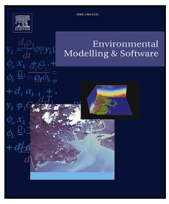
Environmental Modelling and Software 191 (2025) 106488

Contents lists available at ScienceDirect

Environmental Modelling and Software

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



Uncertainty quantification for LiDAR-based maps of ditches and natural streams

Florian Westphal [a,*], William Lidberg [b], Mariana Dos Santos Toledo Busarello [b], Anneli M. Ågren [b]

[a] Jönköping AI Lab, School of Engineering, Jönköping University, Gjuterigatan 5, Jönköping, 551 11, Sweden
[b] Department of Forest Ecology and Management, Swedish University of Agricultural Sciences, Skogsmarksgränd 17, Umeå, 901 83, Sweden

ARTICLE INFO

Dataset link: Uncertainty Quantification for LiDAR-based Maps of Ditches and Natural Streams (Original data), Automatic Detection of Ditches and Natural Streams from Digital Elevation Models Using Deep Learning (Reference data)

Keywords:
Semantic segmentation
Uncertainty quantification
Monte Carlo dropout
Conformal prediction
Small-scale hydrology
LiDAR

ABSTRACT

This article compares novel and existing uncertainty quantification approaches for semantic segmentation used in remote sensing applications. We compare the probability estimates produced by a neural network with Monte Carlo dropout-based approaches, including predictive entropy and mutual information, and conformal prediction-based approaches, including feature conformal prediction (FCP) and a novel approach based on conformal regression. The chosen task focuses on identifying ditches and natural streams based on LiDAR derived digital elevation models. We found that FCP's uncertainty estimates aligned best with the neural network's prediction performance, leading to the lowest Area Under the Sparsification Error curve of 0.09. For finding misclassified instances, the network probability was most suitable, requiring a correction of only 3% of the test instances to achieve a Matthews Correlation Coefficient (MCC) of 0.95. Conformal regression produced the best confident maps, which, at 90% confidence, covered 60% of the area and achieved an MCC of 0.82.

1. Introduction

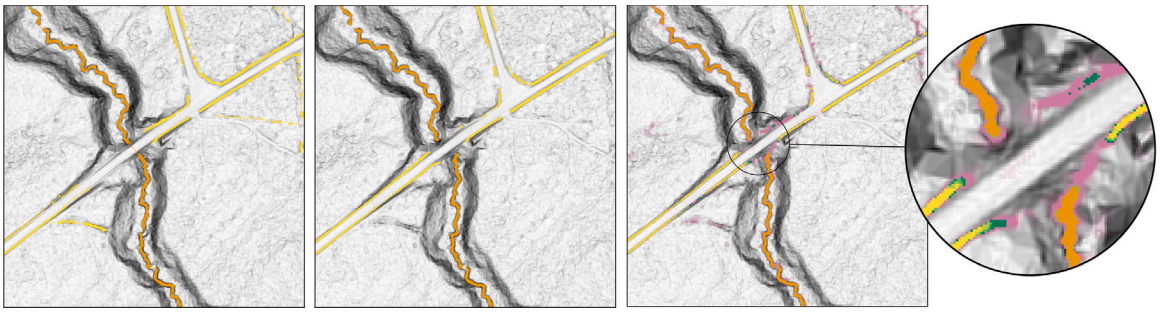
Having accurate maps of a landscape is crucial for supporting informed decisions in various applications, including sustainable land-use management (Pagella and Sinclair, 2014). Creating large-scale maps, such as those covering an entire country, is a labor-intensive process that requires significant human effort. Consequently, the automated analysis of remote sensing data has become a common solution (Blaschke, 2010). This involves the analysis of data from sources such as optical images, synthetic aperture radar, hyperspectral imaging, and Light Detection and Ranging (LiDAR) (Toth and Jóźków, 2016). Historically, traditional computer graphics-based approaches have been used for remote sensing applications (Savelonas et al., 2022), but more recently, deep learning-based methods have been used successfully in these applications (Yuan et al., 2020). Deep learning-based approaches tend to convert the remote sensing data into images, and apply semantic segmentation to assign one of the classes of interest to every pixel of the image. For example, O'Neil et al. (2020) have mapped wetlands based on aerial images and topographic indices calculated based on a LiDAR derived digital elevation model (DEM). Similarly, Busarello et al. (2025) have investigated the use of different topographic indices as representation of a DEM derived from LiDAR data. Based on these rasterized representations, they trained a neural network to detect ditches and natural streams.

One challenge when working with automatically generated maps is assessing their reliability. A common approach to estimating the quality of these maps is by comparing them with a representative portion of the actual landscape, which provides a good general estimate as long as the evaluated landscape is representative of the overall terrain. However, the actual quality can vary significantly depending on location, with some parts being more accurate and others less so (Kasraei et al., 2021). For decision-making purposes, it is important to have an estimate of reliability at specific locations, which can be achieved by quantifying the uncertainty of the used model at the point of interest (Xu et al., 2022).

Quantifying uncertainty in deep learning models initially appears straightforward, as they typically provide class-wise probabilities for each pixel. However, research has shown that these estimates tend to be overconfident, due to the training process rewarding overconfident predictions (Guo et al., 2017; Sensoy et al., 2018). In response, various methods have been developed to quantify neural network uncertainty, which Gawlikowski et al. (2023) categorize into four primary

* Corresponding author.
E-mail address: florian.westphal@ju.se (F. Westphal).

https://doi.org/10.1016/j.envsoft.2025.106488
Received 17 December 2024; Received in revised form 14 April 2025; Accepted 21 April 2025
Available online 2 May 2025



(a) Groundtruth Labels **(b)** Predicted Labels **(c)** Visualized Uncertainty

Fig. 1. Illustration of the semantic segmentation task. Ditches (yellow) and natural streams (orange) should be identified in a given chip based on the slope image derived from the digital elevation model at a 0.5 m resolution. The uncertainty of the 5% most uncertain pixels, as quantified by Feature Conformal Prediction is displayed using pink for background pixels, green for ditches and blue for streams. The strength of the color is determined by the uncertainty value.

directions: single network deterministic approaches, Bayesian methods, ensemble techniques, and test-time augmentation methods.

Deterministic methods, such as Dirichlet prior networks (Gawlikowski et al., 2022), have been used in remote sensing applications, as well as ensemble techniques, such as deep ensembles (Lakshminarayanan et al., 2017). For example, Chaudhary et al. (2022) utilized deep ensembles to quantify uncertainty in generated maximum water depth hazard maps, which aid in estimating the risk of flooding. Additionally, deep ensembles have been leveraged to estimate the uncertainty in wavelength bands from Sentinel-2 whose spatial resolution had been enhanced to a resolution of 10 m (Iagaru and Gottschling, 2023).

However, the primary focus has centered on Bayesian methods. The most prevalent approach among these Bayesian methods is Monte Carlo dropout (MC dropout) (Gal and Ghahramani, 2016), which has been used, for example, by Kampffmeyer et al. (2016) to quantify the uncertainty of their method on an urban object classification task based on a digital surface model (DSM). MC dropout has also been used by Martínez-Ferrer et al. (2022) for uncertainty quantification of their approach to retrieve different biophysical variables, such as leaf area index and canopy water content from surface reflectance data. Another notable Bayesian approach involves the application of Bayesian neural networks (Blundell et al., 2015; Goan and Fookes, 2020). Hertel et al. (2023) have conducted a comparative analysis of both methodologies and advocate for the use of Bayesian neural networks, as they tend to be less likely to indicate high confidence in incorrect predictions.

One other approach to uncertainty quantification is the conformal prediction framework (Vovk et al., 2005), which has been primarily applied to simple classification and regression tasks, but more recently was adapted to semantic segmentation. For example, Wieslander et al. (2021) have used conformal prediction for medical image segmentation, while Labuzzetta (2022) has applied subsample conformal prediction to the task of surface water and grassed waterway segmentation. Additionally, Singh et al. (2024) have demonstrated how conformal prediction can be applied to different tasks in earth observation, such as tree species mapping, land cover classification and canopy height estimation, and advocate for its more widespread use. While these works are based on more traditional formulations of conformal prediction, Teng et al. (2023) have proposed Feature Conformal Prediction (FCP), which is particularly adjusted to the use with deep neural networks, and has been shown to be more effective at quantifying the uncertainty of a neural network in general semantic segmentation tasks.

This article compares uncertainty estimates derived from the predictions of a neural network (network probability) with mutual information and predictive entropy — two uncertainty metrics calculated through MC dropout — to those obtained via conformal regression and FCP. We focus on these methods, in contrast to Bayesian neural networks (Blundell et al., 2015) or deep ensembles (Lakshminarayanan et al., 2017), since they can be integrated into existing network architectures for semantic segmentation tasks, and do not incur extensive training times, due to the need to train multiple models. Notably, conformal prediction-based methods enable the production of predictions with a specified confidence level. Ideally, this would result in a map featuring only confident predictions, such as those above a 90% confidence level. Therefore, we investigate the usefulness of those confident maps.

For our comparison, we select the remote sensing task of detecting ditches and natural streams from a DEM (Fig. 1), which has been derived from LiDAR data. In particular, we perform this detection task on data derived from a DEM at 1 m resolution, as well as at 0.5 m resolution. This task is especially challenging due to the narrowness of the objects of interest, requiring high detection precision. In contrast to other semantic segmentation problems, most pixels are background pixels, while only few represent ditches and even fewer represent natural streams, leading to a significant class imbalance. Additionally, distinguishing between streams and ditches in a DEM can be difficult, as they often appear similar. These challenges contribute to uncertainty in predictions, which we aim to estimate.

Uncertainty quantification is crucial in this context because it could help identify natural streams that have been erroneously predicted as ditches. This distinction is significant, as natural streams require distinct management strategies to preserve their ecological integrity (Swedish PEFC, 2023). For example, avoiding the crossing of these streams with heavy machinery can prevent soil disturbance, which otherwise can exacerbate sedimentation and disrupt ecological functions (Bishop et al., 2009). In contrast, ditches can be more easily cleaned or maintained without needing permits.

This article addresses the following research questions:

1. Which of the investigated uncertainty quantification approaches, i.e., network probability, mutual information, predictive entropy, conformal regression, and FCP produces the most reliable uncertainty estimates?
2. To what degree does the resolution of the DEM impact the uncertainty estimates?
3. To what extent is it possible to generate useful maps with a specific confidence level using conformal regression and FCP?

2. Methodology

2.1. Mapping ditches and streams: Network probability

For mapping ditches and streams, our approach employs a U-Net architecture (Ronneberger et al., 2015) similar to that used by Busarello et al. (2025) (Fig. 2), which has been demonstrated to be effective for this task. The U-Net takes as input a 500×500 pixels large chip of the landscape represented by the local slope derived from a DEM. This

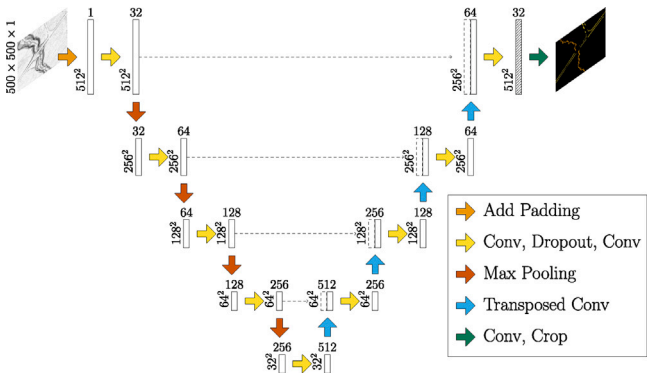


Fig. 2. U-Net architecture for mapping streams and ditches. The colored arrows show different processing steps, the dashed arrows indicate concatenation of feature maps, and the shaded feature maps indicate the ones being used for Feature Conformal Prediction.

input is then downsampled through a series of convolutional, dropout, and max pooling layers. Notably, our approach differs from Busarello et al. (2025) in that we utilize concrete dropout (Gal et al., 2017), which has been shown to improve uncertainty estimates obtained through MC dropout (Mukhoti and Gal, 2018).

After four downsampling steps, the extracted feature maps are upsampled using transposed convolutions, and processed by convolution and dropout layers to reach the original input size. At each upsampling step, the feature maps of the corresponding downsampling step are concatenated to ensure that no relevant information is lost. The final output is produced by applying a convolutional layer to the last feature maps (shaded feature maps in Fig. 2) . The output consists of three bands, each representing one of the considered classes: background, ditch, and natural stream.

