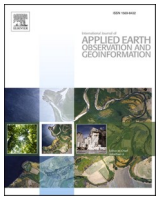
International Journal of Applied Earth Observation and Geoinformation 143 (2025) 104759

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



BayeSiamMTL: Uncertainty-aware multitask learning for post-disaster building damage assessment

Victor Hertel [a,*], Omar Wani [b], Christian Geiß [a,c], Marc Wieland [a], Hannes Taubenböck [a,d]

[a] German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Germany
[b] Tandon School of Engineering, New York University (NYU), USA
[c] Department of Geography, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
[d] Earth Observation Research Cluster (EORC), Julius-Maximilians-Universität Würzburg, Germany

ARTICLE INFO

Keywords:
Building damage assessment
Uncertainty quantification
Bayesian deep neural network
Semantic segmentation
Rapid disaster response
Crisis information management

ABSTRACT

Accurate and timely building damage assessment (BDA) is critical for effective disaster response and recovery. However, existing machine learning approaches in this context do mostly not account for uncertainties, which are essential for ensuring trustworthy and transparent results. This study introduces a hybrid Bayesian deep learning framework with integrated uncertainty quantification to enhance BDA, thereby making model predictions more reliable and interpretable. We propose BayeSiamMTL, a novel Bayesian Siamese multitask learning architecture that combines deterministic segmentation of building footprints with probabilistic change detection for damage level classification. By encoding model parameters as probability distributions and utilizing variational inference with Monte Carlo approximation, BayeSiamMTL produces pixelwise posterior predictive distributions, providing detailed insights into both damage predictions and their associated uncertainties. Our analysis explores key aspects of Bayesian modeling and, to our knowledge, is the first to provide quantified insights into the model's classification dynamics, revealing internal decision-making tendencies and sources of uncertainty. Additionally, we introduce confidence-informed damage maps in the form of stratified probabilities of damage clusters and minimum/maximum damage extents delineated from confidence intervals. Model performance is evaluated across multiple datasets to assess the impact of domain shifts and out-of-distribution samples. Experimental results show that BayeSiamMTL not only achieves a performance advantage over its deterministic counterpart but also exhibits significantly better generalization capabilities under domain shifts with a relative performance improvement of 42 %. While background pixels represent the primary source of confusion across all damage levels, our findings indicate that building destructions are more frequently confused with intact buildings rather than among varying degrees of damage.

1. Introduction

In recent decades, machine learning (ML) techniques have become indispensable across various scientific disciplines and practical applications. These technological advances have also transformed the process of post-disaster building damage assessment (BDA) from air- and spaceborne remote sensing data. The extent of building damage is a critical indicator of disaster impact, offering important insights into the affected population and the economic damage (Geiß et al., 2023). Consequently, timely and accurate assessments are essential for effective disaster response and humanitarian assistance. While in-situ surveys remain fundamental, the acquisition of supplementary information has shifted from manual inspection of aerial imagery to sophisticated, data-driven ML approaches (Deng and Wang, 2022; Ge et al., 2023). In particular, deep learning has emerged as a powerful tool for automating the detection and quantification of building damage, enabling rapid assessments that generally surpass the accuracy and efficiency of traditional methods (Zheng et al., 2021; Ge et al., 2020).

As deep neural networks (DNNs) become more prevalent, ensuring confidence in their predictions becomes crucial, especially in safety–critical applications such as rapid disaster response and crisis information management. In these contexts, map-based information derived

* Corresponding author at: German Aerospace Center (DLR), Münchener Straße 20, 82234 Weßling, Germany.
E-mail address: victor.hertel@dlr.de (V. Hertel).

https://doi.org/10.1016/j.jag.2025.104759
Received 4 April 2025; Received in revised form 16 July 2025; Accepted 27 July 2025
Available online 16 August 2025
1569-8432/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

from ML methods plays a key role in guiding urgent decisions where incorrect or misinformed actions can have severe consequences. This applies specifically to the context of BDA, since the extent of damage to buildings stands out as a key indicator of disaster impact, directly influencing resource allocation, relief strategies, and guidance of rescue teams toward the affected population. Misestimation or inadequate handling of uncertainty in damage assessments can result in the misallocation of limited resources, delayed interventions, and increased fatalities. Therefore, trustworthy representations of uncertainty should be considered a key feature of any ML-based BDA algorithm (Hüllermeier and Waegeman, 2021). However, conventional deterministic DNNs typically encode model parameters as single values and thus fail to provide reliable measures of uncertainty, leading to either overconfidence or underconfidence in their predictions (Gawlikowski et al., 2023). To overcome this constraint, it is vital to equip DNNs with mechanisms that quantify uncertainty, ensuring that predictions flagged as highly uncertain receive further scrutiny or are deferred to human experts (Gal and Ghahramani, 2016).

Predictive uncertainty generally arises from two main sources: epistemic (systematic) uncertainty and aleatoric (statistical) uncertainty. Epistemic uncertainty stems from limited knowledge and can be reduced by improving model design or training data. In contrast, aleatoric uncertainty arises from inherent variability in the data and cannot be eliminated (Kendall and Gal, 2017; Kiureghian and Ditlevsen, 2009). Numerous uncertainty quantification (UQ) techniques have been developed to enable DNNs to assess the reliability of their outputs. These techniques can be grouped according to whether they rely on single or multiple networks, and whether those networks are deterministic or probabilistic in nature. Ensemble methods, for instance, involve multiple deterministic networks and combine their predictions to leverage their collective diversity for uncertainty estimation (Lakshminarayanan et al., 2017). Test-time augmentation techniques, on the other hand, rely on a single deterministic model and apply various augmentations to the input data during inference in order to produce multiple predictions (Shorten and Khoshgoftaar, 2019). Another category includes single deterministic networks that make predictions via one single forward pass. Here, uncertainty is either quantified by external methods or directly predicted by the network (Malinin and Gales, 2019, 2018). Finally, Bayesian methods employ probabilistic DNNs, where conditional probability distributions over model parameters yield slightly different results across multiple forward passes. This integrated, scalable, and resource-efficient approach to capturing predictive uncertainty makes Bayesian methods especially well-suited for applications in the remote sensing domain (Gal and Ghahramani, 2016; Mobiny et al., 2021).

Within the Bayesian framework, all inference about unknown quantities involves computing posterior distributions (Blei et al., 2017). This process typically begins with assuming a prior distribution over model parameters and subsequently applies Bayes' theorem to estimate their posterior distributions (Jospin et al., 2022). Since computing the exact posterior is intractable, variational inference techniques approximate it by optimizing over a set of more tractable distributions (Gawlikowski et al., 2023). For instance, Blundell et al. (2015) introduced Bayes by Backprop, a backpropagation-compatible algorithm that learns a probability distribution over DNN weights. Nevertheless, defining meaningful weight priors remains challenging, especially for deep architectures with high-dimensional weight spaces. To address this challenge, Krishnan et al. (2020) proposed MOPED (model priors with empirical Bayes using DNN) to determine more informed weight priors for Bayesian DNNs. Another widely used technique for approximating the posterior distribution is Monte Carlo (MC) dropout. Dropout randomly deactivates certain model neurons during training to improve generalization and reduce co-tuning (Abdar et al., 2021). When applied both during training and inference, dropout acts as an approximate Bayesian variational inference method for deep Gaussian processes (Gal and Ghahramani, 2016; Kingma et al., 2015). However, Hertel et al. (2023) found that MC dropout can yield overconfident prediction intervals, whereas Bayesian variational inference is generally more flexible in learning both the mean and the spread of the parameter posterior.

In disaster response and humanitarian assistance, decisions often must be made under considerable uncertainty. In these high-stakes settings, reliable uncertainty estimation is essential for guiding rapid assessments that inform intervention strategies responsibly. Current ML-based BDA approaches generally fall into two categories: cascade-based and multitask network architectures. In cascade-based strategies, building localization is performed on pre-disaster imagery, and the resulting footprints are subsequently used to support damage assessment in post-disaster scenes. However, treating these tasks independently can introduce knowledge gaps by overlooking interdependencies (Zheng et al., 2021). To address these limitations, recent studies (Gholami et al., 2022; Hao et al., 2021) have proposed Siamese networks that perform bi-temporal building localization alongside integrated damage classification. Despite this progress, existing methods remain almost exclusively deterministic and lack comprehensive uncertainty quantification. Meanwhile, Bayesian approaches to uncertainty quantification have been successfully applied in various fields, including remote sensing (Dera et al., 2020; Hertel et al., 2023; Lee and Li, 2024; Zhang and Diao, 2023) and medical applications (Herzog et al., 2020; Thiagarajan et al., 2022). Within the specific context of BDA, Bin et al. (2022) employed Monte Carlo dropout for uncertainty estimation, although their method primarily relied on variance as the sole measure of predictive uncertainty. There remains a gap in statistically sound quantification and advanced Bayesian evaluation of uncertainties within the context of BDA. Additionally, despite emergency response scenarios necessitating lightweight and computationally efficient solutions, many state-of-the-art BDA studies leverage complex and extensively parameterized models optimized specifically for particular datasets (Chen et al., 2024; Kaur et al., 2023; Yu et al., 2025). Consequently, the practical deployment of such complex and tailored models in emergency response contexts faces significant challenges related to generalization capabilities and computational demands (Hertel et al., 2025).

