



Integrating predictive uncertainty into satellite-based flood mapping

Chi-Ju Chen¹, Chun-Hao Yang², Wei Weng³, and Li-Pen Wang^{1,4}

¹Department of Civil Engineering, National Taiwan University, Taipei, 106, Taiwan

²Institute of Statistics and Data Science, National Taiwan University, Taipei, 106, Taiwan

³Department of Department of Geography, National Taiwan University, Taipei, 106, Taiwan

⁴Department of Civil and Environmental Engineering, Imperial College London, London, SW7 2AZ, UK

Correspondence: Chun-Hao Yang (chunhaoy@ntu.edu.tw), Wei Weng (weiweng@ntu.edu.tw), and Li-Pen Wang (lpwang@ntu.edu.tw)

Abstract. Flood extent maps derived from satellite imagery are increasingly important for flood response and risk management. Although recent deep learning models have substantially improved segmentation accuracy, they typically provide deterministic predictions without indicating their reliability. This limits their operational value in regions affected by cloud contamination, spectrally ambiguous floodwater, or environmental conditions insufficiently represented in the training data. This study develops an uncertainty-aware flood mapping framework by integrating Evidential Deep Learning (EDL) into an operational cloud-aware mapping workflow. The proposed framework simultaneously generates flood extent and pixel-level predictive uncertainty, enabling uncertainty information to be incorporated directly into the mapping process. Its performance is evaluated against a deterministic UNet++ baseline and two established uncertainty estimation approaches, Deep Ensemble and MC Dropout, using the extended WorldFloods dataset. Across the test dataset, the three uncertainty-aware methods achieved comparable IoU values, all consistently outperforming the deterministic baseline, indicating that predictive uncertainty can be incorporated without degrading segmentation performance. Their uncertainty behaviour and computational requirements, however, differed substantially. Among the evaluated approaches, the evidential formulation provided the most favourable operational trade-off by preserving segmentation performance, producing stable epistemic uncertainty estimates across diverse flood events, and requiring only a single model evaluation per image tile. These findings demonstrate that uncertainty quality should be regarded as a complementary evaluation criterion alongside segmentation accuracy. By combining competitive segmentation performance, informative uncertainty estimates, and computational efficiency, the proposed framework provides a practical framework for operational uncertainty-aware satellite-based flood mapping.

1 Introduction

Floods represent one of the most destructive natural hazards globally, posing severe threats to human life, infrastructure, and economic development (Doocy et al., 2013; Merz et al., 2021). To support flood-risk assessment, emergency response, and long-term adaptation planning, considerable effort has been devoted to the development of hydrological and hydrodynamic flood modelling systems capable of simulating flood generation and inundation processes (Bates and De Roo, 2000; Teng et al., 2017). At the same time, advances in Earth observation have enabled increasingly large-scale satellite-based flood



mapping and global flood databases derived directly from remote sensing observations (Tellman et al., 2021). Rather than replacing flood modelling, satellite-based flood mapping and flood models are often complementary, with remotely sensed inundation products supporting rapid situational awareness, model validation, calibration, and data assimilation within flood forecasting and inundation modelling frameworks (Bates, 2012). In this context, optical satellite imagery, particularly from Sentinel-2 and Landsat missions, has become an important data source for flood extent mapping because it provides relatively high spatial resolution, broad spatial coverage, and frequent revisit capability (Portalés-Julià et al., 2023; Tulbure et al., 2022). These characteristics make optical imagery especially valuable for rapid large-scale flood assessment and disaster-response applications.

Earlier optical flood mapping methods often relied on spectral-index approaches, such as the Normalized Difference Water Index (NDWI) and the Modified NDWI (McFEETERS, 1996; Xu, 2006). These methods are computationally efficient and easy to implement, but their performance may degrade under varying illumination conditions, terrain shadows, cloud contamination, dark surfaces, or spectrally complex floodwater conditions (Verpoorter et al., 2012; Liu et al., 2022; Memon et al., 2015). Many of these situations not only increase mapping difficulty but also introduce substantial ambiguity in the interpretation of optical observations. More recently, deep learning-based semantic segmentation approaches have emerged as a central framework for flood mapping because they can jointly exploit spectral information and spatial context to delineate flooded regions more effectively in complex scenes (Portalés-Julià et al., 2023; Mateo-Garcia et al., 2021; Isikdogan et al., 2017; Ronneberger et al., 2015; Zhou et al., 2018). However, despite these advances, uncertainty associated with flood predictions remains insufficiently addressed in most operational flood mapping tasks.

The development of large-scale datasets has further accelerated the adoption of deep learning for operational flood mapping. In particular, WorldFloods and the associated ml4floods framework have substantially improved the ability of segmentation models to generalize across diverse flood events, landscapes, and environmental conditions (Portalés-Julià et al., 2023; Mateo-Garcia et al., 2021). This progress has made cloud-aware global flood mapping increasingly feasible and has expanded the practical applicability of automated flood delineation systems. However, the increasing geographic and environmental diversity represented in global flood archives also increases the likelihood that models encounter conditions that are insufficiently represented during training. Such situations may lead to unreliable or overconfident predictions, particularly in visually ambiguous or previously unseen environments.

Despite recent progress, most existing flood segmentation models remain fundamentally deterministic. They typically provide a single prediction without explicitly quantifying predictive reliability, while modern deep neural networks are known to produce overconfident predictions even when they are incorrect or operating outside the training distribution (Guo et al., 2017; Gawlikowski et al., 2023; Sensoy et al., 2018). In operational flood-response settings, such behaviour may reduce the practical usefulness of automated flood maps because false alarms and missed inundation areas may lead to different downstream consequences for emergency response, manual verification, and risk assessment (Nadav-Greenberg and Joslyn, 2009; Schreck et al., 2024). Consequently, uncertainty quantification should not be regarded merely as an auxiliary model property, but rather as an important component of operational decision support.



Predictive uncertainty is commonly interpreted as two complementary components: aleatoric and epistemic uncertainty. The former represents ambiguity or noise inherent in the observations, while the latter reflects limitations in model knowledge arising from insufficient or unrepresentative training data (Nix and Weigend, 1994; Kendall and Gal, 2017; Gawlikowski et al., 2023; Schreck et al., 2024). In optical flood mapping, for example, aleatoric uncertainty may arise from factors such as cloud contamination, terrain shadows, or mixed land-water pixels, whereas epistemic uncertainty may become more critical in regions or environmental conditions insufficiently represented during training. Distinguishing these uncertainty sources can therefore provide additional insight into the reliability and limitations of flood segmentation models.

A range of methods has been proposed for uncertainty quantification in deep learning, among which Deep Ensembles and Monte Carlo Dropout (MC Dropout) are two widely used approaches in practice (Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016; Gawlikowski et al., 2023; Gao et al., 2024). Despite providing useful uncertainty estimates, these methods typically require either multiple independently trained models or repeated stochastic model evaluations during flood-map production. In large-scale flood monitoring systems that process large volumes of high-resolution satellite imagery, such requirements may substantially increase computational cost, processing time, memory demand, and operational implementation complexity (Portalés-Julià et al., 2023; Gao et al., 2024; Schreck et al., 2024). These limitations motivate the development of uncertainty-aware flood mapping approaches that require only one trained model and one model evaluation per image tile while providing effective uncertainty information.

To address this challenge, this study introduces a dense evidential framework for uncertainty-aware global optical flood extent segmentation. The main idea is to reformulate the existing cloud-aware flood mapping from deterministic pixel classification into pixel-level evidential inference. For each valid pixel and each binary prediction task, non-negative *evidence* is used to parametrise a task-specific Beta distribution, from which class probabilities and uncertainty measures can be derived after one evaluation of the trained model. UNet++ is used as the segmentation architecture for dense feature extraction, while the methodological contribution lies in the evidential probabilistic formulation that converts optical flood segmentation into a reliability-aware mapping product. To better understand the practical behaviour of evidential uncertainty in operational flood mapping, the proposed framework is evaluated against Deep Ensembles and MC Dropout methods under a unified experimental setting. Particular attention is given not only to segmentation performance, but also to uncertainty quality, uncertainty decomposition behaviour, as well as the practical trade-off between uncertainty-estimation strategies requiring one or repeated model evaluations.

The remainder of this paper is organised as follows. Section 2 describes the methodology, including the proposed uncertainty-aware framework, dataset, models and uncertainty decomposition. Section 3 presents the results and discussion. Section 4 concludes the findings of this work and the future direction.



2 Methodology

2.1 Uncertainty-aware flood mapping framework

90 Figure 1 illustrates the proposed uncertainty-aware flood mapping framework for optical satellite imagery. The framework extends the operational cloud-aware flood mapping workflow of Portalés-Julià et al. (2023) by incorporating Evidential Deep Learning (EDL) to estimate predictive uncertainty alongside flood segmentation. Starting from Sentinel-2 image tiles, the workflow first generates cloud and water probability maps together with their corresponding pixel-level uncertainty maps using a multi-output UNet++ model. A brightness reflectance map is subsequently derived from the visible bands and combined with
95 the predicted cloud and water probabilities through a deterministic post-processing procedure to produce a mutually exclusive land–water–cloud classification. In this way, the framework provides not only the final flood map but also spatially explicit uncertainty information, allowing users to assess the reliability of the predicted flood extent alongside the segmentation result.

The deterministic post-processing procedure is required because cloud and water are predicted independently through two binary segmentation tasks. Consequently, a pixel may simultaneously exhibit high cloud and high water probabilities, particularly
100 in visually ambiguous regions such as thin clouds, cloud shadows, or highly reflective floodwater surfaces. To ensure physically consistent flood maps, the independently predicted cloud and water probability maps are reconciled using the rule-based post-processing procedure proposed by Portalés-Julià et al. (2023), producing the final mutually exclusive land–water–cloud classification.

Beyond the final categorical flood map, the proposed framework retains the uncertainty information associated with the cloud
105 and water probability maps as additional spatial raster products. This uncertainty-aware design enables regions with potentially unreliable predictions, such as flood boundaries, cloud shadows, thin-cloud regions, spectrally ambiguous floodwater, or scenes insufficiently represented in the training data, to be distinguished from regions where the model is highly confident.

