a binary classification problem, u^{mis} and u^{Gap} are identical up to a scaling factor but provide complementary information to u^{Var}. The former shows high uncertainty values if there is a considerable confusion among the classes in the estimated posterior. In the case of an ensemble and other sampling-based UQ approaches, there are two possible reasons for this: either, each ensemble member estimates a class posterior with the same confusion among different classes, or each ensemble member estimates different posterior distributions, which only in the combined average show a strong confusion among classes. In the first case, u^{Var} would be low (even if all estimated posteriors would be the uniform distribution, their standard deviation would be minimal), while it would

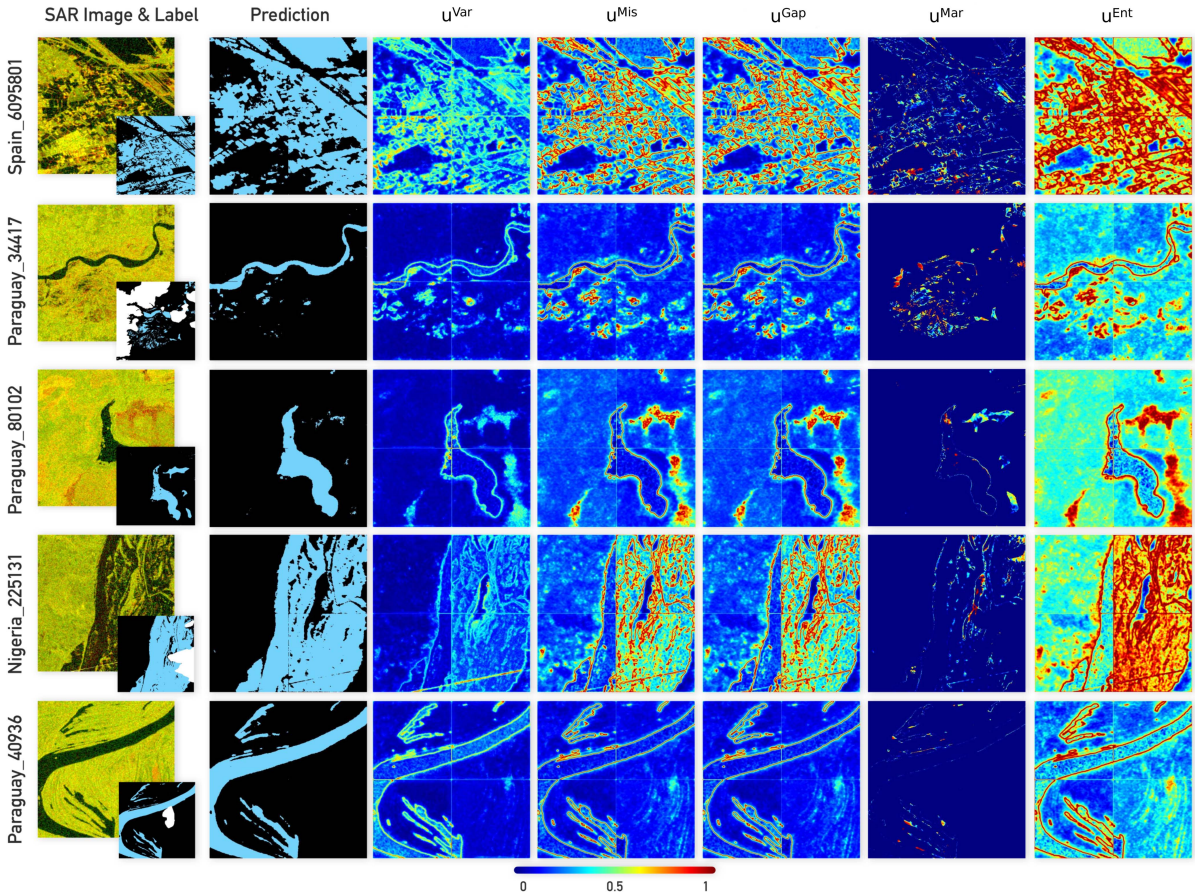


Fig. 6. Example SAR images, reference labels, predictions, and their uncertainty heatmaps for various uncertainty metrics and the student ensemble method.

be large in the second case. Thus, the combined use of both metrics is to some extent able to illustrate the contributions of different uncertainty sources.

Fig. 7 shows example results of the various uncertainty estimation methods. As it appears to be most informative, we leverage u^{Var} apart from the simple softmax baseline (for which it cannot be applied and we use u^{Mis} instead). While all methods are able to observe uncertainty at semantic boundaries, the student ensemble shows the most distinct lines. Patch borders are visible within all uncertainty maps. The student ensemble and the MC dropout method highlight the artifact in the fourth row SAR image the most. Areas with high VV response (shown as red in the pseudo-RGB SAR images) that have a tendency to be misclassified as water are highlighted by all methods. All methods have large regions of high uncertainty within the flood pixels. In particular, TTDA that estimates aleatoric uncertainty shows the highest uncertainty there. This is in-line with the intuition that errors in the training dataset tend to lead to increased data uncertainty.