In this paper, we introduce BayeSiamMTL, a novel Bayesian Siamese multitask learning architecture that fuses deterministic binary semantic segmentation with probabilistic multiclass change detection. This integrated design enables the model to simultaneously identify bi-temporal building footprints and evaluate their corresponding damage levels with variational inference-based UQ. BayeSiamMTL is optimized for efficient operation on large datasets and facilitates effective and transparent building damage assessment in humanitarian disaster response. We explain the belonging Bayesian statistical framework to derive and interpret posterior predictive distributions (PPDs) on a pixel-level. Based on these distributions, we analyze key aspects of Bayesian modeling, including the number of MC samples required for PPD convergence and the influence of probabilistic parameter initialization. To our knowledge, this study is the first to provide quantified insights into the model's classification dynamics, revealing internal decision-making tendencies and sources of confusion based on approximately 235 billion pixel evaluations. Furthermore, we evaluate the model performance across multiple datasets to assess the impact of domain shifts and out-of-distribution (OOD) samples in the Bayesian framework. Finally, we propose confidence-informed damage maps that leverage predictive posteriors to produce integrated prediction-certainty visualizations. These maps provide disaggregated and stratified probabilities of distinct damage clusters as well as the minimum and maximum extents of building damage delineated from confidence intervals.

The structure of this paper is as follows. Section 2 describes the datasets and pre-processing steps. Section 3 elaborates on the proposed methods. Results and findings are presented in Section 4. Section 5 contains the discussion and Section 6 concludes the paper.

2. Data

Building damage assessment can be separated into two sequential tasks: building segmentation and damage level classification. In this study, multiple datasets specific to each task are used. By integrating data from diverse spatial resolutions, geographical contexts, and disaster events, we aim to provide a comprehensive basis for robust model training and uncertainty quantification.

2.1. Building segmentation

OpenEarthMap (OEM) is a comprehensive dataset designed for global high-resolution land cover mapping (Xia et al., 2023). It comprises 2.2 million annotated segments derived from 5,000 air- and spaceborne images, spanning 97 regions across 44 countries on 6 continents. The dataset provides manually annotated land cover labels across eight classes, with a ground sampling distance (GSD) ranging from 25 cm to 50 cm. We use the class 'building' to train the building segmentation module of BayeSiamMTL.

2.2. Building damage classification

Damage severity is typically assessed using ordinal classification scales, which rely on qualitative descriptors that must be precisely defined to differentiate highly heterogeneous damage patterns. This challenge has spurred the development of hazard-specific frameworks such as the European Macroseismic Scale 1998 (EMS-98) for earthquake damage. Yet, no universally recognized standard currently exists for remote sensing-based BDA (Cotrufo et al., 2018). In this context, Hertel et al. (2025) proposed a framework that harmonizes damage descriptors from both engineering and remote sensing domains, aligning them with internationally accepted standards. The classification scheme in Table 1 is adopted in the present study to ensure consistency and interoperability across different datasets.

2.2.1. xBD dataset

The xBD dataset is one of the largest publicly available resources for assessing building damage using satellite imagery (Gupta et al., 2019). It provides annotations for over 850,000 buildings across more than 45,000 km^2 of remote sensing imagery, at a GSD of 80 cm. Covering pre- and post-disaster imagery from 19 distinct events, xBD serves as the primary training source for the damage classification module of BayeSiamMTL. The official test split of the corresponding xView2 challenge consists of 10 initially published events (tier1). During the challenge, an additional nine events (tier3) were made available. In order to obtain representative results and assess generalization capabilities, we create a new dataset split which fills up the official xView2 test set such that 20 % of the data per event is used for testing. This data split can be found in the Supplementary materials. For comparison across different studies, results are also reported using the original xView2 test split. Following Hertel et al. (2025), the 'Minor damage' and 'Major damage' categories are combined into a single 'Damaged' category to maintain consistency with the above-mentioned classification scheme.

2.2.2. Ahr valley dataset

In 2021, Europe experienced catastrophic flood events that caused extensive damage to its built environment and significant harm to the population. This event is regarded as one of the most lethal European flood events in nearly three decades and among the costliest on record (Szymczak et al., 2022). In response, the German Aerospace Center (DLR) conducted an aerial survey of the severely impacted Ahr valley in Germany, capturing post-disaster RGB imagery at a spatial resolution of 7 cm GSD. Building on this data, Hertel et al. (2025) introduced a bi-temporal BDA dataset by integrating pre-disaster aerial imagery with a 20 cm GSD and annotating nearly 10,000 buildings. This dataset is used to assess the impact of domain shifts and out-of-distribution samples.

2.3. Pre-processing

All imagery is normalized using the mean and standard deviation per channel of the entire respective dataset. The data is then divided into non-overlapping tiles of 256 × 256 pixels. To increase data volume and minimize boundary artefacts, a second layer of identically sized tiles is generated, offset by 128 pixels in both the vertical and horizontal directions. This offset creates overlapping regions between adjacent tiles, thereby preserving contextual continuity at the tile boundaries and centering edge features within neighboring tiles. To address the reso-

Table 1
Engineering-based building damage classification criteria with remote sensing data (Hertel et al., 2025).

	Damage		Description
	Structural	Non-structural	
No visible damage	None	None	• Structure appears to have complete structural integrity • Walls remain standing • Roof is virtually undamaged
Damaged	None	Light	• Only wetting through • Dirt • Light cracking to loadbearing walls • Doors and windows pushed in • Washing out of foundations • Larger cracking in loadbearing walls and slabs • Settlement • Collapse of non-loadbearing walls
	Light	Moderate	
	Moderate	Heavy	
Destroyed	Heavy	Very heavy	• Collapse of loadbearing walls, slab • Collapse of larger parts of building • Dislocation: building completely washed away, toppled or displaced from foundation
	Very heavy	Very heavy	
Invalid	–	–	• Uncertain interpretation due to image quality (e.g. shadow or degraded resolution due to high off-nadir angle)

lution discrepancy between pre- and post-disaster imagery in the Ahr valley dataset, the post-disaster imagery is resampled to a GSD of 20 cm to match the pre-disaster imagery.

3. Methods

3.1. Model setup and training procedure

In this study, we present BayeSiamMTL, a novel Bayesian Siamese multitask learning architecture that integrates deterministic binary segmentation with probabilistic multiclass change detection. As illustrated in Fig. 1, BayeSiamMTL employs two deterministic, weight-shared UNets to segment building footprints from pre- and post-disaster images. The segmented patches are subsequently processed by a Bayesian Siamese difference-based decoder, which performs probabilistic damage level classification. By encoding the segmentation task deterministically, the model ensures that its probabilistic UQ capabilities are exclusively focused on the damage level classification task, eliminating any interference from building localization. This design aligns with emergency response priorities, where accurately identifying damage severity is more critical than precisely delineating building contours. Training is conducted in two sequential stages to enhance efficiency and optimize task-specific performance. In the first stage, the building segmentation modules are trained using the OEM dataset. In the second stage, the damage classification module is trained on the xBD dataset. Once the segmentation training is complete, the outputs of the Bayesian Siamese decoder are masked using the pre-disaster building footprints. This approach ensures that predictions are focused on existing structures, effectively eliminating changes unrelated to buildings. After the initial sequential training, BayeSiamMTL can be further