Because uncertainty is estimated directly within the segmentation model, the proposed framework requires only one trained model and one model evaluation per image tile while providing explicit uncertainty estimation. In contrast, MC Dropout
110 estimates uncertainty from repeated stochastic model evaluations, whereas Deep Ensemble combines multiple independently trained models; the proposed framework simultaneously generates both segmentation predictions and uncertainty estimates from a single model evaluation. As a result, the computational structure remains close to that of conventional deterministic inference while providing additional information regarding predictive reliability.

For each binary prediction task, the proposed framework derives three complementary uncertainty measures: (i) Dempster–
115 Shafer Theory (DST) uncertainty mass, representing the overall predictive uncertainty; (ii) aleatoric uncertainty, representing ambiguity arising from the observations; and (iii) epistemic uncertainty, representing uncertainty associated with model knowledge. Consequently, each Sentinel-2 tile produces two probability maps and six uncertainty maps, which are retained alongside the final segmentation output to support uncertainty-aware flood mapping and subsequent decision-support analysis.

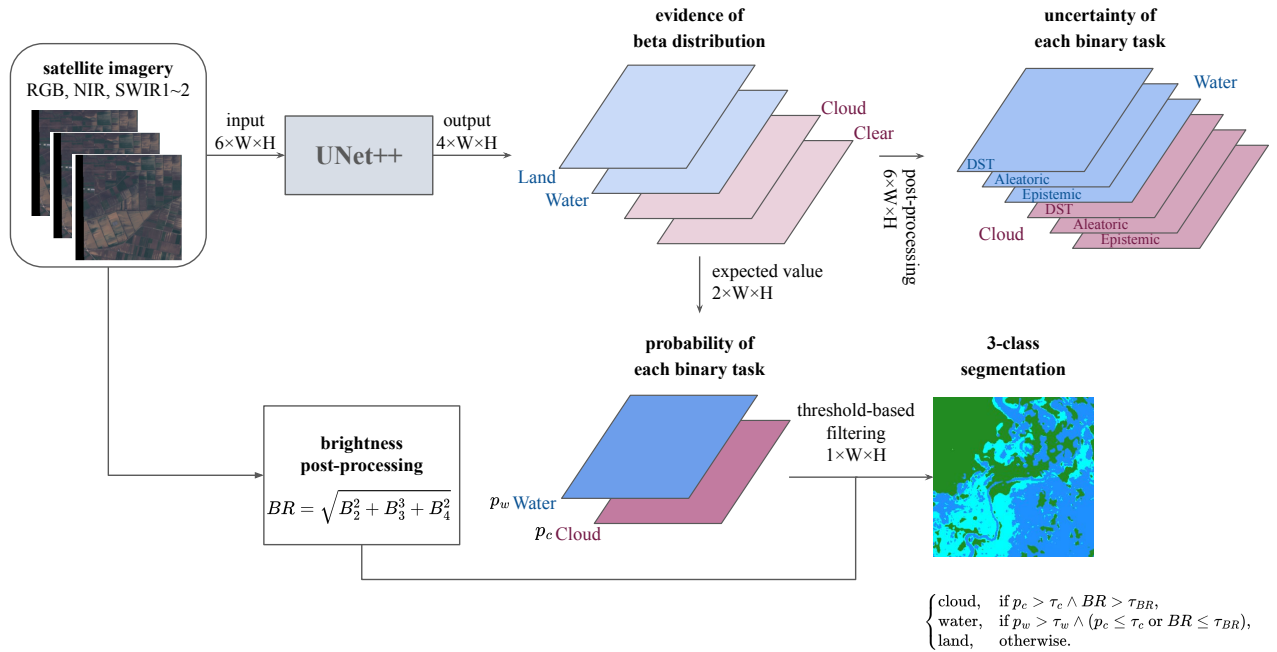


Figure 1. Processing workflow for flood mapping and uncertainty estimation from Sentinel-2 imagery. Sentinel-2 tiles are processed by a multi-output UNet++ to produce per-pixel water and cloud probabilities, which are combined to obtain mutually exclusive land/water/cloud classes; the evidential formulation additionally yields uncertainty measures (DST mass u , aleatoric, and epistemic uncertainty).

2.2 Dataset

To support uncertainty-aware flood mapping under diverse environmental and geographic conditions, this study uses the extended WorldFloods dataset for model development (Portalés-Julià et al., 2023; Mateo-Garcia et al., 2021; Image and Signal Processing (ISP), 2024). As described by Portalés-Julià et al. (2023), the extended archive builds upon the original WorldFloods dataset through additional event collection, quality control, and manual curation. The dataset contains 13-band Sentinel-2 Level-1C image patches co-located with reference flood masks from more than 180 flood events distributed across different geographic and hydrological settings. When tiled into 256×256 patches, the archive yields 73,809 image-mask pairs.

For model development and evaluation, this study follows the event-based train/validation/test split proposed by Portalés-Julià et al. (2023), such that the training, validation, and test subsets remain spatially and temporally independent. This strategy reduces data leakage across events while providing a more realistic assessment of model generalization to previously unseen flood conditions. The split statistics used in this study are summarised in Table 1.

In the extended WorldFloods labeling scheme, cloud and water are represented as two overlapping binary tasks rather than a single mutually exclusive multi-class mask (Portalés-Julià et al., 2023). Accordingly, separate reference masks are provided for

**Table 1.** Summary of the event-based split of the extended WorldFloods dataset used in this study (adapted from Portalés-Julià et al. (2023)).

Split	Flood events	Flood maps	256×256 patches
Training	119	475	65,582
Validation	14	16	3,524
Test	11	18	4,703
All	144	509	73,809

the clear/cloud and land/water tasks. This labelling strategy is consistent with the two-head segmentation framework adopted in this study, in which cloud and water predictions are learned independently before being combined through the post-processing procedure described in Section 2.1.

135 2.3 Evaluation framework for pixel-level uncertainty-aware flood mapping

This section defines the controlled evaluation framework used to evaluate pixel-level uncertainty-aware flood mapping. The framework is designed to isolate how alternative probabilistic formulations affect satellite-derived flood maps under the same map-production setting, rather than to assess different segmentation architectures. All models are trained and evaluated on the extended WorldFloods dataset described in Sect. 2.2 and use the same Sentinel-2 input configuration, event-based train–
 140 validation–test split, cloud-aware post-processing procedure, and evaluation metrics. This controlled design ensures that differences in segmentation performance, uncertainty reliability, and inference requirements can be attributed primarily to how each method represents predictive probabilities and uncertainty, rather than to changes in data handling, network architecture, or downstream processing.

Four inference strategies are evaluated: (i) a deterministic UNet++ baseline, (ii) the proposed dense evidential model based
 145 on Evidential Deep Learning (EDL) (Sensoy et al., 2018), (iii) a Deep Ensemble model (Lakshminarayanan et al., 2017), and (iv) an MC Dropout model (Gal and Ghahramani, 2016). These methods represent distinct approaches to predictive uncertainty in dense segmentation: deterministic point-probability prediction, evidential distributional prediction from a single trained model, ensemble inference combining multiple independently trained models, and stochastic inference based on repeated model evaluations. They were selected because deterministic baselines, ensembles, dropout-based sampling, and evi-
 150 dential models are widely used reference strategies for uncertainty estimation in deep learning (Gawlikowski et al., 2023). The evaluation is organised around three operationally relevant aspects: (i) the representation of water/land and cloud/clear probabilities, (ii) the availability and interpretation of pixel-level uncertainty maps, and (iii) the computational cost implications of requiring one or repeated model evaluations during flood-map production.

UNet++ is adopted as a shared feature-extraction architecture across all evaluated models to control for architectural dif-
 155 ferences and ensure a consistent evaluation setting for uncertainty estimation approaches. Consequently, the methodological contribution of this study lies not in the segmentation architecture itself, but in the evidential probabilistic formulation, uncertainty decomposition, and training strategy that enable uncertainty-aware flood mapping. The following subsections first



describe the common segmentation architecture, followed by the evidential output representation, the corresponding training objective, the decomposition of predictive uncertainty into aleatoric and epistemic components, and the model configuration and training settings of the evaluated uncertainty estimation methods.

2.3.1 Shared segmentation architecture: UNet++

This study adopts UNet++ (Nested U-Net) as the common segmentation architecture for all evaluated models because of its demonstrated effectiveness for optical flood mapping and dense image segmentation tasks (Zhou et al., 2018; Portalés-Julià et al., 2023). UNet++ extends the original U-Net architecture through redesigned skip connections that improve multi-scale feature fusion and segmentation accuracy (Zhou et al., 2018). U-Net-family architectures have been successfully applied to cloud-aware optical flood mapping, particularly in multi-output settings that separately predict water and cloud probabilities (Portalés-Julià et al., 2023).

The model inputs consist of six Sentinel-2 spectral bands, namely the red, green, blue (RGB), near-infrared (NIR), and two short-wave infrared (SWIR1 and SWIR2) bands. These spectral channels jointly capture surface reflectance characteristics associated with water, vegetation, clouds, and other land-cover types, providing complementary information for flood extent delineation (McFEETERS, 1996; Xu, 2006; Portalés-Julià et al., 2023). Following the cloud-aware segmentation strategy of Portalés-Julià et al. (2023), the network employs two independent output heads corresponding to the clear/cloud and land/water binary classification tasks.

Across the deterministic, EDL, Deep Ensemble, and MC Dropout models, the UNet++ architecture and input-output configuration are kept identical. Consequently, differences in segmentation performance and uncertainty characteristics can be attributed primarily to the adopted uncertainty modelling strategy rather than architectural variations. The following subsections describe the evidential formulation used to parametrise predictive probabilities and uncertainty within this common segmentation framework.

2.3.2 Evidential output representation

Conventional segmentation models directly predict a class probability for each pixel but provide little indication of how reliable that probability is. Consequently, two pixels with the same predicted flood probability may represent very different levels of model confidence. The proposed Evidential Deep Learning (EDL) formulation addresses this limitation by representing each pixel-level prediction as a probability distribution rather than a single probability value. This probabilistic representation enables both the predicted class probability and the associated predictive uncertainty to be inferred from a single model evaluation. In this study, the evidential formulation is applied independently to the two binary segmentation tasks, namely the water/land and cloud/clear prediction heads.