In summary, all presented UQ methods and metrics capture the spatial distribution of uncertainty. While different in the estimation of the spatial variation of the uncertainty, most methods are consistent for several types of underlying image content, e.g., boundaries of class segments. However, there are also significant differences among the methods as for example the interior of class regions. In general, the student ensemble in combination with the standard deviation appears to return the most distinct uncertainty maps.

C. Quantitative Results

A quantitative evaluation of uncertainty estimation is a challenging task since there exist no reference uncertainty values. A first approach are calibration curves: a machine learning model is well calibrated if the distribution of the model's decision probabilities does correspond to the fraction of actual correct predictions. This relationship leads to a *calibration curve* (or reliability diagram [84], [85]) where a model with perfect calibration would show a straight, diagonal line: a model estimate of the class posterior with 0% (100%) for the positive class should never (always) happen for samples of this class. On the other hand, if the class posterior is 50%, the model should be correct half of the time.

In Fig. 8(a), all uncertainty estimation methods show a relatively high calibration. This is surprising in particular for using the softmax of the model as it goes against the common assumption that deep learning models are generally badly calibrated. The student ensemble is slightly better calibrated especially for probability values smaller than 0.6, but only by a small margin.

The high calibration when directly using the softmax probability is rather uncommon. To test whether the student ensemble

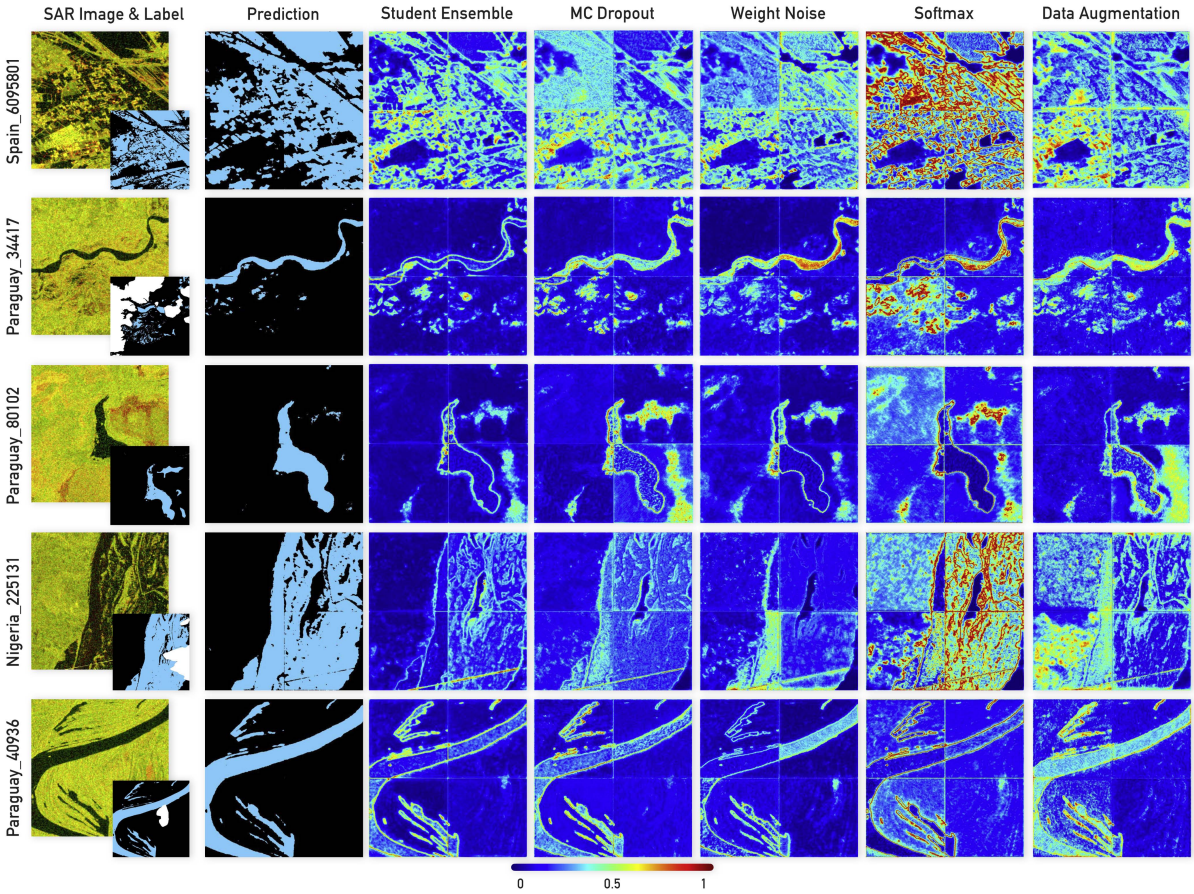


Fig. 7. Example SAR images, reference labels, predictions and their uncertainty heatmaps for various methods. The uncertainty metric is u^{Var} for all methods apart from the softmax baseline where it is u^{Mis}.