In contrast to most U-Net architectures, our approach does not utilize a softmax layer, which would map the output at each pixel to a probability distribution over the three classes and be trained using cross entropy loss. Instead, we employ a linear activation function in the last convolutional layer and train the network using mean squared error, as proposed by Teng et al. (2023) to improve uncertainty estimates of FCP. Labels are mapped farther apart using a double log transform, resulting in large positive and negative values. Unlike Teng et al. (2023), who applied a Gaussian blur to the labels, we found this approach to be detrimental to performance, likely due to the narrow nature of our objects of interest, i.e., ditches and streams. To address class imbalance, we implement median frequency balancing (Eigen and Fergus, 2015) as suggested by Busarello et al. (2025).

Uncertainty estimates are derived from predicted network probabilities. This involves reversing the double log transform to obtain probabilities between 0 and 1 for each pixel and class. It should be noted that these probabilities are not calibrated in any way. The class with the highest probability is selected for each pixel. Uncertainty values are then calculated as the difference between the predicted probability and 1. This approach assumes that high confidence predictions yield probabilities close to 1, whereas low confidence predictions result in lower probabilities and thus higher uncertainty values.

2.2. MC dropout: Predictive entropy and mutual information

MC dropout has been proposed by Gal and Ghahramani (2016) as a method for estimating the uncertainty of a neural network. The main idea behind MC dropout is that if a neural network is certain about its prediction, introducing small random changes in its execution will not affect its prediction. Conversely, when a network is uncertain about its prediction, these small changes will lead to large variations in the predicted outcome. Thus, the network's uncertainty can be estimated by observing the variability in its predicted output when run multiple times. MC dropout introduces small random changes using dropout layers within the network architecture.

In a dropout layer (Srivastava et al., 2014), a randomly selected subset of neurons has its output set to zero. At each new input, a predefined probability determines which neurons are dropped. This probability is learned in concrete dropout (Gal et al., 2017), which we use in this study. Unlike the traditional use of dropout layers, which typically activates them only during training to promote robustness, MC dropout keeps those layers active during inference, resulting in varying outputs for identical inputs processed multiple times.

MC dropout estimates the uncertainty by using these varying outputs to compute two different metrics: predictive entropy and mutual information. These metrics measure different types of uncertainty, viz. aleatoric and epistemic uncertainty. Aleatoric uncertainty captures uncertainty caused by the data, such as ambiguity at the border between ditch and background, whereas epistemic uncertainty captures uncertainty caused by the model itself, for example, due to insufficient training data.

Predictive entropy captures both aleatoric and epistemic uncertainty and is approximated for a given input $\boldsymbol{x}$ and a given training set D_{train} as:

$$\hat{\mathbb{H}}[y|\boldsymbol{x}, D_{train}] = - \sum_{c \in C} \left(\frac{1}{T} \sum_{t=1}^{T} p\left(y = c|\boldsymbol{x}, \hat{w}_t\right) \right) \ln \left(\frac{1}{T} \sum_{t=1}^{T} p\left(y = c|\boldsymbol{x}, \hat{w}_t\right) \right) \quad (1)$$

Here, C is the set of classes, T is the number of outputs y to collect for variations of the neural network $\hat{w}_t$, which are produced by the dropout layers, and $p\left(y = c|\boldsymbol{x}, \hat{w}_t\right)$ is the probability of input $\boldsymbol{x}$ being in class c. In contrast, mutual information measures only the epistemic uncertainty and is approximated as:

$$\hat{\mathbb{I}}[y, w|\boldsymbol{x}, D_{train}] = \hat{\mathbb{H}}[y|\boldsymbol{x}, D_{train}] + \frac{1}{T} \sum_{c \in C} \sum_{t=1}^{T} \left(p\left(y = c|\boldsymbol{x}, \hat{w}_t\right) \ln p\left(y = c|\boldsymbol{x}, \hat{w}_t\right) \right) \quad (2)$$

This study computes predictive entropy and mutual information values for each pixel within every output chip, based on 1 000 outputs collected for each chip.

2.3. Conformal regression

Conformal regression is a part of the conformal prediction framework (Vovk et al., 2005), offering guarantees for machine learning model predictions. Unlike standard regression, conformal regression generates prediction intervals rather than single numerical values. The framework ensures that, for a pre-defined percentage of predictions (e.g., 90%), the true value lies within the provided interval. While this can be achieved easily by making this interval arbitrarily large, the challenge lies in finding a narrow yet guarantee-ensuring interval.

While there are two types of conformal regression, this article focuses on the inductive case, as it does not require frequent re-training. Inductive conformal regression estimates the size of the prediction interval based on a calibration set, which is separate from the training, validation, and test datasets. The interval is derived by measuring the difference between the predicted value and the true value for all instances of the calibration set, using a non-conformity function, such as mean absolute error (MAE), resulting in a non-conformity score. Based on a pre-defined confidence-level, e.g., 90%, the difference or non-conformity score of the 90th percentile is selected, and the interval is set as the value predicted by the machine learning model plus or minus the selected value. This ensures that the true value of 90% of instances in the calibration set lies within the produced interval, since their prediction errors were smaller than the one chosen. Because the calibration set is required to be exchangeable with the test set, i.e., they

both come from the same distribution, it can be expected that this guarantee will hold also for unseen instances from the test set.

One issue with the described approach is that it assigns the same interval to all instances, leading to overly large intervals for most of them. This can be addressed by normalizing non-conformity scores through instance difficulty estimation. For example, Cortés-Ciriano and Bender (2019) estimate instance difficulty using MC dropout, recording predicted outputs for the same instance i multiple times with enabled dropout layers and calculating mean μ_i and standard deviation σ_i over those outputs. The non-conformity score α_i is then computed based on the corresponding true value y_i over all instances in the calibration set $\mathcal{D}_{cal}$, resulting in a list of non-conformity scores S, which is then sorted in ascending order.

$$\alpha_i = \frac{|y_i - \mu_i|}{e^{\sigma_i}}$$
$$S = \alpha_1, \ldots, \alpha_q, \text{ with } q = |\mathcal{D}_{cal}| \tag{3}$$

Based on this list, the non-conformity score α_p is selected, which corresponds to the chosen confidence level $1 - \epsilon$ (e.g., 0.9 for $\epsilon = 0.1$). For a new instance j, the prediction interval around the mean of the MC dropout samples μ_j is then derived by multiplying the selected α_p with the instance's difficulty, as measured by the standard deviation over the MC dropout samples σ_j (Cortés-Ciriano and Bender, 2019).

$$p = \lceil (1 - \epsilon)(q + 1) \rceil, \text{ for } \alpha_p$$
$$\mu_j \pm \alpha_p \cdot e^{\sigma_j} \tag{4}$$

Another challenge in deriving regression intervals is that the distribution of non-conformity scores may vary depending on certain properties of the instances. For example, when dealing with instances having large true values, the error may be greater than for those with small true values. If this difference in distribution is not taken into account, the derived regression intervals will be larger than necessary for instances with small true values and possibly too narrow for instances with large true values, depending on their prevalence in the calibration set.

For classification problems, Mondrian conformal prediction (Vovk et al., 2005) addresses these issues by categorizing instances based on a Mondrian taxonomy that considers certain properties of each instance. A separate conformal predictor is then built for each category. Mondrian regression, proposed by Boström and Johansson (2020), follows a similar approach. It divides the calibration instances into different categories based on a Mondrian taxonomy, specifically an estimate of difficulty. The prediction interval within each category is derived from the non-conformity score at a specific percentile. This methodology allows for more tailored prediction intervals that are narrower for instances belonging to simpler categories and wider for those in harder categories. Since simpler categories typically have low errors and thus low non-conformity scores, their prediction intervals can be narrower. In contrast, harder categories will have higher non-conformity scores, leading to broader prediction intervals.

In our implementation, each pixel in an input chip is associated with three real values indicating to which of the three classes it belongs. After reverting the double log transform, we perform conformal regression to derive a prediction interval for the three class values of each pixel. Since the class values can be seen as the probability of the pixel to belong to each of the classes, the estimated intervals can be interpreted as probability ranges. The estimation of these intervals involves calculating non-conformity scores per class for every pixel in all calibration set chips, followed by normalization using 100 Monte Carlo samples as proposed by Cortés-Ciriano and Bender (2019).

While we record non-conformity scores per class, we also employ Mondrian conformal regression to obtain more tailored intervals. This approach differs from the original Mondrian taxonomy by Boström and Johansson (2020), which utilized estimated instance difficulty. In contrast, our taxonomy categorizes predictions for each class into two categories: pixels with predicted probabilities close to zero and those near one. This distinction is important because we observed in initial experiments the tendency of classes with few pixels to have most commonly a predicted probability value of zero with a low non-conformity score. Conversely, when the actual class is predicted (i.e., the predicted probability exceeds 0.5), the non-conformity scores tend to be substantially higher. Given this observation, it is reasonable to create categories based on the predicted values.

Thus, we group the non-conformity scores of instances from the calibration set $\mathcal{D}_{cal}$ for each class individually into two lists, one for which the predicted probability is lower than 0.5, $S^{<0.5}$, and one for which the predicted probability is larger or equal, $S^{\geq 0.5}$. Those lists are then sorted in ascending order, and the non-conformity scores corresponding to the chosen confidence-level $1 - \epsilon$ are selected as before.

$$S^{\geq 0.5} = \alpha_1^{\geq 0.5}, \ldots, \alpha_r^{\geq 0.5}$$
$$S^{<0.5} = \alpha_1^{<0.5}, \ldots, \alpha_s^{<0.5}, \text{ with } r + s = |\mathcal{D}_{cal}|$$
$$t = \lceil (1 - \epsilon)(r + 1) \rceil, \text{ for } \alpha_t^{\geq 0.5}$$
$$u = \lceil (1 - \epsilon)(s + 1) \rceil, \text{ for } \alpha_u^{<0.5} \tag{5}$$

We then calculate intervals for each pixel j in a new chip by collecting 100 Monte Carlo samples of output predictions for the pixel and computing the respective mean μ_j and standard deviation σ_j. Given the selected non-conformity scores and the estimated means and standard deviations, the interval for one of the possible classes for pixel j is derived as follows:

$$\mu_j \pm \left(\mu_j \alpha_t^{\geq 0.5} + (1 - \mu_j) \alpha_u^{<0.5} \right) \cdot e^{\sigma_j} \tag{6}$$

By multiplying the selected non-conformity scores with the probability mean and its inverse respectively, the final interval is derived as combination of both scores depending on how much the pixel's prediction agrees with the respective categories. This way of assigning the corresponding non-conformity score to a pixel is computationally more efficient than having to find the applicable score based on some other feature of the pixel, such as difficulty, via a look-up, as it is the case in the Mondrian approaches by Boström and Johansson (2020), Wieslander et al. (2021), and Labuzzetta (2022).

The uncertainty value for each class is determined by the size of the interval, where a larger interval indicates greater uncertainty in the prediction. Unlike MC dropout, which produces uncertainty values per pixel, the conformal regression approach derives an uncertainty value per pixel per class.

2.4. Feature conformal prediction (FCP)

In contrast to conformal regression, which computes non-conformity scores based on the output of a machine learning model, FCP (Teng et al., 2023) calculates these scores based on an intermediate feature representation of a neural network. This feature representation can be, for example, the feature maps produced by a convolutional layer. These feature maps are then converted into a single vector by flattening the corresponding tensor, enabling FCP to obtain a predicted output for an input instance as a point in a high-dimensional vector space.

When applying conformal regression, it is clear what constitutes a true value for computing the non-conformity score, i.e., the target value of an instance. In contrast, identifying the true feature representation of an instance is not straightforward. FCP assumes this true representation to be the infimum, which corresponds to the feature representation with the smallest numerical values, which produces the correct output. However, finding this optimal representation is challenging. As a result, FCP approximates the infimum by optimizing the original feature representation for a given input instance to produce the correct output using gradient descent. It should be noted that this approach modifies the values of the feature representation rather than adjusting neural network weights. The non-conformity score is then computed using a norm distance, such as the infinity norm, between the vector of the

original representation and the one derived through gradient descent. This yields a single non-conformity score per instance, differing from the conformal regression case where multiple scores are generated corresponding to each output.