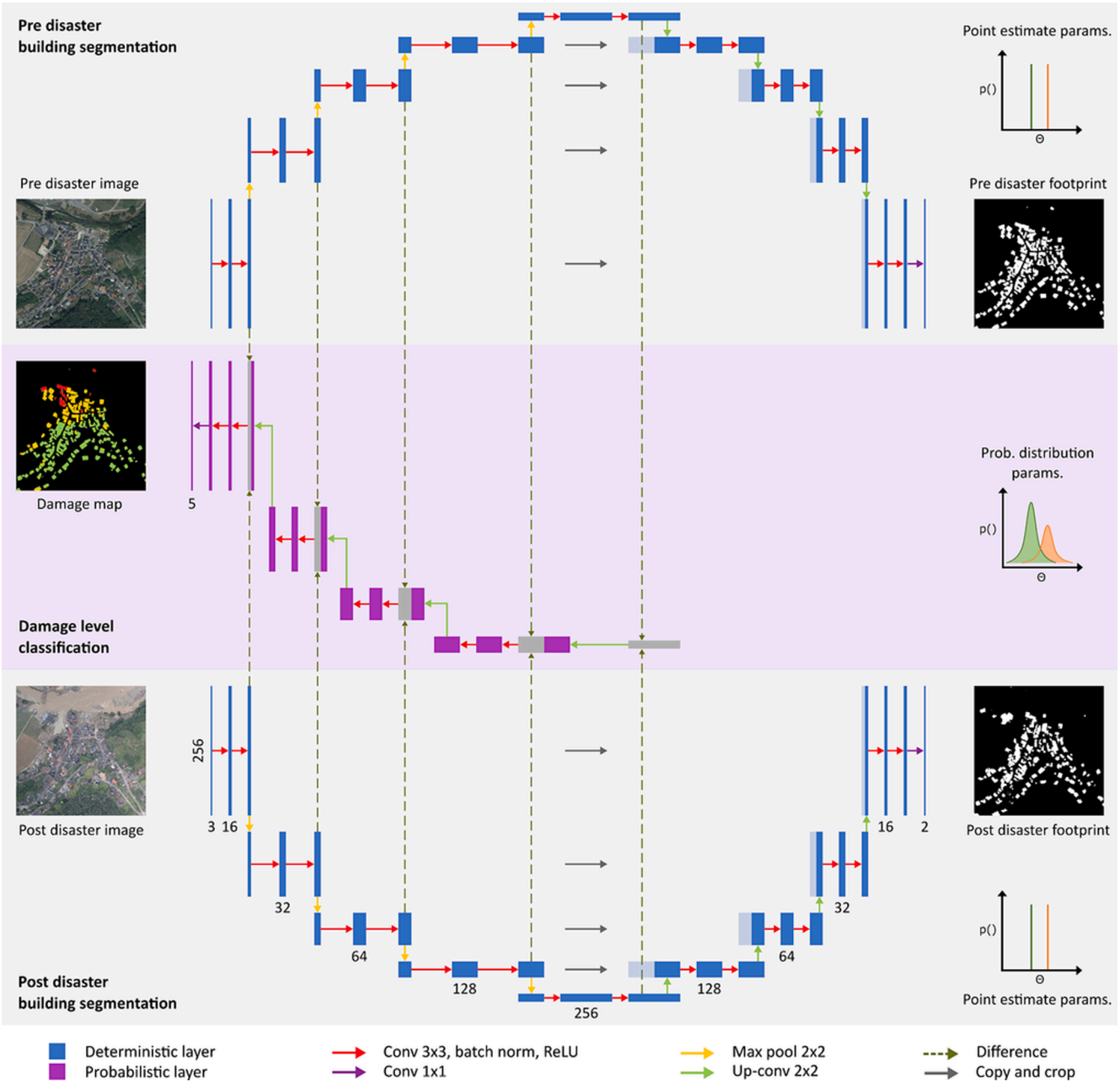


Fig. 1. Bayesian Siamese multitask learning architecture (BayeSiamMTL). The two weight-shared UNets with point estimate model parameters (highlighted in grey) simultaneously segment building footprints from pre- and post-disaster imagery. The Bayesian Siamese difference-based decoder with probability distributions as model parameters (highlighted in purple) classifies damage levels probabilistically based on the segmented patches.

adapted to new domains via multitask fine-tuning. Implementation details are given in Section 3.5.

3.2. Experimental setup

To benchmark and compare performance with conventional BDA models, BayeSiamMTL is evaluated against a purely deterministic counterpart that retains the same architecture but employs only deterministic layers (1, Table 2). This deterministic baseline is trained sequentially as described in Section 3.1. To address the challenge of Bayesian weight initialization, two scenarios are considered: random weight initialization (2) and advanced weight initialization using MOPED (3). The probabilistic baseline (2) is again trained sequentially. For Experiment 3, the probabilistic weights in the Bayesian layers are initialized with parameters from the previously trained deterministic model (1). This initialization is followed by multitask fine-tuning to refine the Bayesian layers, enabling the model to leverage prior knowledge while effectively adapting to the probabilistic task and capturing predictive uncertainties.

3.3. Bayesian deep neural networks and variational inference

In conventional deep learning, model parameters $\theta = (w_1, \cdots, w_K)$ are optimized by maximizing the likelihood of the observed data or, equivalently, minimizing a corresponding loss function. These methods typically handle model parameters as point estimates, which do not capture the inherent uncertainties in the model, data, and parameter estimation (Gawlikowski et al., 2023). Bayesian DNNs address this limitation by describing model parameters as probability distributions rather than fixed values. These distributions encode aggregated uncertainties about the parameters and allow the model to quantify its confidence. In Bayesian DNNs, during each forward pass, the model parameters θ are sampled from their respective posterior distributions. This stochastic sampling leverages the diversity of outputs generated across multiple forward passes for robust uncertainty estimation and improved generalization (Hüllermeier and Waegeman, 2021). Fig. 2 illustrates the distinction between deterministic and probabilistic convolutional layers in BayeSiamMTL, and demonstrates how an example input image is processed through each variant.

Given a training input-target pair (x, y), the posterior distribution over the space of parameters $p(\theta|x, y)$ represents the updated belief about θ after observing the training data (x, y). This posterior is obtained by assuming a prior distribution $p(\theta)$ over the model parameters θ and then applying Bayes' theorem:

$$p(\theta|x, y) = \frac{p(y|x, \theta) \cdot p(\theta)}{p(y|x)} \propto p(y|x, \theta) \cdot p(\theta) \tag{1}$$

Here, the prior distribution $p(\theta)$ represents prior beliefs about θ before observing (x, y). The likelihood $p(y|x, \theta)$ quantifies how well the model with parameters θ explains the observed data (x, y). The evidence $p(y|x)$ acts as a normalization constant to ensure that the posterior distribution integrates to one. It aggregates over all possible values of θ, weighted by their prior probabilities, to capture the full range of plausible parameter values consistent with the observed data (Gawlikowski et al., 2023). The evidence is defined as:

$$p(y|x) = \int p(y|x, \theta) \cdot p(\theta) d\theta \tag{2}$$

Table 2
Overview of the experimental setup.

Experiment		OEM	xBD	Mode
(1)	Deterministic baseline	✓	✓	Sequential training
(2)	Probabilistic baseline	✓	✓	Sequential training
(3)	MOPED based on (1)	×	✓	Multitask fine-tuning

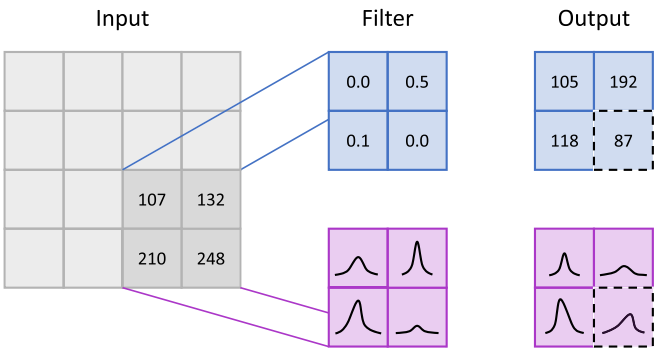


Fig. 2. Input with exemplary pixel values and corresponding output, resulting from i) a deterministic CNN filter with point estimates (highlighted in blue), and ii) a probabilistic CNN filter with probability distributions over weights.

The posterior distribution $p(\theta|x, y)$ is generally intractable due to the high dimensionality and complex nature of DNNs. To address this challenge, variational inference is commonly employed as an approximation technique. Variational inference approximates the true posterior $p(\theta|x, y)$ with a simpler, parameterized distribution $q(\theta)$. This is achieved by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}(q(\theta) \| p(\theta|x, y))$. Since the true posterior $p(\theta|x, y)$ depends on the intractable evidence $p(y|x)$, direct computation of the KL divergence is not feasible. Instead, the evidence lower bound (ELBO) function is optimized, which is equivalent to the KL divergence up to a constant (Abdar et al., 2021; Graves, 2011).

3.4. Bayesian model averaging and Monte Carlo approximation

Once the posterior distribution over the space of parameters $p(\theta|x, y)$ has been approximated, predictions y^* for new, unseen data x^* can be made by calculating the posterior predictive distribution (PPD) $p(y^*|x^*, x, y)$. This distribution accounts for the uncertainty in the model's prediction, capturing both the variability in the model parameters θ and the data (Lynch, 2005). The posterior predictive distribution can be obtained through Bayesian model averaging, which involves marginalizing the likelihood $p(y|x, \theta)$ over the posterior distribution of the model parameters:

$$p(y^*|x^*, x, y) = \int p(y^*|x^*, \theta) \cdot p(\theta|x, y) d\theta \tag{3}$$

The posterior predictive distribution $p(y^*|x^*, x, y)$ represents the likelihood of the prediction y^* given the new input x^* and the model parameters θ. The uncertainty about the model parameters θ given the training data (x, y) is reflected by the posterior distribution $p(\theta|x, y)$.

Since the integral in (3) is typically intractable for the most common prior-posterior pairs, approximation techniques such as Monte Carlo approximation are employed (Gawlikowski et al., 2023). According to the law of large numbers, the expected values of the PPD can be approximated by the mean of the predictions from N stochastic models, $f_{\theta_1}, \cdots, f_{\theta_N}$, each parameterized by samples $\theta_1, \cdots, \theta_N$ drawn from the posterior distributions of the model parameters. The estimate will become more accurate as the value of N increases. This yields the following approximation:

$$y^* \approx \frac{1}{N} \sum_{i=1}^{N} y_i^* = \frac{1}{N} \sum_{i=1}^{N} f_{\theta_i}(x^*) \tag{4}$$

Fig. 3 illustrates the probabilistic outputs of a Bayesian DNN in the context of building damage assessment. The histograms show the distributions of softmax outputs for a single pixel across 100 MC samples. In contrast to deterministic models, which yield a single, constant softmax vector per pixel regardless of repeated evaluations, the Bayesian

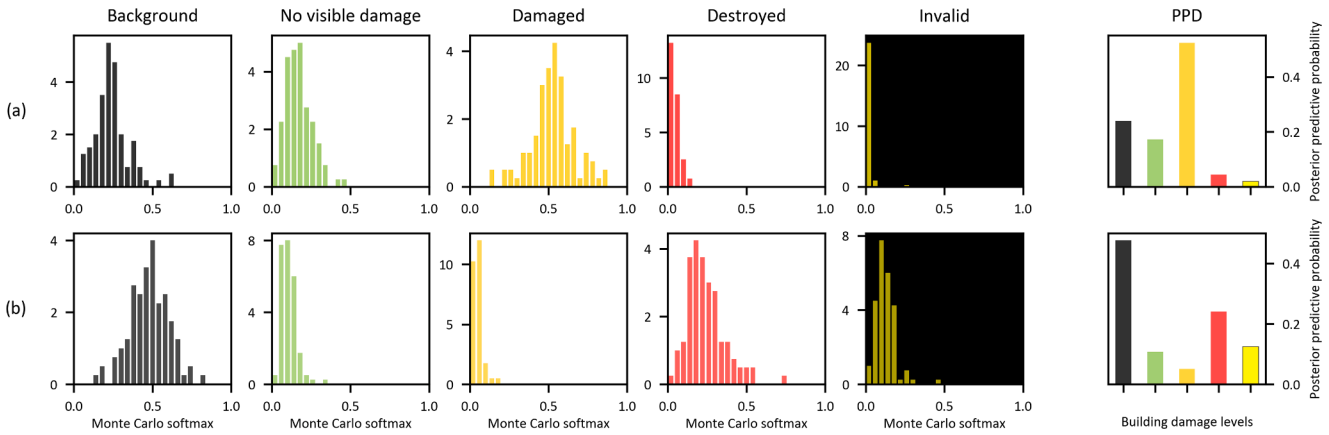


Fig. 3. Probabilistic output of BayeSiamMTL with the respective posterior predictive distribution (PPD): (a) shows a correctly classified 'Damaged' pixel. (b) shows a 'Destroyed' pixel incorrectly classified as 'Background'.