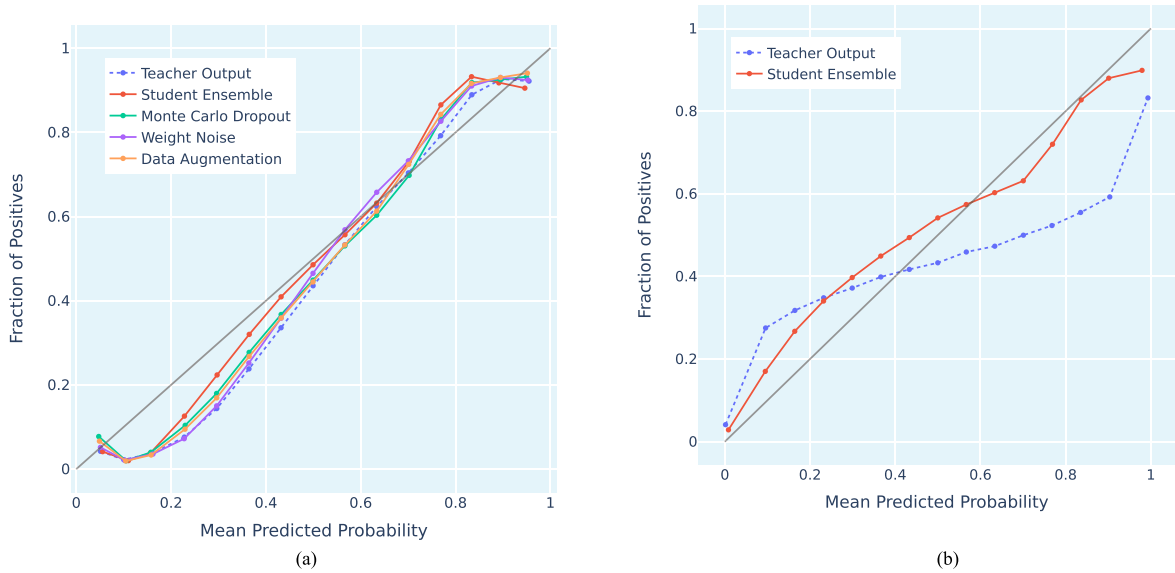


Fig. 8. Calibration curves illustrating the relation between estimated class probability and actual occurrence of the class for samples with this probability. The gray diagonals represent perfect model calibration. Other traces correspond to the uncertainty estimators. (a) Softmax not temperature scaled, i.e., $T = 1.0$. The student ensemble is trained on the logits of the teacher model. (b) Softmax of the teacher model is temperature scaled with $T = 0.2$. The student ensemble is trained on the temperature scaled output values of the teacher.