The base score is derived, similar to conformal regression, by computing the non-conformity scores for the calibration set and selecting, for example, the 90th percentile. Given a test instance, FCP derives its corresponding feature representation and applies perturbations to this representation, ensuring that the resulting new feature representations do not deviate beyond the distance indicated by the base score. These perturbations are achieved using Linear Relaxation based Perturbation Analysis (LiRPA) (Xu et al., 2020). Subsequently, FCP estimates the resulting output intervals by applying the neural network to the perturbed feature representations. In summary, FCP performs conformal regression in feature space and derives output prediction intervals through perturbation analysis. Mathematical proofs of the correctness and efficiency of the method have been derived by Teng et al. (2023).

Our implementation utilizes feature maps generated prior to the output layer (shaded feature maps in Fig. 2) for FCP. In contrast to Teng et al. (2023), who found that features can be extracted from various layers without altering the prediction intervals, our findings suggest that using feature maps from any other layer results in unreasonably large prediction intervals for our task and network architecture. This may be because the skip connections in our U-Net architecture interfered presumably with the perturbation step, as the perturbations were applied only to the feature maps of the upsampling path and not those of the downsampling path. We employ perturbation analysis to derive prediction intervals for every pixel and class. Similar to our conformal regression implementation, the size of the interval is interpreted as uncertainty, where larger intervals indicate higher uncertainty.

3. Experiments

3.1. Dataset

For this article, we used a dataset provided by Busarello et al. (2025)., consisting of LiDAR-derived DEMs for 12 distinct regions in Sweden, further described by Lidberg et al. (2023). The dataset is available in two resolutions, 0.5 m and 1 m, corresponding to input chips of 500 × 500 pixels representing areas of 250 m×250 m and 500 m×500 m, respectively. To address class imbalance, chips with less than 250 ditch or stream pixels were removed, resulting in a dataset where still only 1.1% and 0.1% of all pixels belong to the ditch and natural stream class, respectively (Busarello et al., 2025).

Topographic indices are utilized to provide a rasterized representation of the DEM. In our experiments, the local slope was used, which signifies the change in elevation between every pixel in the DEM, with inclination displayed in degrees (Florinsky, 2016). This index was chosen due to its superior performance in stream detection and satisfactory results for ditch detection (Busarello et al., 2025). To reduce execution time, we focused on a single index; however, all uncertainty quantification methods remain applicable when multiple indices are considered.

To evaluate the chosen uncertainty quantification methods, we employed 10-fold cross-validation to facilitate statistical analysis. However, since conformal regression and FCP require a calibration set, the dataset was divided into 11 folds: nine for training, one for calibration, and one for testing to ensure exchangeability between folds. Stratified sampling by region ensured that chips in each fold cover the 12 distinct regions similarly well, preserving representativeness throughout training, calibration, and test set.

Apart from ensuring exchangeability, we needed to prevent information about the test set from leaking into the training and calibration set to avoid biasing the evaluation and obtaining miscalibrated uncertainty estimates. This was achieved using the following partitioning strategy. The dataset was divided into chips without overlap, ensuring

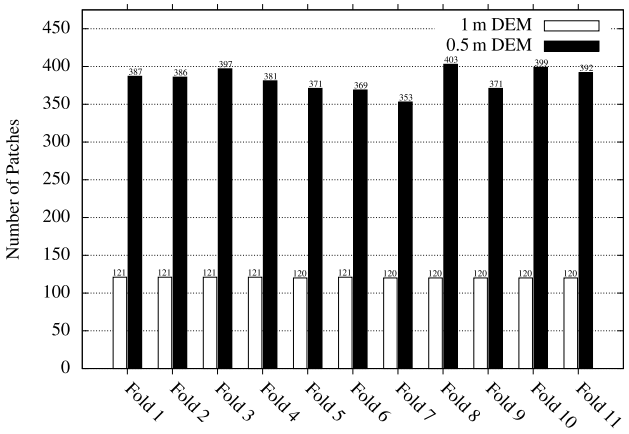


Fig. 3. Number of chips in each of the 11 folds for the digital elevation model (DEM) with resolution 1 m and 0.5 m.

that no chip's information was shared between training, calibration and test set. Within each region, chips were grouped to minimize borders with adjacent chips in other folds. To optimize this grouping, a heuristic algorithm was used due to the NP-hard nature of the problem[1], yielding an approximate optimal solution for partitioning.

After splitting the chips from the 1 m DEM into 11 folds, the corresponding chips were then selected for the 0.5 m DEM, ensuring that both resolutions contained the same ditches and streams within each fold. This design prevented differences in performance between the two resolutions being attributed to varying levels of complexity, rather than resolution itself. While the number of chips for the 1 m DEM was nearly the same for all folds, this number varied more for the 0.5 m DEM (Fig. 3). The reason for this variation was that a different number of chips was dropped in each fold, depending on the number of 0.5 m DEM chips containing at least 250 ditch or stream pixels.

3.2. Performance metrics

The neural network's performance in classifying pixels as background, ditch, or natural stream was evaluated using the Matthews Correlation Coefficient (MCC) (Matthews, 1975; Yule, 1912) and F_1 score. Given that there were more than two classes, we used the multiclass version of MCC proposed by Gorodkin (2004). MCC provides a balanced view of the classification performance across all classes, while F_1 score focuses on the performance for a specific class, making it particularly suitable for investigating the network's performance for one class of interest (Chicco et al., 2021).

To evaluate the performance of uncertainty quantification approaches, we utilized the Area Under the Sparsification Error Curve (AUSE) (Ilg et al., 2018). Unlike the commonly used Patch Accuracy vs. Patch Uncertainty (PAvPU) (Mukhoti and Gal, 2018), AUSE also considers the uncertainty estimates for accurate predictions and does not require parameter tuning (Dreissig et al., 2023). Furthermore, AUSE is more suitable than the Expected Calibration Error (ECE) (Pakdaman Naeini et al., 2015) because ECE tends to overestimate calibration performance on imbalanced datasets (Dreissig et al., 2023). In contrast, AUSE can be combined with a performance metric that is robust to imbalanced data, such as MCC (Chicco et al., 2021). The main idea behind AUSE is that network outputs should be correct when

[1] NP-hard problems are computational problems for which there is no known algorithm which finds a solution in a number of steps polynomial in its input (Garey and Johnson, 1979). There is no efficient algorithm to solve them.

estimated to have low uncertainty, but may be incorrect when their uncertainty is high.

The sparsification curve is obtained by sorting pixels by their uncertainty and removing a fraction of the most uncertain pixels. Then, classification performance is measured on the remaining pixels. Here, we used MCC for multi-class evaluation and F_1 score for single-class evaluation. This process is repeated for increasing fractions of pixels. The resulting performance curve should gradually increase if uncertainty aligns with correctness.

The sparsification error curve is obtained by subtracting the sparsification curve for one uncertainty quantification approach from the oracle curve, i.e., the sparsification curve derived by sorting and removing pixels by actual distance between predicted and true values. This optimal sorting removes the most incorrect predictions first and is thus the best an uncertainty quantification method can achieve. For a good uncertainty quantification method, there will be a small area under the sparsification error curve, which can be used as single measure to compare between uncertainty quantification approaches.

Furthermore, we evaluated the practical use of those approaches using a correction curve, which we propose for this evaluation. This curve illustrates the impact different uncertainty quantification methods would have when used for correcting uncertain pixels, rather than removing them as is done for the sparsification curve. This correction curve shows how many pixels would need manual investigation to achieve a specified MCC value or F_1 score, facilitating informed decision-making. The correction error curve can be obtained by subtracting the correction curve of a particular uncertainty quantification method from the oracle correction curve. Based on this, we define the Area Under the Correction Error Curve (AUCE) as a metric for evaluating how well an uncertainty quantification approach identifies pixels that require correction relative to the optimal solution.

3.3. Experiment design

In our experiments, 10 U-Net models were trained on different fold combinations using a unique calibration and test set for each model. The implementation utilized pytorch 2.0.1 (Ansel et al., 2024) with training performed on a computer equipped with approximately 1 TB of RAM, two Intel Xeon Platinum processor with 32 cores each, and one 40 GB partition of an NVIDIA A100 GPU. We performed training using the Adam optimizer (Kingma and Ba, 2015) and a batch size of 16. Furthermore, each model was trained for 300 epochs in case of the 1 m DEM, and for 165 epochs, in case of the 0.5 m DEM, as these values were determined to be optimal based on validation loss performance. Given the reduced instance count for the 1 m DEM, training for more epochs was reasonable since there were fewer weight update steps per epoch.

After training the models, their performance was evaluated using MCC and F_1 score on the respective test sets. A Bayesian t-test for correlated observations (Corani and Benavoli, 2015) was conducted to determine if there were significant differences between the models' performance on the 1 m and 0.5 m DEM data. This statistical test was chosen, since it avoids the shortcomings of more traditional null hypothesis significance tests (Benavoli et al., 2017). Basically, it computes the probability of the performance difference between two approaches to lie within or outside of a pre-defined region of practical equivalence (ROPE). In our evaluation, we chose the ROPE to be a difference in MCC value of 0.05, meaning that the performance difference of two methods would have to be at least 0.05, for us to consider one method significantly better or worse than the other. Given that this test is a paired test, we paired the MCC result on one test fold from the 1 m DEM with its corresponding test fold from the 0.5 m DEM, i.e., the fold which covers the same areas, just at a higher resolution.

Given the trained models, we calibrated the conformal regression and FCP approaches on the respective calibration sets. We then derived uncertainty estimates for the chips in the corresponding test sets using the investigated approaches, i.e., network probability, mutual information, predictive entropy, conformal regression, and FCP. The execution time was measured for each approach. We then calculated the AUSE for all approaches on each test fold and both resolutions. This allowed us to investigate whether a lower resolution lead to poorer uncertainty estimates by comparing the AUSE scores between resolutions using the Bayesian t-test. Specifically, we paired the scores for each test fold and method of one resolution with those of the other resolution to determine if there were significant differences in uncertainty estimation quality.

Furthermore, we compared the AUSE scores for different uncertainty quantification methods using the Bayesian t-test to determine which method performed best. This comparison involved pairing the AUSE score of each two methods based on the corresponding folds and resolution. When comparing the AUSE, we considered a ROPE of 0.05 sufficient to identify practically relevant differences in performance among the evaluated methods. To facilitate efficient comparison of methods, high-density intervals (HDIs) were derived using the Bayesian t-test. The HDI plot displays the 95% probability intervals in which performance differences between methods lie, as well as the ROPE. By focusing on intervals not overlapping with the ROPE, statistically significant differences can be identified between methods.

To illustrate the practicality of these methods, we derived correction curves considering all classes, as well as curves focusing solely on predicted ditch and stream pixels. This allowed us to investigate the effort required to correct errors where natural streams were mistakenly predicted to be ditches or vice versa. Since, for illustrative purposes only, sparsification and correction curves displaying the performance of a single model had to be selected, the model with AUSE and AUCE values closest to the mean performance at both 1 m and 0.5 m resolutions was selected. The chip used for illustration was chosen as the one containing the most ditch and stream pixels from the test set of this model.

Lastly, we explored the possibility of generating reliable prediction maps using conformal regression and FCP. To this end, we calibrated these methods for various confidence thresholds, spanning from 50% to 90%, and included only pixels for which the probability interval of the most probable class did not overlap with those of any other class. We then computed the recall for each class, as well as the average recall over all classes. The recall was derived by dividing the number of confidently and correctly predicted pixels of a class by the total number of pixels of that class in the test set. Thus, giving an indication of the percentage of classified pixels in those confident maps. We also evaluated the classification performance on only those pixels classified with high confidence, excluding the ground truth of all pixels to which no single class was assigned. This gave an indication of the correctness of those confident maps.

4. Results

4.1. Mapping performance

Our analysis of the mapping performance revealed that all trained models performed best on the background and second best on the ditch class, but struggled with natural streams (Table 1). Models trained on the 0.5 m DEM outperformed those on the 1 m DEM in terms of MCC. A Bayesian t-test confirmed a significant advantage for the 0.5 m DEM models, estimating that with a probability of 100% they yielded a 0.05 points higher MCC than their 1 m DEM counterparts. This result remained the same even when increasing the ROPE to 0.1.

Table 1
Mapping performance on the 1 m and 0.5 m resolution data as measured by the Matthews Correlation Coefficient (MCC) for all classes, and the F_1 score for the background ($F_1^{(b)}$), ditches ($F_1^{(d)}$), and natural streams ($F_1^{(s)}$). The reported values indicate the mean and standard deviation over 10 test folds. Best performance indicated in bold.