Rather than directly predicting sigmoid probabilities, each prediction head first estimates non-negative *evidence* for the two candidate classes. Within the EDL framework, evidence represents the amount of support the model assigns to each class based on the input image. Larger evidence indicates that the model has stronger confidence in the corresponding prediction, whereas little or no evidence indicates that the available image information provides only weak support. For example, pixels affected by



cloud shadows, thin clouds, or spectrally ambiguous floodwater are expected to accumulate less evidence than pixels containing clearly identifiable floodwater. Let z_{ik} denote the raw output of the prediction head for pixel i and class k . Following Sensoy et al. (2018), non-negative evidence is obtained using a ReLU activation,

$$e_{ik} = \text{ReLU}(z_{ik}) \geq 0, \quad (1)$$

195 where larger evidence values indicate stronger support for the corresponding class.

The predicted evidence is subsequently converted into the parameters of a Beta distribution,

$$\alpha_{ik} = e_{ik} + 1. \quad (2)$$

Adding one to each evidence value represents the absence of prior class-specific evidence. Consequently, when no evidence is accumulated ($e_{ik} = 0$ for all classes), the resulting Beta distribution becomes uniform, indicating complete uncertainty because
200 neither class is favoured. Since both segmentation tasks considered in this study are binary, the general Dirichlet formulation adopted by EDL reduces naturally to the Beta distribution. The complete Dirichlet probability density function is provided in Appendix A for completeness.

For each pixel, the total amount of accumulated evidence is defined as

$$S_i = \sum_{k=1}^K \alpha_{ik}, \quad (3)$$

205 where S_i (also referred to as the Dirichlet strength) quantifies how strongly the predicted probabilities are supported by the model. The predicted probability of class k is obtained directly from the corresponding Beta distribution as

$$\hat{p}_{ik} = \frac{\alpha_{ik}}{S_i}. \quad (4)$$

Unlike deterministic segmentation models, where every predicted probability is treated equally regardless of confidence, the evidential formulation explicitly quantifies how strongly each probability is supported by the available image evidence. A
210 larger value of S_i indicates that the predicted probability is supported by stronger evidence, whereas smaller values indicate that the prediction is associated with limited model confidence.

Within the EDL framework, the overall predictive uncertainty is quantified using the Dempster–Shafer Theory (DST) uncertainty mass (Jsang, 2018; Sensoy et al., 2018; Schreck et al., 2024),

$$u_i = \frac{K}{S_i}, \quad (5)$$



215 where K is the number of classes ($K = 2$ for both binary segmentation tasks). When no evidence is accumulated ($S_i = K$), the uncertainty mass reaches its maximum value ($u_i = 1$). As the amount of accumulated evidence increases, S_i becomes larger and the uncertainty mass decreases accordingly. Consequently, pixels with weak image support, such as those located near flood boundaries or within cloud-contaminated regions, are naturally associated with higher predictive uncertainty than pixels with clear spectral signatures. In the proposed framework, this uncertainty mass provides the overall predictive uncertainty, 220 which is further decomposed into aleatoric and epistemic components in Sect. 2.3.4.

2.3.3 Evidential training objective

The evidential output representation described above enables the model to estimate both predictive probabilities and uncertainty from a single model evaluation. However, this representation is only useful if the predicted uncertainty is consistent with the available image evidence. In other words, the model should produce confident predictions when the image provides clear 225 information, while remaining uncertain when the available evidence is ambiguous or insufficient. To achieve this behaviour, the proposed framework adopts the Evidential Deep Learning (EDL) training objective introduced by Sensoy et al. (2018). The objective combines a data-fitting term that encourages accurate predictions with a regularization term that discourages the model from becoming overconfident when the supporting evidence is weak.

The first component is the Bayes-risk (squared-error) loss, which measures the discrepancy between the expected class 230 probabilities induced by the Beta distribution and the reference labels. Following Sensoy et al. (2018), the resulting objective has the closed-form expression

$$L_{\text{fit}}(\Theta) = \sum_{j=1}^K \left[(y_{ij} - \hat{p}_{ij})^2 + \frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{S_i + 1} \right], \quad (6)$$

where y_{ij} denotes the one-hot reference label. The first term encourages the predicted probabilities to agree with the reference labels, while the second term accounts for the predictive variance of the evidential distribution. Consequently, the model 235 is encouraged not only to make correct predictions, but also to avoid expressing high confidence when the available image evidence is limited. This property is particularly desirable in flood mapping, where cloud contamination, mixed land–water pixels, or spectrally ambiguous floodwater may prevent confident classification.

The second component is a Kullback–Leibler (KL) divergence regularization term,

$$L_{\text{KL}}(\Theta) = KL[D(\mathbf{p}_i | \tilde{\alpha}_i) \parallel D(\mathbf{p}_i | \mathbf{1})], \quad (7)$$

240 where $\tilde{\alpha}_i$ denotes the evidential parameters after removing the evidence associated with the ground-truth class. Intuitively, this regularization suppresses evidence that is not supported by the observations. For incorrectly classified pixels, it encourages the predicted distribution to move towards the non-informative uniform prior rather than maintaining unjustified confidence in the wrong class. As a result, the model learns to distinguish between being confidently correct and appropriately uncertain, thus improving the calibration of the predicted uncertainty.



245 The overall evidential training objective is therefore defined as

$$L(\Theta) = L_{\text{fit}}(\Theta) + \lambda_t L_{\text{KL}}(\Theta), \quad (8)$$

where λ_t is an annealing coefficient that gradually increases during training. This progressive weighting prevents the regularization term from dominating the early optimisation stages, allowing the model to first learn discriminative image features before progressively refining the associated uncertainty estimates. In practice, the model therefore first learns *what* to predict
250 before learning *how confident* it should be in those predictions.

The original EDL objective was developed for independent classification problems. In this study, it is extended to dense cloud-aware flood segmentation by treating every valid pixel as an individual evidential training sample within the cloud/clear and water/land prediction tasks. The Bayes-risk and KL-divergence terms are computed independently for the two prediction heads using their respective Beta parameters and reference labels. Pixels without valid annotations are excluded through va-
255 lidity masks, while class imbalance is addressed using task-specific class weights. This adaptation enables the original EDL formulation to be applied directly to pixel-level uncertainty-aware flood mapping while preserving the probabilistic interpretation of the evidential framework.

2.3.4 Uncertainty decomposition

The uncertainty mass introduced in Sect. 2.3.2 provides a single measure of the overall predictive uncertainty associated
260 with each pixel. However, it does not distinguish whether the uncertainty arises from ambiguity in the observations or from limitations in the model itself. For uncertainty-aware flood mapping, this distinction is valuable because these two sources of uncertainty have different physical interpretations and suggest different operational responses. Aleatoric uncertainty indicates that the available satellite observations are inherently ambiguous, whereas epistemic uncertainty indicates that the model lacks sufficient knowledge to make a reliable prediction. Distinguishing between these two uncertainty sources therefore provides
265 more informative guidance for subsequent decision making than a single overall uncertainty measure. To provide further insight into the origin of predictive uncertainty, this study adopts the law of total variance (LoTV) to decompose the overall predictive uncertainty into aleatoric and epistemic components (Casella and Berger, 2024; Schreck et al., 2024).

The decomposition follows naturally from the law of total variance,

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X]), \quad (9)$$

270 where the first term represents the expected conditional variance and the second term represents the variance of the conditional expectation. Within the evidential framework, these two components are naturally interpreted as aleatoric uncertainty and epistemic uncertainty, respectively (Sensoy et al., 2018; Schreck et al., 2024). In the context of flood mapping, aleatoric uncertainty reflects ambiguity inherent in the satellite observations, such as cloud contamination, mixed land–water pixels, or spectrally ambiguous floodwater, whereas epistemic uncertainty reflects uncertainty arising from limited model knowledge,
275 for example when encountering environmental conditions that are insufficiently represented in the training data.



Under the evidential formulation, let $y_{i,k}$ denote the one-hot reference label for class k at pixel i , $p_{i,k}$ the corresponding class probability, $\alpha_{i,k}$ the Beta/Dirichlet concentration parameter, and $S_i = \sum_k \alpha_{i,k}$ the total accumulated evidence. The two uncertainty components are then obtained directly from the predicted evidential parameters as

$$\mathbb{E}[\text{Var}(y_{i,k} | p_{i,k})] = \mathbb{E}[p_{i,k}(1 - p_{i,k})] \quad (10)$$

$$= \frac{\alpha_{i,k}}{S_i} \left(1 - \frac{\alpha_{i,k}}{S_i}\right) \frac{S_i}{S_i + 1}, \quad (11)$$

$$\text{Var}(\mathbb{E}[y_{i,k} | p_{i,k}]) = \text{Var}(p_{i,k}) \quad (12)$$

$$= \frac{\alpha_{i,k}}{S_i} \left(1 - \frac{\alpha_{i,k}}{S_i}\right) \frac{1}{S_i + 1}, \quad (13)$$

which correspond to the pixel-level estimates of aleatoric and epistemic uncertainty, respectively.

In the proposed framework, these two uncertainty components are computed independently for both the cloud/clear and water/land prediction tasks and retained together with the overall DST uncertainty mass. Consequently, each pixel is characterised not only by its predicted flood probability and overall uncertainty, but also by the dominant source of that uncertainty. This decomposition enables regions affected primarily by ambiguous image observations to be distinguished from those where the model lacks sufficient knowledge, thereby providing additional information for interpreting model behaviour and supporting uncertainty-aware operational flood mapping.

2.3.5 Model configuration and training settings

To ensure a controlled evaluation of the different uncertainty estimation approaches, all models are implemented under a common experimental setting. Unless otherwise stated, the deterministic baseline, EDL, Deep Ensemble, and MC Dropout models share the same UNet++ architecture, Sentinel-2 input configuration, data augmentation procedure, event-based train-validation-test split, and post-processing strategy described in Sections 2.1 and 2.3. Consequently, differences in segmentation performance and uncertainty characteristics can be attributed primarily to the adopted uncertainty modelling strategy rather than differences in network architecture or data processing.

