Pontifícia Universidade Católica
do Rio de Janeiro

Jorge Andres Chamorro Martinez

Semi-Automatic Monitoring of Deforestation in the Brazilian Amazon and Cerrado Biomes: Uncertainty Estimation and Characterization of High Uncertainty Areas

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica.

Advisor : Prof. Raul Queiroz Feitosa
Co-advisor: Prof. Gilson Alexandre Ostwald Pedro da Costa

Rio de Janeiro
December 2023

Pontifícia Universidade Católica
do Rio de Janeiro

Jorge Andres Chamorro Martinez

Semi-Automatic Monitoring of Deforestation in the Brazilian Amazon and Cerrado Biomes: Uncertainty Estimation and Characterization of High Uncertainty Areas

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica. Approved by the Examination Committee.

Prof. Raul Queiroz Feitosa
Advisor
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Gilson Alexandre Ostwald Pedro da Costa
Co-advisor
Universidade do Estado do Rio de Janeiro – UERJ

Gilberto Câmara Neto
Instituto Nacional de Pesquisas Espaciais – INPE

Cláudio Aparecido Almeida
Instituto Nacional de Pesquisas Espaciais – INPE

Prof. Wesley Nunes Goncalves
Universidade Federal de Mato Grosso do Sul – UFMGS

Gilson Antonio Giraldi
Laboratório Nacional de Computação Científica – LNCC

Rio de Janeiro, December the 18th, 2023

Jorge Andres Chamorro Martinez

The author received his bachelor's degree in Electronic Engineering at the University of Nariño in 2015. He obtained his master's degree in Electrical Engineering with emphasis on Signal Processing and Control at the Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) in 2019. Since then, he has worked in the fields of Machine Learning, Digital Image Processing and Remote Sensing.

Bibliographic data

Chamorro J., A.

Semi-Automatic Monitoring of Deforestation in the Brazilian Amazon and Cerrado Biomes: Uncertainty Estimation and Characterization of High Uncertainty Areas / Jorge Andres Chamorro Martinez; advisor: Raul Queiroz Feitosa; co-advisor: Gilson Alexandre Ostwald Pedro da Costa. – Rio de janeiro: PUC-Rio, Departamento de Engenharia Elétrica, 2023.

v., 145 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Estimativa de Incerteza;. 3. Detecção de Desmatamento;. 4. Interpretação de Incerteza;. 5. Amazônia Brasileira;. 6. Cerrado Brasileiro;. 7. Imagens Ópticas;. 8. Aprendizado Profundo;. 9. Sensoriamento Remoto. I. Feitosa, R. Q.. II. Costa, G. A. O. P.. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Acknowledgments

I would like to thank my advisors, Profs. Raul Feitosa and Gilson Costa, for all the help, support, and friendship. To my family for their unconditional support. I would also like to thank all the members of the Computer Vision Lab (LVC) research group for their help and their friendship. Special thanks to Cassiano Gustavo Messias, Luciana Soler, Thaíse Azevedo and the PRODES Amazônia and Cerrado team at INPE, who actively supported this research. To the members of the FORSTCARe project for their constant feedback. Special thanks to Mabel Ortega and Gustavo Nascimento for their help. This study was financed in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Chamorro J., A.; Feitosa, R. Q. (Advisor); Costa, G. A. O. P. (Co-Advisor). **Semi-Automatic Monitoring of Deforestation in the Brazilian Amazon and Cerrado Biomes: Uncertainty Estimation and Characterization of High Uncertainty Areas**. Rio de Janeiro, 2023. 145p. Tese de doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Official monitoring of deforestation in the Brazilian Amazon has relied traditionally on human experts who visually evaluate remote sensing images and label each individual pixel as deforestation or no deforestation. That methodology is obviously costly and time-consuming due to the vast monitored area. The reason for not using fully automatic methods for the task is the need for the highest possible accuracies in the authoritative deforestation figures. In this work, a semi-automatic, deep learning-based alternative is proposed, in which a deep neural network is first trained with existing images and references from previous years, and employed to perform deforestation detection on recent images. After inference, the uncertainty in the network's pixel-level results is estimated, and it is assumed that low-uncertainty classification results can be trusted. The remaining high-uncertainty regions, which correspond to a small percentage of the test area, are then submitted to post classification, e.g., an auditing procedure carried out visually by a human specialist. In this way, the manual labeling effort is greatly reduced.

We investigate various uncertainty estimation strategies, including confidence-based approaches, Monte Carlo Dropout (MCD), deep ensembles and evidential learning, and evaluate different uncertainty metrics. Furthermore, we conduct a comprehensive analysis to identify the characteristics of forest areas that contribute to high uncertainty. We illustrate the main conclusions of the analysis upon 25 selected polygons on four target sites, which exemplify common causes of uncertainty. The target sites are located in challenging study areas in the Brazilian Amazon and Cerrado biomes. Through experimental evaluation on those sites, we demonstrate that the proposed semi-automated methodology achieves impressive F1-score values which exceeds 97%, while reducing the visual auditing workload to just 3% of the target area. The current code is available at `https://github.com/DiMorten/deforestation_uncertainty`.

Keywords

Uncertainty Estimation; Deforestation Detection; Uncertainty Interpretation; Brazilian Amazon; Brazilian Cerrado; Optical Imagery; Deep Learning; Remote Sensing

Resumo

Chamorro J., A.; Feitosa, R. Q.; Costa, G. A. O. P.. **Monitoramento Semiautomático do Desmatamento nos Biomas Brasileiros Amazônia e Cerrado: Estimativa de Incerteza e Caracterização de Áreas de Alta Incerteza**. Rio de Janeiro, 2023. 145p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

O monitoramento oficial do desmatamento na Amazônia brasileira tem dependido tradicionalmente de especialistas humanos que avaliam visualmente as imagens de sensoriamento remoto e rotulam cada pixel individual como desmatamento ou não desmatamento. Essa metodologia é obviamente cara e demorada devido à vasta área monitorada. A razão para não utilizar métodos totalmente automáticos para a tarefa é a necessidade da maior precisão possível nos números oficiais de desmatamento. Neste trabalho é proposta uma alternativa semi-automática baseada em aprendizagem profunda, na qual uma rede neural profunda é primeiro treinada com imagens existentes e referências de anos anteriores, e empregada para realizar detecção de desmatamento em imagens recentes. Após a inferência, a incerteza nos resultados em nível de pixel da rede é estimada e assume-se que os resultados da classificação com baixa incerteza podem ser confiáveis. As demais regiões de alta incerteza, que correspondem a uma pequena porcentagem da área de teste, são então submetidas à pós-classificação, por exemplo, um procedimento de auditoria realizado visualmente por um especialista humano. Desta forma, o esforço de etiquetagem manual é bastante reduzido.

Investigamos várias estratégias de estimativa de incerteza, incluindo abordagens baseadas em confiança, Monte Carlo Dropout (MCD), conjuntos profundos e aprendizagem evidencial, e avaliamos diferentes métricas de incerteza. Além disso, conduzimos uma análise abrangente para identificar as características das áreas florestais que contribuem para a elevada incerteza. Ilustramos as principais conclusões da análise em 25 polígonos selecionados em quatro locais-alvo, que exemplificam causas comuns de incerteza. Os sítios-alvo estão localizados em áreas de estudo desafiadoras nos biomas brasileiros da Amazônia e do Cerrado. Através da avaliação experimental nesses locais, demonstramos que a metodologia semi-automática proposta atinge valores impressionantes de pontuação F1 que excedem 97%, ao mesmo tempo que reduz a carga de trabalho de auditoria visual para apenas 3% da área alvo. O código desenvolvido para este estudo está disponível em `https://github.com/DiMorten/deforestation_uncertainty`.

Palavras-chave

Estimativa de Incerteza; Detecção de Desmatamento; Interpretação de Incerteza; Amazônia Brasileira; Cerrado Brasileiro; Imagens Ópticas; Aprendizado Profundo; Sensoriamento Remoto

Table of contents

List of figures

List of tables

List of Abreviations

PRODES	Amazon Deforestation Monitoring Project
INPE	National Space Research Institute
MCD	Monte Carlo Dropout
AL	Active Learning
Swin	Shifted Windows
SAR	Synthetic Aperture Radar
CNN	Convolutional Neural Network
ROI	Region Of Interest
NLP	Natural Language Processing
MNIST	Modified National Institute of Standards and Technology
CIFAR	Canadian Institute for Advanced Research
SVHN	Street View House Numbers
FCN	Fully Convolutional Network
OOD	Out-Of-Distribution
MHA	Multi-Head Attention
TTA	Test-Time Augmentation
MI	Mutual Information
KL	Kullback-Leibler
ReLU	Rectified Linear Unit
MLP	Multi-Layer Perceptron
DETR	End-to-End Object Detection with Transformers
MT	Mato Grosso
PA	Para
MS	Mato Grosso do Sul
PI	Piauí
AA	Alert Area
SITS	Satellite Image Time Series
BDC	Brazilian Data Cube

1 INTRODUCTION

1.1 Motivation

The Brazilian Amazon biome boasts a remarkable diversity of vegetation types, encompassing various forest typologies, including but not limited to traditional forest, dense forest, forests adorned with vines, bamboo-infused forests, campinarana, dry forest, Várzea forest, Igapó forest, Mangrove forest, and meadow [6]. Likewise, the Brazilian Cerrado biome is the second largest biogeographic region in South America and is considered the savanna formation with most biodiversity in the world [7]. This rich tapestry of ecosystems renders the task of mapping deforestation particularly intricate and demanding. The Amazon Deforestation Monitoring Project (PRODES), developed and operated by the Brazilian National Institute for Space Research (INPE), has provided consistent annual deforestation reports since 1988 [8]. Likewise, the entire Cerrado biome is monitored by PRODES [7]. In the current PRODES methodology, deforestation is visually inspected and manually annotated over satellite images covering the entire Brazilian Legal Amazon, which spans an area of 5 million km^2 (∼60% of the Brazilian territory). However, PRODES detects deforestation only in the areas originally covered by forest phytophysiognomies, corresponding to 4 million km^2. In the case of the Cerrado biome, PRODES monitoring manually inspects and annotates the entire biome area, corresponding to approximately 2 million km^2 (∼24% of the Brazilian territory) [9]. Although costly and time-consuming, the option for such a methodology derives from the high accuracies requirement – the 2022 PRODES report achieved an overall accuracy of 98.8% for the Brazilian Legal Amazon, corresponding to an $F1$-score of 87.1% for the deforestation class [8, 10, 11]. Likewise, high accuracy requirements occur for deforestation detection in the Brazilian Cerrado biome, where PRODES report achieved an accuracy of 93.4, corresponding to an $F1$-score of 87.4% for the deforestation class [11]. It must be noted that according to the PRODES methodology, after the interpreters have identified the deforestation polygons, a team of expert auditors checks the mapped polygons and looks for omissions in the visual classification.

Notwithstanding, many studies have focused on automatic deforestation detection from remote sensing imagery. Recent works used fully convolutional, deep neural networks for deforestation detection, having as inputs pairs of Sentinel-2 or Landsat-8 multispectral images from consecutive years (e.g., [12–14]). Although those works delivered considerably high accuracies, in the order of 79.6 to 81.7% in terms of F1-Score, they do not indicate the uncertainties in their predictions, which may be over or under confident.

Modern deep neural networks tend to produce over-confident outcomes at their output due to the networks' increasing complexity and parameter count. Such uncalibrated confidence is not an adequate measure of the network's reliability [15]. Uncertainty estimation enables the evaluation of the network's trustworthiness during inference.

Nevertheless, various uncertainty estimation techniques have been proposed in the context of deep learning-based methods [16]. Monte Carlo Dropout (MCD) [17] has been employed in many of such techniques. Dropout is commonly used during the training phase of deep neural networks to prevent overfitting and co-adaptation. It works by randomly turning off some of the network neurons during the training iterations. MCD prescribes using dropout at test time to produce slightly different outcomes for the same input. In that way, statistical uncertainty measures can be computed from a predefined number of inference runs. The method is particularly efficient because the deep neural network needs to be trained only once. Another alternative is to derive uncertainty values from the outcomes of an ensemble of networks trained with different random weight initializations and minibatch selections. In that case, however, the training procedure must be carried out multiple times. A more recent alternative is evidential deep learning [18]. Committee-based approaches such as MCD and ensembles are computationally expensive due to the need to train or run inference multiple times. Conversely, evidential learning estimates uncertainty using a single training run and a single forward pass. It considers the predictions of a single deterministic network as a subjective opinion and explicitly models the weight uncertainties.

This thesis addresses the challenge of substituting manual pixel-wise reports, such as the deforestation mapping procedure carried out by PRODES in the Brazilian Amazon and Cerrado biomes since 1988, with a semi-automatic procedure. The aim is to explore the feasibility of using semantic segmentation models for this task. However, it is acknowledged that the current models fall short of achieving the necessary accuracies for such a demanding application, where the implications of inaccurate outcomes could have significant operational and economic consequences. First, we propose a way to use uncertainty estimation to boost the accuracy of semantic segmentation tasks through a semi-automatic methodology. Although the methodology was assessed in the context of deforestation mapping from remote sensing imagery, it can be applied to any semantic segmentation task. In short, it is assumed that most classification errors concentrate on high-uncertainty predictions. So, an uncertainty map can be used to separate the pixel-level classification outcomes into low-uncertainty and high-uncertainty predictions. The low-uncertainty predictions are considered definitive, whereas expert human auditors should revise only the high-uncertainty predictions, which are expected to correctly classify the respective regions. With that approach, it is intended to expressively reduce the need for human intervention in the deforestation mapping processes while achieving a very high classification accuracy. Indeed, in our experimental results for deforestation detection, the auditing process was restricted to a mere 3% of the total area of the study areas, and, assuming that the human specialist is always right, F_1 scores considering the whole test areas were of up to 97.2%.

The above-mentioned results were obtained in a real-world operational setup, in which images and references from the past years were used for training the classifier, which was then applied to a pair of recent anniversary images. Results are presented for two sites in the Brazilian Amazon biome, in the states of Pará and Mato Grosso, and for two additional sites in the Cerrado biome, in the states of Mato Grosso do Sul and Piauí.

The second objective of this study is to investigate the factors contributing to the high uncertainty observed in some predictions. That involves examining the characteristics of forest fragments that exhibit such uncertainty levels. In other words, by exploring the uncertainty estimates, this study seeks to extract pertinent information that surpasses the conventional binary forest/non-forest classification, ultimately improving the understanding of different types of deforestation.

In conclusion, this work proposed a way to exploit uncertainty estimation in semantic segmentation tasks, and was specifically assessed for deforestation detection using satellite imagery. Multiple uncertainty estimation methods were assessed, including confidence-based approaches, MCD, ensembles, and evidential learning. For the dense classification model, a fully convolutional architecture model was used.

1.2
Objectives

1.2.1 General Objective

The general objective of this work is to propose a methodology that exploits uncertainty measures derived from deep learning-based semantic segmentation models, which can deliver very high accuracies in semantic segmentation tasks through a semi-automatic procedure that relies on little human intervention. Particularly, it is proposed to reduce human intervention by indicating the areas to be audited, reducing auditing efforts from 100% of the test area to significantly smaller percentages. Additionally, it is proposed to produce an analysis to determine the causes of uncertainty. The proposed methodology is assessed for deforestation detection from satellite images in the Brazilian Amazon and Cerrado biomes.

1.2.2 Specific Objectives

The specific objectives of this work are the following:

1. Propose an operational methodology that exploits uncertainty estimation in semantic segmentation tasks, aiming at strongly reducing human intervention in the execution of those tasks.
2. Evaluate uncertainty estimation methods such as confidence-based approaches, multi-output approaches, and evidential learning, and compare different uncertainty metrics such as predictive entropy, predictive variance, mutual information, and Kullback-Leibler divergence.
3. Evaluate the proposed methodology for deforestation detection from satellite images in the Brazilian Amazon and Cerrado biomes.
4. Provide a detailed interpretation analysis of the estimated uncertainty maps from the point of view of the user (i.e., the annotating and auditing experts).

1.3 Contributions

The main contributions of this work are the following:

1. A semi-automated methodology for semantic segmentation tasks, which significantly reduces the burden on human interpreters while maintaining high accuracy, comparable to traditional visual interpretation.

2. An evaluation of the proposed methodology in deforestation mapping, using four distinct study areas within the Brazilian Amazon rainforest and the Cerrado biome.

3. A comprehensive evaluation of different uncertainty estimation techniques specifically tailored to achieve the aforementioned objectives.

4. An empirical study that investigates the characteristics of forest areas exhibiting high degrees of uncertainty, according to the semi-automated deforestation detection method.

1.4 Organization of the remaining parts of this thesis

Chapter 2 describes the related work available in the literature for uncertainty estimation using confidence-based approaches, multi-output approaches and evidential learning, and the use of uncertainty maps for aiding the annotating procedure.

Chapter 3 provides the fundamental concepts and theory for a better understanding of the proposed method.

Chapter 4 introduces and explains the preliminary proposed method for aiding the annotating and auditing process using MCD and ensembles as uncertainty estimation techniques.

Chapter 5 presents the datasets employed in the preliminary study, the experimental protocol followed in the experiments, and the preliminary results obtained for the different uncertainty methods in deforestation detection.

Chapter 6 summarizes the conclusions derived from the performed experiments and provides directions for the development of the proposed method.

2
RELATED WORK

Recent works have used fully convolutional neural networks (CNNs) for deforestation mapping [12–14, 19]. In those works, the input to the CNNs was the concatenation of the optical images acquired at different dates, i.e., T_{-1} and T_0, in a so-called early fusion scheme. In [12], an encoder-decoder fully convolutional network model based on the ResUnet [20] delivered the best results when compared to different CNN architectures used for deforestation detection in the Brazilian Amazon. The preceding studies have shown promising accuracies ranging from 79.6% to 81.7% in terms of $F1$-Score. However, when considering the overarching goal of replacing manual reporting processes across the Brazilian Amazon and Cerrado biomes, these accuracies still fall short in comparison to the benchmark set by PRODES reports, which accuracies were estimated as 87.1% and 87.4% for the Amazon and Cerrado biomes, respectively [8, 10, 11]. Unlike prior research endeavors, this study proposes the utilization of uncertainty estimation techniques to streamline manual auditing efforts. The primary objective is to achieve accuracies matching or surpassing the PRODES standards, while concurrently reducing the auditing effort from 100% to significantly lower percentages.

Uncertainty in machine learning models can be divided into aleatoric or data uncertainty and epistemic or model uncertainty. Data uncertainty describes the confidence of the data, and it is related to the inherent randomness of the input data. Data uncertainty cannot be reduced by increasing the amount of training samples. Model uncertainty describes the confidence of the prediction, and it can be reduced by collecting more training data.

2.1
Confidence-based approaches

Various approaches have been proposed to estimate uncertainty in deep learning models. The first group is confidence-based approaches, which estimate uncertainty directly on the outcome of a single inference run. Multiple works used confidence-based approaches for uncertainty estimation [21–23]. In [21], entropy was used as an uncertainty measure for the prediction of a hydrologic variable. In [22], a comparison of confidence-based approaches was made for entropy, maximum margin, and least confidence in an active learning setting, where the three approaches performed similarly for image classification.

2.2 Multiple-outcome approaches

Alternatively, multiple-outcome uncertainty estimation methods rely on multiple training or inference runs to estimate the uncertainty related to a neural network's weights. These approaches measure the level of disagreement among multiple outcomes for the same input.