Resolution	$F_1^{(b)}$	$F_1^{(d)}$	$F_1^{(s)}$	MCC
1 m	**1.00 ± 0.00**	0.62 ± 0.02	0.39 ± 0.06	0.61 ± 0.02
0.5 m	**1.00 ± 0.00**	**0.77 ± 0.03**	**0.43 ± 0.08**	**0.76 ± 0.03**

Table 2
Area Under the Sparsification Error Curve (AUSE) for the 1 m and 0.5 m resolution data derived for the background ($AUSE^{(b)}$), ditch ($AUSE^{(d)}$), and natural stream ($AUSE^{(s)}$) class using F_1 score as performance metric, and the overall AUSE score using the Matthews Correlation Coefficient for network probability ($\mathcal{U}_{prob}$), predictive entropy ($\mathcal{U}_{pe}$), mutual information ($\mathcal{U}_{mi}$), conformal regression ($\mathcal{U}_{cr}$), and feature conformal prediction ($\mathcal{U}_{fcp}$). The reported values indicate the mean and standard deviation over 10 test folds. Best result indicated in bold.

	$AUSE^{(b)}$	$AUSE^{(d)}$	$AUSE^{(s)}$	AUSE
	1 m			
$\mathcal{U}_{prob}$	**0.00 ± 0.00**	0.46 ± 0.23	0.58 ± 0.22	0.42 ± 0.19
$\mathcal{U}_{pe}$	**0.00 ± 0.00**	0.96 ± 0.01	0.97 ± 0.03	0.95 ± 0.03
$\mathcal{U}_{mi}$	**0.00 ± 0.00**	0.95 ± 0.01	0.98 ± 0.00	0.95 ± 0.01
$\mathcal{U}_{cr}$	0.02 ± 0.00	0.33 ± 0.03	0.52 ± 0.07	0.35 ± 0.03
$\mathcal{U}_{fcp}$	**0.00 ± 0.00**	**0.20 ± 0.10**	**0.39 ± 0.11**	**0.20 ± 0.10**
	0.5 m			
$\mathcal{U}_{prob}$	**0.00 ± 0.00**	0.61 ± 0.28	0.64 ± 0.22	0.51 ± 0.21
$\mathcal{U}_{pe}$	**0.00 ± 0.00**	0.73 ± 0.16	0.84 ± 0.18	0.65 ± 0.14
$\mathcal{U}_{mi}$	**0.00 ± 0.00**	0.90 ± 0.07	0.96 ± 0.06	0.84 ± 0.09
$\mathcal{U}_{cr}$	0.02 ± 0.00	0.20 ± 0.03	0.51 ± 0.09	0.23 ± 0.03
$\mathcal{U}_{fcp}$	**0.00 ± 0.00**	**0.09 ± 0.04**	**0.34 ± 0.12**	**0.09 ± 0.04**

4.2. Uncertainty quantification performance

MCC values increased faster for models trained on the 0.5 m DEM compared to those on the 1 m DEM when removing the most uncertain pixels, as indicated by the sparsification curves (Figs. 4(a) and 4(b)). This suggests that uncertainty quantification methods are more effective in identifying misclassified pixels for the 0.5 m DEM than the 1 m DEM. Consequently, areas between sparsification curves and the oracle curve were smaller for the 0.5 m DEM (Table 2).

The trend of improved identification of incorrect pixels with higher resolution did not hold for network probability ($\mathcal{U}_{prob}$), where higher resolution resulted in worse identification. Nonetheless, the Bayesian t-test found that a higher resolution (0.5 m DEM) led to better uncertainty estimates than a lower resolution (1 m DEM) with a probability of 83% (ROPE=0.05). Excluding $\mathcal{U}_{prob}$ increased this probability to 99% (ROPE=0.05).

While the uncertainty quantification performance varied between resolutions for sparsification curves and AUSE, it showed mostly minor differences for correction curves (Figs. 4(c) and 4(d)) and AUCE scores (Table 3). The only exception was conformal regression ($\mathcal{U}_{cr}$) for which correction curves and AUCE scores improved with higher resolution. A Bayesian t-test revealed that, with a probability of 85% (ROPE=0.05), the performances at different resolutions were practically equivalent, i.e, the performance differences lay within the ROPE. Without $\mathcal{U}_{cr}$, this probability rose to 98% (ROPE=0.05).

Comparative analysis of uncertainty quantification methods revealed distinct differences in their sparsification curves (Figs. 4(a) and 4(b)). Notably, the MCC scores for methods, such as mutual information ($\mathcal{U}_{mi}$), predictive entropy ($\mathcal{U}_{pe}$), and network probability ($\mathcal{U}_{prob}$), decreased significantly, especially when the first 5% of uncertain pixels were removed (Fig. 4(b)). This drop in performance was caused by the fact that these methods assigned high uncertainty values to correctly classified pixels, particularly ditch and natural stream pixels (Fig. 5). This tendency is reflected in the higher AUSE scores for these classes (Table 2). The HDIs (Fig. 6), derived from the Bayesian t-test, confirmed that FCP ($\mathcal{U}_{fcp}$) outperformed MC Dropout based approaches, such as predictive entropy ($\mathcal{U}_{pe}$) and mutual information ($\mathcal{U}_{mi}$) with a 100% probability, even when assuming a ROPE of 0.4. Furthermore, $\mathcal{U}_{fcp}$ was estimated to perform better than network probability ($\mathcal{U}_{prob}$) with a probability of 99.4%, and better than conformal regression with a probability of 99.6% (ROPE=0.05).

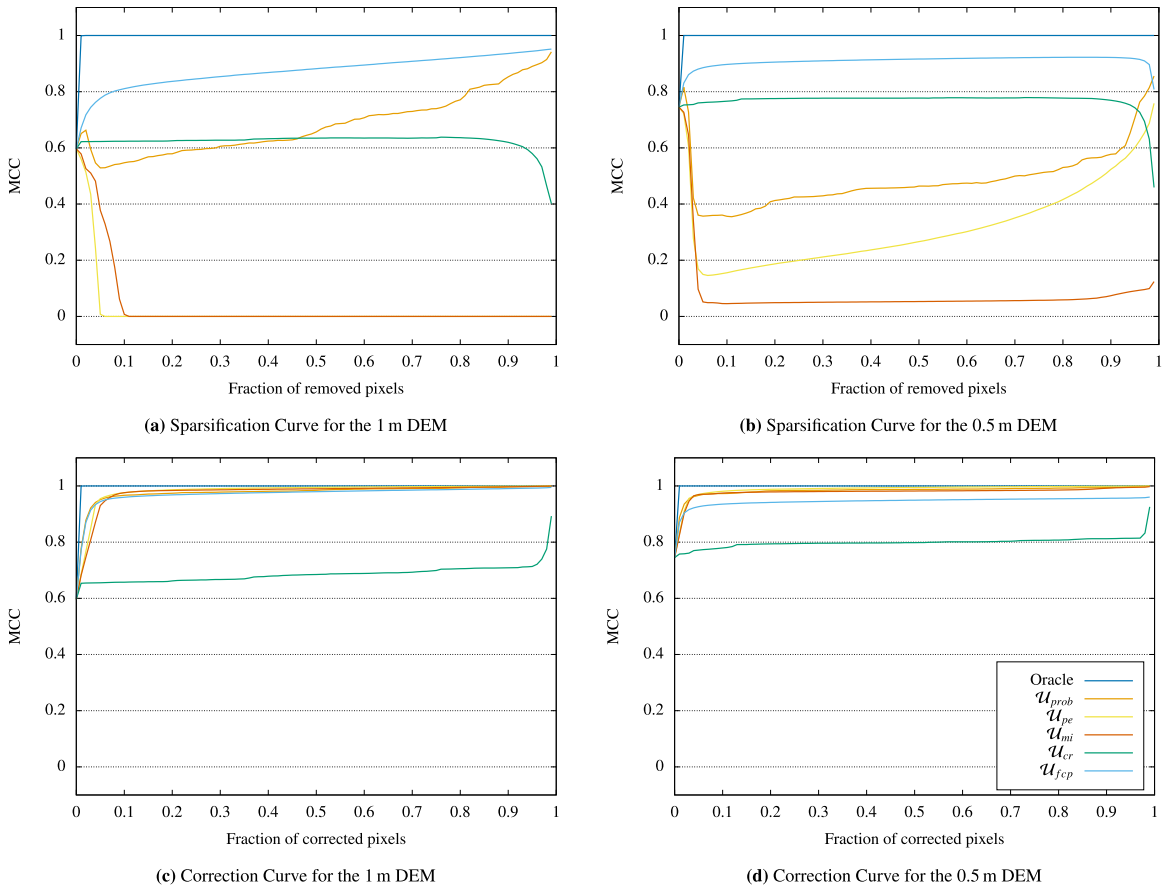
Table 3
Area Under the Correction Error Curve (AUCE) for the 1 m and 0.5 m resolution data derived for the background ($AUCE^{(b)}$), ditch ($AUCE^{(d)}$), and natural stream ($AUCE^{(s)}$) class using F_1 score as performance metric, and the overall AUCE score using the Matthews Correlation Coefficient for network probability ($\mathcal{U}_{prob}$), predictive entropy ($\mathcal{U}_{pe}$), mutual information ($\mathcal{U}_{mi}$), conformal regression ($\mathcal{U}_{cr}$), and feature conformal prediction ($\mathcal{U}_{fcp}$). The reported values indicate the mean and standard deviation over 10 test folds. Best result indicated in bold.

	$AUCE^{(b)}$	$AUCE^{(d)}$	$AUCE^{(s)}$	AUCE
	1 m			
$\mathcal{U}_{prob}$	**0.00 ± 0.00**	**0.01 ± 0.00**	0.04 ± 0.01	0.02 ± 0.00
$\mathcal{U}_{pe}$	**0.00 ± 0.00**	**0.01 ± 0.00**	**0.03 ± 0.01**	**0.01 ± 0.00**
$\mathcal{U}_{mi}$	**0.00 ± 0.00**	0.02 ± 0.00	0.04 ± 0.01	0.02 ± 0.00
$\mathcal{U}_{cr}$	**0.00 ± 0.00**	0.29 ± 0.04	0.38 ± 0.06	0.29 ± 0.04
$\mathcal{U}_{fcp}$	**0.00 ± 0.00**	0.10 ± 0.07	0.17 ± 0.12	0.10 ± 0.07
	0.5 m			
$\mathcal{U}_{prob}$	**0.00 ± 0.00**	**0.01 ± 0.00**	0.04 ± 0.01	0.01 ± 0.00
$\mathcal{U}_{pe}$	**0.00 ± 0.00**	**0.01 ± 0.00**	**0.03 ± 0.01**	**0.01 ± 0.00**
$\mathcal{U}_{mi}$	**0.00 ± 0.00**	0.01 ± 0.01	0.05 ± 0.02	0.02 ± 0.01
$\mathcal{U}_{cr}$	**0.00 ± 0.00**	0.19 ± 0.02	0.41 ± 0.07	0.20 ± 0.02
$\mathcal{U}_{fcp}$	**0.00 ± 0.00**	0.06 ± 0.04	0.16 ± 0.11	0.06 ± 0.04

The correction curves (Figs. 4(c) and 4(d)) revealed that $\mathcal{U}_{cr}$ and $\mathcal{U}_{fcp}$ exhibited inferior performance compared to $\mathcal{U}_{prob}$, $\mathcal{U}_{pe}$, and $\mathcal{U}_{mi}$. This indicates that correcting pixels identified by the latter enables faster achievement of higher performance. This is likely caused by their strong focus on ditches and natural streams (Fig. 5), which make up only a small portion of the dataset, but are frequently misclassified (Table 1). Specifically, an MCC of 0.95 was attainable with an average correction rate of 3% (approximately 2.87 million pixels) using $\mathcal{U}_{pe}$. Using the Bayesian t-test, we found that the probability of $\mathcal{U}_{pe}$, $\mathcal{U}_{mi}$, and $\mathcal{U}_{prob}$ being practically equivalent to be 100% (ROPE=0.05). Furthermore, the test suggested that $\mathcal{U}_{cr}$ performed significantly worse than all other methods with a probability of 100% (ROPE=0.05). $\mathcal{U}_{fcp}$ was found to perform significantly worse than $\mathcal{U}_{pe}$, $\mathcal{U}_{prob}$, and $\mathcal{U}_{mi}$ with a probability of 78.1%, 73.5%, and 69.1% (ROPE=0.05) respectively.