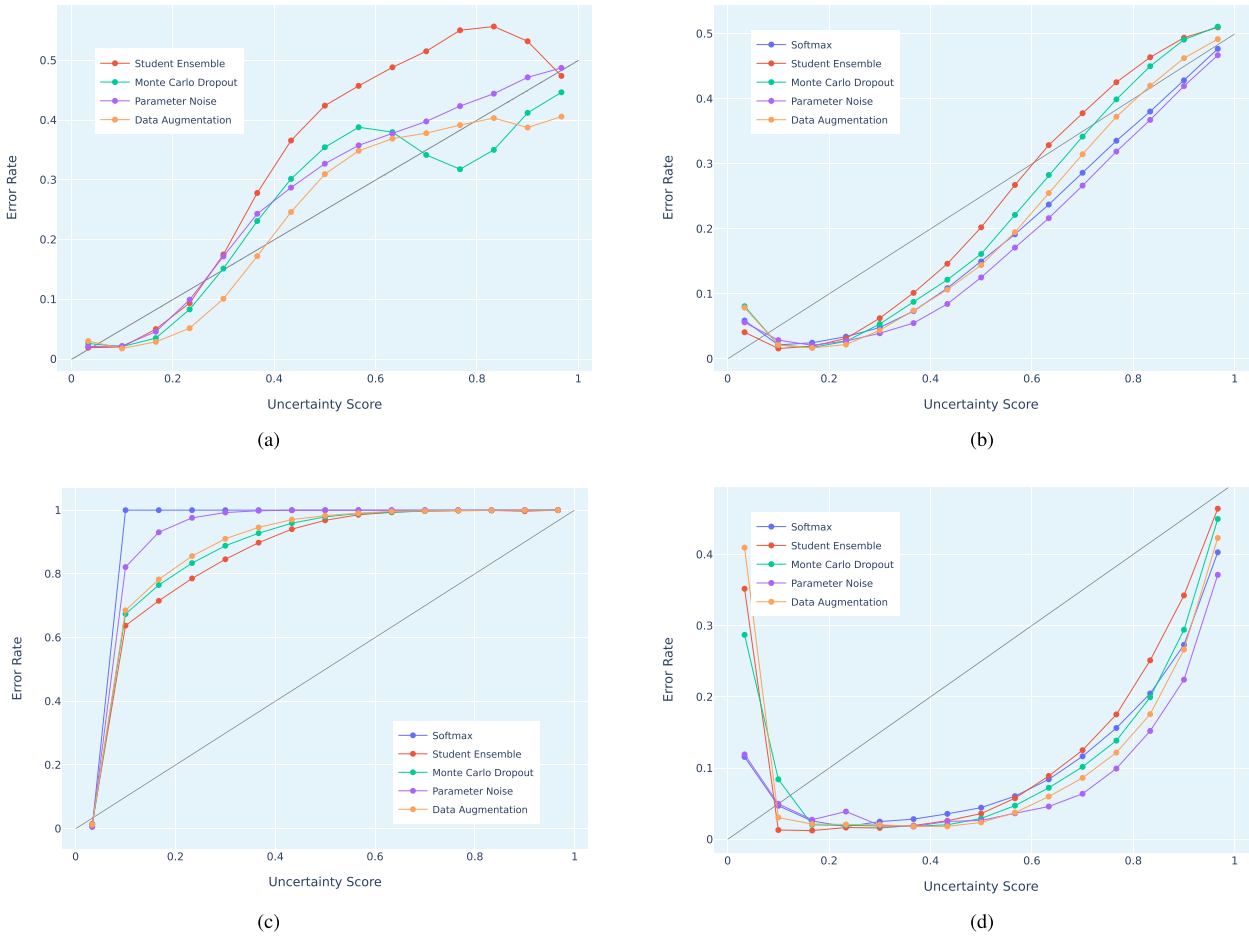


Fig. 9. Uncertainty versus error rate plots for different uncertainty estimation metrics. The curves for u^{Gap} are identical to the ones of u^{Mis} and thus not shown. The gray diagonal refers to the optimal case when the uncertainty score is directly proportional to the average error. (a) u^{Var} (Note: Not applicable to simple softmax). (b) u^{Mis}. (c) u^{Mar}. (d) u^{Ent}.

will lead to an improvement in the case of a less well calibrated model, we artificially decreased the teachers calibration by changing the temperature scaling in the softmax function and training the students on the softmax output instead of the logits. The resulting calibration curves are shown in Fig. 8(b) and indicate that the student ensemble uncertainty estimation method is able to achieve a better calibration.

As a second approach to quantify the performance of uncertainty estimation, we propose to relate uncertainty estimates directly with error rates. Intuitively, larger error rates should occur more frequently with larger uncertainty scores, i.e., if the model is very certain, it should be correct with higher probability, and if it is uncertain, more errors should exist on average. The value range of the uncertainty is divided into linearly spaced bins. For each bin, the average error rate of the base model is computed over samples with this uncertainty value. In comparison to calibration curves where the "raw" decision probability values of the uncertainty estimation methods are assessed, we measure here the quality of the processed, final uncertainty scores. The optimum is again a straight diagonal line: If the model is completely certain in its decision ($u = 0$), it should never be wrong while if it is completely uncertain ($u = 1$), any class assignment should be more or less random leading to the maximum error rate (e.g., 0.5 for a binary classification problem).

Fig. 9 shows the corresponding uncertainty-error curves for different uncertainty metrics. For the u^{Var} [i.e., Fig. 9(a)] the curve of the student ensemble is the most distant from the diagonal showing the worst performance despite the overall good qualitative performance discussed above. For this metric, a high uncertainty value refers to a large variance across model predictions. A possible reason for the discrepancy between the quantitative and qualitative results of the student ensemble with u^{Var} is that the variance of the probabilistic predictions in the ensemble can be small (i.e., in the worst case every member estimates a uniform class distribution), yet still lead to wrong class estimates (i.e., high error rates). This differs from the other approaches, where variations in the model (MC dropout, parameter noise) or data (data augmentation) create predictions that assign more or less probability mass to one of the possible classes, i.e., create outputs with higher variance that shift the erroneous samples to the right side of the plot.

Using the margin metric u^{Mar} [see Fig. 9(c)] does not lead to very meaning full error curves: As soon as the probability of the

true class is smaller than the maximum probability, errors are unavoidable.[2] The entropy metric u^{Ent} [see Fig. 9(d)] shows a narrow range of low error rates over a wide range of uncertainty values and a wide range of error rates for a narrow range of high uncertainty values. Thus, it overpredicts uncertainty since even for moderate to slightly high uncertainty values, the actual error rate is low. The best behavior shows u^{Mis} where all curves lie closest to the diagonal with a small advantage for the student ensemble.