2.2.1 Ensembles

The first multiple-outcome approach trains an ensemble of networks and computes statistical measures that consider the different predictions of the individual networks that compose the ensemble [24–31]. In [24], the authors first used ensembles as an uncertainty method. Their method expressed higher uncertainty for samples outside the training distribution for multiple regression and classification datasets. They also found that random weight initialization and training dataset shuffling introduced sufficient variability for reliable uncertainty estimates. In [26], a comparison was made between multiple uncertainty estimation methods, including deep ensembles and MCD under dataset shift. They found that both accuracy and the quality of uncertainty degraded with the increase of dataset shift. They found that traditional calibration methods, such as temperature calibration, were significantly outperformed by methods that estimated epistemic uncertainty under increased shift. They also concluded that deep ensembles performed the best and most robustly under increasing dataset shift across multiple datasets such as CIFAR-10 and ImageNet. The superiority of ensembles to other techniques was further studied in [29], where it was found that ensembles with random initialization produced a better diversity-accuracy trade-off, resulting in more meaningful uncertainty compared to single training approaches for well-known datasets such as CIFAR-100 and ImageNet. Similarly, ensembles outperformed MCD in multiple classification and regression tasks in [27].

Multiple works have used deep ensembles for semantic segmentation [25, 27, 30–33]. In [32], a comparison between multiple uncertainty estimation methods was made for 3D semantic segmentation of point clouds. The authors concluded that deep ensembles outperformed the remaining methods in terms of classification accuracy and calibration, defined as a measure of how reliable the prediction probability was. In [30], a deep ensemble was used for uncertainty estimation on the Cityscapes autonomous driving dataset. They used a student-teacher distillation training approach. They trained a teacher model with higher complexity and parameter count, and then the student ensemble members were trained to replicate the teacher's outcomes while having lower complexity and parameter count. The resulting ensemble outperformed the teacher network in terms of accuracy while producing similar uncertainty results in terms of mIoU for the low-uncertainty samples under multiple uncertainty thresholds. Similarly, in [31], a deep ensemble with predictive variance was used for uncertainty estimation in autonomous vehicle scene understanding using a teacher-student approach. First, a teacher ensemble of networks was trained. Then, a single student network was trained using the dataset ground truth, the teacher's classification outcome, and the teacher's uncertainty estimates as inputs to the training loss. The student network had two output heads: A classification and an uncertainty estimation head, allowing to estimate uncertainty using a single forward pass. They found that their approach was a good predictor of incorrectly labeled pixels and that uncertainty robustly detected out-of-distribution samples. In [33], deep ensembles were used for uncertainty estimation in road segmentation from Synthetic Aperture Radar (SAR) data. Ensembles outperformed the compared methods in terms of accuracy and uncertainty usefulness. Although deep ensembles have consistently produced the best metrics across the multiple-outcome approaches, training multiple models is computationally expensive.

2.2.2 Monte Carlo Dropout (MCD)

A Bayesian network learns the posterior distribution for a network's trainable weights, allowing it to compute the principled predictive uncertainty. However, Bayesian techniques have been proven unpractical for deep neural networks due to the large amount of data needed, proportional to the number of network parameters [16].

The most common approximation for Bayesian networks is Monte Carlo Dropout (MCD). It has been demonstrated that MCD can be viewed as having a mathematical equivalence to Bayesian networks [1, 17]. Dropout is commonly used during training as a regularization technique. MCD additionally uses dropout at inference, producing a different outcome in each inference run. Uncertainty is then estimated by calculating statistical measures like variance and entropy over a predefined number of inference runs.

Multiple works have used MCD in classification problems [24, 34–39]. In [34], MCD was used for uncertainty estimation in the classification of electroencephalogram (EEG) signals. Uncertainty histograms showed a correlation between high uncertainty and incorrect samples for the Brain-Computer Interface (BCI) Competition IV dataset. They proposed to reject high uncertainty predictions, obtaining lower error rates for classification. In [24], a comparison was made for MCD and deep ensembles in the image classification datasets SVHN, MNIST, and ImageNet. In all cases, deep ensembles outperformed MCD regarding classification accuracy, network calibration, and out-of-distribution detection. In [36], MCD was used for medical image classification. They proposed an uncertainty metric based on the overlap between the top-2 predicted class distributions, which outperformed the predictive variance metric. In [35], MCD was used to select high-uncertainty samples for further inspection in medical image classification.

Recently, various works have used MCD for uncertainty estimation in semantic segmentation applications that rely on deep learning models. Many of those works used U-Net [40] based networks, and employed MCD at inference time to approximate a Bayesian network. Such an approach has been employed in many application areas such as urban mapping with aerial and satellite images [33, 41], medical image segmentation [42, 43], and fingerprint ROI segmentation [44]. MCD was also used in [45] to estimate uncertainty in the semantic segmentation of video frames. In all the above-mentioned works, however, the estimated uncertainty maps were only used in the analysis of the corresponding models' predictions, i.e., they were not employed in further processing steps.

In the medical image segmentation field, multiple works have used MCD [25, 46–58]. In [48], MCD with predictive variance as uncertainty metric was used for tumor volume estimation. They compared using dropout at different stages of the network and found that using dropout in every layer produced the most stable variance estimates. In [51], MCD was used to obtain uncertainty for a safety-critical image segmentation task. They managed to detect segmentation errors that demanded expert review automatically.

In [52], uncertainty was estimated using MCD for Optical Coherence Tomography (OCT). They found a correspondence between high-uncertainty regions and inaccurate segmentations. In [53], a relationship was found between high uncertainty estimates from MCD and regions of high inter-observer variability, which measures the disagreement occurring among multiple expert annotators. Inter-observer variability is inherent in medical image segmentation applications. The results were assessed for brain tumor cavity segmentation. Similarly, in [50], MCD was used to obtain uncertainty estimates for multiple sclerosis lesion segmentation. They found high uncertainty values in small lesions and lesion boundaries, which corresponded with typical human-annotator variability sources.

The authors in [54] used uncertainty from MCD as an input to a post-processing refinement strategy and found that it outperformed Conditional Random Fields (CRF). The refinement strategy trains an additional Graph-Convolutional Network (GCN) on the high-confidence voxels to reclassify and refine the outcome. In [55], uncertainty from MCD was used for anatomic anomaly detection in retinal tomography segmentation. They found that epistemic uncertainty was higher for regions whose appearance differed significantly from training data, achieving high accuracy in the anomaly detection task, with anomalies corresponding to diseased samples. Furthermore, they found high uncertainty values in other deviations in normal scans, such as cut edge artifacts. In [25], deep ensembles and MCD were compared for brain, heart, and prostate magnetic resonance imaging (MRI) image segmentation. They used average entropy as an uncertainty metric. They found a correlation between segmentation quality and uncertainty values, and they also found that the average entropy could be used for effectively detecting out-of-distribution samples. For comparison, they employed scoring rules, which assess the quality of uncertainty estimation by rewarding properly calibrated probabilistic forecasts. They found that ensembles outperformed MCD.

In [58], uncertainty was estimated using MCD for MRI image segmentation. Consistently with other works, they found a correlation between high uncertainty and erroneous areas. In [57], epistemic uncertainty was estimated by using MCD at inference, and stochastic Gaussian noise was added to the input image for additional variability. They used uncertainty to guide the selection of pseudo-labels in an unsupervised setting, disregarding the pseudo-labels with high uncertainty. Additionally, they used uncertainty in the unsupervised training loss to select the most certain samples for optimization. Similarly, [56] incorporated uncertainty using MCD to a semi-supervised training loss by using a student-teacher model and allowing the student model to learn from the more reliable targets guided by the teacher model's uncertainty.

In [59], the resulting uncertainty map estimated using MCD was used as an additional input to train a second U-Net network. The map was concatenated with the original input image for the semantic segmentation of scanned historical maps.

2.2.3

Test-Time Augmentations (TTA)

Alternatively, Test-Time Augmentations (TTA) trains a single deterministic network and uses random data augmentation operations at test time, allowing to obtain a different outcome in each inference run. As in MCD and ensembles, uncertainty is obtained by calculating a statistic over the inference repetitions [60]. TTA differs from the previous multiple-outcome approaches in that it exclusively calculates the data uncertainty. In [37], TTA was used in combination with MCD for skin lesion classification. They found that combining both approaches improved the classification metrics compared to only using MCD.

2.3 Single-Outcome Deterministic Approaches

Multiple-outcome methods are computationally expensive because they require training or inferring multiple times. Single-run deterministic methods have been proposed as an alternative to multi-output approaches. In these approaches, uncertainty is determined using a single forward pass. Such methods are divided into two groups: In the first group, a single network is explicitly designed and trained with the aim of quantifying uncertainty [18, 61–68]. In the second group, additional components are used to compute uncertainty, generally as a post-processing step for already trained networks, having no effect on the network predictions [69–72]. Both groups are called internal and external, respectively.

2.3.1 Internal Deterministic Approaches

Many internal methods followed the idea of predicting a distribution over the possible outcomes instead of a point-wise estimation. Evidential deep learning [18] is an alternative that estimates model uncertainty with a single training and inference run. It is based on the Dempster-Shafer theory of evidence [73]. It assumes that the outcome of a single deterministic network is a subjective opinion and learns the function leading to those opinions as a Dirichlet distribution by directly estimating the Dirichlet parameter α at the output of the network. Their work performed similarly to multiple-outcome approaches while being less computationally expensive. In [74], evidential learning was used for action recognition in an open set problem, where new classes not observed during training come about at inference, whose uncertainty is expected to be high. Their results outperformed the remaining approaches. In [75], evidential learning was used for chest radiograph image classification. Similar to our work, they assessed classification metrics for varying levels of high-uncertainty coverage. In [76], evidential learning was extended by separating the output evidence into vacuity and dissonance to better separate between in-distribution and Out-Of-Distribution (OOD) samples. Two datasets containing overlapping classes and OOD samples were integrated into the framework to achieve this. In [66], a modification of evidential learning was proposed to maximize the representation gap in OOD samples by additionally using OOD samples during training. More recently, [77] demonstrated the limitations of evidential learning for unbalanced datasets. They suggested the implementation of a data augmentation method during the training phase, along with a post-hoc calibration process on a validation dataset, in order to mitigate bias stemming from the imbalanced data distribution.

Similarly, in [61], a contemporary variation of evidential learning called prior networks was used to estimate uncertainty modeling the point-wise predictions' distribution as a Dirichlet distribution. Different from evidential learning, they used OOD samples during training. They trained the network to produce a sharp Dirichlet distribution focused on the correct class for in-distribution samples and a flat Dirichlet distribution for OOD samples by minimizing the expected KL-divergence. They explicitly calculated the distributional uncertainty, representing the difference between the training and testing distributions. In contrast, in the Bayesian networks, distributional uncertainty is considered a part of model uncertainty. As an extension, in [62] they argued that when data uncertainty is high, the posterior behaves as an undesired multi-modal distribution. As a solution, they formulated the loss as a reverse KL-divergence. In [63], they assumed that the point-wise estimates were sampled from an unknown distribution, and they argued that using a mixture of Dirichlet distributions offered greater flexibility for approximating this unknown distribution. In [64], they eliminated the need for OOD samples during training by learning the classes' distribution over a latent space. During inference, the sample was mapped into this latent space, and class-specific densities in the latent space were utilized to parameterize a Dirichlet distribution.

Besides Dirichlet distribution-based approaches, other internal deterministic methods have been proposed. In [65], the authors proposed a modified softmax called *inhibited softmax*, which introduced an inhibition parameter to calibrate the softmax outcomes, compensating the deep networks' typical over-confidence and producing more accurate uncertainty values. Their results were similar to MCD, requiring a single training and inference run. In [78], Radial Basis Functions (RBF) achieved competitive accuracy and good uncertainty results. RBFs learn a linear transformation on the logits and classify a sample based on the distance between the transformed logits and the centroids associated with learned classes.

Recent works have used evidential learning for uncertainty estimation in semantic segmentation [79–85]. In [79], an FCN is used for evidential learning by replacing the softmax layer with ReLU and attaching a Dirichlet layer at the outcome, which produced the Dirichlet parameters α at the output. They estimated uncertainty for the semantic segmentation of underwater imagery. Their approach was the same as in this work, although the application field differed. Although they qualitatively observed a correspondence between low accuracy and high uncertainty, they did not quantify the benefits of using uncertainty. In [82], evidential learning was used for uncertainty quantification in medical image segmentation. They argued that despite recent advances in semantic segmentation, clinicians remain skeptical of its uses due to its black-box nature. Estimating uncertainty enhanced the reliability and explainability of their system. They found a correspondence between out-of-distribution samples and high uncertainty values. They also applied their approach to real-world clinical safety applications.

In [83], evidential learning was applied for brain tumor 3D segmentation, and compared with ensemble and MCD. They concluded that ensembles and MCD produced higher accuracies and better calibration outcomes in raw images. Even so, evidential learning resulted in a higher overlap between uncertainty and error areas. Similarly, in [84], a comparison of evidential learning was made with other uncertainty estimation methods, including MCD, ensembles, and probabilistic U-Net, which learns a conditional probability of the train data with a variational autoencoder and obtains uncertainty values as distances to the train distribution. The methods were assessed for 3D brain tumor segmentation. They observed the performance of all uncertainty methods to decay in the presence of increasing levels of Gaussian noise. As expected, evidential learning produced the lowest inference times due to the need for multiple sampling at inference by the MCD, ensembles, and probabilistic methods. They also found their framework to be competitive with the compared approaches. In [85], evidential deep learning was compared with other uncertainty estimation methods such as MCD and ensembles for CityScapes and KITTI datasets. They compared in terms of classification accuracies, qualitative uncertainty maps, and out-of-distribution detection. They separated uncertainty into aleatoric and epistemic components. They concluded that evidential learning produced inferior results compared to MCD in terms of epistemic uncertainty estimation, and they argued that it was difficult to approach a rich Dirichlet distribution only by encouraging the network to produce a uniform distribution for low-accuracy samples. Their best accuracy outcomes were obtained with deep ensembles in terms of mIoU. They also found the deep ensembles to be the best choice for epistemic uncertainty estimation. In [80], evidential learning was used for class incremental learning, where a previously trained network is extended with new classes. The authors modeled the occurrence of an unknown class (background) as the estimated uncertainty. The proposed approach outperformed other state-of-the-art methods for incremental learning. In [81], instead of estimating evidence from a learned distribution (model-based), they used a distance-based approach, where the mass functions were obtained based on distances to prototypes. In sum, the network summarizes the training set by a small number of prototypes and calculates evidence as the proximity from an input vector to such prototypes.

2.3.2 External Deterministic Approaches

Among the external deterministic approaches, which don't modify the network training procedure, [69] argued that estimating the prediction and uncertainty from the same network results in biased outcomes. Thus, they used two networks: One to produce the predictions, and another to produce the uncertainty values using the first network's outcomes as input. In this instance, uncertainty was not assessed concerning the classification outcome but rather focused on discerning variations in opinions among multiple annotators. In other words, they estimated the annotators' uncertainty with respect to the ground truth reference. In the second network, they proposed to estimate uncertainty as a supervised task in a problem with noisy labels. They used the level of disagreement among multiple annotators as a ground truth reference for the supervised training of the uncertainty values. Similarly, [70] used two networks to separately obtain predictions and uncertainty. The uncertainty network was trained to detect the prediction network's errors and takes as input the prediction statistics of neighboring points in the representation space. In [68], the softmax outcome was decomposed as the quotient of the class probability and the domain probability, allowing them to address the network's over-confidence problem for OOD detection.

2.4

Active Learning

Active learning aims to select the most relevant samples from annotated data to be annotated by a human oracle. The AL solution consists of a loop of actions executed iteratively until some quality criterion is achieved. At each AL cycle, a small subset of the non-annotated image pool is selected based on the outcome of a given segmentation module. The selected images are then added to the training set after being annotated by an analyst. Finally, the segmentation module is retrained upon the new training set. This closes one AL loop, which can run iteratively until a stopping criterion is reached. Incrementing the training set makes the segmentation module more accurate with each AL cycle run, to the point where the performance gains become so small that the effort involved in more AL cycles is no longer worth it.

In a way similar to our work, multiple methods have used confidence-based approaches, MCD, ensembles and evidential learning for uncertainty estimation in an active learning scheme, although they were devised for different application fields [86–99]. Besides, the majority of their approaches corresponded to image-level classification. Instead, our work addresses semantic segmentation (i.e., pixel-level classification). In [86], maximum margin was proposed as an uncertainty measure in AL for multi-class image classification. Their approach outperformed random sampling. Similarly, in [87], the most informative samples were queried using confidence-based uncertainty methods, including least confidence, maximum margin, and entropy from a single inference run for image classification using CNNs. In [88], the most informative samples were selected in each AL loop using uncertainty from ensembles and Mutual Information (MI) as uncertainty metric.

In the context of active learning, a two-step procedure has been used. First, the most informative samples (i.e., those with the highest uncertainty) are usually preselected. However, the group of samples with the highest uncertainty is likely to contain too similar samples. Thus, a second step selects the most representative (or diverse) samples within the preselected group, which are sent to the oracle for annotation in each AL cycle.

In [89], a confidence-based approach (maximum margin) was used to select the most relevant samples in Natural Language Processing (NLP) and image classification tasks. The authors did not assess other uncertainty estimation methods. In that work, first, the samples with the highest uncertainty were preselected. Then, the preselected set was applied to a clustering K-Means algorithm with the number of clusters equal to the desired amount of selected samples. The final selection corresponded to the samples closest to each cluster center. In [92], the most informative samples were selected using uncertainty from MCD. Diversity across the selected samples was achieved by considering the correlations between data points in each acquisition batch. In [90], the authors inferred on test data, estimated the uncertainty maps using MCD, and sent samples with the highest uncertainty to an expert who manually annotated them. Then, the network was retrained, and the cycle was repeated for multiple iterations. They also proposed a method to reduce uncertainty in the predicted polygons' borders. That work was employed in the semantic segmentation of histology data.

Similarly, [91] used MCD to compute uncertainty metrics for active learning in the context of image classification; the proposed method was evaluated on the MNIST dataset and biomedical data. In [93], MCD was used to compute a guiding metric for active learning in object detection tasks by annotating and fine-tuning the network with the highest-ranked test samples according to their uncertainty scores. They evaluated the method using pedestrian detection datasets. In [100], deep ensembles and MCD were compared for uncertainty estimation in an AL setting. Deep ensembles consistently outperformed MCD in MNIST, CIFAR-10, and a medical image classification application. They argued that the performance of MCD was lower due to reduced outcome variability and model complexity.

In [95], a two-step procedure was used in an AL scheme for biomedical image segmentation. First, ensembles were used for preselecting the most informative images, measured by the highest uncertainty. Then, a subgroup of the preselected samples was obtained, corresponding to the group with the highest similarity to the preselected group. Different from our work, they selected a specific number of images for an auditor to annotate. Instead, we select image regions within a remote sensing raster for an auditor to re-annotate. In [96], uncertainty was measured through gradient embeddings, and diversity was calculated through the k-MEANS++ sampling technique, which computes the distance from the unlabeled samples to the already annotated ones and gives a higher selection probability to samples with maximum distance to the annotated set. In [97], evidential deep learning was used for active learning. Their approach outperformed multiple-outcome approaches like MCD and ensembles. However, they only did experiments with image classification. Instead, our work estimates uncertainty for semantic segmentation.

In [23], a comparison of multiple uncertainty estimation techniques, including confidence-based methods, MCD, and ensembles, was made for active learning in image classification tasks. They found that none of the methods performed significantly better than the others.

Table 1 presents a summary of the related works grouped by classification task (i.e., sequence classification, image classification and semantic segmentation), uncertainty method, and publication year. Diverging from prior efforts, this thesis specifically tackles the challenge of semantic segmentation, tailor-made for scenarios demanding the replacement of entirely manually generated reports, such as deforestation mapping in the Brazilian Amazon and Cerrado biomes. The objective is to employ a semantic segmentation network, albeit one that has yet to attain the precision levels achieved through meticulous manual inspection. In this context, our approach significantly diminishes the auditing effort from a full 100% to substantially lower percentage values.