When focusing solely on pixel classifications predicted to be ditches or streams, overall $\mathcal{U}_{mi}$ was found to be most effective in identifying misclassified streams and ditches (Figs. 7(a) and 7(b)). A Bayesian t-test revealed that for ditch pixels incorrectly classified as stream pixels, $\mathcal{U}_{mi}$ had a significantly higher AUCE score with a probability greater than 95% (ROPE=0.05) when compared to $\mathcal{U}_{prob}$, $\mathcal{U}_{cr}$, and $\mathcal{U}_{fcp}$. Using $\mathcal{U}_{mi}$ to correct these errors, on average 70.6% of stream pixels (≈ 40 000) needed to be corrected to achieve an F_1 score of 0.95 for ditches. For correcting pixels classified as ditch, the Bayesian t-test revealed that $\mathcal{U}_{mi}$ had a significantly higher AUCE score than $\mathcal{U}_{cr}$ with a probability of 99.1% (ROPE=0.05). However, we found that $\mathcal{U}_{fcp}$ and $\mathcal{U}_{prob}$ lead to achieving an F_1 score of 0.95 for the stream pixels with fewer corrections than $\mathcal{U}_{mi}$. Both required on average the correction of 75% pixels (≈ 714 000). In contrast, $\mathcal{U}_{mi}$ required a correction of 79.7%. It should be noted that these F_1 scores were calculated not on all pixels, but only on those initially classified as ditch or natural stream.

$\mathcal{U}_{fcp}$ had significantly faster inference times compared to $\mathcal{U}_{cr}$ and the MC dropout-based $\mathcal{U}_{pe}$ and $\mathcal{U}_{mi}$ (Table 4). Specifically, processing the entire surface area of Sweden at a 0.5 m resolution using $\mathcal{U}_{fcp}$, producing both the actual prediction and the uncertainty estimates, would take approximately 80 h, whereas an MC dropout-based approach would require around 3 years on the same hardware. It should be noted that both MC dropout-based approaches have the same execution time,



(a) Sparsification Curve for the 1 m DEM

(b) Sparsification Curve for the 0.5 m DEM

(c) Correction Curve for the 1 m DEM

(d) Correction Curve for the 0.5 m DEM

Fig. 4. Sparsification and correction curves for the oracle, network probability ($\mathcal{U}_{prob}$), predictive entropy ($\mathcal{U}_{pe}$), mutual information ($\mathcal{U}_{mi}$), conformal regression ($\mathcal{U}_{cr}$), and feature conformal prediction ($\mathcal{U}_{fcp}$) computed on one test fold. The curves are shown for both resolutions of the digital elevation model (DEM), 1 m and 0.5 m. The classification performance was measured across all classes using the Matthews Correlation Coefficient (MCC).

Table 4
Execution times in seconds for predictive entropy ($\mathcal{U}_{pe}$), mutual information ($\mathcal{U}_{mi}$), conformal regression ($\mathcal{U}_{cr}$), and feature conformal prediction ($\mathcal{U}_{fcp}$) on one chip covering an area of 500 m ×500 m (1 m resolution) or 250 m ×250 m (0.5 m resolution). The reported values indicate the mean and standard deviation over all chips in the 10 test sets. Fastest execution time indicated in bold.

	$t_{1\,\mathrm{m}}$ (s)	$t_{0.5\,\mathrm{m}}$ (s)
$\mathcal{U}_{pe}/\mathcal{U}_{mi}$	14.13 ± 0.78	14.00 ± 0.72
$\mathcal{U}_{cr}$	1.75 ± 1.53	1.49 ± 0.22
$\mathcal{U}_{fcp}$	$\mathbf{0.06 \pm 0.01}$	$\mathbf{0.04 \pm 0.01}$

Table 5
Recall for the confident maps generated from the 0.5 m resolution data using conformal regression for different confidence levels, measured for the background ($Recall^{(b)}$), ditches ($Recall^{(d)}$), natural streams ($Recall^{(s)}$), and the class average ($Recall$). The reported values indicate the mean and standard deviation over 10 test folds.

Confidence	$Recall^{(b)}$	$Recall^{(d)}$	$Recall^{(s)}$	$Recall$
90.0%	1.00 ± 0.00	0.56 ± 0.04	0.24 ± 0.08	0.60 ± 0.03
80.0%	1.00 ± 0.00	0.59 ± 0.04	0.25 ± 0.08	0.61 ± 0.04
70.0%	1.00 ± 0.00	0.62 ± 0.05	0.27 ± 0.08	0.63 ± 0.04
60.0%	1.00 ± 0.00	0.65 ± 0.05	0.29 ± 0.09	0.64 ± 0.04
50.0%	1.00 ± 0.00	0.67 ± 0.04	0.30 ± 0.09	0.66 ± 0.04

since that time is dominated by the sampling process, which is the same for both approaches.

4.3. Conformal prediction performance

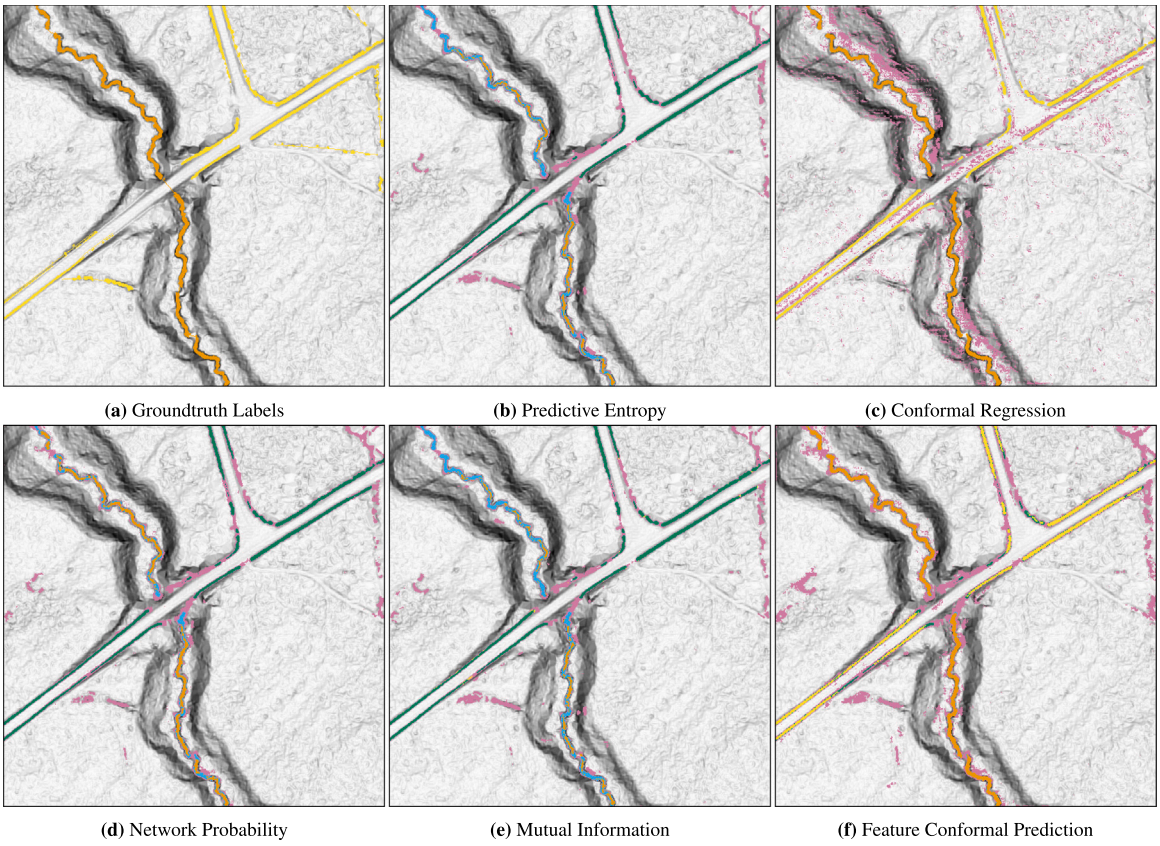
When generating confident maps using the conformal prediction approaches, FCP resulted in significantly lower recall for all confidence levels than conformal regression ($\mathcal{U}_{fcp}$: 0.12–0.13; $\mathcal{U}_{cr}$: 0.60–0.66), prompting a focus on maps generated using the latter. As expected, recall increased with decreasing confidence (Table 5). However, even highly confident maps covered a sizeable portion of background (100%), ditch (56%), and natural stream pixels (24%).

Similarly to expectation, classification performance degraded with decreasing confidence levels, with one notable exception being the background class, whose performance remained stable (Table 6). However, even at 50% confidence, the performance on confidently classified pixels, as measured by MCC, surpassed the overall performance on all pixels (Table 1).

5. Discussion

5.1. Choice of uncertainty quantification method

When comparing the evaluated uncertainty quantification approaches, FCP outperformed others in terms of AUSE (Table 2) but not in terms of AUCE (Table 3). This discrepancy stems from AUSE and AUCE addressing different questions. AUSE assesses alignment between predictions and uncertainty estimates (Dreissig et al., 2023), while AUCE evaluates the ability to identify misclassified pixels. The choice of method depends on the goal: AUSE is more informative for creating



(a) Groundtruth Labels **(b)** Predictive Entropy **(c)** Conformal Regression

(d) Network Probability **(e)** Mutual Information **(f)** Feature Conformal Prediction

Fig. 5. Illustration of the groundtruth map, as well as the uncertainty maps for the 0.5 m resolution showing the 5% most uncertain pixels as estimated by the evaluated uncertainty quantification approaches. The maps show the local slope image for certain background pixels and uncertain ones in **pink**. Furthermore, the maps show certain (**yellow**) and uncertain (**green**) ditches, as well as certain (**orange**) and uncertain (**blue**) streams.

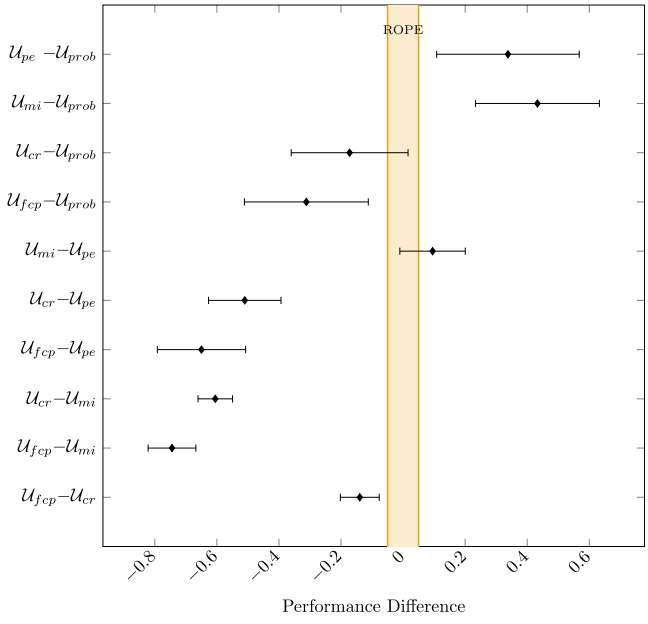


Fig. 6. High-density intervals derived using a Bayesian t-test for correlated observations indicating the intervals in which the performance differences between the compared methods, network probability ($\mathcal{U}_{prob}$), predictive entropy ($\mathcal{U}_{pe}$), mutual information ($\mathcal{U}_{mi}$), conformal regression ($\mathcal{U}_{cr}$), and feature conformal prediction ($\mathcal{U}_{fcp}$), lie with a probability of 95%. The performance is measured as area under the sparsification error curve for all classes, and the Region of Practical Equivalence (ROPE) indicates a performance difference of 0.05.