framework produces varying softmax outputs for the same pixel across multiple forward passes. This variability directly captures the predictive uncertainty associated with the pixel's classification. The rightmost column shows the PPD, which is computed using Monte Carlo approximation and the respective Bayesian equation in (4). The damage class with the highest posterior predictive probability (PPP) is selected as the classification output, while dispersions in the PPD serve as a measure to quantify the predictive uncertainty. Fig. 3a illustrates a pixel correctly classified as 'Damaged'. In this case, the MC output distribution for 'Damaged' exhibits a probability mass concentrated toward higher confidence, whereas distributions for the incorrect classes shift toward lower probabilities. These observations are reflected and quantified in the respective PPD. In contrast, Fig. 3b shows a 'Destroyed' pixel that has been misclassified as 'Background'. Here, the MC output distributions are slightly more dispersed, reflecting the uncertainty in the model's prediction.

3.5. Implementation details

BayeSiamMTL is implemented in PyTorch, utilizing the open-source BayesianTorch library (Krishnan et al., 2022) for the Bayesian Siamese decoder. A notable challenge in employing probabilistic weights arises from the similarity of weights drawn within a mini-batch, which constrains the variance reduction effect of larger batches. To overcome this limitation, the Flipout estimator (Wen et al., 2018) is employed, providing effective variance reduction by pseudo-independently sampling weights for each individual sample.

The probabilistic decoder processes the latent tensors derived from two input images and calculates their element-wise differences at the bottleneck level. This difference-based design enables the network to explicitly focus on scene changes, which is critical in BDA tasks. The resulting feature maps are subsequently upsampled to the original image resolution, producing a pixel-wise damage classification map via a multi-class softmax activation function. Model training leverages the Bayes by Backprop algorithm (Blundell et al., 2015), in combination with the Adam optimizer, to minimize the evidence lower bound (ELBO). The ELBO loss includes a cross-entropy term for damage classification accuracy and a scaled KL divergence term that regularizes the variational posterior (see Section 3.3 for details). Bayes by Backprop integrates variational inference into standard backpropagation, enabling efficient, tractable approximation of the posterior distribution over model weights by iteratively updating the variational parameters.

All hyperparameters employed in this study (cf. in Table 3) are determined through preliminary experiments. Section 4.1 provides an in-depth analysis of the influence of the number of MC samples. A standard normal prior, $N(0,1)$, is commonly adopted in Bayesian deep learning and has been found to provide the best balance between weight initialization and weight exploration (LaBonte et al., 2020). Early stopping criteria and batch sizes are chosen to optimize computational efficiency within the constraints of available hardware resources. To ensure consistency and reproducibility across experiments, random seeds are fixed for all stochastic processes, and deterministic algorithms are enforced in PyTorch. Experiments are performed on a Dell Precision 5820 Tower equipped with 64 GB RAM, an Intel Xeon W-2235 CPU, and an NVIDIA RTX A4000 GPU.

Table 3
Hyperparameters and model setup.

Parameter	
Initial learning rate	1×10^{-3}
Learning rate decay	0.5 / 5 epochs validation loss stagnation
Maximum epochs	50 per task
Early stopping	After 10 epochs validation loss stagnation
Batch size	16
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Prior distribution	$N(0,1)$
Monte Carlo samples	100
MOPED perturbation δ	0.5
Seed	11

3.6. Accuracy assessment

The F_1 score is a widely used metric for evaluating performance in BDA. In this study, both the weighted F_1^{weighted} score and the macro-averaged F_1^{macro} score are computed based on an aggregated confusion matrix. The F_1^{weighted} score accounts for class imbalance by weighting the contribution of each class according to its prevalence. In contrast, the F_1^{macro} score treats all classes equally, regardless of their frequency, ensuring that equal importance is assigned to all damage levels. This dual evaluation approach provides a comprehensive assessment of the models' performance, capturing both their ability to handle imbalanced datasets and their capacity to detect rare but critical damage levels. Additionally, class-wise F_1 scores are reported to provide a more detailed analysis of the models' performance for individual classes. For probabilistic outputs, the class with highest probability of the PPD, i.e. the highest posterior predictive probability (PPP), determines the prediction.

4. Results

4.1. Learning behavior on xBD

An important yet often overlooked hyperparameter for Bayesian

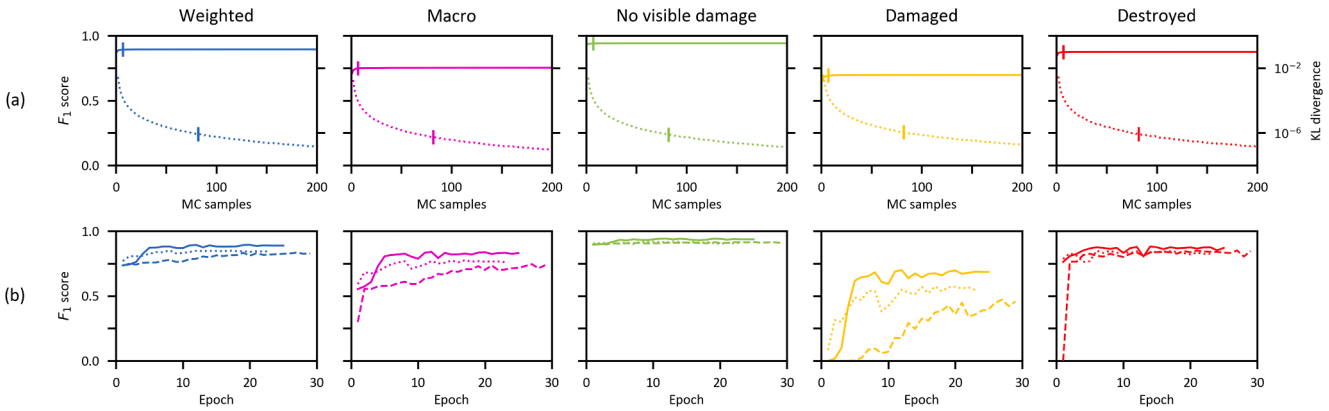


Fig. 4. Average and class-specific learning behavior: (a) illustrates the convergence of F_1 scores (solid) and KL divergence (dotted) as the number of Monte Carlo (MC) samples increases. The epoch of convergence is highlighted by vertical marks. (b) shows the training progression of Experiment 1 (dotted), Experiment 2 (dashed), and Experiment 3 (solid).

DNNs is the number of MC samples needed to ensure reliable predictions. Fig. 4a illustrates the average and class-specific F_1 scores as a function of MC samples, plotted as solid lines. A relative convergence criterion is defined as $\frac{|x_{k+1}-x_k|}{|x_k|} < \delta$, where x_k and x_{k+1} represent values from consecutive iterations, and $\delta = 1 \times 10^{-3}$. The F_1 scores meet the convergence criterion after seven MC samples. However, accuracy metrics only require convergence of the highest PPP, as this determines the predicted class. For reliable uncertainty quantification, convergence of the whole PPD is required. The dotted lines in Fig. 4a illustrate how the KL divergence of the PPDs decreases as the number of MC samples increases. The convergence criterion for the KL divergence is met after 82 MC samples. For this analysis, KL divergence is calculated pixelwise and averaged over the entire test split of the xBD dataset. All calculations are performed for up to 200 MC samples, ensuring a robust convergence assessment of both classification accuracy and uncertainty quantification. 'Invalid' pixels are excluded from Fig. 4 as they are not present in the xBD dataset.