Table 1: Summary of related works grouped by classification task, uncertainty method and publication year.

Task	**Method**	**Year**
Sequence classification	Confidence	1973 [21]
	MCD	2021 [34]
Image classification	Confidence	2014 [22], 2022 [23]
	Ensembles	2017 [24], 2019 [26, 29], 2020 [27]
	MCD	2016 [17], 2017 [24, 35], 2019 [36], 2020 [37, 38], 2023 [39]
	TTA	2020 [37]
	Deterministic	2018 [18, 61, 65], 2019 [62, 63, 69], 2020 [64, 66, 68, 70, 76, 78], 2021 [74, 75], 2022 [77]
	AL	2009 [86], 2011 [88], 2016 [87], 2017 [91], 2019 [89, 90, 92, 93, 96], 2021 [94], 2022 [97]
Semantic segmentation	Ensembles	2020 [25, 27, 32], 2021 [31, 33], 2023 [30]
	MCD	2017 [1, 46], 2018 [45, 47–49, 51–53], 2019 [55, 56], 2020 [25, 43, 50, 54, 57], 2021 [33, 41, 42, 44], 2022 [58]
	TTA	2019 [60]
	Deterministic	2020 [85], 2021 [81], 2022 [79, 83, 84], 2023 [80, 82]
	AL	2017 [95], 2020 [99], 2023 [98]

3
FUNDAMENTALS

This chapter presents the theoretical foundations on which the proposals presented in subsequent chapters are based. It is organized into two sections. The first section presents the concept of uncertainty. The second section presents the literature's most widely used uncertainty assessment approaches.

3.1
Uncertainty

Although deep neural networks have produced high accuracies in a vast range of applications, as a consequence of their increasing complexity and parameter count, their predictions tend to suffer from over or under-confidence, meaning they are badly calibrated. As a result, neural networks do not offer a method to determine whether a test outcome is reliable or not [15]. Uncertainty estimation allows us to assess the network's reliability at inference.

There are multiple sources of uncertainty in a classification task. Depending on the input data domain, uncertainty can be divided into three groups: *In-domain*, *Domain-shift* and *Out-of-domain* uncertainty [2]. *In-domain* refers to uncertainty for an input drawn from within the training distribution. Such *In-domain* uncertainty may also be divided into two sub-groups: Model (Also known as epistemic) and data (Also known as aleatory). Model uncertainty refers to shortcomings during the modeling process, such as errors during training (e.g., an inadequate selection of batch size, optimizer, learning rate, dropout rate, stopping criteria, regularization), a model with an inadequate complexity (over- or under-fitting), or a lack of knowledge due to an insufficient amount or relevance of the training samples, resulting in a bad coverage of the training distribution. Model uncertainty may be improved by increasing the training set as a decreasing number of possible models become a plausible fit. Model uncertainty may also be reduced by modifying the model architecture and training strategy.

On the other hand, data uncertainty represents the inherent fluctuation present in the input data. This variability is inherent to the data and cannot be reduced by increasing the amount of training samples. Data uncertainty is higher when the input data is noisy. In the context of remote sensing, data uncertainty may occur due to low spatial resolution, which may make it challenging to recognize fine details and smaller objects, due to ambiguity at the pixels corresponding to object boundaries, due to labeling noise, due to data obstructions such as cloud-covered portions in optical images, among others.

Figure 1 presents sample aleatoric and epistemic uncertainty for a sample image in semantic segmentation for autonomous driving [1]. Aleatoric uncertainty is higher in objects far from the camera and in boundaries between class predictions. Polygon C represents a region far from the camera, presenting a high aleatoric uncertainty. Epistemic uncertainty is higher in semantically and visually challenging regions, as illustrated in polygons A and B. Polygon B is difficult to classify due to the color similarity between the footpath and street classes. In polygon A, the network produced an error, which resulted in high model uncertainty. Such error may be reduced by increasing training data for its input pattern.

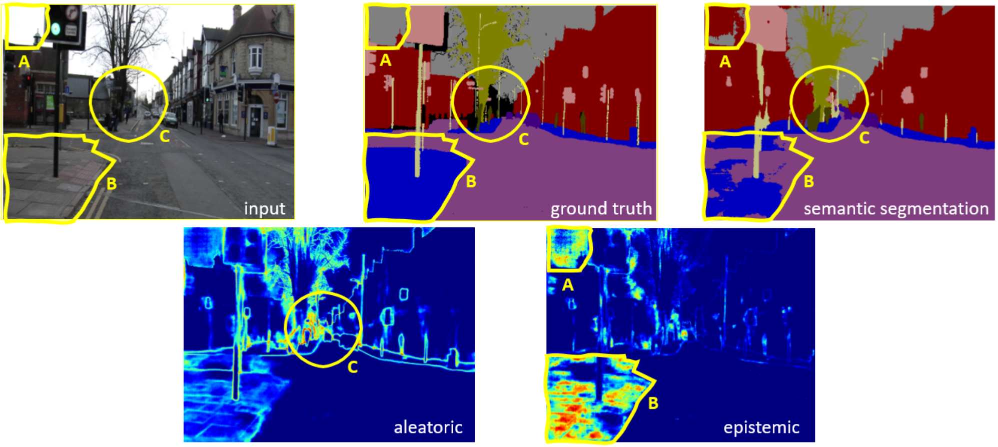


Figure 1: Visualization of data and model uncertainty for a sample image in autonomous driving. A, B and C are polygons of interest. A and B: High model uncertainty. C: High data uncertainty. Taken from [1]

Figures 2 and 3 illustrate data and model sources of uncertainty for a binary classification task. Data uncertainty occurs when samples from different classes overlap in the representation space, making the overlapping area difficult to classify correctly. Model uncertainty may be measured by training multiple models (Two models in the example) and measuring their level of disagreement. If both models agree on the classification outcome, model uncertainty is low. If each model produces a different outcome, model uncertainty is high. It is important to note that in the assessed application of deforestation detection, the data exhibits a high degree of imbalance, introducing an additional challenge when identifying different sources of uncertainty.

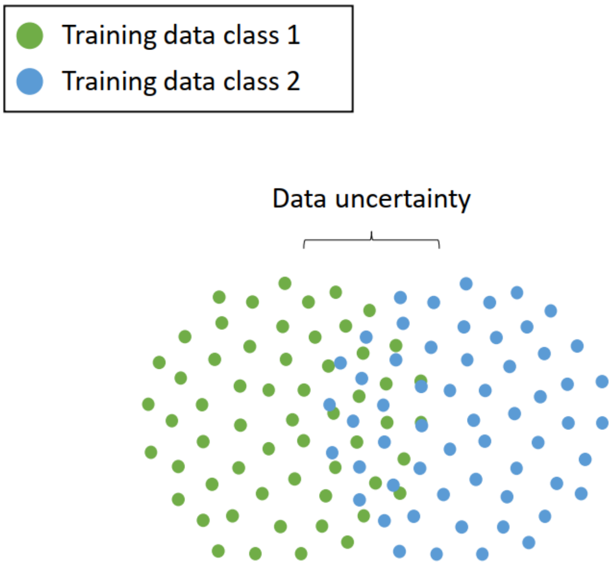


Figure 2: Visualization of data uncertainty. Samples from contiguous classes overlap in the intermediate region, resulting in higher data uncertainty. Taken from [2]

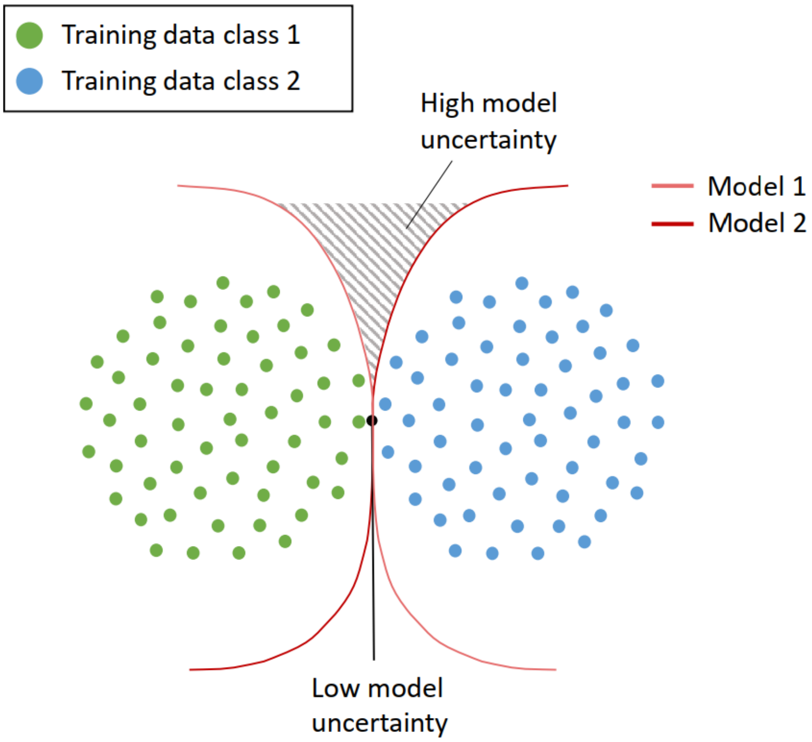


Figure 3: Visualization of model uncertainty. Areas where multiple models disagree imply higher model uncertainty. Taken from [2]

Domain-shift uncertainty occurs when the input is drawn from a shifted distribution compared to the training distribution. This shift occurs due to the inherent variability of the real-world scenarios. In remote sensing applications, domain shift in the temporal dimension occurs when training in an image from a particular date and testing in images from future dates, with a larger domain shift being expected for larger time differences. Domain shift also occurs in the spatial dimension when training in a specific geographical region and testing in places different from the training area, with distribution shifts occurring due to multiple factors, including changes in ecological conditions such as types of vegetation and biome types.

Out-of-domain uncertainty represents the uncertainty from an input drawn from an unknown distribution, where the input distribution is different and far from the training distribution. Different from domain-shifted data, where useful information can still be learned in the presence of a distribution shift, in-domain knowledge cannot be extracted from out-of-domain samples. For example, in a network trained to classify between cat and dog images, a domain-shifted sample may be a blurred image of a dog. In contrast, an out-of-domain sample may be a bird, which the network is unable to explain. Figure 4 presents sample out-of-domain images for an autonomous driving application (e.g., the network was not trained on anomalous samples such as unwanted animals or airplanes on the road). Figure 5 presents out-of-distribution uncertainty for the binary classification problem in the feature representation space. Out-of-distribution data is outside the original training distribution without the occurrence of any overlapping.

3.2 Uncertainty Assessment Approaches

We group uncertainty estimation into

- confidence-based,
- multiple-outcome-based, and
- evidential deep learning (EDL),

methods as detailed in the following.

3.2.1 Confidence-Based Approaches

The first group of methods depends on the model's confidence in its prediction. Below, we briefly present the most widely used metrics computed from the discrete probability distribution at a model's output.

In the following, we denote with $\mathbf{y} = [y_1, ..., y_K]$ the K-dimensional vector representing the discrete probability distribution assigned by the model for a

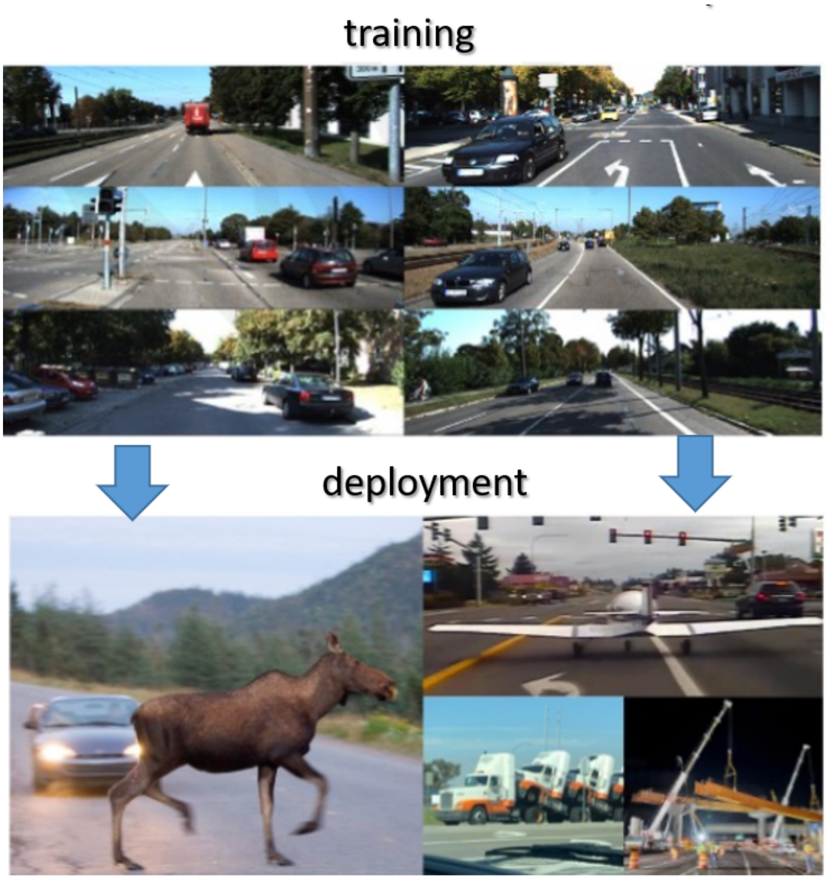


Figure 4: Sample out-of-distribution occurrences in an autonomous driving application, corresponding to unseen examples during training such as unexpected animals and airplanes on the road [3].

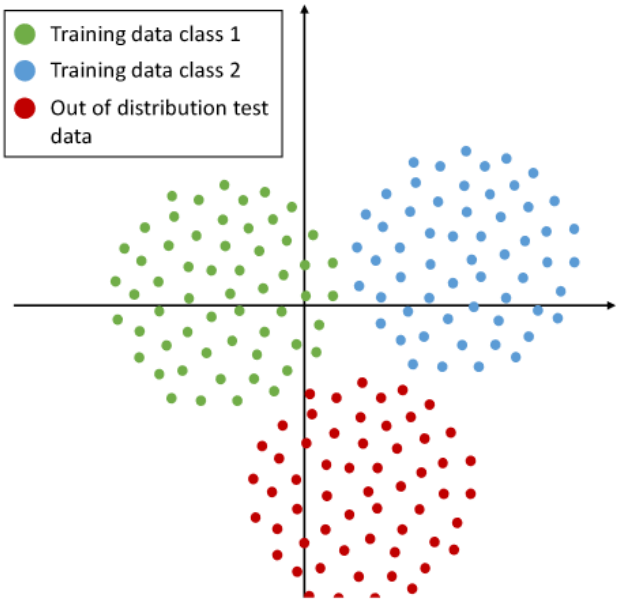


Figure 5: Visualization of out-of-distribution uncertainty. Taken from [2]

given input. It is assumed that $\mathbf{y}$ is a simplex, i.e., $y_k \geq 0$ for all $k \in \{1, ..., K\}$ and that $\{y_k\}_{k=1}^{K}$ add up to one.

Entropy:

The most widely used metric for a classifier's confidence is entropy defined as

$$h(\mathbf{y}) = -\sum_k y_k * log(y_k). \tag{3-1}$$

The highest model confidence corresponds to the lowest entropy value ($h(\mathbf{y}) = 0$) when the model assigns probability equal to one to one particular class and probability zero to all other classes. Conversely, the highest entropy value is reached when the model gives equal probabilities to all classes, corresponding to the lowest confidence.

Figure 6 illustrates the confidence estimation for two sample softmax outcomes. Confidence is lower in the rightmost entropy, where the class probabilities are closer to being equal.

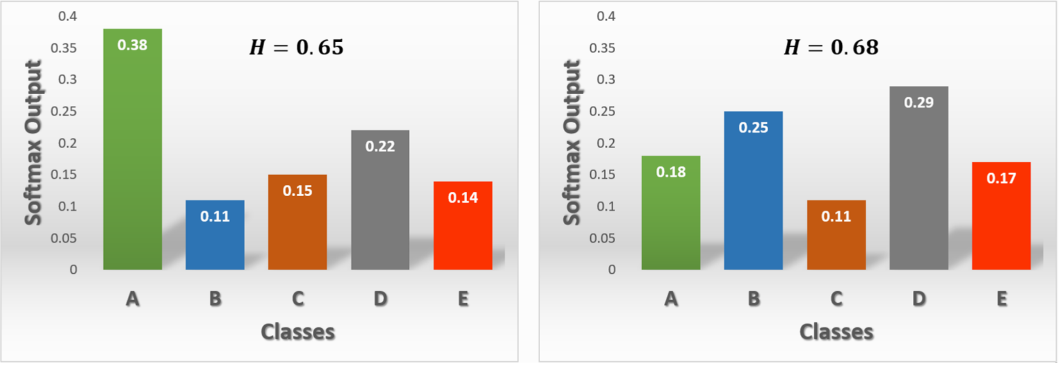


Figure 6: Visualization of two sample softmax outputs and their confidence values obtained as entropy [3].

Maximum Margin:

Margin-based confidence [101] corresponds to the difference between the highest and the second most likely class, formally:

$$MaxMargin(\mathbf{y}) = y_a - y_b. \tag{3-2}$$

where y_a and Y_b are the scores assigned to the first and the second most likely class labels, respectively.

The margin-based confidence method is illustrated in Figure 7 for two classification outcomes. The rightmost one presents a lower margin compared to the leftmost example, indicating lower confidence.

Most Likely Label:

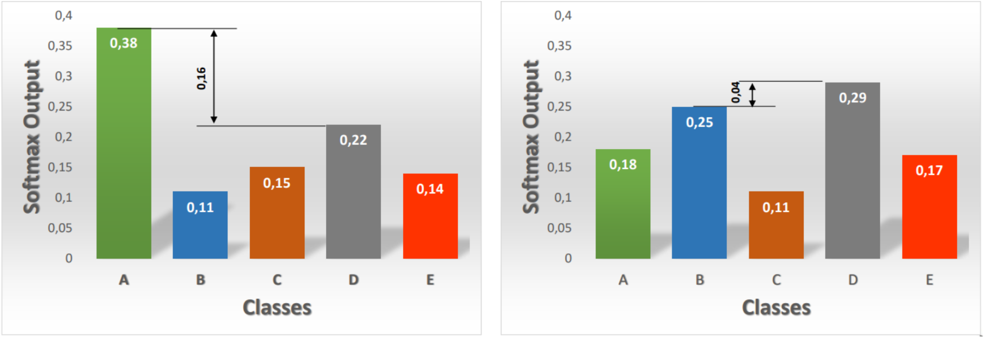


Figure 7: Visualization of the margin-based method. The lowest margin is in the rightmost example, indicating lower confidence [3].

In this approach, one takes as confidence the raw predicted probability value of the most likely class label, formally

$$MaxL(\mathbf{y}) = max(\{y_k\}_{k=1}^{K}). \tag{3-3}$$

Notice that $MaxL(\mathbf{y})$ ranges from $1/K$ to 1.

Figure 8 illustrates the most likely label method. The rightmost example presents the lowest confidence.