V. Conclusion

Machine learning and in particular deep learning has long shown its advantages across multiple domains and applications. Yet, corresponding approaches are usually treated as black-box models that given a certain input provide a prediction but without much further insights in, e.g., why this particular prediction is made or how reliable it is.

UQ aims to provide an answer for the second point. While there are models that offer an inherent uncertainty estimate together with their prediction (e.g., BNNs), post hoc approaches that can be applied to an already trained model allow to use the best model for a given task, yet still obtain an uncertainty estimate of its prediction.

For the use case of flood detection from SAR images, we implement and evaluate several post hoc approaches. Results show that nearly all methods produce similar results in terms of both qualitative and quantitative measures. This includes directly using the softmax of the model as basis to quantify uncertainty, which is in contrast to common assumptions of deep learning models and warrants further analysis. An exception is using metrics based on the decision margin and entropy of the posterior. The former cannot be applied during deployment and is more related to accuracy than uncertainty. The latter does not provide sufficiently distinct uncertainty values, i.e., the same error rate range is spread over a wide range of uncertainty values. At least qualitatively, the student ensemble provides the best results if only by a small margin. In particular in combination with the variation or misclassification metric, it produces uncertainty maps of fine granularity, highlights regions where high uncertainty is expected (false classifications, boundaries of patches and semantic segments, image artifacts, etc.), and is otherwise homogeneous.

While differences to other methods are small, student ensembles offer several advantages. For example, it is a true post hoc method, i.e., it can be applied to any kind of model and makes no assumptions about the architecture or the data. This is in contrast to for example MC dropout and TTDA: The former can only be applied as a post hoc approach if the original model already includes dropout layers. The latter leverages data transformations, which require domain knowledge since not all transformations are reasonable or physically plausible (e.g., image rotation is a questionable transformation for SAR data, which comes with an inherent orientation due to the image acquisition geometry). A disadvantage of student ensembles is that they need to be trained before usage with all the limitations of general neural network training (e.g., appropriate tuning of hyperparameters, etc.). That also means that there needs to be a training dataset for the student ensembles. While this dataset does not need to include reference data of the target variable (since the students are trained on the logits of the teacher) and could be the training dataset of the teacher, it is a requirement that other methods do not have. Since all evaluated approaches are sampling based, the multiple inference runs of student ensembles are a disadvantage shared with all methods. Thus, run time during inference is mainly proportional to the number of runs but otherwise very similar among the approaches. TTDA has a small additional overhead due to applying image transformations that are not required by the other methods. The student ensemble offers some flexibility as the student model can be simpler or more complex than the teacher model, requiring potentially less or more time during inference. Our recommendation, however, is to use the same architecture as the teacher model to ensure that the students can adequately mimic its behavior.

Existing real-world applications leveraging deep neural network-based segmentation for flood monitoring greatly benefit from post hoc uncertainty estimation methods. By incorporating a final uncertainty estimation step in a plug-and-play fashion within the flood prediction pipeline, system users gain additional insights into which floodplains should be prioritized with higher confidence. This enhances decision-making by helping authorities allocate resources efficiently while reducing the risk of misclassification.

UQ improves model reliability by indicating when and where predictions can be trusted, supporting adaptive data collection to refine model performance, especially in regions with domain shifts or varying environmental conditions. In addition, uncertainty estimates can be integrated with model predictions to enhance overall accuracy. This approach not only aids in risk management and policy planning by identifying areas prone to false positives or negatives but also ensures scientific transparency and compliance with regulatory standards [86]. By embedding uncertainty into flood detection models, decision-makers can act with greater confidence, improving disaster response efficiency and system robustness.

Future work should reevaluate whether the assumption that deep neural networks are in general badly calibrated still holds for modern approaches. Also, the relationship between the variation and misclassification metrics should be further analyzed as it has the potential to disentangle different aspects of the predictive uncertainty. Last but not least, there are a variety of options to further improve student ensembles, e.g., optimizing the students with additional uncertainty losses [19] or sacrificing some estimation performance for a better runtime (e.g., via BatchEnsembles [42], [43], [44], [45]).

In summary, this article empirically shows that different UQ approaches behave very similarly in the use case of flood detection from SAR images and differences are much stronger for different metrics. In general, however, all approaches lead to rather consistent results indicating the maturity of the field. Users should prioritize methods based on their individual advantages:

[2]Note, that the error is computed for a single model instance while the uncertainty is computed via sampling and thus based on the posterior estimate averaged over all models.

TTDA for requiring aleatoric uncertainty, softmax for quick evaluation, student ensembles for a slight lead in epistemic estimations.