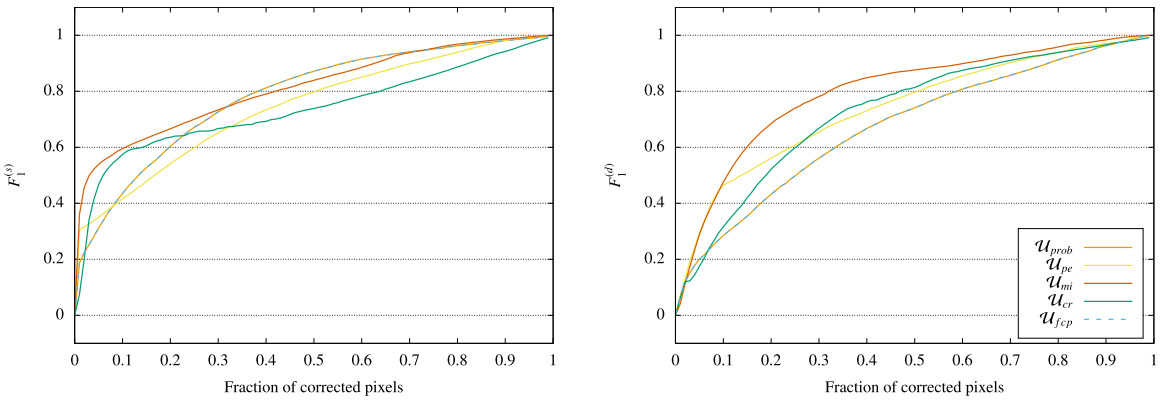
Table 6
Mapping performance for only the pixels included in the confident maps generated using conformal regression on the 0.5 m resolution data as measured by the Matthews Correlation Coefficient (MCC) for all classes, and the F_1 score for the background ($F_1^{(b)}$), ditches ($F_1^{(d)}$), and natural streams ($F_1^{(s)}$). The reported values indicate the mean and standard deviation over 10 test folds.

Confidence	$F_1^{(b)}$	$F_1^{(d)}$	$F_1^{(s)}$	MCC
90.0%	1.00 ± 0.00	0.83 ± 0.03	0.44 ± 0.10	0.82 ± 0.03
80.0%	1.00 ± 0.00	0.82 ± 0.03	0.44 ± 0.10	0.81 ± 0.03
70.0%	1.00 ± 0.00	0.81 ± 0.03	0.44 ± 0.09	0.80 ± 0.03
60.0%	1.00 ± 0.00	0.80 ± 0.03	0.43 ± 0.09	0.79 ± 0.03
50.0%	1.00 ± 0.00	0.79 ± 0.02	0.43 ± 0.09	0.78 ± 0.03

prediction uncertainty maps, whereas AUCE appears to be suitable for pixel-level correction.

Upon examining the uncertainty map generated by FCP for a broader area (Fig. 8(b)), it becomes clear that the model is generally confident in its ditch predictions, except in border regions or where ditches exhibit unusual bends. Additionally, while the two natural streams in the area (the orange lines in Fig. 8(a)) were not well identified by the model, it is relatively straightforward to trace their paths from the uncertainty maps due to the presence of uncertain background pixels on the map. This can help alert a human viewer to the presence of these streams, which would be imperceptible in the prediction map alone.

When examining the top-performing uncertainty quantification methods according to AUCE, we found that network probability, predictive entropy, and mutual information consistently identified predictions on ditch and natural stream pixels as the most uncertain ones, regardless of prediction correctness (Fig. 5). On the other hand, predictive entropy and mutual information tended to exhibit overconfidence in



(a) Correcting predicted ditch pixels

(b) Correcting predicted stream pixels

Fig. 7. Correction curves for the network probability ($\mathcal{U}_{prob}$), predictive entropy ($\mathcal{U}_{pe}$), mutual information ($\mathcal{U}_{mi}$), conformal regression ($\mathcal{U}_{cr}$), and FCP ($\mathcal{U}_{fcp}$) computed only on pixels from one test fold on the 0.5 m resolution data. The curves indicate the F_1 score for the stream class ($F_1^{(s)}$) and the ditch class ($F_1^{(d)}$) respectively, considering only pixels previously classified as ditch or stream.

incorrect predictions, as evidenced by their low AUSE scores. This tendency aligns with findings by Hertel et al. (2023), who also observed this characteristic of MC dropout-based approaches. Given that only about 1% of pixels belong to ditches, and even fewer to natural streams, it is likely that the methods' strong AUCE performance is an artifact of the highly skewed class distribution. This phenomenon arises because fixing a few pixels in classes with low instance counts and generally poorer performance can improve MCC scores more than correcting pixels from the mostly correct majority class (Table 1). As a result, one may find that for more balanced datasets, the AUCE scores of these methods may be lower compared to FCP. Additionally, the tendency to identify correctly classified pixels as uncertain can be problematic for their use in detecting incorrectly classified pixels, since the high false positive rate may lead people to dismiss detections of potentially misclassified pixels (Axelsson, 2000).

When specifically examining corrections of pixels misclassified as ditches or natural streams, we observed that mutual information outperformed other approaches in identifying ditch pixels mistakenly classified as streams. Conversely, FCP and network probability were more effective at identifying stream pixels incorrectly classified as ditches. This disparity may stem from the fact that most ditch pixels were accurately predicted, leaving only few natural stream pixels to be detected. In this scenario, overconfidence in incorrect predictions is more detrimental than when there is a larger number of misclassified pixels, as it was the case for the pixels classified as natural stream. Given that natural streams underlie stronger protections (Swedish PEFC, 2023), it is more important to identify stream pixels misclassified as ditch than vice versa.

One notable finding was that network probability achieved comparable AUCE scores to MC dropout-based approaches, while outperforming them in AUSE scores. The strong performance in identifying stream pixels among those classified as ditch is likely a consequence of that. Thus, it appears that network probability has effectively balanced high uncertainty values for ditches and streams with cautious avoidance of undue certainty in incorrectly classified pixels, at least for this dataset.

In Fig. 8(c), we observe the pixel corrections for pixels marked as most uncertain by network probability. It is evident that all predicted ditch and stream pixels were corrected due to their relatively high uncertainty. However, there are also instances where pixels were not corrected despite being wrongly predicted (stream pixels in Fig. 8(c), zoomed-in region), resulting from the model's undue confidence in its predictions. This confidence can be attributed to the fact that the natural stream is not visible in the DEM, as indicated by the one pixel wide line in the ground truth. Given that the figure showcases the correction of the 5% most uncertain pixels, a significant number of background pixels were also corrected, even though they were correctly predicted. One notable aspect of these corrected background pixels is that they appear to follow a specific pattern. Upon analyzing the slope values of those corrected background pixels, we found them to be significantly higher than average slope values. Furthermore, similar patterns have been observed in data from other regions, but not consistently across all areas, suggesting that these may be caused by minor differences in the data collection process.

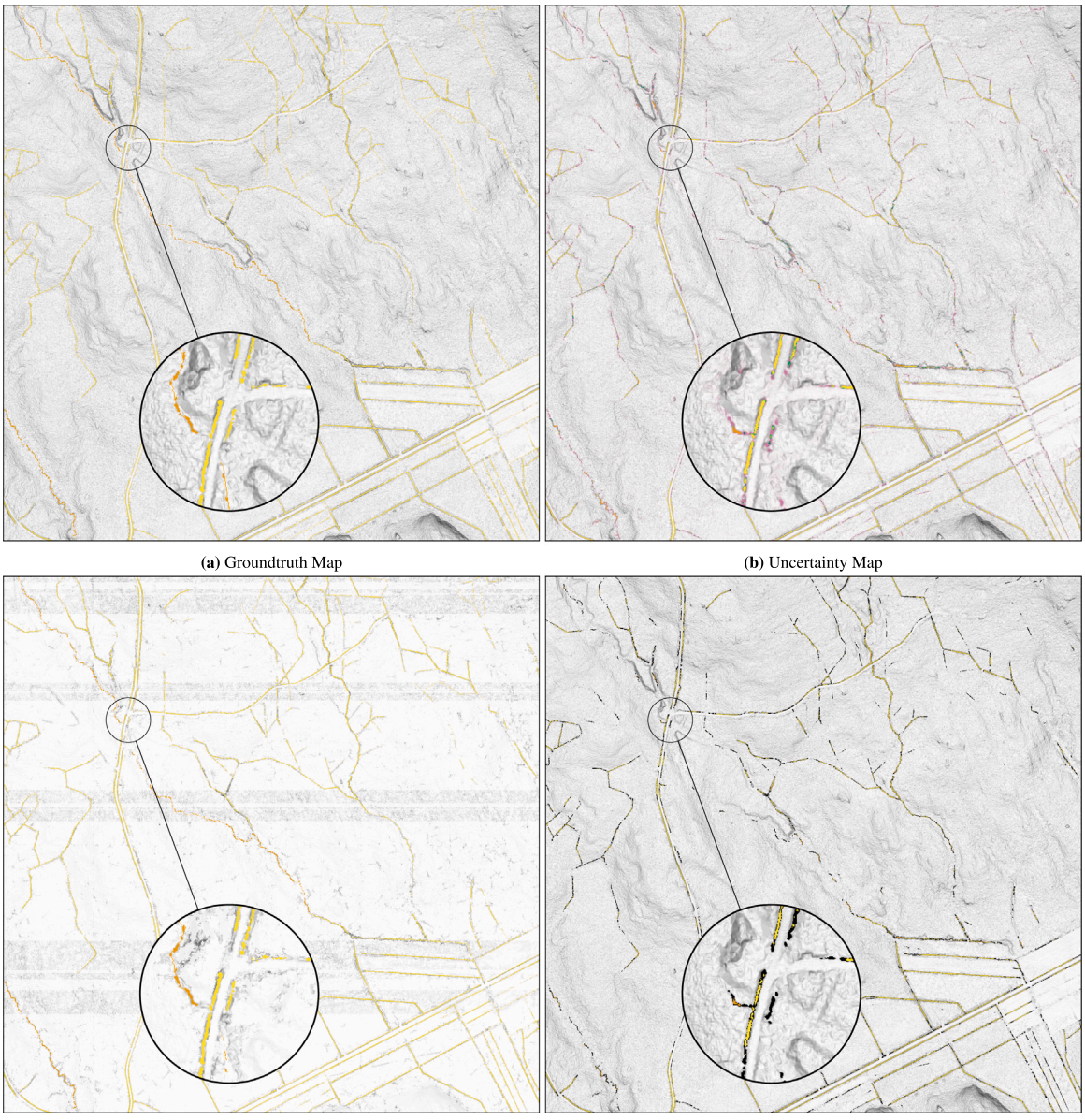
When evaluating execution performance, arguably, the fastest uncertainty estimates were derived using network probability, since it equals the model's inference speed of approximately 0.014 s per chip, resulting in an estimated processing time of 28 h for all of Sweden. While this was significantly shorter than the 80 h required for FCP, we deem FCP still feasible, especially when compared to the execution times for MC dropout-based approaches ($\approx$ 3 years) or conformal regression ($\approx$ 124 days). It is worth noting that these times can be significantly reduced by using fewer Monte Carlo samples. For example, utilizing just 10 samples, as Kampffmeyer et al. (2016), would reduce the time required for MC dropout and conformal regression to 280 h and 298 h, respectively. However, this may come at the cost of reduced uncertainty quantification performance.

In summary, our results show that FCP yielded the most accurate uncertainty estimates at a reasonable processing speed. Therefore, we believe it is well-suited as a method for generating uncertainty maps. However, when attempting to identify which pixels require correction in the generated ditch and stream maps, we found that using network probability was more effective. This approach identified the pixels that needed correction better and resulted in lower execution times.

5.2. Impact of resolution

The classification performance was improved when detecting ditches and streams on higher resolution data (Table 1). This is reasonable since landscape outlines were captured more accurately, which simplified the detection problem. This finding aligns with the findings by Busarello et al. (2025) on mapping ditches and streams, but also with findings on mapping other terrain features, such as ephemeral gullies (Chowdhuri et al., 2021), and rock glaciers (Robson et al., 2020).

Higher resolution DEMs also yielded more accurate uncertainty estimates as indicated by the obtained AUSE scores. While it is unsurprising, that a lower resolution leads to a higher uncertainty (Pogson and Smith, 2015; Wu et al., 2024), the observed reduced alignment between estimated model uncertainty and actual performance is likely due to the network's generally poorer performance on lower resolution data. In contrast to AUSE, the AUCE scores were mostly unaffected by the resolution, presumably since AUCE performance was largely



(a) Groundtruth Map **(b)** Uncertainty Map

(c) Corrected Map **(d)** Confident Map

Fig. 8. Illustration of the groundtruth, uncertainty, corrected, and confident map over an area of 1.5 km × 1.5 km at a 0.5 m resolution. In all maps, certain or correct background pixels are shown by the local slope image, while ditches are shown in **yellow**, and streams in **orange**. The uncertainty map was generated using feature conformal prediction and displays the 5% most uncertain background (**pink**), ditch (**green**), and stream (**blue**) pixels. The corrected map was derived by correcting the 5% most uncertain pixels as estimated by network probability. Corrected pixels are shown with full intensity, while not corrected pixels have low intensity. The confident map was derived using conformal regression at a 90% confidence level, and pixels where the model did not commit to one class are shown in **black**.

improved by ditch and stream detection rather than uncertainty quantification accuracy. Thus, as long as a method could identify most ditch and stream pixels it would get a high AUCE score, even if it marked many correctly classified pixels as uncertain.