Based on these insights, the experiments in Table 2 are conducted with each training epoch being evaluated using 100 MC samples. Fig. 4b illustrates the progression of the average and class-specific F_1 scores during classification training (Experiments 1 and 2) and multiclass fine-tuning (Experiment 3). The deterministic baseline model is represented by dotted lines (Experiment 1), BayeSiamMTL with random weight initialization by dashed lines (Experiment 2), and BayeSiamMTL with MOPED-based weight initialization by solid lines (Experiment 3). The results show that the deterministic baseline model accumulates slightly more knowledge compared to its probabilistic counterpart, with the

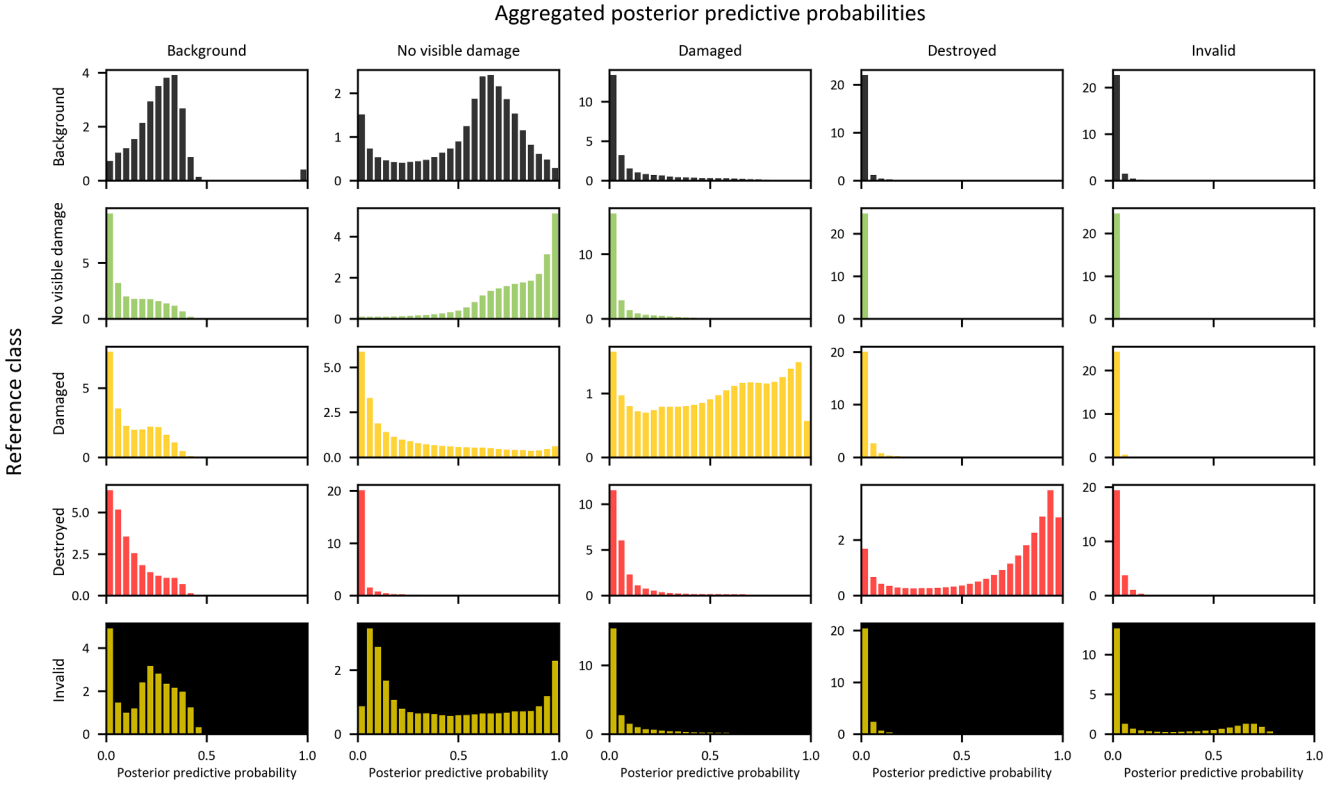


Fig. 5. Distribution of posterior predictive probabilities (PPPs) aggregated over all 2.35 billion pixels in the xBD dataset, each evaluated with 100 Monte Carlo samples. The PPPs are separated class-wise and aggregated in individual plots horizontally, while being organized vertically based on the reference class. Diagonal plots (correct classifications) ideally shift right (high PPPs), whereas off-diagonal plots (misclassifications) shift left (low PPPs).

highest improvement observed for the 'Damaged' class. For the 'No visible damage' and 'Destroyed' classes, the performance differences between the deterministic and probabilistic baseline models are minor. However, Experiment 3 clearly surpasses the performance of both Experiments 1 and 2. This demonstrates the effectiveness of combining deterministic pre-training with Bayesian fine-tuning for improving overall model performance.

4.2. Aggregated uncertainty analysis

Fig. 5 provides detailed insights into the model's classification behavior, shedding light on its internal decision-making tendencies and sources of uncertainty. For this analysis, all approximately 2.35 billion pixels in the xBD test split are evaluated using 100 MC samples, resulting in approximately 235 billion individual pixel evaluations. The corresponding PPPs, as previously illustrated in Fig. 3, are separated by class and aggregated in individual plots horizontally, while being organized vertically based on the reference class. In an ideal classifier, histograms along the diagonal (representing correct classifications) would be concentrated toward the right, indicating high predictive probabilities for the true class. Conversely, off-diagonal histograms (representing misclassifications) would be skewed toward the lower end of the probability spectrum, reflecting low predictive probabilities for incorrect predictions. Fig. 5 is structured as a 5 × 5 grid, with individual histograms referenced using the [row, column] notation. This layout provides a nuanced perspective on the model's ability to differentiate between classes.

For pixels labeled as 'Background' (first row), the model demonstrates high confidence in excluding the categories 'Damaged', 'Destroyed', and 'Invalid', as their PPPs are concentrated toward the lower end of the probability spectrum (see [1, 3], [1, 4], and [1, 5]). Pixels labeled as 'No visible damage' generally receive higher-confidence classifications, with 'Damaged', 'Destroyed', and 'Invalid' categories again clustering near the lower end of the probability spectrum ([2, 3], [2, 4], [2, 5]). Nonetheless, there is noticeable uncertainty between 'Background' and 'No visible damage', as indicated by the relatively broad PPP spectra in [1, 2] and [2, 1]. Turning to 'Damaged' pixels (third row), the respective PPP distribution in [3, 3] is relatively flat, covering a broad range of probabilities from low to high. Despite this variability, the model consistently assigns low probabilities to the 'Destroyed' ([3, 4]) and 'Invalid' ([3, 5]) categories, reflecting confidence in excluding these incorrect classes. The outputs for 'Background' ([3, 1]) and 'No visible damage' ([3, 2]) hover around midrange probabilities but tend to converge toward lower values. 'Destroyed' pixels show more expected behavior, with correct classifications heavily clustered at the high end of the probability spectrum in [4, 4], while misclassifications appear at lower probability values. Finally, for 'Invalid' pixels (bottom row), the model frequently misclassifies them as either 'No visible damage' ([5, 2]) or 'Background' ([5, 1]), yet rarely assigns high predictive probabilities to the 'Damaged', 'Destroyed', or even 'Invalid' classes ([5, 3], [5, 4], [5, 5]).

4.3. Model transferability

Model performance is evaluated on an independent test split of the xBD dataset (see Section 2.2.1) and further assessed via test-only evaluations on the Ahr valley dataset, highlighting the impact of domain shifts and OOD samples. The results in Table 4 align with the trends discussed in Section 4.1. All experiments achieve state-of-the-art performance on the xBD dataset, with particularly strong results in the 'No visible damage' and 'Destroyed' categories. Performance in the 'Damaged' category remains consistently lower. However, BayeSiamMTL initialized with pre-trained deterministic weights using MOPED (Experiment 3) considerably improves performance and outperforms both Experiments 1 and 2, most notably in the challenging 'Damaged' category.

On the unseen Ahr valley dataset, classification performance declines significantly due to the presence of domain shifts. Although the probabilistic baseline (Experiment 2) generally surpasses the deterministic baseline (Experiment 1), it falls short in the 'Destroyed' category. In contrast, Experiment 3 consistently demonstrates substantial performance advantages under these conditions. It achieves significantly higher metrics compared to both baselines, demonstrating its superior generalization capabilities in handling domain shifts and OOD samples. Low performance for the 'Invalid' category is expected due to the absence of samples in the xBD dataset.

5. Discussion

5.1. Model performance and computational efficiency

The performance of BayeSiamMTL is best understood through the lens of its operational design. Unlike many recent BDA models that depend on heavily parameterized architectures fine-tuned to specific datasets, BayeSiamMTL is intentionally designed to be lightweight, generalizable, and uncertainty-aware.

The deterministic baseline features a compact architecture with just 2.7 million parameters and a computational cost of 16.1 billion floating point operations (GFLOPs) per 256 × 256 pixel tile. Its Bayesian extension introduces only 0.6 million additional parameters, yielding a total of 3.3 million parameters and a computational cost of 19.7 GFLOPs. This modest increase demonstrates that uncertainty quantification adds value along an orthogonal axis of model complexity (Wani et al., 2025), refuting the common assumption that trustworthy uncertainty estimates necessitate significantly larger networks. In contrast, state-of-the-art CNN and transformer-based methods such as MambaBDA-Small, DamFormer, ChangeOS-101, and MTF are substantially larger in both parameter count and computational cost (Chen et al., 2024, 2022; Weber and Kané, 2020; Yu et al., 2025; Zheng et al., 2021). BayeSiamMTL, by comparison, is up to an order of magnitude more efficient in terms of both model size and computational demand (cf. in Table 5).

Beyond efficiency, BayeSiamMTL is designed for operational damage assessment and optimized for the internationally recognized three-class damage taxonomy of institutions like the Copernicus Emergency Management Service (Hertel et al., 2025). This contrasts with the xBD

Table 4
Average and class-specific accuracy metrics of experiments on the xBD and Ahr valley datasets.

Experiment		$F_1^{weighted}$(%)	F_1^{macro}(%)	F_1 per class (%)			
				No visible damage	Damaged	Destroyed	Invalid
			xBD dataset				
(1)	Deterministic baseline	85.2	76.8	91.8	53.7	84.9	n/a
(2)	Probabilistic baseline	83.5	74.5	91.3	47.2	85.1	n/a
(3)	MOPED based on (1)	**89.4**	**83.8**	**94.1**	**69.7**	**87.5**	n/a
			Ahr valley dataset				
(1)	Deterministic baseline	37.9	32.8	59.2	22.3	49.7	0.0
(2)	Probabilistic baseline	45.2	35.5	57.6	37.3	47.2	0.0
(3)	MOPED based on (1)	**53.7**	**40.5**	**60.9**	**51.2**	**50.0**	**0.2**

Table 5
Performance of BayeSiamMTL trained on the xBD dataset's original four-class damage scheme and evaluated on the corresponding xBD test split in comparison with current state-of-the-art methods, along with model size and computational cost.