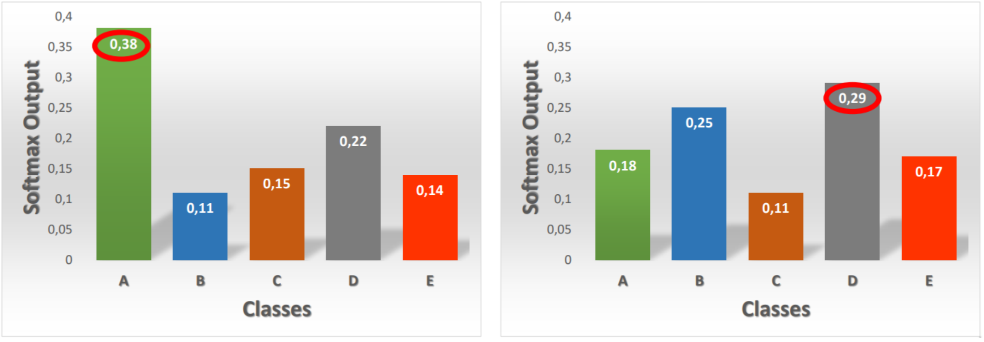


Figure 8: Visualization of the most likely label method. The highest uncertainty is in the rightmost example [3].

Table 2: Comparison of confidence-based approaches. Correlation refers to whether the method has direct or inverse correlation with the uncertainty measure.

Approach	**Equation**	**Ranges**	**Correlation**
Entropy	$-\sum_k y_k * log(y_k)$	-	Direct
Maximum Margin	$y_a - y_b$	$1/K$ to 1	Inverse
Most Likely Label	$max(\{y_k\}_{k=1}^{K})$	0 to 1	Inverse

3.2.2 Multiple-Outcome Approaches

The methods introduced in the previous subsection are simple to implement and require a single training and inference run. Differently, the consistency-based uncertainty estimates rely on outcomes of multiple inference runs corresponding by either considering multiple random initialization values during training (e.g., ensembles), creating randomness at inference (e.g., Monte Carlo Dropout), or using the same network on augmented versions of the same input data (Test-Time Augmentation). In the following, we introduce the consistency-based methods.

To accommodate multiple outcomes for the same input, we denote hereafter with $\mathbf{y}^{(i)} = [y_1^{(i)}, ..., y_K^{(i)}]$ the K-dimensional vector representing the discrete probability distribution assigned by the i-th model for a given input, where $i \in \{1, ..., n\}$, and n is the number of outcomes.

The so-called final prediction is given by the average probability vector computed over the n outcomes, formally:

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}^{(i)}. \tag{3-4}$$

3.2.2.1 Monte Carlo Dropout (MCD)

Transforming a neural network into a Bayesian model involves replacing fixed weight values with random samples drawn from a prior distribution denoted as $p(w)$, with w being the network weights. As data is observed, it provides fresh perspectives on the weights, leading to the creation of a posterior distribution $p(w|\mathcal{D})$, with $\mathcal{D}$ being the training dataset. With this posterior distribution, Bayesian inference computes a predictive distribution for the outputs $\mathbf{y}$ of an unseen data point x^* by integrating over all conceivable weight values.

$$p(\mathbf{y}|x^*, \mathcal{D}) = \int \underbrace{p(\mathbf{y}|x^*, w)}_{\text{Data}} \underbrace{p(w|\mathcal{D})}_{\text{Model}} \, dw \tag{3-5}$$

Equation 3-5 describes the network uncertainty [16]. The expression $p(w|D)$ is termed the posterior distribution for the model parameters, representing the uncertainty associated with these parameters when considering a training dataset $\mathcal{D}$.

In accordance with Bayes' theorem,

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$$

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{\int p(\mathcal{D}|w)p(w)dw}$$

Unfortunately, calculating this value analytically is not feasible because of the intractable integral in the denominator. To tackle this problem, MCD offers a solution by approximating the true posterior $p(w|\mathcal{D})$, with $q(w)$, where the act of sampling weights $w \sim q(w)$ is achieved by implementing dropout within the network. As a consequence, the resultant predictive distribution is as follows:

$$q(\mathbf{y}|x^*) = \int p(\mathbf{y}|x^*, w)q(w)dw \tag{3-6}$$

With the predictive distribution described in Equation 3-6, we can derive an impartial estimator by employing Monte Carlo sampling:

$$\hat{q}(\mathbf{y}|x^*) = \frac{1}{n}\sum_{i=1}^{n} p(\mathbf{y}|x^*, \hat{w}_i), \hat{w}_i \sim q(w)$$

The network parameters, denoted as $\hat{w}_i$, are drawn from the dropout distribution, for a predetermined number of iterations n [17, 36].

Dropout is a technique commonly used in the training phase to reduce overfitting. All the forward and backward connections with a dropped node are temporarily removed, thus creating a new network architecture out of the parent network. The set of dropped-out nodes is randomly selected at each forward/backward pass. Usually, at inference time, all nodes remain active. Figure 9 illustrates dropout for a fully connected neural network. Each dropout calculation randomly drops a predefined percentage of units from the original network, resulting in a different outcome.

MCD differs from the conventional Dropout because the strategy is applied also in the inference step. In MCD, the inference is carried out n times, each time with a different set of dropped-out neurons, resulting in a different outcome at each run. It can be proved that this procedure is an approximation of Bayesian inference [17]. It corresponds to sampling from the posterior distribution in each inference run.

The uncertainty value is obtained by calculating uncertainty metrics over the n predictions. This calculation is carried out in semantic segmentation for each image pixel, producing an uncertainty map. In Subsection 3.2.2.4, we describe different ways to compute uncertainty metrics from these n results.

Conventional dropout's effectiveness diminishes when applied to images due to the strong correlation among adjacent pixels. Even if pixels are

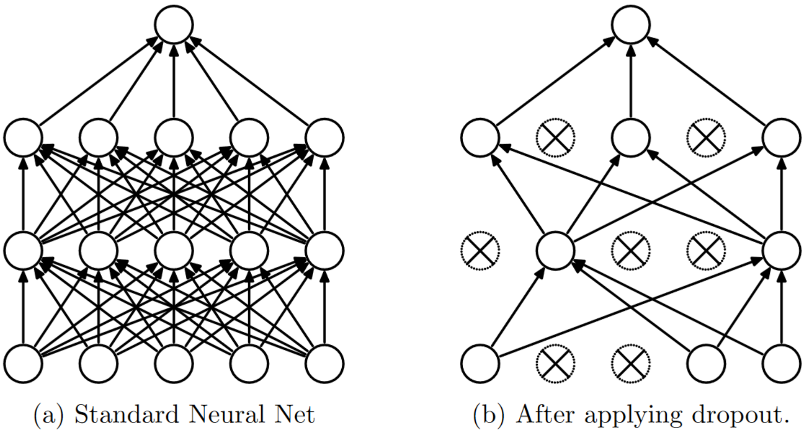


Figure 9: Dropout in a neural network. For simplicity, the figure represents an uni-dimensional feature space. The left figure represents a standard network with two hidden layers. The right figure represents a thinned net by applying dropout to the leftmost network. Removed units are represented as crossed. Taken from [4]

randomly dropped, their likely values can still be reasonably inferred by observing the surrounding pixels. Hence, opting to drop entire feature maps could be more congruent with the initial purpose of the dropout technique. Such a technique is called spatial dropout [102]. Figure 10 illustrates this concept.

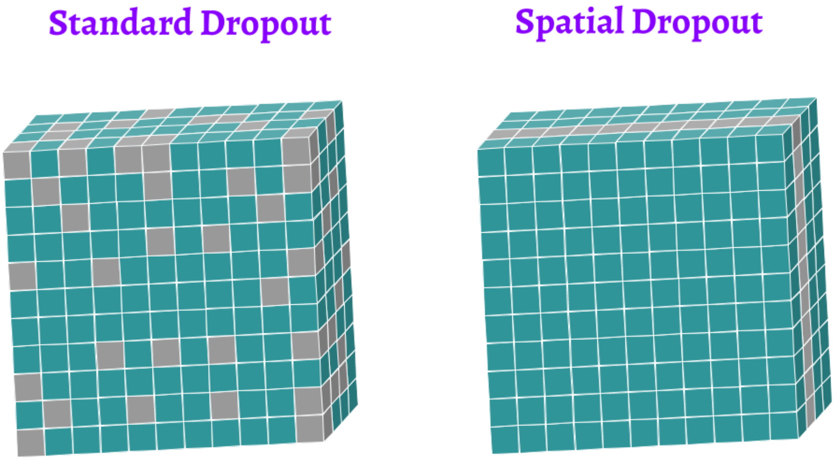


Figure 10: Standard and spatial dropout. Spatial dropout is better suited for images, dropping entire channels instead of individual pixels. Taken from [5].

Figure 11 illustrates MCD. Computing n inference runs over a single network results in n different values using dropout at inference, where uncertainty is calculated as their level of disagreement.

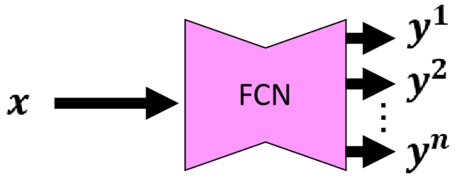


Figure 11: Visualization of MCD. A single network is trained, and n inference runs are obtained using dropout at inference, from which uncertainty is computed [3].

3.2.2.2 Ensembles

Ensemble methods combine the predictions of several different deterministic networks at inference. In line with Equation 3-5, ensembles aim to approximate the posterior distribution for the model weights $p(w|\mathcal{D})$, which represent model uncertainty, by training multiple sets of parameters and then averaging the outcomes from these distinct models.

A key issue in designing network Ensembles is diversity. In other words, the errors of each ensemble member should concentrate on different regions of the feature space. The reason for this is that we want to assess the uncertainty stemming from divergent opinions among various classification models. Diversity can be achieved in different ways. In Deep Learning, it is common practice to create ensembles by training the same basic network architecture starting from different random initializations [24]. So, each training run produces a different ensemble member.

This is the strategy adopted in our research. Therefore, we obtain n distinct networks, obtaining n different results for each input. As with the Dropout strategy, uncertainty is calculated upon the multiple outcomes, using different metrics, as described in Subsection 3.2.2.4.

It is worth noticing that this method is more computationally expensive than MCD, because it involves training n networks, whereas MCD requires training a single network. However, in an operational application inference time is more relevant, where both MCD and Ensembles present the same computational cost.

Figure 12 presents the ensemble method. Multiple repetitions of the same base network are trained using the same input data during training. Then, at inference, each network produces a different output, and uncertainty is calculated as the level of disagreement.

Figure 13 presents the parameter space and the model's convergence areas for different uncertainty estimation methods including single-outcome confidence-based approaches (i.e., deterministic networks), Bayesian networks

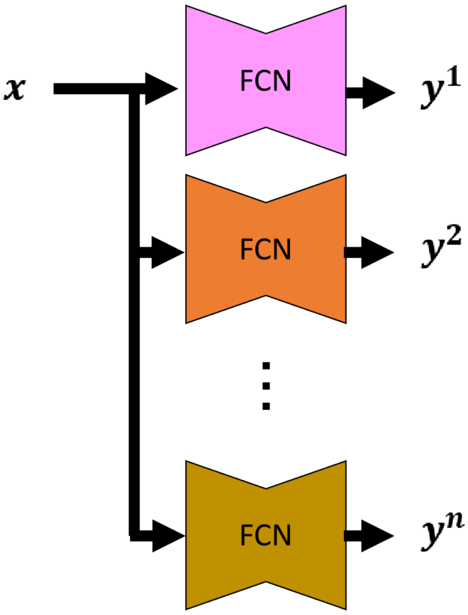


Figure 12: Visualization of ensemble. A defined number n of training repetitions results in n different values at inference, from which uncertainty is computed [3].

(Approximated in this work by MCD) and deep ensembles. For simplicity, the figure represents an uni-dimensional parameter space. The horizontal axis represents the network parameters ω and the vertical axis the loss value. ω_1^*, ω_2^*, and ω_3^* represent multiple optimal parameter values for the training loss. A deterministic network learns a single point estimate in the parameter space. Instead, the Bayesian network also considers the surrounding of a single point. The ensemble networks consider multiple optimal points in the parameter space, each corresponding to a different training run [2].

3.2.2.3
Test-Time Augmentation

Test-Time Augmentation (TTA) [60] is one of the simplest uncertainty estimation techniques, and it relies on a single network. First, a single deterministic network is trained. At test time, data augmentation is applied to each test sample (e.g., rotation, flip, mirror, and scaling). Then, the disagreement between the different outcomes from the sample augmentations is calculated using one of the uncertainty metrics described in Subsection 3.2.2.4.

TTA is different from the previous approaches because it estimates the so-called aleatoric (data) uncertainty caused by variations in the input caused by various noise sources. MCD and ensembles capture uncertainty linked to the models themselves, the so-called epistemic or model uncertainty.

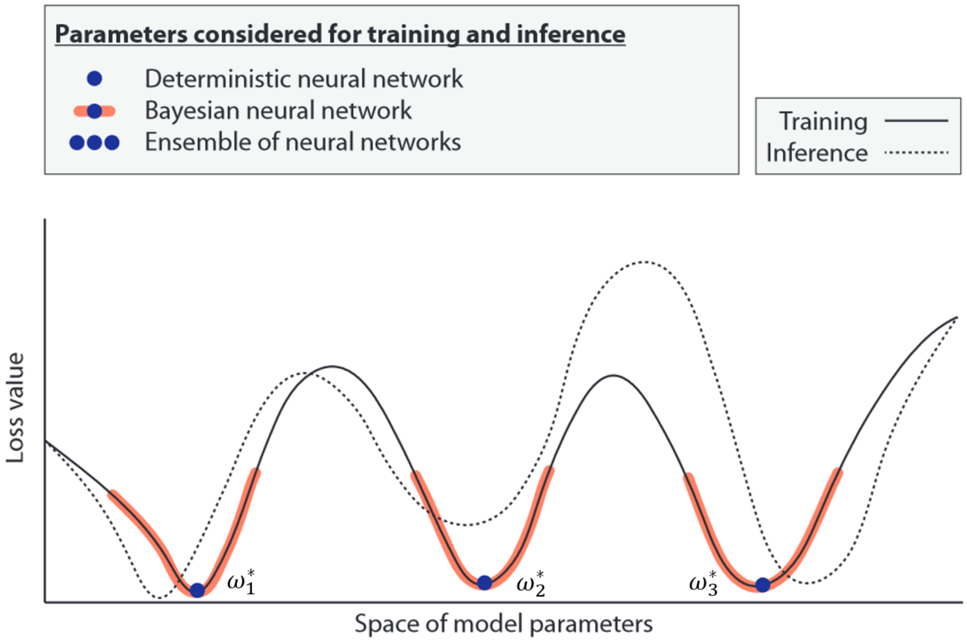

Parameters considered for training and inference
Deterministic neural network
Bayesian neural network
Ensemble of neural networks
Training
Inference
Loss value
ω_1^*
ω_2^*
ω_3^*
Space of model parameters

Figure 13: Visualization of the parameter space for different approaches including deterministic networks (Single point estimate), Bayesian networks (In this work approximated by MCD), and ensembles (Multiple point estimates). For simplicity, the figure represents an uni-dimensional parameter space. Taken from [2].

3.2.2.4 Uncertainty Metrics

This subsection describes the metrics used in this research to estimate the uncertainty from the n results produced by MCD, Ensembles, or TTA.

Predictive Entropy:

In a few words, the Predictive Entropy $H(\mathbf{y})$ is the *entropy*, as defined in eq. 3-1, computed upon the *final prediction* defined in eq. 3-4 [60, 103]. So, the definition of predictive entropy is given by:

$$H(\mathbf{y}) = -\frac{1}{k}\sum_{k=1}^{K} \mu_k log(\mu_k).$$

where μ_k is the mean probability across inference runs for class k, K is the number of classes, and $[\mu_1, \mu_2, ..., \mu_K]$ corresponds to the final prediction $\boldsymbol{\mu}$ (eq. 3-4). Predictive entropy captures the total uncertainty, including the model and data uncertainty [61]. Additionally, we assessed other uncertainty metrics as an ablation study.

Predictive Variance:

The Predictive Variance is the average variance computed over the K classes [95]. Recall that the k-th component μ_k of the final prediction vector $\boldsymbol{\mu}$ (eq. 3-4) is given by

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n} y_k^{(i)}.$$

So, the variance of the n predictions for class k can be computed as

$$\sigma_k^2 = \frac{1}{n}\sum_{i=1}^{n}(y_k^{(i)} - \mu_k)^2.$$

The Predictive Variance $\boldsymbol{\sigma}^2$ is the average of class variances, specifically:

$$\boldsymbol{\sigma}^2 = \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2.$$

Mutual Information:

The Mutual Information $MI(\mathbf{y})$ assesses the similarity between a set of random variables that were sampled simultaneously. It informs how much information from one random variable is present in another [104]. Its value is the difference between the predicted entropy computed on the final prediction and the average of the entropy of each prediction:

$$MI(\mathbf{y}) = H(\mathbf{y}) - \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K} y_k^{(i)} log(y_k^{(i)}).$$

$MI(\mathbf{y})$ can be interpreted as the difference between the total uncertainty, captured by the predictive entropy $H(\mathbf{y})$, and the expected data uncertainty, captured by the expected entropy of each individual inference run. Considering the total uncertainty as the sum of model and data uncertainty, $MI(\mathbf{y})$ captures model uncertainty [61].

Expected Kullback-Leibler Divergence:

This Kullback-Leibler (KL) Divergence measures the divergence between two probability distributions [105]. In this context, the Expected KL Divergence (EKL) measures the average (expected) KL divergence between the final and the individual predictions, formally:

$$EKL(\mathbf{y}, \boldsymbol{\mu}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \mu_k log(\mu_k / y_k^{(i)}).$$

3.3 Evidential Deep Learning (EDL)

Multiple-outcome approaches are computationally expensive due to the need to train or infer multiple times. Evidential Deep Learning (EDL) is an alternative uncertainty estimation method that allows to obtain uncertainty using a single training run and a single inference step [18].

Let us assume that each winning class elected by one of the aforementioned multi-output approaches represents a sample drawn from some underlying categorical distribution parameterized by a probability vector $\boldsymbol{p}$. Intuitively, if one of those classes often comes about, this is an indication of low uncertainty. On the other hand, if more than one class occurs with similar frequency, this is a strong indication of high uncertainty.

Going a step further, we assume that each probability vector $\boldsymbol{p}$ is drawn from some other higher-order distribution.

EDL seeks to learn the parameters of such high-order distributions. So, instead of drawing samples from distributions, EDL aims at learning the distribution over these distributions, called hereafter evidential distribution, directly from the data. Therefore, a sample drawn from the evidential distribution defines itself as a distribution over the data.

Let us assume that a class label $L \in \{1, ..., k\}$ is a sample drawn from a likelihood function of the categorical form parameterized by probability vector $\boldsymbol{p}$, i.e.,

$$L \sim categorical(\boldsymbol{p})$$

As in [18], we further assume that the discrete probability distribution $\boldsymbol{p}$ can be estimated using a Dirichlet prior:

$$\boldsymbol{p} \sim D(\boldsymbol{\alpha})$$

The Dirichlet prior is itself parameterized by a set of K (the number of classes) parameters, here denoted $\boldsymbol{\alpha}$. A sample drawn from the Dirichlet distribution is a realization of the distribution probabilities $\boldsymbol{p}$.

The equation for the Dirichlet distribution is the following:

$$D(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} p_i^{\alpha_i - 1} & \text{for } \mathbf{p} \in S_K, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{S}_K$ is the K-dimensional unit simplex:

$$\mathcal{S}_K = \left\{ \mathbf{p} | \sum_{i=1}^{K} p_i = 1 \;\; and \;\; 0 \leq p_1, ..., p_K \leq 1 \right\}. \tag{3-7}$$

and $B(\boldsymbol{\alpha})$ is the K-dimensional multinomial beta function.

Figure 15 illustrates the K-dimensional unit simplex for a sample $K = 3$ ($\mathcal{S}_3$). It can be observed that every possible value of $\mathbf{y}$ is contained in the simplex defined by Equation 3-7.