Most methods showed increased uncertainty quantification performance with higher resolutions, except network probability, which decreased due to overconfidence in its predictions. This overconfidence was caused by the simplified learning problem, which allowed the model to assign more extreme probability estimates to pixels, as incentivized by the training process. As noted by Guo et al. (2017) and Sensoy et al. (2018), this leads to poorer uncertainty estimates.

There was no difference in processing time for a chip of 1 m resolution versus one with a 0.5 m resolution (Table 4), since both have the same number of pixels. However, four 0.5 m resolution chips are required to cover the same area as one 1 m resolution chip. This results in four times longer processing times for the 0.5 m resolution. As such, it is important to consider whether the gained performance improvements justify the increased processing costs.

In summary, there is a motivation for conducting high-resolution LiDAR scans to improve ditch and stream detection and to obtain more accurate uncertainty estimates. However, this may decrease the accuracy of uncertainty estimates obtained by network probability as performance improves.

5.3. Confident segmentation maps

When generating confident segmentation maps, we found that only $\mathcal{U}_{cr}$ consistently produced a reasonable number of single-class predictions for various confidence levels, ruling out $\mathcal{U}_{fcp}$ from further evaluation. This appears contradictory to the findings by Teng et al. (2023), who showed that FCP produced shorter confidence bands than a baseline conformal prediction approach. It is reasonable to assume

that shorter confidence bands also would lead to a higher number of single-class predictions. However, it should be noted that the conformal prediction approach used by Teng et al. (2023) differs from $\mathcal{U}_{cr}$ used in this article, which is the likely reason for the observed differences.

For $\mathcal{U}_{cr}$, recall improved as the confidence level decreased (Table 5). This was expected since lower confidence thresholds allow $\mathcal{U}_{cr}$ to make more errors and thus commit to single-class predictions for more pixels. Similarly in line with expectations was the observed decrease in precision, indicated by lower MCC and F_1 scores (Table 6). This decrease is caused by $\mathcal{U}_{cr}$ actually making more errors at lower confidence levels.

Compared to the models' results on all pixels (Table 1), we observed improved classification performance for predictions with high confidence levels (Table 6). Specifically, we achieved an MCC of 0.82 for 90% confident predictions, surpassing the MCC of 0.76 obtained on all pixel predictions. This performance difference was largely due to clear improvement in the ditch class, which was attained through $\mathcal{U}_{cr}$ not assigning a class in border regions where it is challenging to determine where the ditch ends and the background begins, or areas where the ditch was not clearly visible in the DEM (Fig. 8(d), zoomed-in region). These observations align well with the findings by Koski et al. (2023), who found that the main causes of error in detecting small watercourses with deep learning were boundary issues and unclear visual expression in the DEM.

Despite committing to a single class with high confidence, it is possible for $\mathcal{U}_{cr}$ to make errors. For example, many natural stream pixels were confidently predicted as background (Fig. 8(d), zoomed-in region), which was not unexpected. This outcome is consistent with the fact that $\mathcal{U}_{cr}$ allows for 10% errors at a 90% confidence level. It is important to note that the guarantees provided by this method apply to probability intervals rather than the classes themselves. A model that consistently missed to predict the natural stream class, would make significantly fewer errors than 10%, due to its low occurrence rate (less than 1%). Instead, it would in over 99% of the cases be correct in predicting the probability for the stream class to be close to 0%. Consequently, $\mathcal{U}_{cr}$ primarily prevented overprediction in minority classes, such as ditch and stream, as observed in Fig. 8(d) and reflected in their low recall values (Table 5).

Our analysis revealed that neither $\mathcal{U}_{cr}$ nor $\mathcal{U}_{fcp}$ are particularly suitable for generating confident maps of ditches and natural streams. Although $\mathcal{U}_{cr}$ produced more confident predictions than $\mathcal{U}_{fcp}$, the generated maps only covered around 60% of all pixels, particularly omitting ditch and stream pixels. This means that the prediction sets for pixels of these classes frequently contained more than one possible prediction. This observation is in line with the findings by Ghosh et al. (2023), who show that conformal prediction tends to result in large prediction sets for challenging datasets, while obtaining narrower sets for simple ones. Apart from this issue, it also took a considerable amount of time to generate the confident maps (Table 4).

5.4. Limitations and future work

This article's evaluation of uncertainty quantification methods is limited to one specific remote sensing task with an extreme class distribution. This may have skewed results, as MC dropout-based solutions likely perform differently in terms of AUCE on tasks with more balanced distributions. Although investigating extreme cases is valuable, given that classes with relatively few instances are not uncommon in remote sensing (Kossmann et al., 2021), it would be interesting to investigate if MC dropout's AUCE performance would decrease when applied to tasks with more balanced distributions.

Furthermore, the dataset used in this study is limited by its two-resolution format (1 m and 0.5 m). As demonstrated, classification and uncertainty quantification performance improve with increasing resolution. However, it is plausible that returns diminish at some point, warranting investigation into the optimal resolution threshold. Additionally, the uncertainty quantification performance of $\mathcal{U}_{prob}$ has been observed to decrease with increased resolution, suggesting a possible trend where higher resolutions lead to overconfident predictions. Higher resolution datasets would aid in investigating this trend as well.

Another limitation of our study is that we have only investigated a restricted set of uncertainty quantification approaches. For example, Bayesian neural networks (Blundell et al., 2015) were excluded from this study since they cannot derive uncertainty estimates from the same model as the other investigated approaches. This would have complicated direct comparisons between the methods, as it is less clear if differences in uncertainty quantification performance are due to differences in the used methods or due to the different models. Nevertheless, exploring Bayesian neural networks would be valuable for future research as they have been shown to outperform MC dropout-based approaches by Hertel et al. (2023). Similarly, deep ensembles have been shown to perform better than MC dropout-based approaches (Lakshminarayanan et al., 2017). Investigating how they compare to the evaluated conformal prediction-based approaches could be worthwhile. However, due to their significant training time requirements, we excluded them from this article; using the recommended number of networks in the ensemble would have quintupled the necessary training time.

It should be noted that none of the investigated uncertainty quantification approaches is able to handle out-of-distribution (OOD) data, i.e., data that is distinctively different from the training data. Alarab et al. (2021) have shown this for network probability and MC dropout-based approaches, while this limitation of conformal prediction has been pointed out, for example, by Angelopoulos et al. (2022). This is not a big problem for the studied dataset, since it has been specifically designed to be representative of the Swedish landscape (Busarello et al., 2025). However, in situations where OOD data is present, the obtained uncertainty estimates may not be reliable. One approach to handle OOD data would be to build on ideas from the "Learn then Test" framework (Angelopoulos et al., 2022).

Our investigation was further limited by focusing solely on conformal regression approaches within either feature space ($\mathcal{U}_{fcp}$) or output space ($\mathcal{U}_{cr}$). The focus on probability ranges rather than actual class predictions may have hindered the utility of generated confidence maps, as they tended to suppress minority class predictions. In the future, this limitation could be addressed by exploring whether the conformal classification approach by Wieslander et al. (2021) can be made more computationally efficient or through further investigation into recent methods proposed by Mossina et al. (2024), Brunekreef et al. (2024). By focusing on conformal classification approaches, the guarantees provided by the conformal predictor would apply directly to the classification outcome, and thus might produce more usable confident maps.

6. Conclusions

In this article, we investigated various uncertainty quantification techniques, including network probability, predictive entropy, mutual information, conformal regression, and feature conformal prediction, and applied them to a specific remote sensing task: identifying ditches and natural streams from elevation data sourced from a digital elevation model (DEM). Additionally, the impact of different DEM resolutions on classification and uncertainty quantification performance was explored. Furthermore, confident maps were generated using conformal prediction methods. Our key findings include:

- Feature conformal prediction (Teng et al., 2023) produces uncertainty estimates most aligned with the actual neural network performance at a reasonable cost to the execution time. However, for correcting misclassified pixels, the network probability output is more suitable, at least for the investigated dataset.
- A higher resolution DEM leads to better classification performance and better uncertainty estimates.

- Conformal regression and feature conformal prediction are not suitable to generate confident maps, since they are overly conservative in their estimates and the model performance is too limited.

CRediT authorship contribution statement

Florian Westphal: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **William Lidberg:** Writing – review & editing, Resources, Funding acquisition, Data curation, Conceptualization. **Mariana Dos Santos Toledo Busarello:** Writing – review & editing, Data curation. **Anneli M. Ågren:** Writing – review & editing, Resources, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Ollama with the llama3.1 model in order to improve the language of the written text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

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.

Acknowledgments

This study was funded by the Swedish research council Formas (proj. no. 2021-00115) and Knut and Alice Wallenberg Foundation (2018.0259 Future Silviculture). It was also partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation. The funding sources had no involvement in study design, collection, analysis and interpretation of data, nor in the writing of the article.

Data availability

I have shared the link to my data/code at the Attach File step.

Uncertainty Quantification for LiDAR-based Maps of Ditches and Natural Streams (Original data) (GitHub)
Automatic Detection of Ditches and Natural Streams from Digital Elevation Models Using Deep Learning (Reference data) (Swedish National Data Service).

References

Alarab, I., Prakoonwit, S., Nacer, M.I., 2021. Illustrative discussion of MC-dropout in general dataset: Uncertainty estimation in bitcoin. Neural Process. Lett. 53, 1001–1011. http://dx.doi.org/10.1007/s11063-021-10424-x, URL http://dx.doi.org/10.1007/s11063-021-10424-x.

Angelopoulos, A.N., Bates, S., Candès, E.J., Jordan, M.I., Lei, L., 2022. Learn then test: Calibrating predictive algorithms to achieve risk control. URL https://arxiv.org/abs/2110.01052. arXiv:2110.01052.

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M.Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., Chintala, S., 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. ASPLOS'24, ACM, http://dx.doi.org/10.1145/3620665.3640366, URL https://pytorch.org/assets/pytorch2-2.pdf.

Axelsson, S., 2000. The base-rate fallacy and the difficulty of intrusion detection. ACM Trans. Inf. Syst. Secur. 3, 186–205. http://dx.doi.org/10.1145/357830.357849, URL http://dx.doi.org/10.1145/357830.357849.

Benavoli, A., Corani, G., Demšar, J., Zaffalon, M., 2017. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. J. Mach. Learn. Res. 18 (1), 2653–2688.

Bishop, K., Allan, C., Bringmark, L., Garcia, E., Hellsten, S., Högbom, L., Johansson, K., Lomander, A., Meili, M., Munthe, J., Nilsson, M., Porvari, P., Skyllberg, U., Sørensen, R., Zetterberg, T., Åkerblom, S., 2009. The effects of forestry on hg bioaccumulation in nemoral/boreal waters and recommendations for good silvicultural practice. AMBIO: A J. Hum. Environ. 38, 373–380. http://dx.doi.org/10.1579/0044-7447-38.7.373.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65, 2–16. http://dx.doi.org/10.1016/j.isprsjprs.2009.06.004, URL http://dx.doi.org/10.1016/j.isprsjprs.2009.06.004.

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network. In: Bach, F., Blei, D. (Eds.), Proceedings of the 32nd International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 37, PMLR, Lille, France, pp. 1613–1622, URL https://proceedings.mlr.press/v37/blundell15.html.

Boström, H., Johansson, U., 2020. Mondrian conformal regressors. In: Gammerman, A., Vovk, V., Luo, Z., Smirnov, E., Cherubin, G. (Eds.), Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications. In: Proceedings of Machine Learning Research, vol. 128, PMLR, pp. 114–133, URL https://proceedings.mlr.press/v128/bostrom20a.html.

Brunekreef, J., Marcus, E., Sheombarsing, R., Sonke, J.-J., Teuwen, J., 2024. Kandinsky conformal prediction: Efficient calibration of image segmentation algorithms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4135–4143.

Busarello, M.D.S.T., Ågren, A.M., Westphal, F., Lidberg, W., 2025. Automatic detection of ditches and natural streams from digital elevation models using deep learning. Comput. Geosci. 196, 105875. http://dx.doi.org/10.1016/j.cageo.2025.105875.