The deterministic baseline, Deep Ensemble, and MC Dropout models are trained using binary cross-entropy (BCE) losses for the cloud/clear and water/land prediction tasks. The proposed EDL model instead employs the evidential training objective described in Sect. 2.3.3. To account for class imbalance, task-specific class weights are applied during training. For the BCE-based models, cloud pixels are weighted six times more heavily than clear pixels, while water pixels are weighted six times more heavily than land pixels. For the EDL model, cloud pixels are weighted twenty times more heavily than clear pixels, whereas the water/land weighting remains unchanged. These class weights affect only the optimisation process; all inference and post-processing settings are kept identical across the four evaluated models.

Method-specific configurations are introduced only where required for uncertainty estimation. The Deep Ensemble model consists of ten independently trained network members, whose predictions are combined during inference to estimate predictive uncertainty. For the MC Dropout model, dropout is retained during model evaluation with a dropout rate of 0.5, and



predictive uncertainty is estimated from ten stochastic model evaluations. In contrast, the proposed EDL model directly produces predictive probabilities and uncertainty estimates from one model evaluation, without requiring repeated evaluations or multiple trained models.

310 3 Results and discussion

3.1 Evaluation methodology and metrics

The evaluation is designed to address three complementary questions. First, can uncertainty-aware prediction be incorporated without reducing flood segmentation performance relative to deterministic inference under both aggregated and event-specific evaluations? Second, do the estimated uncertainty maps provide meaningful information for identifying unreliable predictions?
315 Third, what trade-offs exist between segmentation accuracy, uncertainty quality, and computational efficiency for operational flood mapping?

To answer these questions, the evaluation is conducted at three spatial aggregation levels: the overall test dataset, individual flood events, and representative case studies. Overall statistics provide a summary of average model behaviour across the complete test dataset, while event-level analyses reveal how performance varies under different flood scenarios. Representative
320 case studies are subsequently used to provide qualitative insight into the spatial characteristics of the predicted flood maps and uncertainty distributions. This hierarchical evaluation avoids over-reliance on a single aggregated metric and enables both statistical and visual assessment of model behaviour.

Flood segmentation performance is evaluated using Accuracy, Precision, Recall, F1 score, and Intersection over Union (IoU). Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. These metrics
325 are defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (14)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (15)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (16)$$

$$\text{F1} = \frac{2\text{Precision Recall}}{\text{Precision} + \text{Recall}}, \quad (17)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \quad (18)$$

Among these metrics, the F1 score and IoU are treated as the primary indicators because they directly quantify the agreement between the predicted and reference flood masks, whereas Accuracy can be overly influenced by the predominance of non-flood pixels.

The quality of uncertainty estimation is evaluated using two complementary analyses. First, retention curves assess the
335 uncertainty ranking capability by examining whether pixels assigned higher uncertainty correspond preferentially to less re-



liable predictions. Second, Patch Accuracy versus Patch Uncertainty (PAvPU) quantifies the consistency between prediction correctness and uncertainty, thereby providing a direct measure of the practical usefulness of uncertainty for decision support.

PAvPU is defined as

$$\text{PAvPU} = \frac{n_{ac} + n_{iu}}{n_{ac} + n_{au} + n_{ic} + n_{iu}}, \quad (19)$$

340 where n_{ac} , n_{au} , n_{ic} , and n_{iu} denote accurate-and-certain, accurate-and-uncertain, inaccurate-and-certain, and inaccurate-and-uncertain outcomes, respectively. Rather than adopting universal thresholds, the IoU and uncertainty thresholds are determined independently for each model and uncertainty definition using the point on the corresponding retention curve closest to a retention rate of 0.9. The selected IoU threshold is used to distinguish accurate from inaccurate 3×3 patches, while the corresponding uncertainty threshold separates certain from uncertain predictions. Patches containing invalid pixels or undefined
345 IoU values are excluded from the analysis.

Together, these complementary evaluations provide a comprehensive assessment of the evaluated approaches from three perspectives: segmentation accuracy, uncertainty reliability, and operational applicability.

3.2 Segmentation performance

3.2.1 Overall and cross-event performance

350 We first assess whether uncertainty-aware prediction can be incorporated without reducing flood segmentation performance at both the overall test-set level and the individual event level. The overall metrics provide an aggregated assessment of average model behaviour, whereas the event-level analysis examines how performance varies across different flood scenarios.

Table 2 summarises the overall segmentation performance on the test dataset. The deterministic UNet++ baseline obtains an IoU of 0.7992, while the uncertainty-aware methods all remain within a narrow IoU range above this baseline. These values
355 indicate that, before any uncertainty-based filtering, the three uncertainty-aware methods achieve broadly similar overlap-based segmentation performance. Thus, none of the evaluated uncertainty-aware strategies degrades overall IoU relative to the deterministic baseline. Under the one-evaluation setting, the proposed EDL model also improves the F1 score from 0.8884 to 0.8961, indicating that uncertainty estimation can be incorporated without reducing overlap-based segmentation performance in this experimental setting.

360 The event-level metrics further demonstrate that segmentation performance varies across flood events (Appendix Table A1). More importantly, EDL records IoU values above the deterministic baseline in 13 of the 18 events, indicating that the evidential formulation generally preserves or improves segmentation performance rather than introducing a systematic IoU penalty.

3.2.2 Representative case study

To complement the quantitative evaluation, we examine the representative flood event EMSR466_AOI01_DEL_PRODUCT.
365 This event contains a complex combination of river channels, fragmented inundation patterns, and surrounding land-cover



Table 2. Overall segmentation performance on the test dataset.

Model evaluation scheme	Method	Accuracy	Precision	Recall	F1	IoU
One evaluation	Deterministic UNet++ baseline	0.9269	0.8303	0.9553	0.8884	0.7992
One evaluation	EDL	0.9342	0.8626	0.9323	0.8961	0.8117
Repeated evaluations	Deep Ensemble	0.9433	0.9057	0.9083	0.9070	0.8299
Repeated evaluations	MC Dropout	0.9404	0.9014	0.9029	0.9022	0.8218

heterogeneity, providing a challenging scenario for flood segmentation. Rather than serving as evidence on its own, this case study illustrates the qualitative characteristics underlying the overall statistical patterns.

Figure 2 presents the corresponding confusion maps, while Table 3 summarises the segmentation metrics. The proposed EDL model produces a spatially coherent flood extent along the main inundated region. The case also shows that local omission and
 370 commission patterns can differ from aggregate dataset-level metrics, illustrating that dataset-level metrics do not fully describe local spatial behaviour.

Consistent with these visual observations, Table 3 shows that EDL retains robust Accuracy, F1 score, and IoU for this event. This case is used to illustrate local spatial behaviour rather than to establish a general ranking among uncertainty-estimation strategies.

375 3.2.3 Discussion

Taken together, the overall and event-level evaluations suggest that uncertainty-aware prediction does not necessarily reduce flood-segmentation performance. All three uncertainty-aware methods preserve or improve overall IoU relative to the deterministic baseline, while event-level outcomes vary with flood-scene characteristics. This result is important because uncertainty estimation can sometimes introduce a segmentation-performance penalty, whereas the evaluated methods retain overlap-based
 380 accuracy under the controlled setting used here.

Within the setting of one model evaluation per tile, the proposed EDL framework consistently improves upon the deterministic UNet++ baseline while maintaining an identical network architecture and processing cost. This indicates that the observed performance gains arise primarily from the evidential probabilistic formulation rather than architectural modifications. The representative case study further demonstrates that EDL can provide reliable segmentation quality in spatially complex flood
 385 scenes, motivating the subsequent analysis of whether these improvements are accompanied by more informative uncertainty estimates.

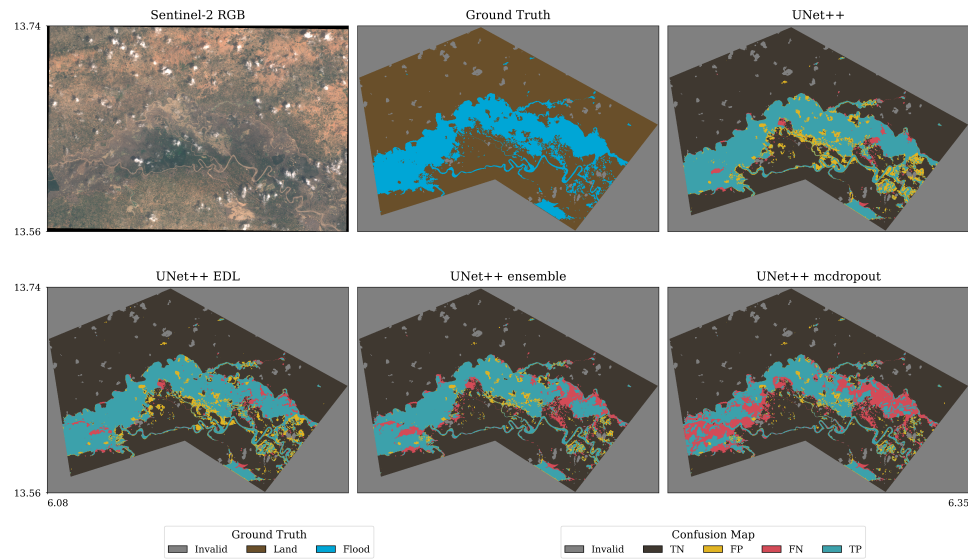


Figure 2. Confusion maps for the representative event EMSR466_AOI01_DEL_PRODUCT.

Table 3. Segmentation performance for the representative event EMSR466_AOI01_DEL_PRODUCT.

Model	Accuracy	Precision	Recall	F1	IoU
Deterministic UNet++ baseline	0.9267	0.8060	0.9442	0.8696	0.7694
EDL	0.9301	0.8301	0.9176	0.8717	0.7725
Deep Ensemble	0.9231	0.8800	0.8142	0.8458	0.7328
MC Dropout	0.8746	0.8632	0.6126	0.7166	0.5583

3.3 Uncertainty evaluation

Beyond segmentation accuracy, the central objective of this study is to evaluate whether the estimated uncertainty maps provide meaningful information regarding prediction reliability. We therefore assess uncertainty quality from two complementary perspectives. Retention curves evaluate the ability of each uncertainty measure to rank unreliable predictions, while PAvPU quantifies the consistency between prediction correctness and uncertainty. Together, these analyses assess whether uncertainty estimates are informative for uncertainty-aware flood mapping and downstream decision support.