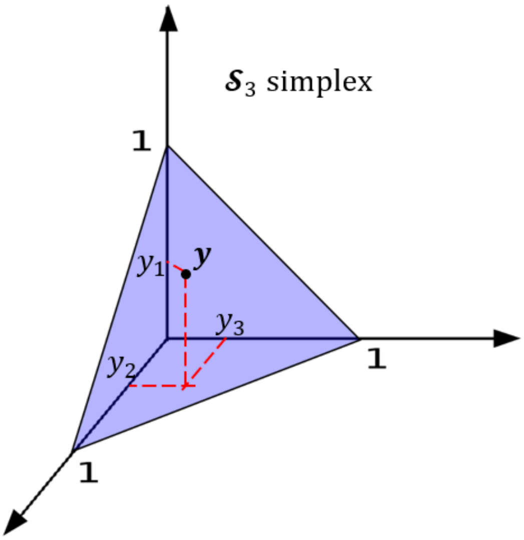


Figure 14: Illustration for the K-dimensional unit simplex defined in Equation 3-7, for $K = 3$. Each node of the simplex represents one class [3].

EDL divides uncertainty into the data, model, and additionally estimates the distributional uncertainty. Distributional uncertainty refers to the overlap between the train and test distributions, and a high value indicates the sample is Out-Of-Distribution (OOD).

Figure 15 presents sample class probabilities for deterministic networks, placed as point samples in the K-th dimensional simplex for $K = 3$, with a point in the corner representing a probability of 100% for the corresponding class and a point in the center meaning equal predicted probability for all classes. Instead of representing a single point (i.e., a single probability vector p), the resulting α_i parameters from the evidential distribution describe all the possible probability vectors at the network's output, which can be represented as a density function in the K-th dimensional simplex, as presented in Figure 16. In that figure, low uncertainty results in high concentrated values near one of the simplex corners (left image). High data uncertainty results in highly concentrated values near the simplex's central region (center image), and high distributional uncertainty results in highly dispersed values covering a large portion of the simplex (right image). In terms of the Dirichlet parameters, high data uncertainty is represented as equal or similar $\boldsymbol{\alpha}$ values for all classes (e.g. $\boldsymbol{\alpha} = [50, 50, 50]$) and high distributional uncertainty is represented as $\boldsymbol{\alpha}$ values close to 1 for all classes (e.g. $\boldsymbol{\alpha} = [1.5, 1.5, 1.5]$), and it indicates an OOD sample due to not having any belief in any of the output classes.

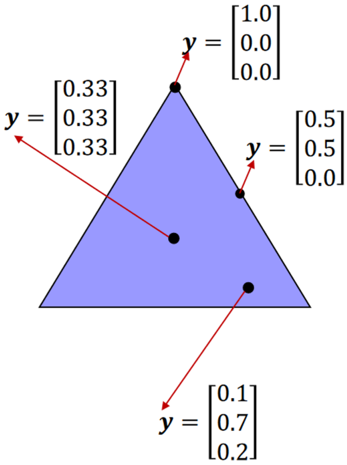

$\boldsymbol{y} = \begin{bmatrix}1.0\\0.0\\0.0\end{bmatrix}$
$\boldsymbol{y} = \begin{bmatrix}0.33\\0.33\\0.33\end{bmatrix}$
$\boldsymbol{y} = \begin{bmatrix}0.5\\0.5\\0.0\end{bmatrix}$
$\boldsymbol{y} = \begin{bmatrix}0.1\\0.7\\0.2\end{bmatrix}$

Figure 15: Class probabilities from deterministic networks are illustrated as point samples from the K-th dimensional simplex ($K = 3$). 4 point samples are presented [3].

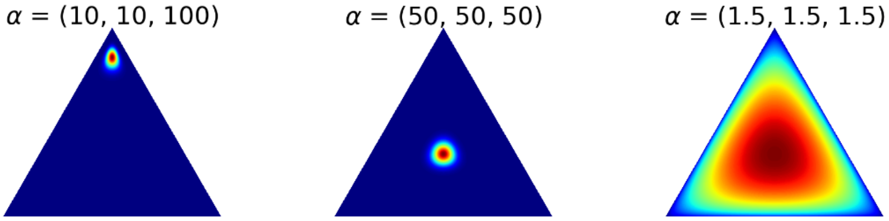


Figure 16: Desired behaviors of a Dirichlet distribution for the K-th dimensional simplex ($K = 3$). The left image represents a sample with low uncertainty. The center image represents high data uncertainty. The right image represents high distributional uncertainty, indicating an OOD sample. Image from [2]

Typical neural networks have a softmax function at the last layer to produce class probabilities $\{y_k\}_{k=1}^{K}$ from the logits at its inputs produced by the network, as shown in Figure 17.

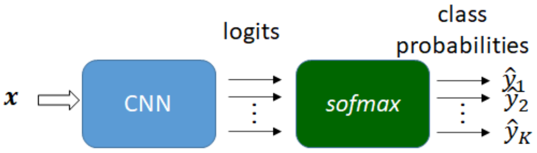


Figure 17: Typical CNN architecture with a softmax attached to the last layer [3].

In an evidential network, instead, the network computes the so-called evidences e_k, with the evidences being obtained by replacing the softmax activation with an activation function (e.g., ReLU) that ensures non-negative values while not applying the softmax constrain. To elaborate, the primary dissimilarity between evidence and softmax lies in the constraints imposed on their values. Softmax constrains its values to represent probabilities, ensuring that the sum of probabilities across all classes equals 1. Conversely, evidence is not bound by such probability normalization constraints; its only requirement is to maintain a non-negative value. Evidence values are applied to the last layer to compute the parameters of the Dirichlet distribution $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_K]$, as $\alpha_k = e_k + 1$, rather than a single probability distribution $\boldsymbol{y} = [y_1, ..., y_K]$ (see Figure 18).

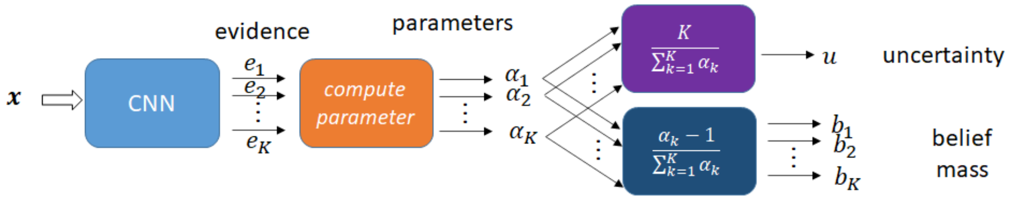


Figure 18: EDL approach: the CNN computes the evidences from which the Dirichlet parameters are computed. The belief mass and the uncertainty are directly computed from the Dirichlet parameters. Image from [3].

The uncertainty and class probabilities can be directly computed from the Dirichlet parameters as follows.

Let b_k, be the belief mass for class $k \in \{1, ..., K\}$ and u the uncertainty mass, where

$$u + \sum_{k=1}^{K} b_k = 1,$$

with $u \geq 0$ and $b_k \geq 0$. The belief mass can be computed from the evidence e_k for class k (provided from the network output), as

$$b_k = \frac{e_k}{S},$$

$$u = \frac{K}{S},$$

where $S = \sum_{k=1}^{K}(e_k + 1)$.

Additionally, the expected probability for the k-th class is the mean of the corresponding Dirichlet distribution and computed as

$$y_k = \frac{\alpha_k}{S},$$

where $\sum_{k=1}^{K}(y_k) = 1$.

The equations above are easily differentiable and bring no difficulty to the backpropagation of the error gradient. The loss function for training the network is composed of up to three terms that take into account the cross-entropy to minimize the prediction error, the variance of the Dirichlet experiment generated by the neural network, and the Kullback-Leibler (KL) divergence as a regularization term.

The training loss is presented in Equation 3-9, where the first term is the sum of squares loss and the second term corresponds to the Dirichlet distribution function.

$$L_i(\Theta) = \int \|\mathbf{y}_i - \mathbf{p}_i\|_2^2 \frac{1}{B(\alpha_i)} \prod_{i=1}^{K} p_{ij}^{\alpha_{ij}-1} d\mathbf{p}_i,$$

$$L_i(\Theta) = \sum_{j=1}^{K} \mathbb{E}\left[y_{ij}^2 - 2y_{ij}p_{ij} + p_{ij}^2\right] = \sum_{j=1}^{K}\left(y_{ij}^2 - 2y_{ij}\mathbb{E}\left[p_{ij}\right] + \mathbb{E}\left[p_{ij}^2\right]\right). \tag{3-8}$$

Then by applying the identity:

$$\mathbb{E}\left[p_{ij}^2\right] = \mathbb{E}\left[p_{ij}\right]^2 + Var(p_{ij}),$$

the training loss is obtained in an interpretable form:

$$\begin{aligned} L_i(\Theta) &= \sum_{j=1}^{K}(y_{ij} - \mathbb{E}\left[p_{ij}\right])^2 + Var(p_{ij}), \\ &= \sum_{j=1}^{K}(y_{ij} - \alpha_{ij}/S_i)^2 + \frac{\alpha_{ij}(S_i - \alpha_{ij})}{S_i^2(S_i+1)}, \\ &= \sum_{j=1}^{K}(y_{ij} - \tilde{p}_{ij})^2 + \frac{\tilde{p}_{ij}(1-\tilde{p}_{ij})}{(S_i+1)}. \end{aligned} \tag{3-9}$$

The loss over a batch of samples is computed by summing the loss for each sample within the batch. During training, the model is expected to generate high evidence for repeating input patterns that match specific ground truth classes. For example, in a digit recognition problem (e.g., MNIST), a circular pattern may produce high evidence for a sample with a class label 0. This means that the outcome of the network, corresponding to the evidence values, should be increased when the network observes such a pattern during training. On the other hand, when counter-examples are observed, such as the same circular pattern for a class label 6, the network's parameters should be tuned to produce smaller amounts of evidence for this pattern (At the output of class 0) while minimizing the training loss, as long as the overall loss also decreases.

However, when counter-examples are limited, decreasing the magnitude of the generated evidence in the counter-examples may increase the overall loss, even if the loss decreases for the individual sample, consequently generating some unwanted evidence for the incorrect class. Such undesired evidence would not be a problem if the sample is correctly classified (if the evidence for the correct class is still larger than the remaining classes). Even so, the authors preferred the evidence of the incorrect samples to shrink to zero. Note that Dirichlet distribution with zero total evidence corresponds to the uniform distribution with an $\boldsymbol{\alpha}$ value of 1 for all classes: $D(\mathbf{p}_i| \langle 1, ..., 1\rangle)$, and total uncertainty ($u = 1$). They achieved this by incorporating a Kullback-Leibler (KL) regularization term in the training loss function, which minimizes the evidence from the losing classes (i.e., the incorrect classes) to 0 by approximating their corresponding Dirichlet parameters $\boldsymbol{\alpha}$ to 1. The resulting training loss function is:

$$L_i(\Theta) = \sum_{j=1}^{K} L_i(\Theta) + \lambda_t \sum_{j=1}^{K} KL\left[D(\mathbf{p}_i|\tilde{\boldsymbol{\alpha}}_i)||D(\mathbf{p}_i)| \langle 1, ..., 1\rangle\right], \tag{3-10}$$

where $\lambda_t = min(1.0, t/10) \in [0, 1]$ is the annealing coefficient, t is the training epoch, $D(\mathbf{p}_i| \langle 1, ..., 1\rangle)$ is the uniform Dirichlet distribution. In this context, $\tilde{\boldsymbol{\alpha}} = \mathbf{y}_i + (1 - \mathbf{y}_i) * \boldsymbol{\alpha}_i$ represents the predicted Dirichlet parameters following the removal of non-misleading evidence from the predicted parameters $\boldsymbol{\alpha}_i$ associated with sample i. In other words, $\tilde{\boldsymbol{\alpha}}_i$ corresponds to $\boldsymbol{\alpha}_i$ after modifying the Dirichlet parameter value for the winning class to 1, with the objective of only modifying the Dirichlet parameters from the non-winning classes in the KL regularization term.

The KL divergence term equation is:

$$KL[D(\mathbf{p}_i|\boldsymbol{\alpha}_i)||D(\mathbf{p}_i)|\mathbf{1}]$$

$$= log\left(\frac{\Gamma(\sum_{k=1}^{K} \tilde{\alpha}_{ik})}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\tilde{\alpha}_{ik})}\right) + \sum_{k=1}^{K}(\tilde{\alpha}_{ik} - 1)\left[\psi(\tilde{\alpha}_{ik}) - \psi\left(\sum_{k=1}^{K}\tilde{\alpha}_{ik}\right)\right], \tag{3-11}$$

where $\mathbf{1}$ denotes a parameter vector consisting of K ones. Additionally, $\Gamma(\cdot)$ refers to the *gamma* function, while $\psi(\cdot)$ represents the *digamma* function.

As indicated at the beginning of this explanation, this approach involves training a single network and a single inference step to compute class probability distributions and uncertainty.

4
UNCERTAINTY ESTIMATION TO REDUCE THE MANUAL ANNOTATION EFFORT

This chapter describes the design of experiments to evaluate the impact of uncertainty estimation to aid the annotation and auditing process in deforestation mapping in the Amazon rainforest, using Monte Carlo Dropout (MCD), ensembles, Evidential Deep Learning (EDL), and a single confidence measure, i.e., the entropy. The next chapter reports and discusses the results of these experiments.

The following sections describe the adopted methodology, which includes adding the distance from the past deforestation maps to the input data. Besides, in a second protocol, the training relies only on labeled training data from earlier dates, avoiding manual labeling of parts of the target image for training the network.

4.1
Deep Neural Network (DNN)

Figure 19 describes the network architecture used in this work. The proposed methodology can be applied to any Deep Neural Network (DNN) designed for semantic segmentation. Noteworthy examples of such networks include UNet [40], ResUNet [12, 20], DeepLabV3+ [106], and Swin UNet [107]. In the case of MCD, following previous works [41–43], it may be beneficial to add dropout layers to each encoding and decoding stage for UNet-like architectures.

To detect the deforestation changes from time T_{-1} to T_0, the input to the network is the concatenation of Sentinel-2 images corresponding to both dates ($S2_{T_{-1}} : S2_{T_0}$), where T_0 is the current date, T_{-1} is the date from one year before the current date, $S2_{T_0}$ is the Sentinel-2 image for the current date, and $S2_{T_{-1}}$ is the Sentinel-2 image for the previous year.

4.2
Temporal distance to past deforestation

A recent study indicated that deforestation tends to occur in spatial proximity to areas that were deforested in preceding years. The study con-

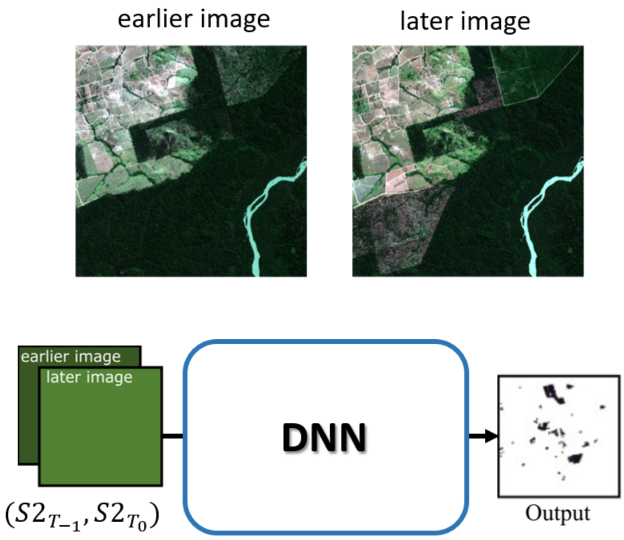


Figure 19: Base architecture for deforestation detection. Input is the concatenation of T_{-1} and T_0. Output is the segmentation map with detected deforestation changes from T_{-1} to T_0.

sistently demonstrated enhancements in results by incorporating information about deforestation in previous years as an input [108].

In this work, we leverage the information about past deforestation taken from PRODES [8] and use it as an additional input feature map concatenated to the input Sentinel-2 image pair. We define the temporal distance to past deforestation map $D_{T_0}(x, y)$ as:

$$D_{T_0}(x, y) = T_0 - deforestation\_year(x, y) + 1$$

where T_0 is the current year, and $deforestation\_year(x, y)$ is the year of deforestation for the pixel at location (x, y). We set $deforestation\_year(x, y)$ to T_0 in pixel locations where deforestation has not occurred. If we trained and tested on the most recent image pair, the input tensor comprised $(D_{T_{-1}}(x, y), S2_{T_{-1}}, S2_{T_0})$, where T_0 and T_{-1} correspond to the current and previous year acquisition dates, and $S2_{T_0}$ and $S2_{T_{-1}}$ represent the corresponding Sentinel-2 images for T_0 and T_{-1} dates, respectively. Figure 20 illustrates the resulting input tensor using temporal distance map to past deforestation as an additional input. The temporal distance map was normalized, ensuring the values were between 0 and 1 before concatenating it to the input image.

4.3 Uncertainty to Aid the Auditing Process

As mentioned, the experiments used as ground truth were the deforestation reports published by PRODES. In PRODES's methodology, 100% of the

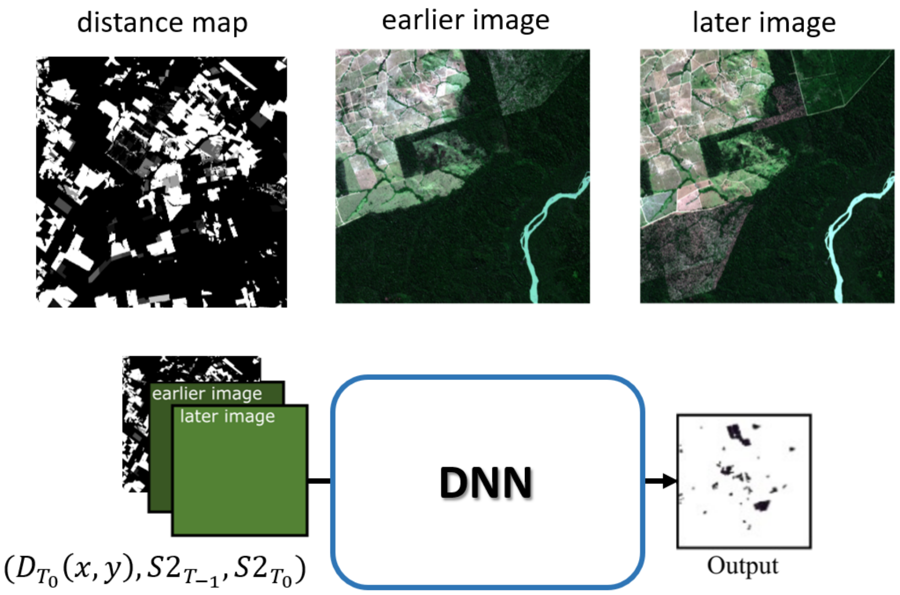


Figure 20: Illustration of the distance to past deforestation map as input for deforestation detection.

non-deforested area until the current year is visually inspected and audited by INPE's personnel in each upcoming year. In this work, we propose an alternate semi-automatic workflow where we first train a neural network on image pairs from previous years and then run the inference on a recent image pair (as explained in Section 4.4).

So, we consider a methodology to reduce the labeling/auditing effort that involves two steps. The first step consists of running an automatic classification using a deep learning model trained as outlined before. Beyond delivering a deforestation map, the first step computes the uncertainty associated with the classification of each pixel. In the second step, the photo interpreter only audits the areas whose uncertainty computed in step 1 exceeds a user-selected threshold. We leave for future works the specific details in which such areas to be audited are presented to the auditor.