Method		F_1^{xView2}(%)	F_1^{loc}(%)	F_1^{dmg}(%)	Param (M)	GFLOPs
(1)	Deterministic baseline	65.4	65.4	65.4	2.7	16.1
(2)	Probabilistic baseline	22.3	48.5	11.1	3.3	19.7
(3)	MOPED based on (1)	78.4	78.2	78.4	3.3	19.7
MambaBDA-Small		**81.1**	86.6	**78.8**	52.1	130.8
DamFormer		77.0	**86.9**	72.8	32.5	169.3
ChangeOS-101		75.5	85.7	71.1	58.1	157.3
MTF		74.1	83.6	70.0	44.4	268.8

dataset's four-class damage scheme, which is more prone to label ambiguity and class imbalance.

These architectural and methodological distinctions limit the information gain of direct metric-to-metric comparisons with dataset-optimized and highly parameterized state-of-the-art models. Nevertheless, to ensure long-term comparability with existing and future benchmarks, all experiments are retrained using the xBD dataset's original four-class damage scheme and evaluated on its corresponding test split. Performance is assessed following the xView2 challenge protocol (Gupta et al., 2019), which combines building localization and damage classification into a single metric: $F_1^{\text{xView2}} = 0.3 \times F_1^{\text{loc}} + 0.7 \times F_1^{\text{dmg}}$, where F_1^{dmg} denotes the harmonic mean of class-wise damage scores. This formulation strongly penalizes underperforming classes, ensuring the metric reflects balanced performance across all classes. Under this evaluation, BayeSiamMTL achieves a score of 78.4 % (cf. in Table 5), placing it in the same performance magnitude as highly parameterized models (Chen et al., 2024, 2022; Weber and Kané, 2020; Yu et al., 2025; Zheng et al., 2021). This is particularly notable given BayeSiamMTL's lightweight architecture and its added capability of providing calibrated uncertainty estimates.

In summary, BayeSiamMTL achieves operationally meaningful classification accuracy, credible uncertainty quantification, and a substantially reduced computational footprint. Its core contribution lies not in marginal performance gains over recent benchmarks, but in introducing a reliable and efficient probabilistic framework tailored for scalable deployment in real-world disaster scenarios. Moreover, this study demonstrates that extending deterministic models into probabilistic counterparts can lead to substantial improvements in performance and generalization capability, all while preserving the original architectural structure.

5.2. Decoding classification dynamics: Insights and uncertainties

The aggregated histograms of PPPs in Fig. 5 provide a detailed perspective into the model's classification dynamic within a probabilistic, uncertainty-aware framework. While previous studies have indicated intermediate damage levels as a key challenge in BDA, this analysis is the first to systematically investigate and quantify these tendencies through probabilistic modeling.

A primary outcome emerging from Fig. 5 is that 'Background' pixels cause the most confusion. The model often assigns comparably high PPPs to 'Background' across all predictions, producing wide-ranging probability spectra in the first column of Fig. 5. This challenge likely stems from inaccurate or incomplete building footprint annotations or from the broad heterogeneity of 'Background' characteristics within the dataset. Addressing these issues could involve refining training datasets, improving segmentation precision, or incorporating additional features to mitigate ambiguity.

Focusing on the actual damage classes and disregarding both 'Background' and 'Invalid' pixels, the internal 3×3 grid in [2–4, 2–4] of Fig. 5 becomes most relevant. Contrary to the commonly held view that separating intermediate damage levels is especially difficult, the model confidently differentiates 'Damaged' from 'Destroyed' buildings ([3, 4]) and vice versa ([4, 3]). Nonetheless, the relatively broad PPP distribution for 'Damaged' instances in [3, 3] suggests significant variability and elevated predictive uncertainty within this specific class. A similar, though less pronounced, pattern appears for 'Destroyed' pixels, which exhibit a bimodal, U-shaped distribution in [4, 4]. These outcomes imply that while the model can be highly confident in correct classifications, it can also occasionally misclassify certain instances with equally high certainty. Such high-confidence misclassifications carry significant risk in practical applications and underscore the need for strategies to mitigate their occurrence.

Examining PPDs of individual pixels (see Fig. 3) reveals a deeper limitation of framing BDA strictly as a semantic segmentation task. Here, Fig. 3a shows a unimodal distribution across damage levels, whereas Fig. 3b displays a bimodal distribution. This bimodality contradicts the physics-based expectation that 'No visible damage' and 'Damaged' would lie closer on the damage spectrum than 'No visible damage' and 'Destroyed'. Such inconsistencies could be alleviated by adopting custom loss functions that encode the ordinal nature of damage levels, ensuring that slight differences in damage severity are penalized less than large differences. Alternatively, a regression-based approach might align more naturally with the continuous spectrum of damage severity, better capturing gradual transitions from minor to severe building damage.

5.3. Generalization and domain adaptation capabilities

Bayesian models are widely recognized for their ability to mitigate overfitting and enhance generalization, making them a robust foundation for techniques in domain adaptation and domain generalization (Liu et al., 2021; Xiao et al., 2021). Our results in Table 4 illustrate this property: although the Bayesian models already show a notable improvement over deterministic approaches on the training dataset (source domain), they display a markedly stronger performance advantage on the unseen Ahr valley dataset (target domain).

To further quantify this improvement, we train BayeSiamMTL on the target domain as a benchmark and report the average and class-wise gains of our Bayesian approaches relative to this benchmark in Table 6. The results show that BayeSiamMTL substantially reduces the performance gap, with the most pronounced benefits observed in the 'Damaged' category, where 48 % of the potential performance gain are achieved. Moreover, Hertel et al. (2025) employed an identical but deterministic model architecture trained exclusively on the xBD dataset and later incorporated semi-supervised and supervised domain adaptation techniques to adapt to the Ahr valley dataset. Notably, our Experiment 3 outperforms several of these adaptation strategies, despite the fact that BayeSiamMTL in Experiment 3 was never exposed to Ahr valley data, whereas the domain adaptation methods either directly or

Table 6
Average and class-specific accuracy metrics of our experiments compared to supervised learning on the Ahr valley (benchmark). The percentages show the performance gain compared to the upper performance limit.

Experiment		$F_1^{weighted}$ (%)		F_1^{macro} (%)		F_1 per class (%)							
						No visible damage		Damaged		Destroyed		Invalid	
Benchmark		75.3	+ 100 %	54.7	+ 100 %	72.0	+ 100 %	81.9	+ 100 %	65.0	+ 100 %	0.0	+ 100 %
(3)	MOPED based on (1)	53.7	+ 42 %	40.5	+ 35 %	60.9	+ 13 %	51.2	+ 48 %	50.0	+ 2 %	0.2	NaN
(2)	Probabilistic baseline	45.2	+ 20 %	35.5	+ 12 %	57.6	− 13 %	37.3	+ 25 %	47.2	− 16 %	0.0	NaN
(1)	Deterministic baseline	37.9	+ 0 %	32.8	+ 0 %	59.2	+ 0 %	22.3	+ 0 %	49.7	+ 0 %	0.0	+ 0 %

indirectly leveraged target domain information.

This capacity to generalize across domains without explicit adaptation can be attributed to the marginalization property of BayeSiamMTL, which involves integration over the model parameter space. By effectively marginalizing over model parameters, Bayesian models benefit from an implicit form of regularization, thereby reducing overreliance on specific training samples and producing more robust predictions on unseen data (Calvetti and Somersalo, 2018; Wilson and Izmailov, 2020).

5.4. Uncertainty visualization and confidence-informed damage maps

Algorithmic developments and performance improvements are central to BDA. However, effective communication and visualization of model predictions are equally important, particularly in rapid response scenarios characterized by limited resource availability. In such contexts, explicitly incorporating prediction uncertainty into the decision-making process is critical for evaluating potentially adverse consequences. Special attention should be given to tail probabilities, as low-probability but high-impact events may require response strategies that deviate from those suggested by the most probable model output alone. Fig. 6 illustrates several confidence-informed damage maps applied to the Ahr valley region, highlighting how explicitly integrating uncertainty information can enhance transparency and support informed decision-making.

In Fig. 6a, the input data and reference mask are presented alongside the two key outputs of BayeSiamMTL: the predicted damage classification, which assigns a specific damage condition to each individual building, and the corresponding certainty mask, which communicates the degree of confidence in each prediction. The probabilistic certainty output directly informs the final classification by selecting the damage category with the highest PPP. In combination, these outputs provide a richer foundation for informed decision-making, particularly under the constraints of post-disaster emergency response. Fig. 6b merges the classification and certainty layers into a unified visualization by linearly interpolating the damage level colors based on the associated PPPs, such that more saturated colors indicate higher model confidence in the predicted damage class. This approach produces continuous transitions that reflect both the model's prediction and varying confidence levels. Decomposing the pixelwise PPDs into their constituent PPPs further reveals stratified probabilities across damage categories, enabling more precise identification and prioritization of damage clusters. Fig. 6c demonstrates how probabilistic outputs can delineate both the