3.3.1 Uncertainty ranking behaviour

Figure 3 presents the retention curves derived from aleatoric uncertainty. For all three uncertainty-aware models, the segmentation metrics improve as progressively more uncertain pixels are removed, indicating that higher uncertainty is generally associated with less reliable predictions. This behaviour indicates that uncertainty estimates can help identify less reliable predictions under aleatoric uncertainty.



The retention-curve shapes differ across uncertainty-estimation strategies, indicating that uncertainty-ranking behaviour depends on both the uncertainty source and the way uncertainty is estimated. When epistemic uncertainty is considered, the EDL retention curves remain relatively stable across the retention range (Fig. 4), indicating that epistemic uncertainty provides a consistent ranking signal within the evidential formulation.

Figure 5 further examines the overall DST uncertainty of EDL together with the combined aleatoric-plus-epistemic uncertainty adopted for the repeated-evaluation methods. The close agreement between the DST and combined uncertainty curves indicates that the evidential formulation preserves a consistent uncertainty ranking across different uncertainty representations.

3.3.2 Overall PAvPU

Table 4 summarises the PAvPU values computed across the complete test dataset. The PAvPU values vary by uncertainty definition and should therefore be interpreted as diagnostics of uncertainty behaviour rather than as a single overall method ranking. For EDL, the corresponding values remain consistent across DST, aleatoric, epistemic, and combined uncertainty measures. Because the uncertainty-aware methods show broadly similar overall IoU before uncertainty-based filtering, this stability in uncertainty quantification becomes an important operational distinction.

The event-level PAvPU results reinforce this observation by showing stable epistemic uncertainty behaviour for EDL across diverse flood events. These results suggest that the principal strength of EDL lies in providing a stable and reliable characterisation of epistemic uncertainty across diverse flood scenarios.

3.3.3 Representative case study

To better understand the quantitative results, we revisit the representative event EMSR466_AOI01_DEL_PRODUCT. Figure 6 shows the uncertainty maps generated by the three uncertainty-aware models, while Table 5 summarises the corresponding PAvPU values.

The uncertainty distributions differ substantially among the evaluated methods. The EDL uncertainty maps exhibit clear spatial correspondence with prediction errors, particularly along flood boundaries and visually ambiguous regions. Consistent with these observations, EDL maintains similar PAvPU values across DST, aleatoric, epistemic, and combined uncertainty measures in this representative event. This behaviour agrees with the overall quantitative evaluation, indicating that the evidential formulation produces uncertainty estimates that remain consistently aligned with prediction correctness.

3.3.4 Local uncertainty analysis

To further examine uncertainty behaviour at finer spatial scales, we analyse a representative local window extracted from the same flood event (Fig. 7). The enlarged view highlights complex flood boundaries and fragmented inundated regions where uncertainty estimation becomes particularly important.



Figure 7 and Table 6 show that the uncertainty maps produced by EDL remain closely aligned with local prediction errors, particularly for epistemic uncertainty. These observations further support the interpretation that EDL provides a spatially consistent representation of model uncertainty in challenging flood environments.

430 3.3.5 Discussion

The uncertainty evaluation reveals behaviour that depends on the uncertainty source, evaluation metric, and spatial aggregation level. In this study, EDL demonstrates a stable representation of epistemic uncertainty across both quantitative evaluations and representative case studies. Since epistemic uncertainty reflects uncertainty arising from limited model knowledge, this property is particularly relevant when flood mapping models are deployed in previously unseen geographic regions or environmental
435 conditions.

Taken together, the retention-curve analysis, PAVPU evaluation, and spatial case studies indicate that the proposed EDL framework provides consistent and interpretable uncertainty estimates while retaining the computational efficiency obtained by requiring only one model evaluation. This finding provides the foundation for the operational discussion presented in the following subsection.

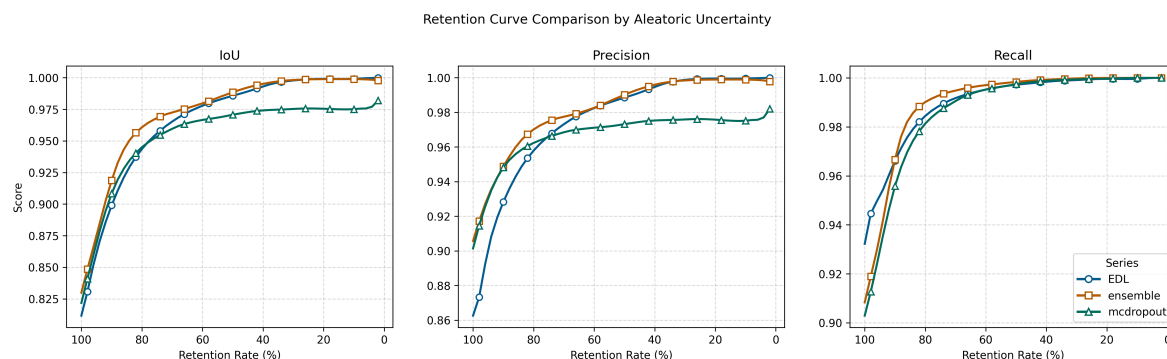


Figure 3. Retention curves under aleatoric uncertainty. From left to right, the panels show IoU, Precision, and Recall.

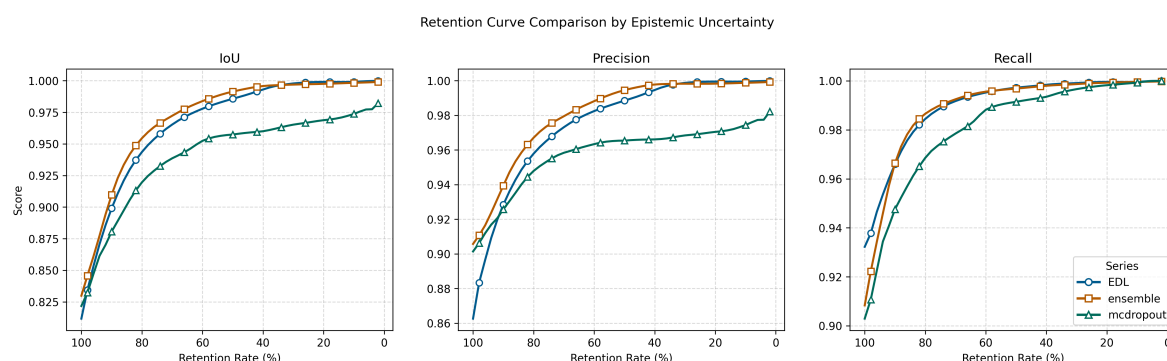


Figure 4. Retention curves under epistemic uncertainty. From left to right, the panels show IoU, Precision, and Recall.



Table 4. Overall PAvPU values across uncertainty definitions.

Model evaluation scheme	Method	Uncertainty type	PAvPU
One evaluation	EDL	DST uncertainty	0.8386
One evaluation	EDL	Aleatoric	0.8388
One evaluation	EDL	Epistemic	0.8392
One evaluation	EDL	Aleatoric + Epistemic	0.8389
Repeated evaluations	Deep Ensemble	Aleatoric	0.8483
Repeated evaluations	Deep Ensemble	Epistemic	0.8091
Repeated evaluations	Deep Ensemble	Aleatoric + Epistemic	0.8461
Repeated evaluations	MC Dropout	Aleatoric	0.8484
Repeated evaluations	MC Dropout	Epistemic	0.8082
Repeated evaluations	MC Dropout	Aleatoric + Epistemic	0.8473

Table 5. Uncertainty reliability for the representative event EMSR466_AOI01_DEL_PRODUCT.

Model	PAvPU (DST)	PAvPU (Aleatoric)	PAvPU (Epistemic)	PAvPU (Aleatoric + Epistemic)
EDL	0.7205	0.7237	0.7211	0.7230
Deep Ensemble	–	0.6036	0.4941	0.5784
MC Dropout	–	0.6980	0.6611	0.7032

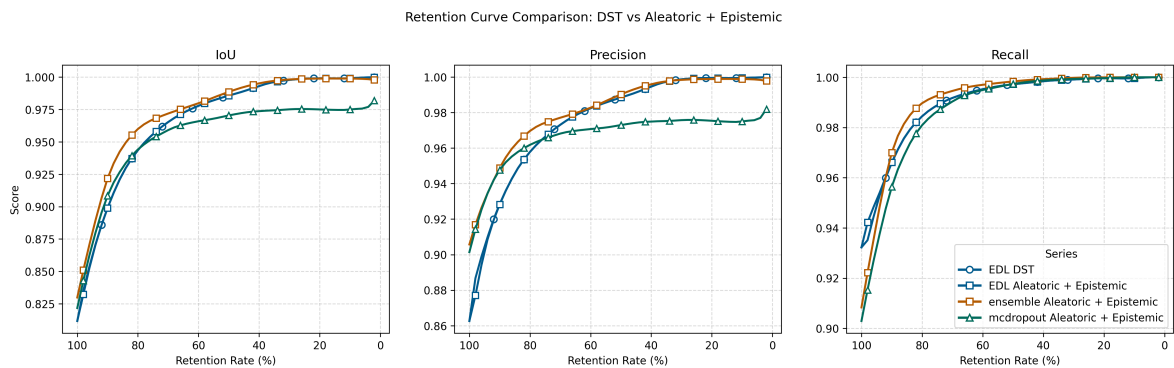


Figure 5. Retention curves for DST uncertainty and combined aleatoric plus epistemic uncertainty. From left to right, the panels show IoU, Precision, and Recall.