The challenge of deforestation detection, where the methodology will be assessed, is a binary classification problem, inherently characterized by a pronounced class imbalance. Typically, the identified deforested areas represent less than 1% of the dataset. This substantial imbalance renders straightforward metrics, such as accuracy, unsuitable. Accuracy, which measures the percentage of correctly classified samples, becomes misleading when applied to imbalanced datasets, as achieving 99% accuracy is possible by simply predicting all outcomes as non-deforestation in the case of a 1% detection rate. To address this issue, our study adopts the F_1-score as a more suitable classification metric for imbalanced scenarios. The F_1-score effectively balances precision and recall, offering a comprehensive evaluation of the model's performance in deforestation detection.

As classification metrics, we obtain $F1_{low}$ as the F_1-score metric for test samples with low uncertainty and $F1_{high}$ as the F_1-score metric for test samples with high uncertainty. We call the percentage of test samples (i.e., test pixels) with high uncertainty the Alert Area (AA). We expect that $F1_{low} > F_1 >> F1_{high}$, where F_1 refers to classification metrics prior to applying the uncertainty methodology.

The final deforestation report results from joining:

– on the pixels with low uncertainty, the classes assigned by the model

– on the pixels with high uncertainty, the classes assigned by the photo-interpreter

$F1_{audit}$ represents the classification metrics (i.e. the F_1-score on the entire test set) after the expert auditor has correctly re-annotated the samples (i.e., the pixels) with high uncertainty corresponding to the AA percentage.

Figure 21 illustrates the threshold selection criteria. The expert photo-interpreter selects the desired percentage of pixels to be audited from the test set (AA), which defines the uncertainty threshold value, considering there is an inverse correlation between AA and uncertainty. Figure 22 illustrates sample classification metrics for multiple uncertainty threshold values. Depending on the selected AA and corresponding uncertainty threshold, samples which uncertainty is lower than the threshold are expected to present higher classification accuracy and samples with higher uncertainty than the threshold are expected to present lower classification accuracy, separating the samples that can be trusted from the samples that need to be re-annotated. Note that the auditing of the high uncertainty pixels may require analyzing their neighborhoods. Future works may also specify a minimum area for the resulting high-uncertainty polygons to be presented to an auditor. Similarly, the auditor could also neglect high uncertainty polygons in the form of thin lines, as occurs on the edges of deforestation polygons. That procedure may be addressed in future works.

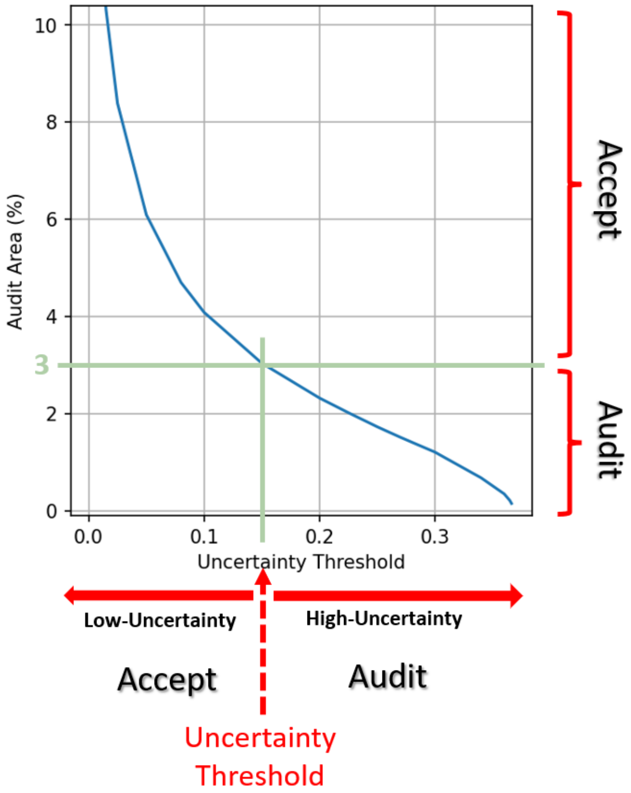


Figure 21: Sample uncertainty threshold vs. Audit Area (AA). The photo-interpreter selects the percentage of test samples (i.e., test pixels) with high uncertainty for auditing (AA), corresponding to an uncertainty threshold value. In this work, results are presented for $AA = 3\%$.

4.4 Assessing Uncertainty For Deforestation Detection

Recent works in deforestation detection used an image from the same date to train and test by dividing the raster into non-overlapping tiles and setting a percentage of those tiles to train and test within the same image [12]. However, in a real-world application, collecting ground truth data for each upcoming date would be costly and unfeasible. More specifically, collecting ground truth data for the target date would be equivalent to performing manual inspection in a large enough percentage of the image in order to perform training, contradicting the primary objective of this work, which is to reduce the need for manual inspection for each upcoming year. In this work, we propose an operational solution where we train the network a date pair from the past and infer on a new upcoming date that was not seen during training. Formally, we trained on $(S2_{T_{-2}}, S2_{T_{-1}}, D_{T_{-2}}(x, y))$ and inferred on $(S2_{T_{-1}}, S2_{T_0}, D_{T_{-1}}(x, y))$, where $S2$ corresponds to a Sentinel-2 image. T_0 corresponds to the image from the current date, T_{-1} to the previous year date, and T_{-2} to the second year before T_0. Notice that this approach does not need annotated ground truth for

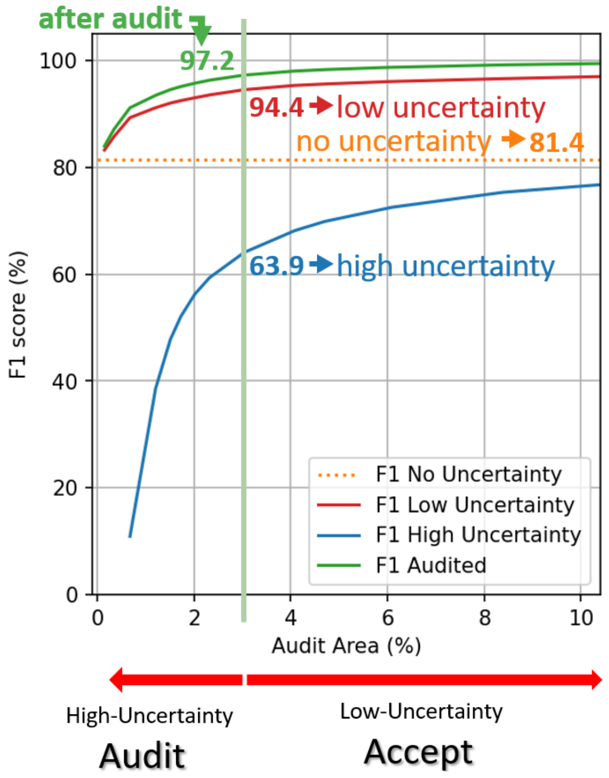


Figure 22: Classification metrics for multiple uncertainty thresholds. $F1_{low}$ corresponds to accepted samples with low uncertainty. $F1_{high}$ corresponds to not accepted samples with high uncertainty. $F1_{audit}$ corresponds to metrics after auditing the high uncertainty samples. In this work, results are presented for $AA = 3$.

the target date T_0. In this way, the proposed approach could directly be applied in a real-world application. The aforementioned training scheme is presented in Figure 23. An operational application could adopt the suggested approach of training on a prior date and inferring on the subsequent date, by executing the training procedure annually. Notice that training in an earlier pair of dates and testing in a current pair of dates may result in a difference between the train and test distributions, which may result in lower classification metrics compared to the ideal case of training and testing on the same date. Future works may further improve classification metrics by using additional training data from additional earlier years, leveraging PRODES reference since 1988.

A comparison of MCD, deep ensembles, and EDL is presented for the ResUnet base network. Besides, entropy from a single forward run is also assessed as a baseline method. In the latter case, we calculated entropy from n different training runs and present the best and worst case in terms of $F1_{low}$. Furthermore, an analysis from the point of view of computational cost is presented by comparing training and inference times on each method and study areas in Subsection 5.3.5. Then, a summary of the quantitative analysis is presented in

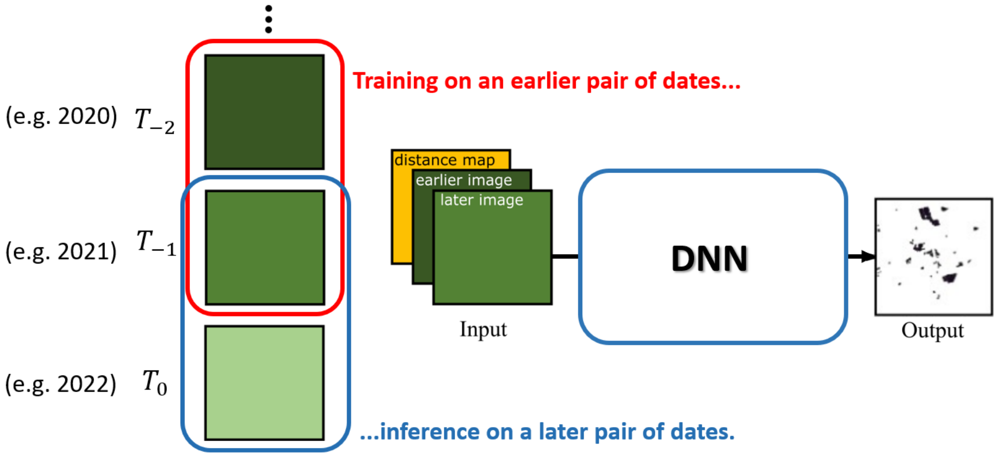


Figure 23: Network scheme training on samples from past date pairs, and inferring on an unseen date.

Subsection 5.3.6 along with a recommendation of which uncertainty method and metric to use for deforestation detection applications according to the results obtained in this work. Finally, uncertainty interpretation is presented correlating the high uncertainty regions with specific forest characteristics and ongoing deforestation processes that could have caused the elevated levels of uncertainty.

5
EXPERIMENTAL ANALYSIS

This chapter reports the experiments carried out in order to validate the method proposed in the previous chapter. First, the datasets used in the experiments for deforestation detection are presented. Then, the experimental protocol followed for the proposed methodology is described, and the parameter setup is detailed. Finally, the results obtained in the experiments are reported.

5.1
Study Areas

We evaluated the proposed method in two study areas in the Brazilian Legal Amazon (Figure 24), and two study areas in the Brazilian Cerrado biome (Figure 25). Both sites in the Amazon have a mixed land cover. The first Amazon site is located in the Para state (PA), with an area of $92 \times 177 Km^2$. This site is mainly composed of dense evergreen forest and pasture. For the single date pair experiments, we used $[T_{-1}, T_0] = [2018, 2019]$. For the experiments training with past dates, we trained with the $[T_{-2}, T_{-1}] = [2017, 2018]$ date pair and tested on a new upcoming image using $[T_{-1}, T_0] = [2018, 2019]$. The second Amazon site is located in the Mato Grosso state (MT), with an area of $134 \times 208 Km^2$. This site is mainly composed of dense forests, soy fields, and pastures. We used $[T_{-1}, T_0] = [2019, 2020]$. For the multiple date pair experiments, we trained with the $[T_{-2}, T_{-1}] = [2018, 2019]$ date pair.

The remaining two sites are located in the Cerrado biome. The first Cerrado site is located in the Mato Grosso do Sul state (MS), with an area of $195 \times 186 Km^2$. This site is mainly composed of wooded savanna. In the single-date experiments, we used $[T_{-1}, T_0] = [2019, 2020]$. For the experiments training on an earlier date, we trained on $[T_{-2}, T_{-1}] = [2018, 2019]$ and tested on an upcoming date $[T_{-1}, T_0] = [2019, 2020]$. The second Cerrado site is located in the Piauí state (PI), with an area of $208 \times 194 Km^2$. The site is mainly composed of wooded savanna. In the single-date experiments, we used $[T_{-1}, T_0] = [2019, 2020]$. In the experiment training with an earlier date, we trained on $[T_{-2}, T_{-1}] = [2018, 2019]$ and tested on $[T_{-1}, T_0] = [2019, 2020]$.

Tables 3, 4, 5, 6 shows the image acquisition dates, vegetation typologies, coordinates, and deforestation pixel counts for the four sites. In the Amazon sites, all but the $60m$ resolution bands of the Sentinel-2 images were used. Due to computational limitations in the Cerrado sites, related to their larger spatial extent, the NIR-R-G-B bands were used in those sites (Bands 8, 4, 3, and 2). The $20m$ resolution bands were up-sampled to $10m$ with the nearest neighbor method in all cases.

Table 3: Detailed PA study area information: location coordinates, vegetation typology, acquisition dates, and class distribution. Previous deforestation pixels were not considered.

Study areas	**PA**
Coordinates	6° 27’ 10.512” S - 8° 3’ 33.192” S
	55° 36’ 47.4084” W - 54° 46’ 24.3624” W
Vegetation	Dense ombrophilous forests and pasture
Date T_{-2}	July [21, 26], 2017
Date T_{-1}	July [21, 26], 2018
Date T_0	July [21, 26], 2019
Deforestation pixels $(\mathrm{T}_{-2}, T_{-1})$	1861735 (1.1%)
No-deforestation pixels $(\mathrm{T}_{-2}, T_{-1})$	103173071 (63.3%)
Previous deforestation pixels $(\mathrm{T}_{-2}, T_{-1})$	58081194 (35.6%)
Deforestation pixels (T_{-1}, T_0)	1838508 (1.1%)
No-deforestation pixels (T_{-1}, T_0)	100903598 (61.9%)
Previous deforestation pixels (T_{-1}, T_0)	60373894 (37%)

Table 4: Detailed MT study area information: location coordinates, vegetation typology, acquisition dates, and class distribution. Previous deforestation pixels were not considered.

Study areas	**MT**
Coordinates	10° 48’ 43.8012” S - 12° 42’ 5.976” S
	55° 12’ 39.384” W - 53° 57’ 49.7916” W
Vegetation	Sparse ombrophilous forests, soy fields and pastures
Date T_{-2}	July [26, 28, 31], 2018
Date T_{-1}	August [02, 05] 2019
Date T_0	August [04, 06, 09, 11], 2020
Deforestation pixels $(\mathrm{T}_{-2}, T_{-1})$	1900166 (1.1%)
No-deforestation pixels $(\mathrm{T}_{-2}, T_{-1})$	99070749 (56.6%)
Previous deforestation pixels $(\mathrm{T}_{-2}, T_{-1})$	74032985 (42.3%)
Deforestation pixels (T_{-1}, T_0)	2271496 (1.3%)
No-deforestation pixels (T_{-1}, T_0)	109996296 (62.9%)
Previous deforestation pixels (T_{-1}, T_0)	62736108 (35.8%)

5.2 Experimental Protocol

The train, validation, and test areas were selected by splitting the site into non-overlapping tiles and selecting 40% for training, 10% for validation, and 50% for testing (See Figures 26 and 27). The tile size was $23 \times 35.4Km^2$, $26.8 \times 41.6Km^2$, $39 \times 37.2Km^2$, and $41.6 \times 38.8Km^2$ for the PA, MT, MS and

Table 5: Detailed MS study area information: location coordinates, vegetation typology, acquisition dates, and class distribution. Previous deforestation pixels were not considered.

Study areas	**MS**
Coordinates	17° 55' 50.952" S - 19° 37' 58.7316" S 54° 10' 50.916" W - 52° 18' 40.6872" W
Vegetation	Wooded savanna
Date T_{-2}	July [23, 25], 2018
Date T_{-1}	August [07, 09] 2019
Date T_0	July [22, 24], 2020
Deforestation pixels (T_{-2}, T_{-1})	473296 (0.1%)
No-deforestation pixels (T_{-2}, T_{-1})	99618511 (27.5%)
Previous deforestation pixels (T_{-2}, T_{-1})	262333593 (72.4%)
Deforestation pixels (T_{-1}, T_0)	506597 (0.1%)
No-deforestation pixels (T_{-1}, T_0)	99008239 (27.3%)
Previous deforestation pixels (T_{-1}, T_0)	262910564 (72.5%)

Table 6: Detailed PI study area information: location coordinates, vegetation typology, acquisition dates, and class distribution. Previous deforestation pixels were not considered.

Study areas	**PI**
Coordinates	9° 14' 33.3024" S - 10° 59' 37.7484" S 46° 4' 54.174" W - 44° 10' 26.7348" W
Vegetation	Wooded savanna
Date T_{-2}	July [16 - 31], 2018
Date T_{-1}	August [03, 08] 2019
Date T_0	August [07 - 22], 2020
Deforestation pixels (T_{-2}, T_{-1})	2816544 (0.7%)
No-deforestation pixels (T_{-2}, T_{-1})	256247610 (63.5%)
Previous deforestation pixels (T_{-2}, T_{-1})	144294046 (35.8%)
Deforestation pixels (T_{-1}, T_0)	2619484 (0.6%)
No-deforestation pixels (T_{-1}, T_0)	253287547 (62.8%)
Previous deforestation pixels (T_{-1}, T_0)	147451169 (36.6%)

PI sites respectively. For training, we extracted overlapping sub-images of size 128 × 128 with 70% overlap from the training regions. We selected only the sub-images with at least 2% of the deforestation class for training. As explained in Section 4.4, results are presented for the ideal case of training and testing on the same pair of dates and for the more operational case of training on an earlier pair of dates. We used online data augmentation by randomly applying rotations and horizontal and vertical flips on each training batch.

Following recent works on deforestation detection in the Brazilian Amazon, in this work a modified version of the fully convolutional ResUnet from [12] was implemented. We used the same parameter configuration as in [12] for the network architecture, and we added dropout in each decoder stage following recent works [41–43]. We used spatial dropout in all cases (Subsection

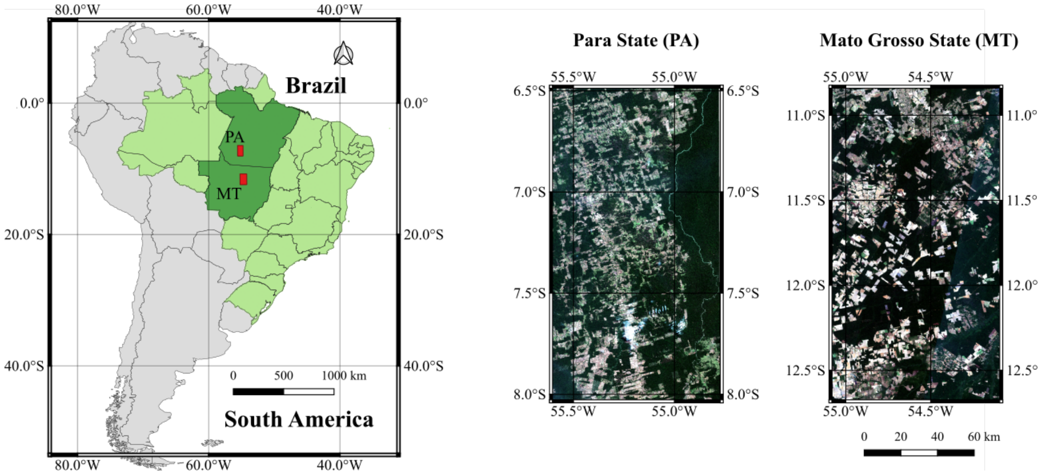


Figure 24: Geographical location of the study areas in the Amazon biome, and RGB composition of the corresponding Sentinel-2 images acquired at T_{-1}.