Chaudhary, P., Leitão, J.P., Donauer, T., D'Aronco, S., Perraudin, N., Obozinski, G., Perez-Cruz, F., Schindler, K., Wegner, J.D., Russo, S., 2022. Flood uncertainty estimation using deep ensembles. Water 14, 2980. http://dx.doi.org/10.3390/w14192980.

Chicco, D., Tötsch, N., Jurman, G., 2021. The matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 14, http://dx.doi.org/10.1186/s13040-021-00244-z, URL http://dx.doi.org/10.1186/s13040-021-00244-z.

Chowdhuri, I., Pal, S.C., Saha, A., Chakrabortty, R., Roy, P., 2021. Evaluation of different DEMs for gully erosion susceptibility mapping using in-situ field measurement and validation. Ecol. Informatics 65, 101425. http://dx.doi.org/10.1016/j.ecoinf.2021.101425, URL https://www.sciencedirect.com/science/article/pii/S1574954121002168.

Corani, G., Benavoli, A., 2015. A Bayesian approach for comparing cross-validated algorithms on multiple data sets. Mach. Learn. 100, 285–304. http://dx.doi.org/10.1007/s10994-015-5486-z, URL http://dx.doi.org/10.1007/s10994-015-5486-z.

Cortés-Ciriano, I., Bender, A., 2019. Reliable prediction errors for deep neural networks using test-time dropout. J. Chem. Inf. Model. 59, 3330–3339. http://dx.doi.org/10.1021/acs.jcim.9b00297, URL http://dx.doi.org/10.1021/acs.jcim.9b00297.

Dreissig, M., Piewak, F., Boedecker, J., 2023. On the calibration of uncertainty estimation in lidar-based semantic segmentation. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems. ITSC, IEEE, http://dx.doi.org/10.1109/itsc57777.2023.10422384, URL http://dx.doi.org/10.1109/itsc57777.2023.10422384,

Eigen, D., Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV.

Florinsky, I.V., 2016. Chapter 2 - topographic surface and its characterization. In: Florinsky, I.V. (Ed.), Digital Terrain Analysis in Soil Science and Geology (Second Edition), second ed. Academic Press, pp. 7–76. http://dx.doi.org/10.1016/B978-0-12-804632-6.00002-X, URL https://www.sciencedirect.com/science/article/pii/B978012804632600002X.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (Eds.), Proceedings of the 33rd International Conference on Machine Learning. In: Proceedings of Machine Learning Research, 48, PMLR, New York, New York, USA, pp. 1050–1059, URL https://proceedings.mlr.press/v48/gal16.html.

Gal, Y., Hron, J., Kendall, A., 2017. Concrete dropout. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. vol. 30, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2017/file/84ddfb34126fc3a48ee38d7044e87276-Paper.pdf.

Garey, M.R., Johnson, D.S., 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman and Company.

Gawlikowski, J., Saha, S., Kruspe, A., Zhu, X.X., 2022. An advanced Dirichlet prior network for out-of-distribution detection in remote sensing. IEEE Trans. Geosci. Remote Sens. 60, 1–19. http://dx.doi.org/10.1109/tgrs.2022.3140324, URL http://dx.doi.org/10.1109/tgrs.2022.3140324.

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. http://dx.doi.org/10.1007/s10462-023-10562-9, URL http://dx.doi.org/10.1007/s10462-023-10562-9. arXiv:2107.03342v3.

Ghosh, S., Belkhouja, T., Yan, Y., Doppa, J.R., 2023. Improving uncertainty quantification of deep classifiers via neighborhood conformal prediction: Novel algorithm and theoretical analysis. Proc. the AAAI Conf. Artif. Intell. 37, 7722–7730. http://dx.doi.org/10.1609/aaai.v37i6.25936.

Goan, E., Fookes, C., 2020. Bayesian neural networks: An introduction and survey. Springer International Publishing, pp. 45–87. http://dx.doi.org/10.1007/978-3-030-42553-1_3, URL http://dx.doi.org/10.1007/978-3-030-42553-1_3,

Gorodkin, J., 2004. Comparing two K-category assignments by a K-category correlation coefficient. Comput. Biol. Chem. 28, 367–374. http://dx.doi.org/10.1016/j.compbiolchem.2004.09.006, URL http://dx.doi.org/10.1016/j.compbiolchem.2004.09.006.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 70, PMLR, pp. 1321–1330, URL https://proceedings.mlr.press/v70/guo17a.html.

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. http://dx.doi.org/10.1016/j.rse.2022.113388, URL http://dx.doi.org/10.1016/j.rse.2022.113388.

Iagaru, D., Gottschling, N.M., 2023. Uncertainty quantification with deep ensemble methods for super-resolution of sentinel 2 satellite images. In: International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. MDPI, p. 4. http://dx.doi.org/10.3390/psf2023009004,

Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., Brox, T., 2018. Uncertainty estimates and multi-hypotheses networks for optical flow. In: Proceedings of the European Conference on Computer Vision. ECCV.

Kampffmeyer, M., Salberg, A.B., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., Bethel, W., 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environ. Model. Softw. 144, 105139. http://dx.doi.org/10.1016/j.envsoft.2021.105139.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations. URL http://arxiv.org/abs/1412.6980.

Koski, C., Kettunen, P., Poutanen, J., Zhu, L., Oksanen, J., 2023. Mapping small watercourses from DEMs with deep learning—exploring the causes of false predictions. Remote. Sens. 15, 2776. http://dx.doi.org/10.3390/rs15112776, URL http://dx.doi.org/10.3390/rs15112776.

Kossmann, D., Wilhelm, T., Fink, G.A., 2021. Towards tackling multi-label imbalances in remote sensing imagery. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE, pp. 5782–5789. http://dx.doi.org/10.1109/icpr48806.2021.9412588, URL http://dx.doi.org/10.1109/icpr48806.2021.9412588,

Labuzzetta, C.J., 2022. Practical Methods for the Advancement of Precision Conservation Via Land Cover Classification and Conformal Prediction (Ph.D. thesis). Iowa State University.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. vol. 30, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Lidberg, W., Paul, S.S., Westphal, F., Richter, K.F., Lavesson, N., Melniks, R., Ivanovs, J., Ciesielski, M., Leinonen, A., Ågren, A.M., 2023. Mapping drainage ditches in forested landscapes using deep learning and aerial laser scanning. J. Irrig. Drain. Eng. 149, http://dx.doi.org/10.1061/jidedh.ireng-9796, URL http://dx.doi.org/10.1061/jidedh.ireng-9796.

Martínez-Ferrer, L., Moreno-Martínez, Á., Campos-Taberner, M., García-Haro, F.J., Muñoz-Marí, J., Running, S.W., Kimball, J., Clinton, N., Camps-Valls, G., 2022. Quantifying uncertainty in high resolution biophysical variable retrieval with machine learning. Remote Sens. Environ. 280, 113199. http://dx.doi.org/10.1016/j.rse.2022.113199, URL http://dx.doi.org/10.1016/j.rse.2022.113199.

Matthews, B., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. et Biophys. Acta (BBA) - Protein Struct. 405, 442–451. http://dx.doi.org/10.1016/0005-2795(75)90109-9, URL http://dx.doi.org/10.1016/0005-2795(75)90109-9.

Mossina, L., Dalmau, J., Andéol, L., 2024. Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3574–3584.

Mukhoti, J., Gal, Y., 2018. Evaluating Bayesian deep learning methods for semantic segmentation. URL http://arxiv.org/abs/1811.12709v2. arXiv:1811.12709v2.

O'Neil, G.L., Goodall, J.L., Behl, M., Saby, L., 2020. Deep learning using physically-informed input data for wetland identification. Environ. Model. Softw. 126, 104665. http://dx.doi.org/10.1016/j.envsoft.2020.104665.

Pagella, T.F., Sinclair, F.L., 2014. Development and use of a typology of mapping tools to assess their fitness for supporting management of ecosystem service provision. Landsc. Ecol. 29, 383–399. http://dx.doi.org/10.1007/s10980-013-9983-9, URL http://dx.doi.org/10.1007/s10980-013-9983-9.

Pakdaman Naeini, M., Cooper, G., Hauskrecht, M., 2015. Obtaining well calibrated probabilities using Bayesian binning. Proc. the AAAI Conf. Artif. Intell. 29, http://dx.doi.org/10.1609/aaai.v29i1.9602.

Pogson, M., Smith, P., 2015. Effect of spatial data resolution on uncertainty. Environ. Model. Softw. 63, 87–96. http://dx.doi.org/10.1016/j.envsoft.2014.09.021, URL http://dx.doi.org/10.1016/j.envsoft.2014.09.021.

Robson, B.A., Bolch, T., MacDonell, S., Hölbling, D., Rastner, P., Schaffer, N., 2020. Automated detection of rock glaciers using deep learning and object-based image analysis. Remote Sens. Environ. 250, 112033. http://dx.doi.org/10.1016/j.rse.2020.112033, URL https://www.sciencedirect.com/science/article/pii/S003442572030403X.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. Springer International Publishing, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28, URL http://dx.doi.org/10.1007/978-3-319-24574-4_28,

Savelonas, M.A., Veinidis, C.N., Bartsokas, T.K., 2022. Computer vision and pattern recognition for the analysis of 2D/3D remote sensing data in geoscience: A survey. Remote. Sens. 14, 6017. http://dx.doi.org/10.3390/rs14236017, URL http://dx.doi.org/10.3390/rs14236017.

Sensoy, M., Kaplan, L., Kandemir, M., 2018. Evidential deep learning to quantify classification uncertainty. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. 31, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf.

Singh, G., Moncrieff, G., Venter, Z., Cawse-Nicholson, K., Slingsby, J., Robinson, T.B., 2024. Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction. Sci. Rep. 14, http://dx.doi.org/10.1038/s41598-024-65954-w, URL http://dx.doi.org/10.1038/s41598-024-65954-w.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Swedish PEFC, 2023. Forest Use Standard. Technical Report, (PEFC SWE 002:5), Swedish PEFC.

Teng, J., Wen, C., Zhang, D., Bengio, Y., Gao, Y., Yuan, Y., 2023. Predictive inference with feature conformal prediction. In: The Eleventh International Conference on Learning Representations. URL https://openreview.net/forum?id=0uRm1YmFTu.

Toth, C., Jóźków, G., 2016. Remote sensing platforms and sensors: A survey. ISPRS J. Photogramm. Remote Sens. 115, 22–36. http://dx.doi.org/10.1016/j.isprsjprs.2015.10.004, URL http://dx.doi.org/10.1016/j.isprsjprs.2015.10.004.

Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic Learning in a Random World. Springer-Verlag, http://dx.doi.org/10.1007/b106715, URL http://dx.doi.org/10.1007/b106715.

Wieslander, H., Harrison, P.J., Skogberg, G., Jackson, S., Friden, M., Karlsson, J., Spjuth, O., Wahlby, C., 2021. Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. IEEE J. Biomed. Heal. Informatics 25, 371–380. http://dx.doi.org/10.1109/jbhi.2020.2996300, URL http://dx.doi.org/10.1109/jbhi.2020.2996300.

Wu, L., Xu, Y., Li, R., 2024. Effects of input data accuracy, catchment threshold areas and calibration algorithms on model uncertainty reduction. Eur. J. Soil Sci. 75, http://dx.doi.org/10.1111/ejss.13519.

Xu, Y., Bai, T., Yu, W., Chang, S., Atkinson, P.M., Ghamisi, P., 2022. AI security for geoscience and remote sensing: Challenges and future trends. http://dx.doi.org/10.1109/MGRS.2023.3272825, URL http://arxiv.org/abs/2212.09360v2. arXiv:2212.09360v2. IEEE Geoscience and Remote Sensing Magazine, Volume 11, Issue 2, Pages 60-85, 2023.

Xu, K., Shi, Z., Zhang, H., Wang, Y., Chang, K.-W., Huang, M., Kailkhura, B., Lin, X., Hsieh, C.-J., 2020. Automatic perturbation analysis for scalable certified robustness and beyond. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems. vol. 33, Curran Associates, Inc., pp. 1129–1141, URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0cbc5671ae26f67871cb914d81ef8fc1-Paper.pdf.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716. http://dx.doi.org/10.1016/j.rse.2020.111716, URL http://dx.doi.org/10.1016/j.rse.2020.111716.

Yule, G.U., 1912. On the methods of measuring association between two attributes. J. R. Stat. Soc. 75, 579. http://dx.doi.org/10.2307/2340126, URL http://dx.doi.org/10.2307/2340126.