440 **3.4 Operational implications of uncertainty-aware flood mapping**

The preceding results demonstrate that uncertainty-aware flood mapping involves segmentation accuracy, uncertainty quality, and computational requirements. These considerations have important implications for operational flood mapping systems,

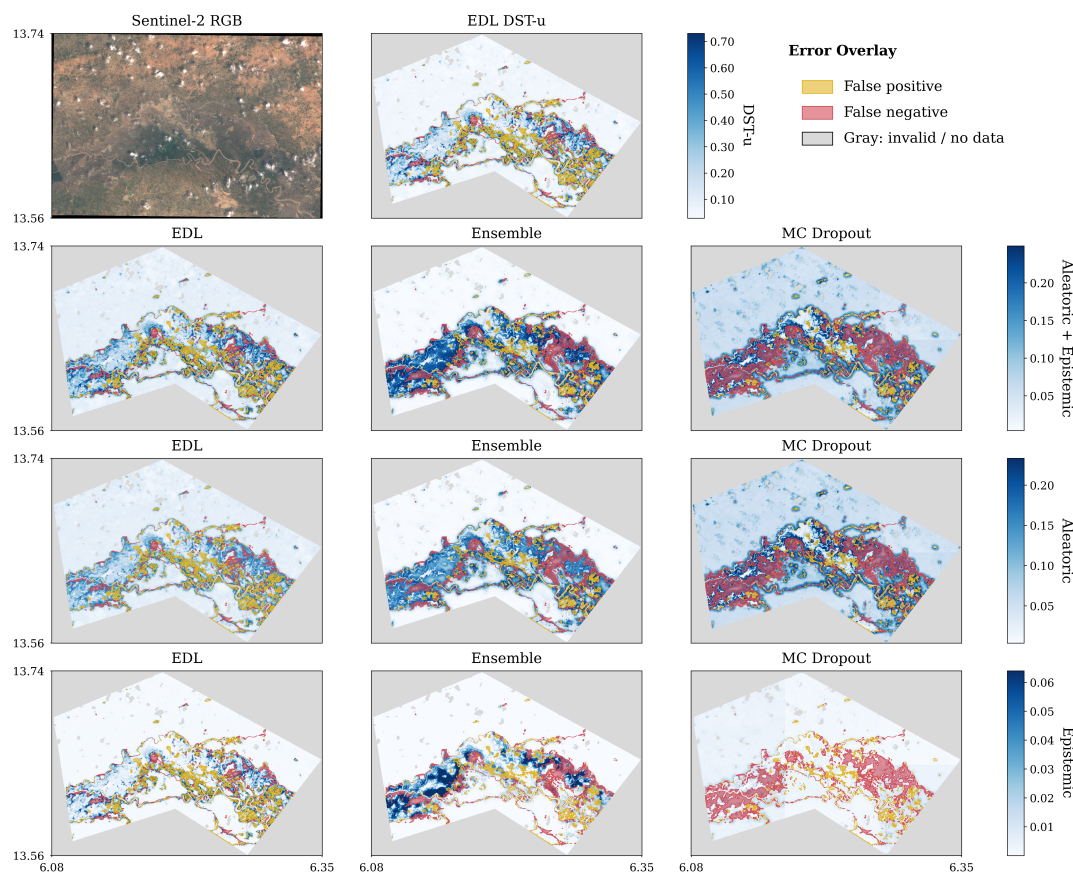


Figure 6. Uncertainty maps for the representative event EMSR466_AOI01_DEL_PRODUCT.

Table 6. Local-window PAvPU values for EMSR466_AOI01_DEL_PRODUCT_r1024_c1024_h0512_w0512.

Uncertainty type	EDL	Deep Ensemble	MC Dropout
DST PAvPU	0.8211	–	–
Aleatoric PAvPU	0.8218	0.7265	0.7437
Epistemic PAvPU	0.8218	0.5286	0.5041
Aleatoric + Epistemic PAvPU	0.8220	0.7301	0.7432

where uncertainty information must often be generated within strict time and resource constraints while remaining sufficiently reliable to support downstream decision making.

445 From a computational perspective, the proposed EDL framework preserves the efficiency of deterministic inference by generating flood predictions and uncertainty estimates simultaneously within one evaluation of the segmentation model. By contrast, both Deep Ensemble and MC Dropout require repeated model evaluations during map production, either through multiple independently trained models or multiple stochastic evaluations. These repeated-evaluation approaches also preserve

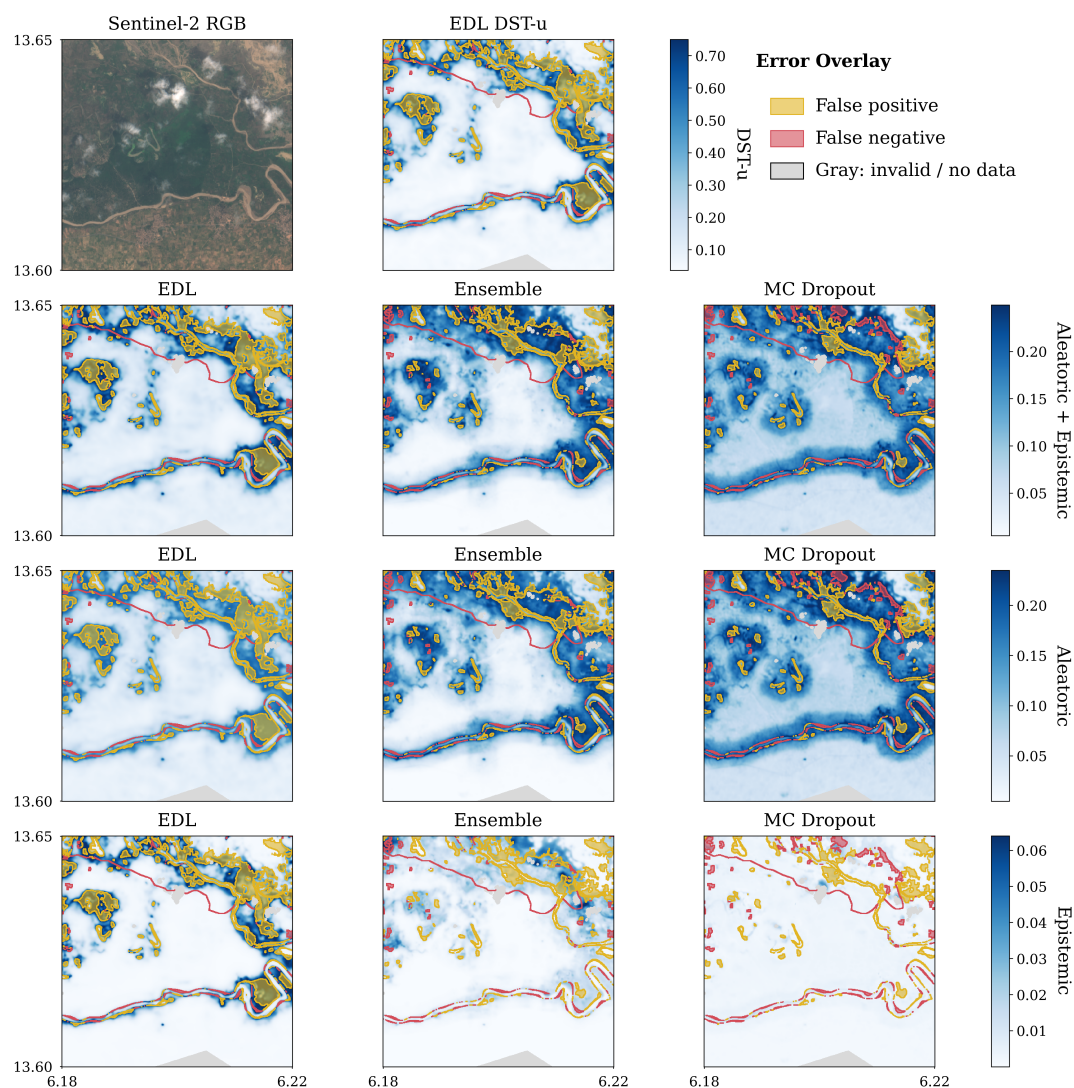


Figure 7. Local-window uncertainty maps for EMSR466_AOI01_DEL_PRODUCT_r1024_c1024_h0512_w0512.



Table 7. Method-level expected computational requirements during flood-map production. Here, P denotes the parameter count of a single UNet++ model, and K denotes the number of stochastic model evaluations or ensemble members. Relative costs are expressed under a sequential processing assumption and are intended as hardware-independent estimates rather than measured processing-time benchmarks.

Method	Stored model parameters	Model evaluations per tile	Expected relative computational cost	Expected operational implication
Deterministic UNet++	P	1	$1 \times$	Reference setting
EDL	$\sim P$	1	$\sim 1 \times$	Similar to deterministic inference with added uncertainty output
MC Dropout (K)	P	K	$\sim K \times$	Longer processing time due to repeated stochastic evaluations
Deep Ensemble (K)	$K \cdot P$	K	$\sim K \times$	Higher storage requirements and repeated model evaluations

or improve overall IoU relative to the deterministic baseline, but their computational requirements increase approximately
 450 linearly with the number of ensemble members or stochastic samples.

Table 7 summarises the expected computational characteristics of the four evaluated methods during flood-map production. Deterministic UNet++ and EDL both require one model evaluation for each Sentinel-2 tile, whereas the computational cost of MC Dropout and Deep Ensemble increases proportionally with the number of repeated model evaluations. Consequently, the proposed EDL framework provides uncertainty estimates with a computational complexity that remains close to conventional
 455 deterministic segmentation, making it particularly attractive for large-scale or time-sensitive flood mapping applications.

Beyond computational efficiency, the uncertainty maps themselves constitute valuable decision-support products rather than merely auxiliary visualisations. Regions associated with high predictive uncertainty can be prioritised for manual inspection, quality assurance, or subsequent refinement using complementary observations such as SAR imagery or hydrodynamic model outputs. In particular, the stable epistemic uncertainty behaviour observed for EDL supports its use for identifying regions
 460 where model knowledge is limited, thereby providing useful information for assessing prediction reliability when the model encounters unfamiliar environmental conditions.

Taken together, the results indicate that the choice of uncertainty estimation strategy depends on the intended operational objective. Although the three uncertainty-aware methods show broadly similar overall IoU values, their uncertainty behaviour and computational requirements differ substantially. In this context, the evidential formulation provides the most operationally
 465 favourable trade-off: it preserves overlap-based segmentation performance, produces stable epistemic uncertainty estimates, and requires only one model evaluation per image tile. For operational flood mapping systems that require reliable uncertainty information with limited additional processing cost, EDL therefore represents a practical and scalable alternative to uncertainty estimation methods that rely on repeated model evaluations.