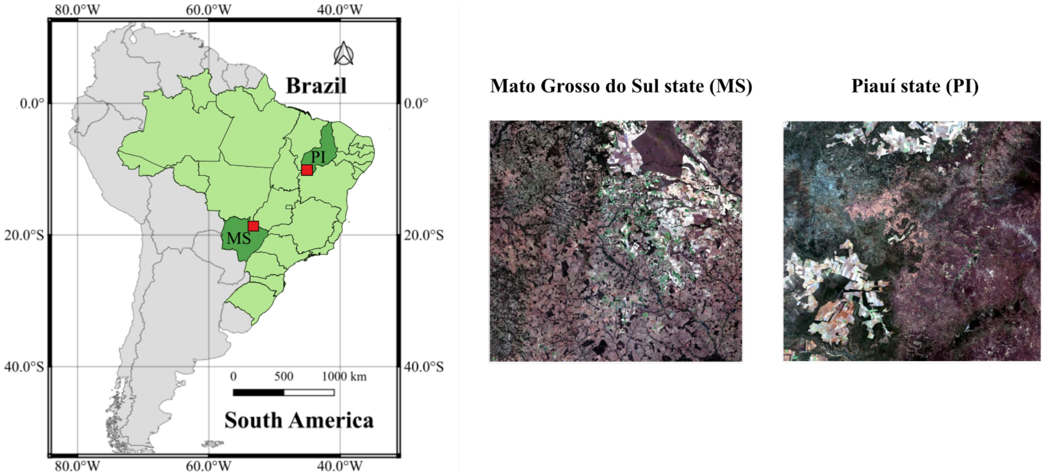
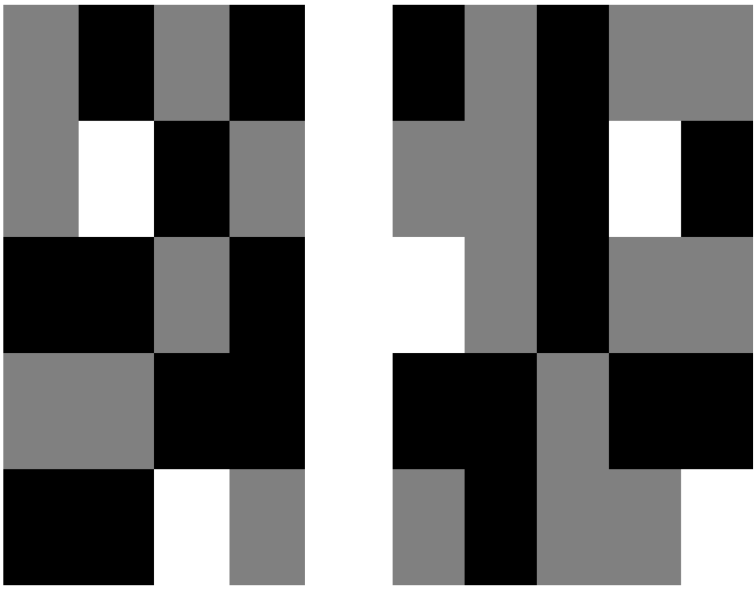


Figure 25: Geographical location of the study areas in the Cerrado biome, and RGB composition of the corresponding Sentinel-2 images acquired at T_{-1}.

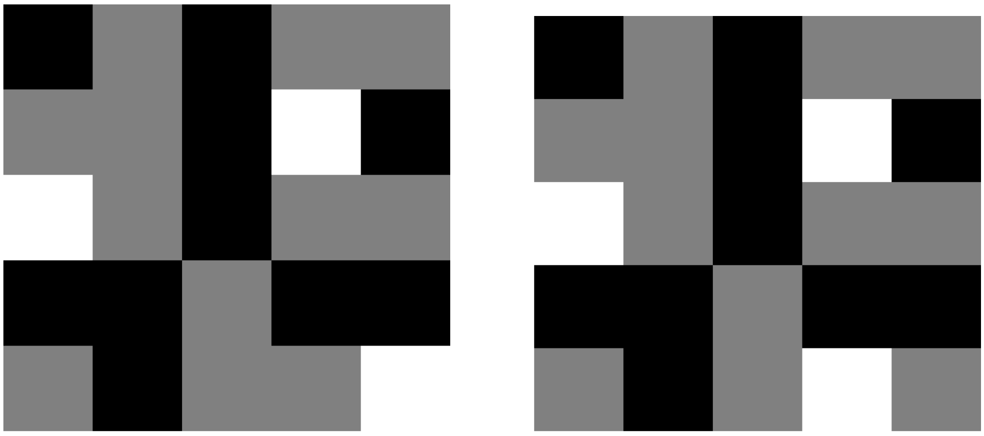
3.2.2.1) [102]. The parameter configuration is presented in Table 7. We used Max-pooling as a down-sampling operator and nearest-neighbor up-sampling. We used a dropout rate of 0.25 in all cases. We used batch size 32. In the multi-output approaches we used Adam optimizer with learning rate 1e-4, and weighted categorical cross entropy with weights of 0.1 for non-deforestation, 0.9 for deforestation, and 0 for not-considered areas, including past deforestation and cloudy regions. Past deforestation refers to areas identified as deforested in years predating the training date. These regions are excluded from both the training and testing phases since their detection is established beforehand. The network's output is focused solely on polygons newly identified on the designated detection date. In EDL, following [18], the network's softmax layer was replaced by ReLU. Samples from the non-considered areas were ignored. For the accuracy assessment, we ignored pixels within a spatial buffer of 2px



(26(a)) PA site

(26(b)) MT site

Figure 26: Train (gray tiles), validation (white tiles) and test (black tiles) mask for (a) PA site and (b) MT site



(27(a)) MS site

(27(b)) PI site

Figure 27: Train (gray tiles), validation (white tiles) and test (black tiles) mask for (a) MS site and (b) PI site

surrounding the ground truth deforestation polygons, and also pixel clusters predicted as deforestation with an area smaller than $6.25ha$ in the Brazilian Amazon sites, and smaller than $4ha$ in the Cerrado sites. The former rule aims to avoid misregistration problems, considering that PRODES references have a spatial resolution equivalent to Landsat multispectral bands. The reason for the later rule was that $6.25ha$ and $4ha$ are PRODES minimum mapping units in the Amazon and Cerrado biomes.

We used ten inference runs ($n = 10$) for uncertainty estimation using MCD. Correspondingly, we used ten training runs for uncertainty estimation

Table 7: Network architecture for the ResUnet-based FCN. C: Convolution, DS: Down-sampling, RB: Residual block, D: Dropout, US: Up-sampling. Convolution layers are parametrized as $(kernel\_width \times kernel\_height, \#filters)$

Encoder	Bottleneck	Decoder	Output

DS(RB(3×3, 32))
DS(RB(3×3, 32))
DS(RB(3×3, 32))
	3× RB(3×3, 128)	D(US(C(3×3, 128)))
		D(US(C(3×3, 64)))
		D(US(C(3×3, 32)))
			Softmax(C(1×1, *#classes*))

using ensembles. We compared results for multiple uncertainty metrics and selected the best overall performing metric. For the EDL method, which requires a single inference run to produce an uncertainty map, we repeated the experiment 10 times and presented results for the average values, as well as the worst and best scenarios with respect to $F_{1_{low}}$. We present entropy from a single inference run for comparison purposes, where we also present results for the averaged values, along with the worst and best performing single inference run from 10 repetitions, with respect to $F_{1_{low}}$. Experiments were carried out in an NVIDIA RTX 2080 Ti GPU.

5.3 Results

In Subsections 5.3.1, 5.3.2, 5.3.3, and 5.3.4, we present results for the proposed methodology in the PA, MT, MS, and PI study areas, respectively.

5.3.1 Uncertainty Estimation Results in PA Site

Table 8 presents results when training and testing on the same date pair, which is a common procedure in recent works [12]. Among the multiple outcome approaches, the best performing method was Ensemble, with an F_1 score of 85.8%. Its improvement over MCD may be related to an enhanced generalization based on multiple locally optimal solutions instead of a single one in the parameter space (Figure 13).

As a baseline comparison, we present results for using entropy from a single inference run. We show the worst and best-performing cases since we repeated this experiment 10 times. Results presented a high variance. Such variability may have occurred due to the single inference run method coming from a single optimal solution during training, different from multi-output solutions which consider either multiple optimal solutions (e.g. ensembles) or the surrounding solutions around a single optimal value (e.g. MCD), as explained in Subection 3.2.2.2. Although the best-case scenario produced results slightly better compared to Ensemble in terms of $F_{1_{low}}$, results from single inference run, the high variance in results from a single inference run makes it unreliable compared to the more robust MCD and Ensemble methods.

Likewise, the worst, mean, and best results for EDL are presented. EDL presented a lower variance in F_1 compared to Single run. The mean performance metrics were similar to Single run, although generally lower. This indicates the need for further research in EDL for deforestation detection. A possible cause for the lower EDL performance may be related to the high-class imbalance in deforestation detection, considering that EDL can produce unwanted evidence for the incorrect class when counter-examples are limited [18].

Table 8: Results for PA site obtained by training and testing on a pair of images from 2018/2019 with $AA = 3\%$. Values are presented in percentages (%). First, second, and third best results in each column are highlighted in **red**, **blue** and **green** respectively

Method	**Uncertainty Metric**	F_1	$F_{1_{low}}$	$F_{1_{high}}$	$F_{1_{audit}}$
MCD	Predictive Entropy	84.1	95.3	61.0	96.9
	Predictive Variance		90.9	**77.5**	96.2
	Mutual Information		88.2	34.0	89.3
	Expected KL		87.2	26.7	88.0
Ensemble	Predictive Entropy	**85.8**	**96.9**	62.3	**97.9**
	Predictive Variance		95.6	**72.3**	97.5
	Mutual Information		92.2	31.3	93.1
	Expected KL		90.9	25.6	91.7
Single run	Entropy (Worst)	83.4	94.1	66.7	96.5
	Entropy (Mean)	**86.2**	**95.9**	68.7	**97.5**
	Entropy (Best)	**87.4**	**97.0**	**74.7**	**98.3**
EDL	Worst	80.0	91.4	51.3	93.3
	Mean	83.1	93.1	59.9	94.9
	Best	83.6	94.8	59.5	96.2

We also present results for multiple uncertainty metrics in MCD and Ensemble approaches. In both cases, the best-performing metric was the predictive entropy, with increase in $F_{1_{low}}$ of up to 8.1% compared to the other metrics. The predictive variance was the second-best metric in both MCD and Ensemble.

In the case of the best-performing multiple-output method (Ensemble) and the best metric (Predictive entropy), we obtained an average F_1 score of 85.8%. After applying the proposed uncertainty-based methodology for an Audit Area (AA) of 3%, we obtained an increased F_1 of 96.9% for the samples with low uncertainty. At the same time, $F_{1_{high}}$ was much lower (62.3%), which indicates that our methodology succeeded in separating the samples whose results we can trust from the samples that we do not know if they are correct.

If we audited the samples with high uncertainty, we would obtain a $F1_{audit}$ of 97.9%, with a significantly smaller auditing effort (3% of the image) compared to the current auditing procedure, where 100% of the image needs to be audited.

Figure 28 presents results for different uncertainty threshold values in the PA site when training and testing in the same date pair, corresponding to a range of AA from 0% to about 10%. Results correspond to the best-performing multiple-output method and the corresponding best uncertainty metric. F_1 before applying the uncertainty methodology is presented in yellow for comparison. $F1_{low}$ increased when increasing AA until a peak value for an AA of approximately 6%, where its value started to decline. This suggests that the uncertainty threshold needs to be adequately selected to obtain the desired outcomes in the proposed methodology. In contrast, $F1_{audit}$ increased monotonically when increasing AA. Both $F1_{low}$ and $F1_{audit}$ got close to 100% even for small values of AA.

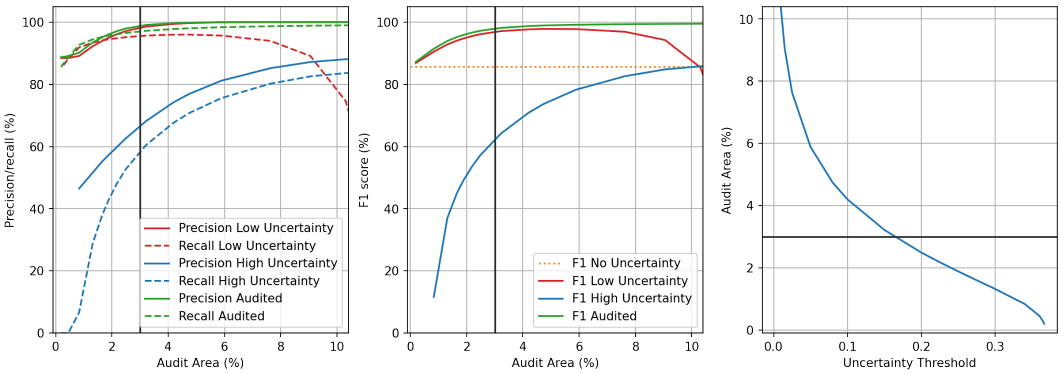


Figure 28: Classification metrics for multiple uncertainty threshold values in PA site. Training and testing in [2018, 2019]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy. A sample AA threshold of 3% is highlighted in gray.

The results in Table 9 correspond to training with date pairs from the past and inferring on a new date unseen during training, representing a more realistic operational setting. In this case, the best multiple-output uncertainty method was also Ensemble, and the best-performing uncertainty metric was also the predictive entropy.

In terms of F_1, results were comparable to the upper bound case of training and testing on the same date, with an F_1 of 81.4%. For an Audit Area (AA) of 3.0%, $F1_{low}$ improved to 94.4%, while $F1_{high}$ was 63.9%, which indicates that the proposed uncertainty methodology was also capable of discerning between samples whose results we can trust and samples we do not know if they are correct in the case of training with past date pairs and inferring on

Table 9: Results for PA site obtained by training on a pair of images from 2017/2018 and testing on a pair of images from 2018/2019 with $AA = 3\%$. Values are presented in percentages (%). First, second, and third best results in each column are highlighted in **red**, **blue** and **green** respectively

Method	**Uncertainty Metric**	F_1	$F_{1_{low}}$	$F_{1_{high}}$	$F_{1_{audit}}$
MCD	Predictive Entropy	78.0	92.0	60.3	96.0
	Predictive Variance		84.8	**73.7**	95.0
	Mutual Information		86.0	34.2	88.6
	Expected KL		84.6	29.4	87.0
Ensemble	Predictive Entropy	**81.4**	**94.4**	63.9	**97.2**
	Predictive Variance		91.7	**72.8**	96.6
	Mutual Information		90.7	46.6	93.3
	Expected KL		88.5	38.7	90.8
Single run	Entropy (Worst)	74.5	83.5	52.4	89.0
	Entropy (Mean)	80.4	91.2	65.9	**97.5**
	Entropy (Best)	**84.9**	**94.9**	**68.0**	**97.1**
EDL	Worst	74.6	81.2	57.4	86.2
	Mean	79.6	89.8	57.3	92.9
	Best	82.6	**92.7**	61.5	95.1

a new upcoming date. If we audited the high uncertainty samples, we could get an $F_{1_{audit}}$ of 97.2%, which is close to what we obtained in the ideal case of training and testing on the same date, with a slight difference of 0.7%.

In this case, entropy results from a single inference run produced high variance, with the worst and best $F_{1_{audit}}$ being 83.5% and 94.9%. The mean value was similar to the best-performing method. However, such variability is undesirable in an operational application. Similar results were obtained with EDL, with slightly lower mean values compared to a Single run, consistently with training and testing on the same date. Unlike previous works, we also assessed the variability of EDL for multiple repetition runs. As in Single run, EDL presented high variability, with the worst and best $F_{1_{audit}}$ being 86.2% and 95.1%. Such variability may indicate that a single point estimate was not sufficient to estimate the underlying distribution of the predicted probabilities in this case, which is the assumption in EDL [61, 109]. Such an assumption was made given a sufficient amount of training data, which indicates that increasing the training dataset with images from additional dates or regions may improve EDL outcomes.

Figure 29 presents results for different uncertainty thresholds in the PA site in the more operational case of training with a past date and testing on an upcoming date. In this case, $F_{1_{low}}$ and $F_{1_{audit}}$ increased monotonically when increasing AA, as expected. Compared to F_1 before applying the uncertainty methodology, $F_{1_{low}}$ and $F_{1_{audit}}$ were significantly higher even for small values of AA.

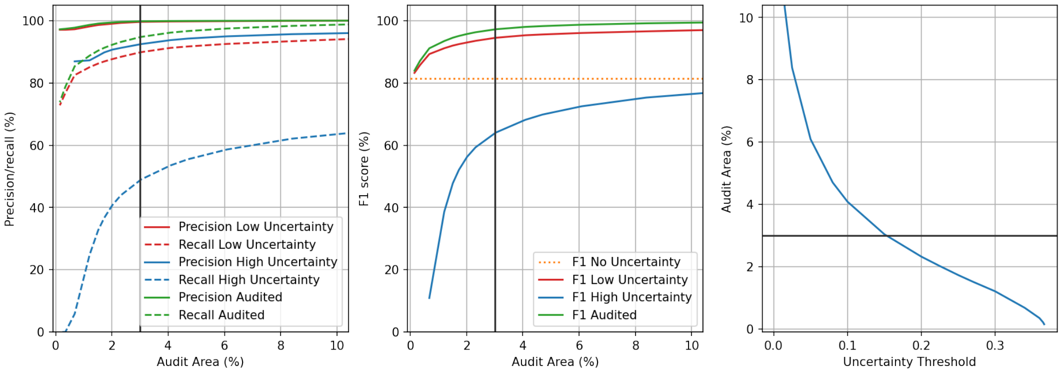


Figure 29: Classification metrics for multiple uncertainty threshold values in PA site. Training in [2017, 2018] and testing in [2018, 2019]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy. A sample AA threshold of 3% is highlighted in gray.

Figures 30 and 31 present classification results and the corresponding uncertainty map for PA site when training in an earlier date for the entire study area. Most of the deforestation polygons were correctly classified. Error areas usually correspond to high uncertainty values. A high uncertainty accumulation in the lower right area which does not correspond to error areas, corresponded with an occlusion in the T_0 image. Further qualitative analysis will be presented in Section 5.3.7.

5.3.2 Uncertainty Estimation Results in MT Site

Table 10 presents classification results for MT site in the traditional scenario, where the network is trained and tested on the same date. As in the PA site, the best multiple-output uncertainty method was Ensemble, with an increase of 2.2% in F_1 compared to MCD. In this case, the best uncertainty metric was the predictive variance in terms of $F_{1_{low}}$ and $F_{1_{audit}}$. For the Ensemble method and the predictive variance metric, applying the proposed uncertainty methodology, for an AA of 3%, the $F_{1_{low}}$ improved to 91.7%, while $F_{1_{high}}$ produced a significantly lower value, indicating that the methodology was able to separate between predictions whose results we can trust and predictions that the network does not know, similarly to PA. If we audited the samples with high uncertainty, we could get a $F_{1_{audit}}$ of 94.8% with minimal auditing effort. Consistently with the PA site, entropy from a single inference run produced results with high variance, with the worst case scenario being 7% lower compared to MCD and the best outcome 0.2% higher compared to Ensemble in terms of $F_{1_{low}}$. Such a result suggests that using a

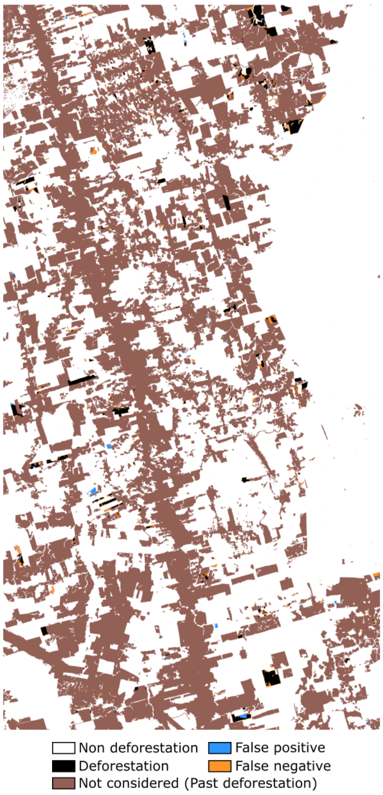


Figure 30: Classification results for PA site. Training in [2017, 2018] and testing in [2018, 2019]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

single inference run may be less reliable and robust compared to the MCD and Ensemble approaches. Consistently with the PA site, EDL performed similarly to Single run with slight decreases in all metrics. Reasons may also be related to the large class imbalance, which was not addressed in [18]. EDL again presented high variance among repetition runs, with a standard deviation in the same range as single run. As explained in the PA analysis, this may indicate that more than a single point estimate may be needed in EDL.