REFERENCES

[1] Y. Ban, P. Zhang, A. Nascetti, A. R. Bevington, and M. A. Wulder, "Near real-time wildfire progression monitoring with Sentinel-1 SAR time series and deep learning," *Sci. Rep.*, vol. 10, no. 1, 2020, Art. no. 1322.

[2] D. Rashkovetsky, F. Mauracher, M. Langer, and M. Schmitt, "Wildfire detection from multisensor satellite imagery using deep semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7001–7016, 2021.

[3] P. Zhang, Y. Ban, and A. Nascetti, "Learning U-net without forgetting for near real-time wildfire monitoring by the fusion of SAR and optical time series," *Remote Sens. Environ.*, vol. 261, 2021, Art. no. 112467.

[4] S. Karimzadeh, M. Ghasemi, M. Matsuoka, K. Yagi, and A. C. Zulfikar, "A deep learning model for road damage detection after an earthquake based on synthetic aperture radar (SAR) and field datasets," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5753–5765, 2022.

[5] E. Nemni, J. Bullock, S. Belabbes, and L. Bromley, "Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery," *Remote Sens.*, vol. 12, no. 16, 2020, Art. no. 2532.

[6] S. Paul and S. Ganju, "Flood segmentation on sentinel-1 SAR imagery with semi-supervised learning," 2021, *arXiv:2107.08369*.

[7] M. Helleis, M. Wieland, C. Krullikowski, S. Martinis, and S. Plank, "Sentinel-1-based water and flood mapping: Benchmarking convolutional neural networks against an operational rule-based processing chain," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2023–2036, 2022.

[8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[9] A. Höhl et al., "Opening the black-box: A systematic review on explainable ai in remote sensing," 2024, *arXiv:2402.13791*.

[10] G. Taskin, E. Aptoula, and A. Ertürk, "Chapter 7 - explainable AI for Earth observation: Current methods, open challenges, and opportunities," in *Advances in Machine Learning and Image Analysis for GeoAI*, S. Prasad, J. Chanussot, and J. Li, Eds. Amsterdam, The Netherlands: Elsevier, 2024, pp. 115–152. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443190773000122

[11] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, pp. 1513–1589, 2023.

[12] R. M. Neal, *Bayesian Learning for Neural Networks*. New York, NY, USA: Springer, 2012.

[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6405–6416.

[14] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[15] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[17] J. Mukhoti, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deep deterministic uncertainty for semantic segmentation," 2021, *arXiv:2111.00079*.

[18] K. Hoebel et al., "An exploration of uncertainty information for segmentation quality assessment," in *Medical Imaging 2020: Image Processing*, vol. 11313. Bellingham, WA, USA: SPIE, 2020, pp. 381–390.

[19] C. J. Holder and M. Shafique, "Efficient uncertainty estimation in semantic segmentation via distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3087–3094.

[20] A. Hein et al., "A comparison of uncertainty quantification methods for active learning in image classification," in *Proc. Int. Joint Conf. Neural Netw.*, 2022, pp. 1–8.

[21] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9690–9700.

[22] J. Postels et al., "On the practicality of deterministic epistemic uncertainty," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 17870–17909.

[23] Y. Liu, S. Yan, L. Leal-Taixé, J. Hays, and D. Ramanan, "Soft augmentation for image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16241–16250.

[24] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deep deterministic uncertainty: A simple baseline," 2021, *arXiv:2102.11582*.

[25] M. Ng et al., "Estimating uncertainty in neural networks for cardiac MRI segmentation: A benchmark study," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 6, pp. 1955–1966, Jun. 2023.

[26] F. L. Jacobsen, S. A. Hicks, P. Halvorsen, and M. A. Riegler, "Estimating predictive uncertainty in gastrointestinal polyp segmentation," in *Proc. IEEE 35th Int. Symp. Comput.-Based Med. Syst.*, 2022, pp. 44–49.

[27] A. Sagar, "Uncertainty quantification using variational inference for biomedical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 44–51.

[28] H. P. Do, Y. Guo, A. J. Yoon, and K. S. Nayak, "Accuracy, uncertainty, and adaptability of automatic myocardial ASL segmentation using deep CNN," *Magn. Reson. Med.*, vol. 83, no. 5, pp. 1863–1874, 2020.

[29] Q. Lyu, C. T. Whitlow, and G. Wang, "Softdropconnect (SDC)–effective and efficient quantification of the network uncertainty in deep mr image analysis," 2022, *arXiv:2201.08418*.

[30] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *Proc. Brit. Mach. Vis. Conf.*, 2017.

[31] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proc. 28th Int. Conf. Int. Conf. Mach. Learn.*, Madison, WI, USA 2011, pp. 681–688.