4 Conclusions

470 This study presents an uncertainty-aware framework for optical satellite flood mapping by integrating Evidential Deep Learning (EDL) into an operational cloud-aware segmentation workflow. Unlike conventional deterministic flood mapping approaches, the proposed framework simultaneously produces flood segmentation and pixel-level uncertainty estimates from one model evaluation, enabling uncertainty information to be incorporated into the flood mapping process without increasing the compu-



tational cost of map production. The proposed framework was evaluated against deterministic UNet++, Deep Ensemble, and
475 MC Dropout under a unified experimental setting using the extended WorldFloods dataset.

The experimental results demonstrate that uncertainty estimation can be incorporated without degrading overlap-based flood
segmentation performance. At the overall test-dataset level, all uncertainty-aware methods preserve or improve IoU relative to
the deterministic baseline, with broadly similar IoU values among the three uncertainty-aware strategies before uncertainty-
based filtering. More importantly, the uncertainty evaluation shows that segmentation accuracy alone does not fully reflect the
480 usefulness of uncertainty estimation. Retention-curve analysis, PAvPU evaluation, and representative case studies collectively
indicate that the evidential formulation provides a stable and informative representation of epistemic uncertainty across diverse
flood events.

These findings highlight that segmentation performance and uncertainty quality should be regarded as complementary eval-
uation dimensions for operational flood mapping. Because the evaluated uncertainty-aware methods show broadly similar
485 overall IoU values, the stable uncertainty behaviour and one-evaluation computational structure of EDL become central oper-
ational advantages. Rather than replacing existing uncertainty estimation approaches in all settings, EDL provides a practical
alternative that requires only one model evaluation per image tile and is particularly well suited to large-scale or time-sensitive
flood mapping applications where both prediction reliability and computational efficiency are important. Although this study
demonstrates the framework in optical flood mapping, the pixel-level evidential formulation is not specific to flood extent
490 classification. In principle, the same formulation can be coupled with dense segmentation architectures for other pixel-level
classification and semantic segmentation tasks, including natural-hazard mapping, land-cover classification, and other remote-
sensing applications where both class labels and prediction reliability are required.

Several limitations should be acknowledged. First, the PAvPU analysis depends on the selected uncertainty threshold and
evaluation window size, and alternative settings may influence the absolute metric values. Second, although the relative com-
495 putational characteristics of the evaluated approaches follow directly from their inference strategies, this study does not provide
hardware-specific benchmarks of inference time or memory consumption. Third, the uncertainty information produced by the
proposed framework has not yet been incorporated into downstream flood mapping or decision-support workflows, and its
practical value in these applications remains to be fully evaluated.

Future work will investigate how uncertainty information can be exploited beyond uncertainty visualization. One promising
500 direction is uncertainty-guided multi-source flood mapping, where uncertainty estimates from optical imagery can be combined
with SAR observations, hydrodynamic simulations, or other complementary data sources to support more transparent and
robust probabilistic data fusion (Festa et al., 2026). More broadly, uncertainty-aware flood mapping may support downstream
decision-support workflows by indicating where additional observations, model refinement, or manual verification are most
needed. This provides a foundation for developing trustworthy Earth observation systems in which predictive uncertainty is
505 treated as an integral component of the mapping product rather than an auxiliary diagnostic generated after segmentation.



Appendix A: Full Dirichlet formulation for the evidential output representation

Section 2.3.2 introduces the evidential output representation using the Beta distribution, which is sufficient for the two binary segmentation tasks considered in this study. For completeness, this appendix presents the corresponding general Dirichlet formulation from which the Beta representation is derived.

510 The Dirichlet distribution is the conjugate prior of the categorical distribution and is defined over the probability simplex of a K -class probability vector $\mathbf{p} = (p_1, \dots, p_K)$. It is parameterized by the concentration parameter vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$, whose elements may be interpreted as pseudo-counts or accumulated evidence supporting each class. The probability density function is given by

$$f(\mathbf{p} | \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K p_k^{\alpha_k - 1}, \quad \mathbf{p} \in S_K, \quad (\text{A1})$$

515 where S_K denotes the K -dimensional probability simplex and

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} \quad (\text{A2})$$

is the multivariate Beta function.

The expected probability of class k under the Dirichlet distribution is

$$\mathbb{E}[p_k] = \frac{\alpha_k}{\sum_{j=1}^K \alpha_j}, \quad (\text{A3})$$

520 where

$$S = \sum_{k=1}^K \alpha_k \quad (\text{A4})$$

is commonly referred to as the *Dirichlet strength*. Larger values of S indicate stronger accumulated evidence and correspondingly lower predictive uncertainty.

525 The evidential formulation adopted in this study follows the framework proposed by Sensoy et al. (2018), in which the network predicts non-negative evidence that is subsequently converted into the Dirichlet concentration parameters. Since both prediction heads in the proposed flood mapping framework correspond to binary classification tasks (water/land and cloud/-clear), the general Dirichlet distribution reduces naturally to the Beta distribution. Consequently, the main text presents the equivalent Beta formulation for clarity and computational convenience.

Appendix B: Event-level metrics



Table A1. Event-level IoU and PAvPU values. A, E, and A+E denote aleatoric, epistemic, and aleatoric plus epistemic uncertainty, respectively.

Event	EDL IoU	Ensemble IoU	MC Dropout IoU	EDL A	EDL E	EDL A+E	Ensemble A	Ensemble E	Ensemble A+E	MC Dropout A	MC Dropout E	MC Dropout A+E
EMSR264_08VATOMANDRY_DEL_v2	0.9462	0.9674	0.9678	0.9724	0.9732	0.9727	0.9794	0.9658	0.9773	0.9739	0.9557	0.9736
EMSR264_18MIANDRIVAZODETAIL_DEL_v2	0.5135	0.6162	0.5222	0.7589	0.7600	0.7594	0.7541	0.5751	0.7469	0.7519	0.6161	0.7458
EMSR273_01GRILE_DEL_MONIT04_v2	0.9223	0.9218	0.9067	0.9070	0.9080	0.9072	0.9126	0.8993	0.9137	0.9147	0.9027	0.9132
EMSR286_08ITUANGONORTH_DEL_MONIT02_v1	0.8563	0.8831	0.8365	0.8496	0.8804	0.8577	0.9170	0.8180	0.9065	0.9309	0.8604	0.9327
EMSR286_09ITUANGOSOUTH_DEL_MONIT02_v1	0.7969	0.8220	0.6841	0.8064	0.8487	0.8216	0.8640	0.6485	0.8400	0.8805	0.7906	0.8791
EMSR333_01RATTALORO_DEL_MONIT01_v1	0.7745	0.8586	0.7821	0.9164	0.9197	0.9174	0.9248	0.8357	0.9162	0.9105	0.8098	0.9075
EMSR333_02PORTOPALO_DEL_MONIT01_v1	0.7367	0.8212	0.7189	0.8594	0.8632	0.8604	0.8789	0.6889	0.8735	0.8745	0.6481	0.8730
EMSR342_06NORTHORMANTON_DEL_v1	0.8104	0.8301	0.8259	0.8519	0.8525	0.8521	0.8614	0.8079	0.8564	0.8656	0.7979	0.8638
EMSR342_07SOUTHNORMANTON_DEL_MONIT03_v2	0.7335	0.7373	0.7396	0.7551	0.7544	0.7549	0.7649	0.7364	0.7674	0.7645	0.7286	0.7623
EMSR347_06MIWANZA_DEL_v1	0.5470	0.6339	0.5510	0.5693	0.5748	0.5711	0.6570	0.5345	0.6473	0.5353	0.5084	0.5386
EMSR347_07ZOMBA_DEL_MONIT01_v1	0.8339	0.8660	0.8608	0.8763	0.8759	0.8762	0.8757	0.7976	0.8686	0.8725	0.8362	0.8717
EMSR347_07ZOMBA_DEL_v2	0.4102	0.7730	0.8296	0.4450	0.4486	0.4461	0.6371	0.6632	0.6830	0.7247	0.6766	0.7439
EMSR419_AO101_DEL_MONIT01	0.9307	0.9483	0.8643	0.9346	0.9350	0.9348	0.9490	0.9247	0.9453	0.9402	0.8834	0.9394
EMSR422_AO106_DEL_PRODUCT	0.9729	0.9658	0.9457	0.9698	0.9693	0.9697	0.9760	0.9475	0.9719	0.9706	0.9437	0.9697
EMSR438_AO102_DEL_MONIT03	0.9339	0.9614	0.9674	0.9561	0.9575	0.9565	0.9659	0.9605	0.9639	0.9697	0.9640	0.9700
EMSR438_AO106_DEL_PRODUCT	0.9361	0.9383	0.9403	0.9486	0.9499	0.9492	0.9534	0.9189	0.9475	0.9521	0.9182	0.9513
EMSR466_AO101_DEL_PRODUCT	0.7725	0.7328	0.5583	0.7237	0.7211	0.7230	0.6036	0.4941	0.5784	0.6980	0.6611	0.7032
EMSR9284_01YLITORNIORNORTHERN_DEL_MONIT01_v1	0.8201	0.8405	0.8408	0.8555	0.8627	0.8577	0.8732	0.7322	0.8510	0.8763	0.7569	0.8782



530 *Code and data availability.* The WorldFloodsv2 dataset used in this study is available from Hugging Face (Image and Signal Processing (ISP), 2024). The code used in this study is accessible on Zenodo (<https://doi.org/10.5281/zenodo.20501552>) (Chen, 2026).

Author contributions. CC and LW jointly conceptualised research idea. CC led the model development and implementation, as well as experimental design and result evaluation, with support from LW. CC and LW jointly prepared manuscript. CY provided statistical guidance and contributed to uncertainty analysis. CY and WW contributed to manuscript revision.

535 *Competing interests.* The contact author has declared that none of the authors has any competing interests.

Acknowledgements. The authors express their gratitude for the financial support received from two National Science and Technology Council research projects (NSTC 113-2923-M-002-001-MY4 and 114-2124-M-002-007-). Artificial intelligence (AI) tools were used to assist in improving the language and readability of the manuscript. All scientific content, analyses, interpretations, and conclusions were developed and verified by the authors.