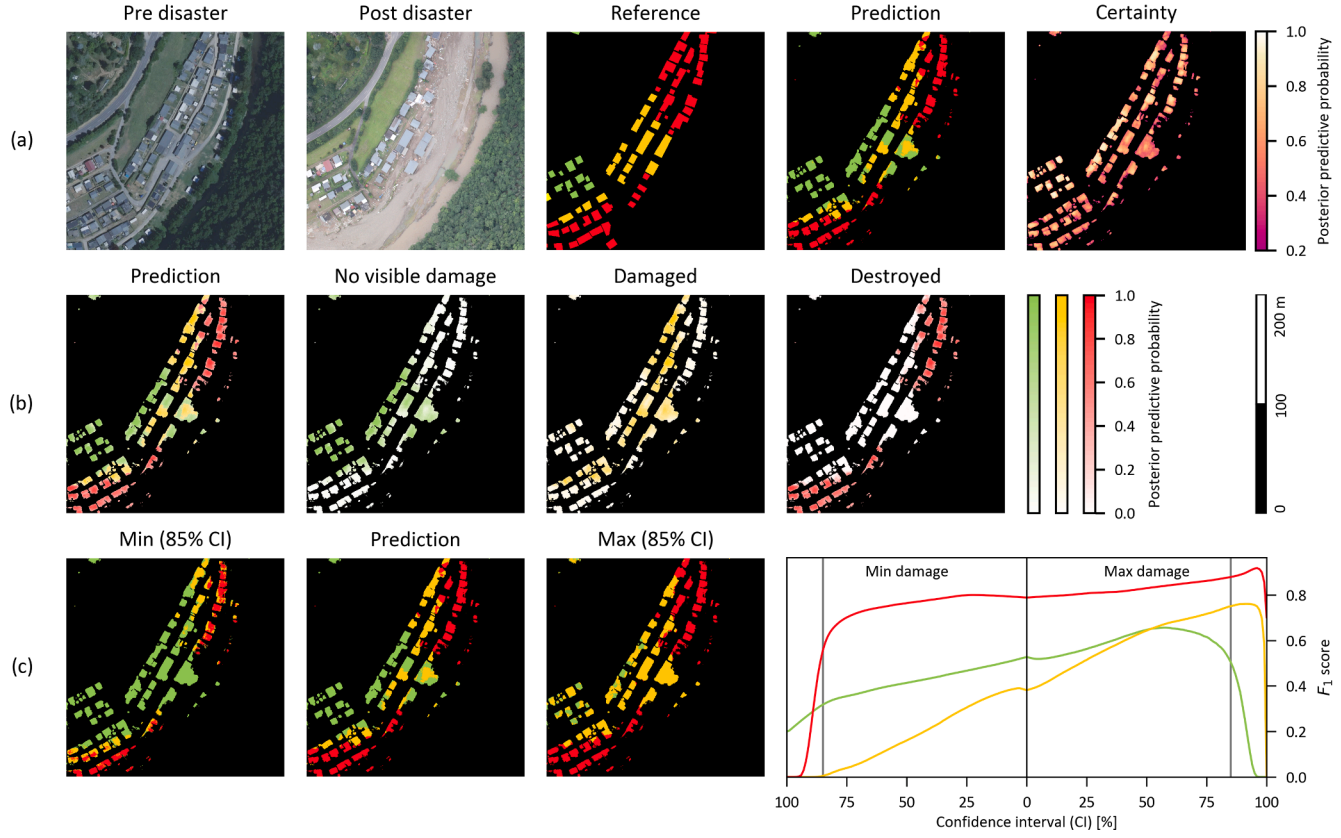


Fig. 6. Confidence-informed damage maps: (a) input data, reference mask, model prediction, and probabilistic certainty mask. (b) integrated prediction-certainty map with disaggregated and stratified probabilities of damage clusters. (c) minimum and maximum damage extent delineated from confidence intervals.

minimum and maximum extent of building damage within a specified confidence interval. Here, an 85 % confidence interval corresponds to a PPP threshold of 7.5 %. The minimum extent is defined by the least severe damage class exceeding this threshold, while the maximum extent reflects the most severe class still deemed plausible. This visualization offers a concise summary of both the best- and worst-case outcomes at the building level. Furthermore, Fig. 6c assesses model performance across confidence intervals, revealing higher performance within the sphere of maximum damage extent. This can be interpreted as a probabilistic tendency analysis toward underestimation of building damage for the illustrated tile.

Overall, these visualizations underscore the value of confidence-informed damage maps. By integrating uncertainty into the final classification, stakeholders are equipped with richer and more actionable information. A forthcoming companion study will further investigate how quantified uncertainties derived from Bayesian modeling can be systematically integrated into maps, with a focus on enhancing decision-making in emergency contexts.

6. Conclusion

This study advances post-disaster building damage assessment (BDA) by integrating uncertainty quantification (UQ) into a hybrid Bayesian deep learning framework. While accurate and reliable damage assessment is essential for guiding humanitarian interventions, UQ ensures that model outputs are both reliable and transparent. To achieve this, we introduced BayeSiamMTL, a Bayesian Siamese multitask learning architecture that combines deterministic binary semantic segmentation for building footprint extraction with probabilistic multiclass change detection for damage level classification. By representing model parameters as probability distributions and applying variational inference with Monte Carlo approximation, we derived pixelwise posterior predictive distributions (PPDs). These PPDs provide a rich depiction of both prediction outcomes and their associated uncertainties, allowing the model to assess its own confidence.

Our results highlighted that Bayesian modeling not only supplies explicit uncertainty estimates but also delivers superior classification performance compared to deterministic baselines. Evaluations on the training xBD dataset and the unseen Ahr valley dataset confirmed BayeSiamMTL's generalization capabilities and robustness under domain shifts, which is a crucial requirement for BDA in real-world disaster scenarios. This behavior can be attributed to the marginalization property of Bayesian methods and is particularly evident when deterministic pre-training is combined with advanced weight initialization strategies such as MOPED. Analyzing over 235 billion individual pixel evaluations revealed the 'Background' class as the primary source of confusion across all damage levels, likely due to imperfect building footprint annotations and the wide variability among non-building pixels. The analysis also indicated that building destructions are more frequently confused with intact buildings rather than among varying degrees of damage. In addition, we illustrated how PPDs can be used to generate confidence-informed damage maps in the form of integrated prediction-certainty visualizations. These maps provide disaggregated and stratified probabilities of distinct damage clusters as well as minimum/maximum damage extents delineated from confidence intervals.

In spite of these promising findings, the study uncovers a limitation inherent in treating BDA strictly as a semantic segmentation task. Specifically, the multimodal nature of certain PPDs contradicts with the physical intuition of an inherent order among damage classes; namely that 'No visible damage' and 'Damaged' would lie closer on the damage spectrum than 'No visible damage' and 'Destroyed'. Addressing this discrepancy may involve adopting custom loss functions that respect the ordinal relationship among damage classes or reframing the problem as a regression task to promote unimodal PPDs.

Overall, the results demonstrate that BayeSiamMTL effectively leverages Bayesian modeling to produce robust and accurate building damage assessments with explicit uncertainty quantification. By outperforming deterministic approaches and demonstrating strong adaptability under domain shifts, this approach provides transparent damage estimation and domain scalability, making it well-suited for real-world disaster management contexts. Future research will further explore unimodal PPDs and advance intuitive visualization strategies that embed uncertainty insights into decision support systems – a critical step in maximizing the potential of Bayesian modeling for disaster response and humanitarian applications.

CRediT authorship contribution statement

Victor Hertel: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Omar Wani:** Writing – review & editing, Validation, Methodology, Conceptualization. **Christian Geiß:** Writing – review & editing, Validation, Conceptualization. **Marc Wieland:** Writing – review & editing, Validation, Conceptualization. **Hannes Taubenböck:** Writing – review & editing, Validation, Conceptualization.

Funding

This work was supported by the German Federal Environmental Foundation (DBU) and DLR internal funding as part of the projects 'KI4FloodDamage 2′ and 'Resiliente Technologien für den Katastrophenschutz' (RESITEK).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2025.104759.

Data availability

The OpenEarthMap and xBD datasets are publicly available. The post-disaster imagery of the Ahr valley dataset can be requested for scientific purposes.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf. Fusion 76, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008.

Bin, J., Zhang, R., Wang, R., Cao, Y., Zheng, Y., Blasch, E., Liu, Z., 2022. An Efficient and Uncertainty-Aware Decision support System for disaster Response using Aerial Imagery. Sensors 22, 7167. https://doi.org/10.3390/s22197167.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational Inference: a Review for Statisticians. J. Am. Stat. Assoc. 112, 859–877. https://doi.org/10.1080/01621459.2017.1285773.

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural networks. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15. JMLR.org, Lille, France, pp. 1613–1622.

Calvetti, D., Somersalo, E., 2018. Inverse problems: from regularization to Bayesian inference. WIREs Comput. Stat. 10, e1427.

Chen, H., Nemni, E., Vallecorsa, S., Li, X., Wu, C., Bromley, L., 2022. Dual-Tasks Siamese Transformer Framework for Building damage Assessment. In: IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium. Presented at the IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, pp. 1600–1603. https://doi.org/10.1109/IGARSS46834.2022.9883139.

Chen, H., Song, J., Han, C., Xia, J., Yokoya, N., 2024. ChangeMamba: Remote Sensing Change Detection with Spatiotemporal State Space Model. IEEE Trans. Geosci. Remote Sens. 62, 1–20. https://doi.org/10.1109/TGRS.2024.3417253.

Cotrufo, S., Sandu, C., Giulio Tonolo, F., Boccardo, P., 2018. Building damage assessment scale tailored to remote sensing vertical imagery. European Journal of Remote Sensing 51, 991–1005. https://doi.org/10.1080/22797254.2018.1527662.

Deng, L., Wang, Y., 2022. Post-disaster building damage assessment based on improved U-Net. Sci. Rep. 12, 15862. https://doi.org/10.1038/s41598-022-20114-w.