[32] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient hamiltonian Monte Carlo," in *Proc. 31st Int. Conf. Mach. Learn.*, Bejing, China, Jun. 2014, pp. 1683–1691. [Online]. Available: https://proceedings.mlr.press/v32/cheni14.html

[33] A. Graves, "Practical variational inference for neural networks," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2011.

[34] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[35] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] N. Hochgeschwender et al., "Evaluating uncertainty estimation methods on 3D semantic segmentation of point clouds," 2020, *arXiv:2007.01787*.

[38] F. Fooladgar et al., "Uncertainty-aware deep ensemble model for targeted ultrasound-guided prostate biopsy," in *Proc. IEEE 19th Int. Symp. Biomed. Imag.*, 2022, pp. 1–5.

[39] Z. Yang et al., "Uncertainty-aware perception models for off-road autonomous unmanned ground vehicles," 2022, *arXiv:2209.11115*.

[40] J. Adams and S. Y. Elhabian, "Benchmarking scalable epistemic uncertainty quantification in organ segmentation," 2023, *arXiv:2308.07506*.

[41] A. Baumann, T. Roßberg, and M. Schmitt, "Probabilistic MIMO U-net: Efficient and accurate uncertainty estimation for pixel-wise regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4498–4506.

[42] O. Laurent et al., "Packed-ensembles for efficient uncertainty estimation," 2022, *arXiv:2210.09184*.

[43] M. Havasi et al., "Training independent subnetworks for robust prediction," 2020, *arXiv:2010.06610*.

[44] Y. Wen, D. Tran, and J. Ba, "Batchensemble: An alternative approach to efficient ensemble and lifelong learning," in *Proc. Int. Conf. Learn. Representations*, 2019.

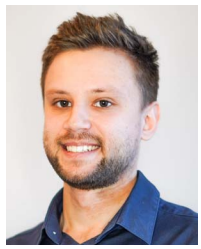
[45] N. Durasov, T. Bagautdinov, P. Baque, and P. Fua, "Masksembles for uncertainty estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13539–13548.

[46] S. Wangiyana, P. Samczyński, and A. Gromek, "Data augmentation for building footprint segmentation in SAR images: An empirical study," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2012.

[47] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation," in *Proc. 4th Int. Workshop Brainlesion: Glioma Mult. Scler. Stroke Traumatic Brain Injuries*, Granada, Spain, 2019, pp. 61–72.

[48] W. He, Z. Jiang, T. Xiao, Z. Xu, and Y. Li, "A survey on uncertainty quantification methods for deep learning," 2025, *arXiv:2302.13425*.