540 References

- Bates, P. D.: Integrating remote sensing data with flood inundation models: How far have we got?, *Hydrological Processes*, 26, 2515–2521, <https://doi.org/10.1002/hyp.9374>, 2012.
- Bates, P. D. and De Roo, A.: A simple raster-based model for flood inundation simulation, *Journal of Hydrology*, 236, 54–77, [https://doi.org/10.1016/S0022-1694\(00\)00278-X](https://doi.org/10.1016/S0022-1694(00)00278-X), 2000.
- 545 Casella, G. and Berger, R.: *Statistical inference*, Chapman and Hall/CRC, 2024.
- Chen, C.-J.: Code for: Evidential deep learning for uncertainty-aware global flood extent segmentation using satellite imagery, <https://doi.org/10.5281/zenodo.20501552>, 2026.
- Doocy, S., Daniels, A., Murray, S., and Kirsch, T. D.: The human impact of floods: a historical review of events 1980–2009 and systematic literature review, *PLoS Currents*, <https://doi.org/10.1371/currents.dis.f4deb457904936b07c09daa98ee8171a>, 2013.
- 550 Festa, D., Hassaan, M., and Wagner, W.: SAR and optical imagery for dynamic global surface water monitoring: addressing sensor-specific uncertainty for data fusion, <https://doi.org/10.2139/ssrn.6382381>, SSRN preprint, 2026.
- Gal, Y. and Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: *Proceedings of the 33rd international conference on machine learning*, edited by Balcan, M. F. and Weinberger, K. Q., vol. 48 of *Proceedings of machine learning research*, pp. 1050–1059, PMLR, New York, New York, USA, <https://proceedings.mlr.press/v48/gal16.html>, 2016.
- 555 Gao, J., Chen, M., Xiang, L., and Xu, C.: A Comprehensive Survey on Evidential Deep Learning and Its Applications, <https://doi.org/10.48550/ARXIV.2409.04720>, version Number: 1 TLDR: This survey provides a comprehensive overview of the current research on Evidential Deep Learning, and delves into the theoretical foundation of EDL, the subjective logic theory, and discusses its distinctions from other uncertainty estimation frameworks., 2024.
- 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., and Zhu, X. X.: A survey of uncertainty in deep neural networks, *Artificial Intelligence Review*, 56, 1513–1589, <https://doi.org/10.1007/s10462-023-10562-9>, tLDR: This work gives a comprehensive overview of uncertainty estimation in neural networks, reviews recent advances in the field, highlights current challenges, and identifies potential research opportunities, and a comprehensive introduction to the most crucial sources of uncertainty is given. remark: survey of UNC, 2023.
- 560 Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q.: On calibration of modern neural networks, in: *Proceedings of the 34th international conference on machine learning - volume 70, ICML'17*, pp. 1321–1330, JMLR.org, Sydney, NSW, Australia, number of pages: 10, 2017.
- 565 Image and Signal Processing (ISP): WorldFloodsv2 (Revision a234ba0), <https://doi.org/10.57967/hf/3149>, 2024.
- Isikdogan, F., Bovik, A. C., and Passalacqua, P.: Surface Water Mapping by Deep Learning, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10, 4909–4918, <https://doi.org/10.1109/JSTARS.2017.2735443>, 2017.
- Jsang, A.: *Subjective Logic: A formalism for reasoning under uncertainty*, Springer Publishing Company, Incorporated, 2018.
- 570 Kendall, A. and Gal, Y.: 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*, pp. 5580–5590, Curran Associates Inc., Long Beach, California, USA, ISBN 978-1-5108-6096-4, number of pages: 11 tex.address: Red Hook, NY, USA, 2017.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles, in: *Advances in neural information processing systems*, edited by Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., vol. 30, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf, 2017.



- Liu, H., Hu, H., Liu, X., Jiang, H., Liu, W., and Yin, X.: A Comparison of Different Water Indices and Band Downscaling Methods for Water Bodies Mapping from Sentinel-2 Imagery at 10-M Resolution, *Water*, 14, 2696, <https://doi.org/10.3390/w14172696>, 2022.
- Mateo-García, G., Veitch-Michaelis, J., Smith, L., Oprea, S. V., Schumann, G., Gal, Y., Baydin, A. G., and Backes, D.: Towards global flood mapping onboard low cost satellites with machine learning, *Scientific Reports*, 11, 7249, <https://doi.org/10.1038/s41598-021-86650-z>, tLDR: A flood segmentation algorithm that produces flood masks to be transmitted instead of the raw images, while running efficiently on the accelerator aboard the PhiSat-1, acting as a proof of concept for this approach., 2021.
- McFEETERS, S. K.: The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features, *International Journal of Remote Sensing*, 17, 1425–1432, <https://doi.org/10.1080/01431169608948714>, 1996.
- 580 Memon, A. A., Muhammad, S., Rahman, S., and Haq, M.: Flood monitoring and damage assessment using water indices: A case study of Pakistan flood-2012, *The Egyptian Journal of Remote Sensing and Space Science*, 18, 99–106, <https://doi.org/10.1016/j.ejrs.2015.03.003>, 2015.
- Merz, B., Blöschl, G., Vorogushyn, S., Dottori, F., Aerts, J. C. J. H., Bates, P., Bertola, M., Kemter, M., Kreibich, H., Lall, U., and Macdonald, E.: Causes, impacts and patterns of disastrous river floods, *Nature Reviews Earth & Environment*, 2, 592–609, <https://doi.org/10.1038/s43017-021-00195-3>, 2021.
- 590 Nadav-Greenberg, L. and Joslyn, S. L.: Uncertainty Forecasts Improve Decision Making Among Nonexperts, *Journal of Cognitive Engineering and Decision Making*, 3, 209–227, <https://doi.org/10.1518/155534309X474460>, 2009.
- Nix, D. and Weigend, A.: Estimating the mean and variance of the target probability distribution, in: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, pp. 55–60 vol.1, IEEE, Orlando, FL, USA, ISBN 978-0-7803-1901-1, <https://doi.org/10.1109/ICNN.1994.374138>, 1994.
- 595 Portalés-Julià, E., Mateo-García, G., Purcell, C., and Gómez-Chova, L.: Global flood extent segmentation in optical satellite images, *Scientific Reports*, 13, 20316, <https://doi.org/10.1038/s41598-023-47595-7>, tLDR: A machine learning framework for operational flood mapping from optical satellite images addressing these problems is presented, based on a clouds-aware segmentation model trained in an extended version of the WorldFloods dataset, providing evidence of better performance than current operational methods based on thresholding spectral indices., 2023.
- 600 Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., vol. 9351, pp. 234–241, Springer International Publishing, Cham, ISBN 978-3-319-24573-7 978-3-319-24574-4, https://doi.org/10.1007/978-3-319-24574-4_28, series Title: Lecture Notes in Computer Science TLDR: It is shown that such a network can be trained end-to-end from very few images and outperforms the prior best method (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks., 2015.
- Schreck, J. S., Gagne, D. J., Becker, C., Chapman, W. E., Elmore, K., Fan, D., Gantos, G., Kim, E., Kimpapa, D., Martin, T., Molina, M. J., Przybylo, V. M., Radford, J., Saavedra, B., Willson, J., and Wirz, C.: Evidential Deep Learning: Enhancing Predictive Uncertainty Estimation for Earth System Science Applications, *Artificial Intelligence for the Earth Systems*, 3, 230093, <https://doi.org/10.1175/AIES-D-23-0093.1>, tLDR: This study compares the uncertainty derived from evidential neural networks to that obtained from ensembles and shows evidential deep learning models attaining predictive accuracy rivaling standard methods, while robustly quantifying both sources of uncertainty., 2024.
- 610



- Sensoy, M., Kaplan, L., and Kandemir, M.: Evidential Deep Learning to Quantify Classification Uncertainty, in: Advances in Neural Information Processing Systems, edited by Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., vol. 31, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf, 2018.
- 615 Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., Erickson, T. A., and Slayback, D. A.: Satellite imaging reveals increased proportion of population exposed to floods, *Nature*, 596, 80–86, <https://doi.org/10.1038/s41586-021-03695-w>, tLDR: The high spatial and temporal resolution of the satellite observations will improve the understanding of where floods are changing and how best to adapt, and the global flood database generated from these observations will help to improve vulnerability assessments, the accuracy of global and local flood models, the efficacy of adaptation interventions and the authors' understanding of the interactions between landcover change, climate and floods., 2021.
- 620 Teng, J., Jakeman, A. J., Vaze, J., Croke, B. F., Dutta, D., and Kim, S.: Flood inundation modelling: A review of methods, recent advances and uncertainty analysis, *Environmental Modelling & Software*, 90, 201–216, <https://doi.org/10.1016/j.envsoft.2017.01.006>, 2017.
- Tulbure, M. G., Broich, M., Perin, V., Gaines, M., Ju, J., Stehman, S. V., Pavelsky, T., Masek, J. G., Yin, S., Mai, J., and Betbeder-Matibet, L.: Can we detect more ephemeral floods with higher density harmonized Landsat Sentinel 2 data compared to Landsat 8 alone?, *ISPRS Journal of Photogrammetry and Remote Sensing*, 185, 232–246, <https://doi.org/10.1016/j.isprsjprs.2022.01.021>, 2022.
- 625 Verpoorter, C., Kutser, T., and Tranvik, L.: Automated mapping of water bodies using Landsat multispectral data, *Limnology and Oceanography: Methods*, 10, 1037–1050, <https://doi.org/10.4319/lom.2012.10.1037>, 2012.
- Xu, H.: Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery, *International Journal of Remote Sensing*, 27, 3025–3033, <https://doi.org/10.1080/01431160600589179>, 2006.
- 630 Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J.: UNet++: A Nested U-Net Architecture for Medical Image Segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, edited by Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J. M. R., Bradley, A., Papa, J. P., Belagiannis, V., Nascimento, J. C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., and Madabhushi, A., vol. 11045, pp. 3–11, Springer International Publishing, Cham, ISBN 978-3-030-00888-8 978-3-030-00889-5, https://doi.org/10.1007/978-3-030-00889-5_1, series Title: Lecture Notes in Computer Science
- 635 TLDR: This paper presents UNet++, a new, more powerful architecture for medical image segmentation where the encoder and decoder sub-networks are connected through a series of nested, dense skip pathways, and argues that the optimizer would deal with an easier learning task when the feature maps from the decoder and encoder networks are semantically similar., 2018.