Dera, D., Rasool, G., Bouaynaya, N.C., Eichen, A., Shanko, S., Cammerata, J., Arnold, S., 2020. Bayes-SAR net: Robust SAR image Classification with uncertainty Estimation using Bayesian Convolutional Neural Network. In: 2020 IEEE International Radar Conference (RADAR). Presented at the 2020 IEEE International Radar Conference (RADAR), pp. 362–367. https://doi.org/10.1109/RADAR42522.2020.9114737.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian Approximation: Representing Model uncertainty in Deep Learning. In: Proceedings of the 33rd International Conference on Machine Learning. Presented at the International Conference on Machine Learning, pp. 1050–1059.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X., 2023. A survey of uncertainty in deep neural networks. Artif. Intell. Rev. 56, 1513–1589. https://doi.org/10.1007/s10462-023-10562-9.

Ge, P., Gokon, H., Meguro, K., 2020. A review on synthetic aperture radar-based building damage assessment in disasters. Remote Sens. Environ. 240, 111693. https://doi.org/10.1016/j.rse.2020.111693.

Ge, J., Tang, H., Yang, N., Hu, Y., 2023. Rapid identification of damaged buildings using incremental learning with transferred data from historical natural disaster cases. ISPRS J. Photogramm. Remote Sens. 195, 105–128. https://doi.org/10.1016/j.isprsjprs.2022.11.010.

Geiß, C., Priesmeier, P., Aravena Pelizari, P., Soto Calderon, A.R., Schoepfer, E., Riedlinger, T., Villar Vega, M., Santa María, H., Gómez Zapata, J.C., Pittore, M., So, E., Fekete, A., Taubenböck, H., 2023. Benefits of global earth observation missions for disaggregation of exposure data and earthquake loss modeling: evidence from Santiago de Chile. Nat. Hazards 119, 779–804. https://doi.org/10.1007/s11069-022-05672-6.

Gholami, S., Robinson, C., Ortiz, A., Yang, S., Margutti, J., Birge, C., Dodhia, R., Ferres, J. L., 2022. On the deployment of post-disaster building damage assessment tools using satellite imagery: a Deep Learning Approach. In: In: IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1029–1036. https://doi.org/10.1109/ICDMW58026.2022.00134.

Graves, A., 2011. Practical Variational Inference for Neural Networks. Advances in Neural Information Processing Systems. Curran Associates, Inc.

Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., Heim, E., Choset, H., Gaston, M., 2019. xBD: A Dataset for Assessing Building Damage from Satellite Imagery. https://doi.org/10.48550/arXiv.1911.09296.

Hao, H., Baireddy, S., Bartusiak, E.R., Konz, L., LaTourette, K., Gribbons, M., Chan, M., Delp, E.J., Comer, M.L., 2021. An Attention-based System for damage Assessment using Satellite Imagery. In: IEEE International Geoscience and Remote Sensing Symposium IGARSS, pp. 4396–4399. https://doi.org/10.1109/IGARSS47720.2021.9554054.

Hertel, V., Chow, C., Wani, O., Wieland, M., Martinis, S., 2023. Probabilistic SAR-based water segmentation with adapted Bayesian convolutional neural network. Remote Sens. Environ. 285, 113388. https://doi.org/10.1016/j.rse.2022.113388.

Hertel, V., Geiß, C., Wani, O., Wieland, M., Taubenböck, H., 2025. Rapid domain adaptation for disaster impact assessment: remote sensing of building damage after the 2021 Germany floods [Manuscript submitted for publication]. Sci. Remote Sens.

Herzog, L., Murina, E., Dürr, O., Wegener, S., Sick, B., 2020. Integrating uncertainty in deep neural networks for MRI based stroke analysis. Med. Image Anal. 65, 101790. https://doi.org/10.1016/j.media.2020.101790.

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. 110, 457–506. https://doi.org/10.1007/s10994-021-05946-3.

Jospin, L.V., Laga, H., Boussaid, F., Buntine, W., Bennamoun, M., 2022. Hands-on Bayesian Neural Networks—A Tutorial for Deep Learning users. IEEE Comput. Intell. Mag. 17, 29–48. https://doi.org/10.1109/MCI.2022.3155327.

Kaur, N., Lee, C.-C., Mostafavi, A., Mahdavi-Amiri, A., 2023. Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images. Comput. Aided Civ. Inf. Eng. 38, 2072–2091. https://doi.org/10.1111/mice.12981.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision?. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 5580–5590.

Kingma, D.P., Salimans, T., Welling, M., 2015. Variational dropout and the local reparameterization trick. Advances in Neural Information Processing Systems. Curran Associates, Inc.

Kiureghian, A.D., Ditlevsen, O., 2009. Aleatory or epistemic? does it matter? structural safety. Risk Acceptance and Risk Communication 31, 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020.

Krishnan, R., Esposito, P., Subedar, M., 2022. Bayesian-Torch: Bayesian neural network layers for uncertainty estimation. https://doi.org/10.5281/ZENODO.5908307.

Krishnan, R., Subedar, M., Tickoo, O., 2020. Specifying Weight Priors in Bayesian Deep Neural Networks with Empirical Bayes. AAAI 34, 4477–4484. https://doi.org/10.1609/aaai.v34i04.5875.

LaBonte, T., Martinez, C., Roberts, S.A., 2020. We Know Where We Don't Know: 3D Bayesian CNNs for credible geometric uncertainty. https://doi.org/10.48550/arXiv.1910.10793.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 6405–6416.

Lee, H., Li, W., 2024. Improving interpretability of deep active learning for flood inundation mapping through class ambiguity indices using multi-spectral satellite imagery. Remote Sens. Environ. 309, 114213. https://doi.org/10.1016/j.rse.2024.114213.

Liu, X., Hu, B., Jin, L., Han, X., Xing, F., Ouyang, J., Lu, J., El Fakhri, G., Woo, J., 2021. Domain Generalization under Conditional and Label Shifts via Variational Bayesian Inference. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. Presented at the Thirtieth International Joint Conference on Artificial Intelligence {IJCAI-21}, International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, pp. 881–887.

Lynch, S.M., 2005. Bayesian Statistics. In: Kempf-Leonard, K. (Ed.), Encyclopedia of Social Measurement. Elsevier, New York, pp. 135–144. https://doi.org/10.1016/B0-12-369398-5/00156-0.

Malinin, A., Gales, M., 2018. Predictive uncertainty estimation via prior networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18. Curran Associates Inc., 7058, p. 7047.

Malinin, A., Gales, M., 2019. Reverse KL-divergence training of prior networks: improved uncertainty and adversarial robustness. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, pp. 14547–14558.

Mobiny, A., Yuan, P., Moulik, S.K., Garg, N., Wu, C.C., Van Nguyen, H., 2021. DropConnect is effective in modeling uncertainty of Bayesian deep networks. Sci. Rep. 11, 5458. https://doi.org/10.1038/s41598-021-84854-x.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image Data Augmentation for Deep Learning. J. Big Data 6, 60. https://doi.org/10.1186/s40537-019-0197-0.

Szymczak, S., Backendorf, F., Bott, F., Fricke, K., Junghänel, T., Walawender, E., 2022. Impacts of Heavy and Persistent Precipitation on Railroad Infrastructure in July 2021: a Case Study from the Ahr Valley, Rhineland-Palatinate, Germany. Atmosphere 13, 1118. https://doi.org/10.3390/atmos13071118.

Thiagarajan, P., Khairnar, P., Ghosh, S., 2022. Explanation and use of uncertainty Quantified by Bayesian Neural Network Classifiers for Breast Histopathology Images. IEEE Trans. Medical Imag. 41, 815–825. https://doi.org/10.1109/TMI.2021.3123300.

Wani, O., Majszak, M., Hertel, V., Geiß, C., 2025. On the safe side: uncertainty awareness for hydroclimatic risk and loss aversion. PLOS Water 4, e0000387. https://doi.org/10.1371/journal.pwat.0000387.

Weber, E., Kané, H., 2020. Building disaster damage assessment in satellite imagery with multi-temporal fusion.

Wen, Y., Vicol, P., Ba, J., Tran, D., Grosse, R., 2018. Flipout: efficient pseudo-independent weight perturbations on mini-batches. https://doi.org/10.48550/arXiv.1803.04386.

Wilson, A.G., Izmailov, P., 2020. Bayesian deep learning and a probabilistic perspective of generalization. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20. Curran Associates Inc, Red Hook, NY, USA, pp. 4697–4708.

Xia, J., Yokoya, N., Adriano, B., Broni-Bediako, C., 2023. OpenEarthMap: a Benchmark Dataset for Global High-Resolution Land Cover Mapping. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Presented at the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 6243–6253. https://doi.org/10.1109/WACV56688.2023.00619.

Xiao, Z., Shen, J., Zhen, X., Shao, L., Snoek, C.G.M., 2021. A bit more Bayesian: Domain-Invariant Learning with uncertainty. Proceedings of Machine Learning Research 139.

Yu, B., Sun, Y., Hu, J., Chen, F., Wang, L., 2025. Post-disaster building damage assessment based on gated adaptive multi-scale spatial-frequency fusion network. Int. J. Appl. Earth Observat. Geoinform. 141, 104629. https://doi.org/10.1016/j.jag.2025.104629.

Zhang, C., Diao, C., 2023. A Phenology-guided Bayesian-CNN (PB-CNN) framework for soybean yield estimation and uncertainty analysis. ISPRS J. Photogram. Remote Sens. 205, 50–73. https://doi.org/10.1016/j.isprsjprs.2023.09.025.

Zheng, Z., Zhong, Y., Wang, J., Ma, A., Zhang, L., 2021. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: from natural disasters to man-made disasters. Remote Sens. Environ. 265, 112636. https://doi.org/10.1016/j.rse.2021.112636.