[49] M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243–297, 2021.
[50] S. Roy and S. Saha, "Chapter 11 - uncertainty quantification in deep neural networks for multi-sensor earth observation," in *Deep Learning for Multi-Sensor Earth Observation*, S. Saha, Ed Amsterdam, Netherland: Elsevier, 2025, pp. 231–247. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978044326484900021X
[51] J. Postels, F. Ferroni, H. Coskun, N. Navab, and F. Tombari, "Sampling-free epistemic uncertainty estimation using approximated variance propagation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2931–2940.
[52] J. Gast and S. Roth, "Lightweight probabilistic deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3369–3378.
[53] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get m for free," 2017, *arXiv:1704.00109*.
[54] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 1789–1819, 2021.
[55] A. Malinin, B. Mlodozeniec, and M. Gales, "Ensemble distribution distillation," 2019, *arXiv:1905.00076*.
[56] E. Englesson and H. Azizpour, "Efficient evaluation-time uncertainty estimation by improved distillation," 2019, *arXiv:1906.05419*.
[57] M. Ferianc and M. Rodrigues, "Simple regularisation for uncertainty-aware knowledge distillation," 2022, *arXiv:2205.09526*.
[58] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608005800231
[59] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, pp. 49–64, 1996. [Online]. Available: https://api.semanticscholar.org/CorpusID
[60] D. Kaplan, "On the quantification of model uncertainty: A Bayesian perspective," *Psychometrika*, vol. 86, pp. 215–238, 2021.
[61] K. Abdelfatah, J. Bao, and G. Terejanu, "Geospatial uncertainty modeling using stacked Gaussian processes," *Environ. Modelling Softw.*, vol. 109, pp. 293–305, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815218304997
[62] R. Hänsch, "Stacked random forests: More accurate and better calibrated," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1751–1754.
[63] T. Ramalho and M. Miranda, "Density estimation in representation space to predict model uncertainty," in *Proc. 3rd Int. Workshop Eng. Dependable Secure Mach. Learn. Syst.*, New York City, NY, USA, 2020, pp. 84–96.
[64] P. Lahoti, K. Gummadi, and G. Weikum, "Responsible model deployment via model-agnostic uncertainty learning," *Mach. Learn.*, vol. 112, no. 3, pp. 939–970, 2023.
[65] M. Shen, Y. Bu, P. Sattigeri, S. Ghosh, S. Das, and G. Wornell, "Post-hoc uncertainty learning using a dirichlet meta-model," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 9772–9781.
[66] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 3868–3878, Dec. 2020.
[67] A. Mobiny, P. Yuan, S. K. Moulik, N. Garg, C. C. Wu, and H. Van Nguyen, "Dropconnect is effective in modeling uncertainty of bayesian deep networks," *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. 5458.
[68] M. Monteiro et al., "Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12756–12767.
[69] B. Mucsányi, M. Kirchhof, and S. J. Oh, "Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks," in *Advances in Neural Information Processing Systems*, A. Globerson et al., Eds. Red Hook, NY, USA: Curran Associates, Inc., 2024, pp. 50972–51038. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/5afa9cb1e917b898ad418216dc726fbd-Paper-Datasets_and_Benchmarks_Track.pdf
[70] D. Dera et al., "Bayes-SAR Net: Robust SAR image classification with uncertainty estimation using Bayesian convolutional neural network," in *Proc. IEEE Int. Radar Conf.*, 2020, pp. 362–367.
[71] X. Chen, K. A. Scott, and D. A. Clausi, "Uncertainty analysis of sea ice and open water classification on SAR imagery using a Bayesian CNN," in *Proc. Int. Conf. Pattern Recognit.*, 2022, pp. 343–356.
[72] X. Chen, K. A. Scott, L. Xu, M. Jiang, Y. Fang, and D. A. Clausi, "Uncertainty-incorporated ice and open water detection on dual-polarized SAR sea ice imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5201213.
[73] J. Haas and B. Rabus, "Uncertainty estimation for deep learning-based segmentation of roads in synthetic aperture radar imagery," *Remote Sens.*, vol. 13, no. 8, 2021, Art. no. 1472. [Online]. Available: https://www.mdpi.com/2072-4292/13/8/1472
[74] Z. Huang, Y. Liu, X. Yao, J. Ren, and J. Han, "Uncertainty exploration: Toward explainable SAR target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4202314.
[75] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.
[76] GitHub, "Sue - segmentation uncertainty estimation," 2020. [Online]. Available: https://github.com/RedekopEP/SUE
[77] R. Hänsch, "Chapter 10 - the power of voting: Ensemble learning in remote sensing," in *Advances in Machine Learning and Image Analysis for GeoAI*, S. Prasad, J. Chanussot, and J. Li, Eds. Amsterdam, The Netherland: Elsevier, 2024, pp. 201–235. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443190773000158
[78] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1607–1616.
[79] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.
[80] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[81] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg, "Sen1FLoods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 210–211.
[82] GitHub, "Plotneuralnet," 2020. [Online]. Available: https://github.com/HarisIqbal88/PlotNeuralNet
[83] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Munich, Germany, 2015, pp. 234–241.
[84] SciKit-Learn, "Probability calibration," 2023. [Online]. Available: https://scikit-learn.org/stable/modules/calibration.html
[85] D. S. Wilks, "On the combination of forecast probabilities for consecutive precipitation periods," *Weather Forecasting*, vol. 5, no. 4, pp. 640–650, 1990.
[86] P. Abbaszadeh, D. Muñoz, H. Moftakhari Rostamkhani, K. Jafarzadegan, and M. Hamid, "Perspective on uncertainty quantification and reduction in compound flood modeling and forecasting," *iScience*, vol. 25 2022, Art. no. 105201.



Jakob Ludwig received the B.Sc. and M.Sc. degree in computer science from the Technische Universität Berlin, Berlin, Germany, in 2020 and 2023. He is currently working toward the Ph.D. degree with the Department of Radiology, Charité – Universitätsmedizin Berlin, Berlin.

The research for this paper was conducted during his time at the SAR Technology department of the German Aerospace Center (DLR). His research interests include deep learning and uncertainty estimation, especially for medical imaging and remote sensing.

Ronny Hänsch (Senior Member, IEEE) received the diploma in computer science and the Ph.D. degree from the TU Berlin, Berlin, Germany, in 2007 and 2014, respectively.

He is a Scientist with the Microwave and Radar Institute, German Aerospace Center (DLR), Oberpfaffenhofen, where he leads the Machine Learning Team with the Signal Processing Group, SAR Technology Department.

Dr. Hänsch served as Chair of the GRSS Image Analysis and Data Fusion Technical Committee and Editor of the GRSS eNewsletter. Currently, he is a Co-Chair with the ISPRS working group on Image Orientation and Sensor Fusion, Editor in Chief of the *Geoscience and Remote Sensing Letters*, Associate Editor of the *ISPRS Journal of Photogrammetry and Remote Sensing*, organizer of the CVPR Workshop EarthVision (2017–2025) and the IGARSS Tutorial on Machine Learning in Remote Sensing (2017–2024).