Table 11 presents results in a more challenging scenario, which is closer to a real operational setting, for the MT site. As in PA, we trained on an image pair from past dates $[T_{-1}, T_0] = [2018, 2019]$ and tested on a new upcoming date $[T_{-1}, T_0] = [2019, 2020]$. Results were similar to previous experiments, with Ensemble producing the best results compared to MCD. The best-performing uncertainty metric was predictive entropy. In this operational setting, we obtained an F_1 of 81.0% with the best-performing approach, which represented a minor drop of 0.4% compared to the ideal case of training and testing case on the same date pair.

Using our proposed methodology to aid the auditing process, for an AA of 3%, we obtained $F_{1_{low}}$ of 93.8%, with a much lower $F_{1_{high}}$, which reinforces

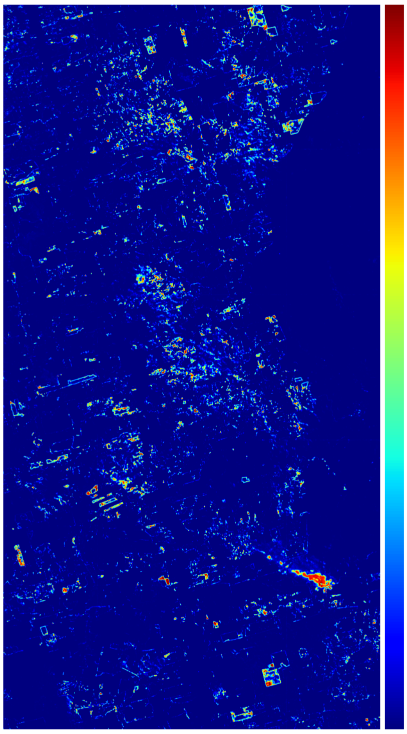


Figure 31: Uncertainty results for PA site. Training in [2017, 2018] and testing in [2018, 2019]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

Table 10: Results for MT site obtained by training and testing on a pair of images from 2019/2020 with $AA = 3\%$. Values are presented in percentages (%). First, second, and third best results in each column are highlighted in **red**, **blue** and **green** respectively

Method	**Uncertainty Metric**	F_1	$F_{1_{low}}$	$F_{1_{high}}$	$F_{1_{audit}}$
MCD	Predictive Entropy	79.2	91.7	57.7	94.8
	Predictive Variance		92.2	64.0	95.9
	Mutual Information		89.3	43.3	92.0
	Expected KL		87.8	39.4	90.4
Ensemble	Predictive Entropy	**81.4**	**94.1**	63.8	**96.6**
	Predictive Variance		**95.1**	**69.0**	**97.7**
	Mutual Information		93.5	61.4	96.0
	Expected KL		93.1	60.3	95.6
Single run	Entropy (Worst)	78.6	84.7	**66.4**	89.8
	Entropy (Mean)	79.9	91.9	62.4	95.0
	Entropy (Best)	**83.0**	**95.3**	66.2	**97.3**
EDL	Worst	78.2	81.8	**66.6**	85.7
	Mean	78.4	85.5	60.3	89.4
	Best	**84.4**	91.0	72.4	94.2

our initial hypothesis that uncertainty estimation can help us separate the samples whose results we can trust from the samples the network does not know if they are correct. After auditing the high uncertainty samples, we got

Table 11: Results for MT site obtained by training on a pair of images from 2018/2019 and testing on a pair of images from 2019/2020 with $AA = 3\%$. Values are presented in percentages (%). First, second, and third best results in each column are highlighted in **red**, **blue** and **green** respectively

Method	**Uncertainty Metric**	F_1	$F_{1_{low}}$	$F_{1_{high}}$	$F_{1_{audit}}$
MCD	Predictive Entropy	77.4	90.2	54.4	94.3
	Predictive Variance		83.7	**70.4**	92.4
	Mutual Information		81.8	39.9	83.8
	Expected KL		79.9	32.5	81.3
Ensemble	Predictive Entropy	**81.0**	**93.8**	**63.0**	**96.7**
	Predictive Variance		**92.6**	**71.0**	**96.9**
	Mutual Information		92.3	59.7	95.4
	Expected KL		91.7	56.6	94.8
Single run	Entropy (Worst)	72.9	87.0	58.7	94.3
	Entropy (Mean)	78.8	90.4	57.3	94.4
	Entropy (Best)	**78.9**	**92.8**	55.6	**96.1**
EDL	Worst	72.4	75.1	55.9	79.2
	Mean	76.7	82.0	56.3	85.9
	Best	**80.6**	87.2	57.7	89.9

an $F_{1_{audit}}$ of 96.7%, which is close to the ideal case where we trained and tested on the same date pair. These results were consistent with the ones from PA site.

In the more operational setting of training and testing with different dates, results for entropy from a single inference run were worse compared to MCD and Ensemble methods, even for the best-performing case. This reinforces the hypothesis that the robustness of multiple-inference approaches was more critical in the operational case.

In Figures 32 and 33 we present the classification metrics for multiple uncertainty threshold values when training and testing on the same date and when training and testing on different dates, respectively. These results correspond to the best-performing approach in each case. We present F_1 before applying the uncertainty methodology in yellow for comparison. We observe a similar behavior to PA site, with $F_{1_{low}}$ and $F_{1_{audit}}$ increasing when increasing AA. The auditor might select a low AA such as 3% and still get high F_1 values, or use a more conservative approach and select a higher AA with correspondingly higher F_1 metrics.

Figures 34 and 35 present classification results and the uncertainty map for MT site in the entire study area, respectively. Across the entire site, the uncertainty map presented high uncertainty values in the error areas. Two main regions with error areas can be observed in the upper right corner and the image center. Correspondingly, both cases represent high uncertainty regions.

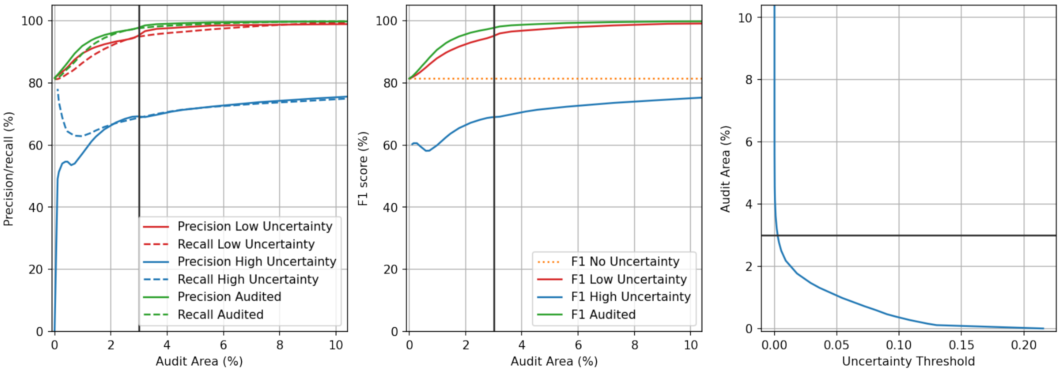


Figure 32: Classification metrics for multiple uncertainty threshold values in MT site. Training and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Variance. A sample AA threshold of 3% is highlighted in gray.

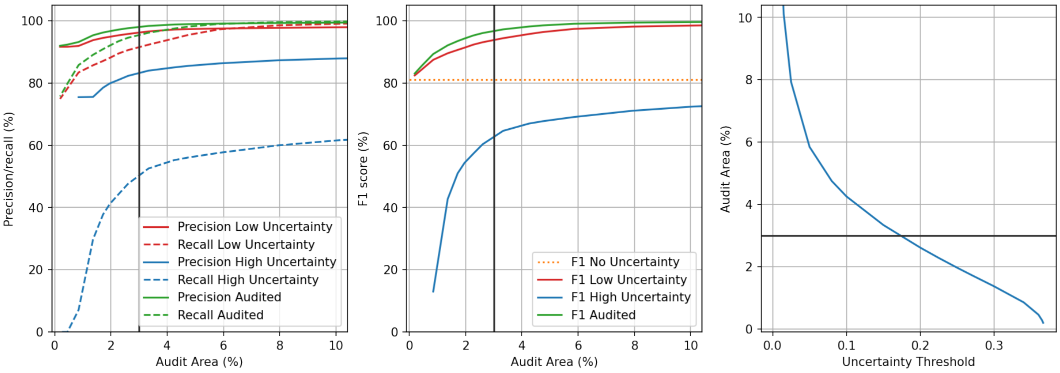


Figure 33: Classification metrics for multiple uncertainty threshold values in MT site. Training in [2018, 2019] and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy. A sample AA threshold of 3% is highlighted in gray.

5.3.3 Uncertainty Estimation Results in MS Site

The Cerrado areas were characterized for having a much lower percentage of deforestation polygons compared to the Amazon sites. For comparison, the Amazon sites had a percentage of deforestation pixels of 1.1% and 1.3% for the target date T_0 in the PA and MT sites, while the Cerrado sites had a percentage of deforestation pixels of 0.1% and 0.6% for the target date in the MS and PI sites.. Despite that, results in MS site were similar to the Amazon regions. Table 12 presents results for all uncertainty methods when training and testing on the same date pair (2019/2020). In general, all methods presented similar results, with Ensemble and Single run producing the highest outcomes. Particularly, Single run produced slightly better outcomes than Ensemble for the best case. However, on average, Ensemble produced the best results with values of 96.9% $F1_{low}$ and 97.9% $F1_{audit}$, which are higher than the current

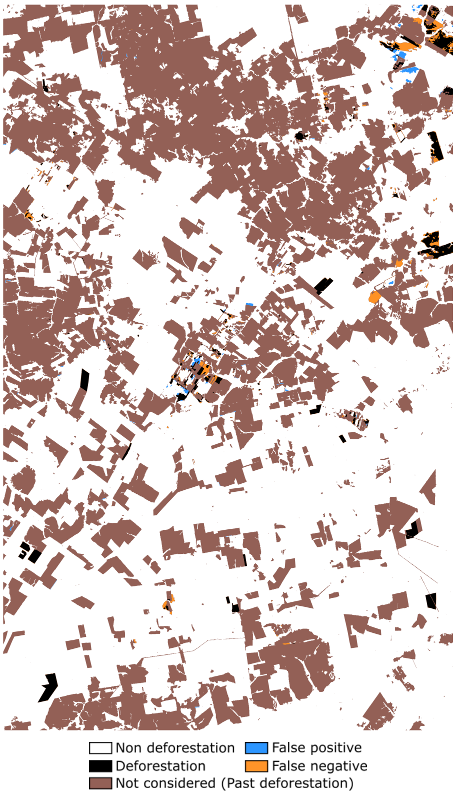


Figure 34: Classification results for MT site. Training in [2018, 2019] and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

metrics obtained by manual inspection. Consistently with previous results, EDL performed similarly to Single run with slight decreases.

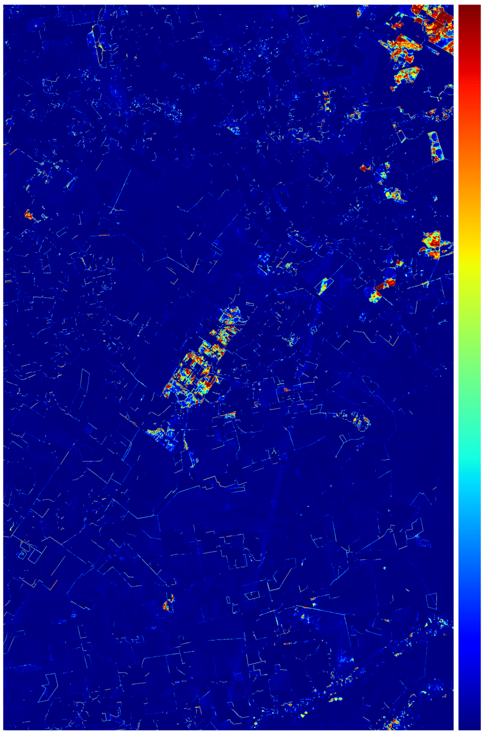


Figure 35: Uncertainty results for MT site. Training in [2018, 2019] and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

Table 12: Results for MS site obtained by training and testing on a pair of images from 2019/2020 with $AA = 3\%$. Values are presented in percentages (%). First, second, and third best results in each column are highlighted in **red**, **blue** and **green** respectively.

Method	**Uncertainty Metric**	F_1	$F1_{low}$	$F1_{high}$	$F1_{audit}$
MCD	Predictive Entropy	84.1	95.3	61.0	96.9
	Predictive Variance		90.9	**77.5**	96.2
	Mutual Information		88.2	34.0	89.3
	Expected KL		87.2	26.7	88.0
Ensemble	Predictive Entropy	85.8	**96.9**	62.3	**97.9**
	Predictive Variance		95.6	**72.3**	**97.5**
	Mutual Information		92.2	31.3	93.1
	Expected KL		90.9	25.6	91.7
Single run	Entropy (Worst)	83.4	94.1	66.7	96.5
	Entropy (Mean)	86.2	**95.9**	68.7	**97.5**
	Entropy (Best)	**87.4**	**97.0**	**74.7**	**98.3**
EDL	Worst	80.8	85.9	49.2	87.2
	Mean	**86.9**	92.0	56.1	92.8
	Best	**91.0**	94.5	69.1	95.2

Table 13 presents results for the more realistic scenario of training with an earlier date pair (2018/2019) and testing with an upcoming date pair (2019/2020). In this case, results also surpassed the desired metrics for an operational implementation. In this more challenging scenario, Ensemble produced the best $F1_{low}$ and $F1_{audit}$ outcomes in general, which indicates the importance of multiple training runs when testing in a scenario that is further from the training distribution. Specifically, the MI and Predictive Variance metrics produced the best $F1_{low}$ and $F1_{audit}$ respectively, indicating the importance of measuring model disagreement among the ensemble members. In terms of uncertainty metrics, EDL produced very similar results to Single run. However, EDL produced higher average and best F_1 values compared to Single run and the best F_1 overall. This may indicate the potential of EDL for future research works. Consistently with previous results, Single run and EDL presented high variability in their outcomes, indicating that multiple-output approaches may be more reliable and robust.

Table 13: Results for MS site obtained by training on a pair of images from 2018/2019 and testing on a pair of images from 2019/2020 with $AA = 3\%$. Values are presented in percentages (%). First, second, and third best results in each column are highlighted in **red**, **blue** and **green** respectively.

Method	**Uncertainty Metric**	F_1	$F1_{low}$	$F1_{high}$	$F1_{audit}$
MCD	Predictive Entropy	**83.2**	91.8	**70.0**	95.4
	Predictive Variance		85.2	**79.3**	94.4
	Mutual Information		89.9	44.4	92.2
	Expected KL		89.1	34.6	90.8
Ensemble	Predictive Entropy	**80.4**	**94.0**	62.5	**96.5**
	Predictive Variance		93.0	**69.1**	**96.7**
	Mutual Information		**94.9**	39.3	96.2
	Expected KL		**94.4**	29.3	95.5
Single run	Entropy (Worst)	80.1	88.5	**69.1**	93.4
	Entropy (Mean)	78.2	90.6	62.4	94.4
	Entropy (Best)	78.2	93.5	63.3	**96.6**
EDL	Worst	71.3	88.5	**69.1**	93.4
	Mean	78.9	86.2	52.2	87.9
	Best	**87.4**	91.7	65.2	93.0

Figures 36 and 37 present results for multiple AA values for the best-performing approach, while training with a current or an earlier date pair respectively. In both cases, the largest uncertainty-related improvements can be already achieved with AA values as low as 3%. The auditor may choose to increase the AA threshold for added slight improvements.

Figures 38 and 39 present classification results and the uncertainty map for the entire study area. The amount of deforestation polygons in the MS site was much lower compared to the Amazon sites. For comparison, the percentage of deforestation pixels in the image was 0.1%, compared to 1.1% and 1.3% in the Amazon sites of PA and MT. Due to the uncertainty metric being Mutual

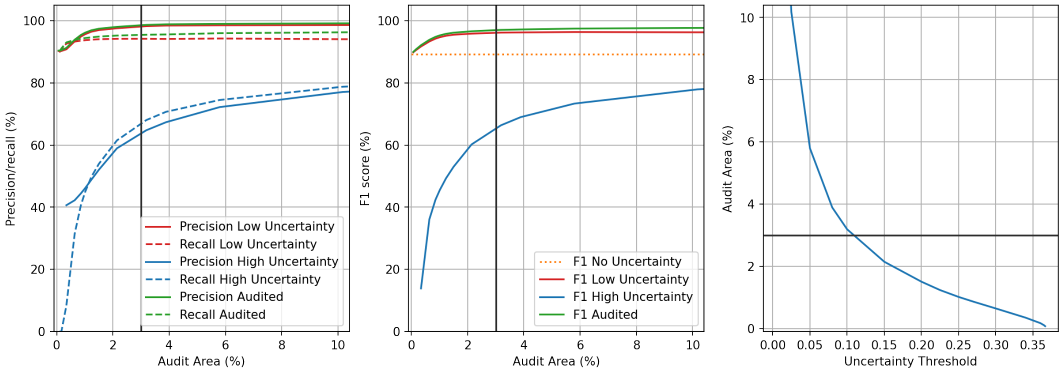


Figure 36: Classification metrics for multiple uncertainty threshold values in MS site. Training and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy. A sample AA threshold of 3% is highlighted in gray.

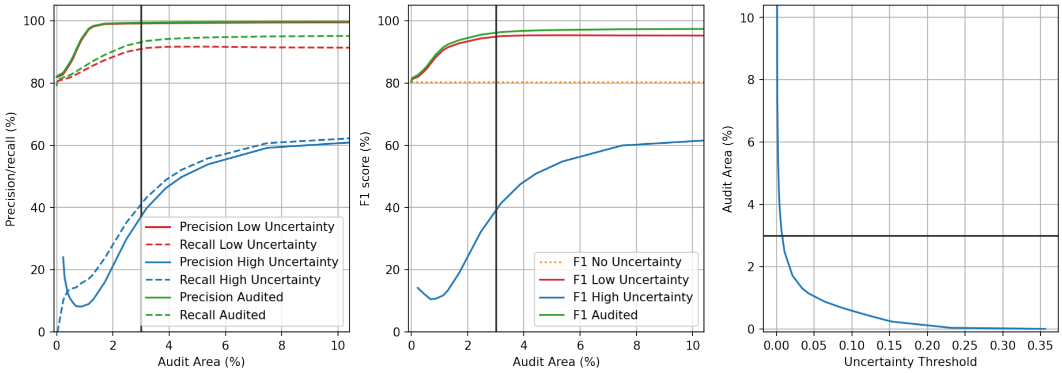


Figure 37: Classification metrics for multiple uncertainty threshold values in MS site. Training in [2018, 2019] and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Mutual Information. A sample AA threshold of 3% is highlighted in gray.

Information, its absolute value is generally lower compared to the predictive entropy-related results. Even so, error areas tend to present higher uncertainty values.

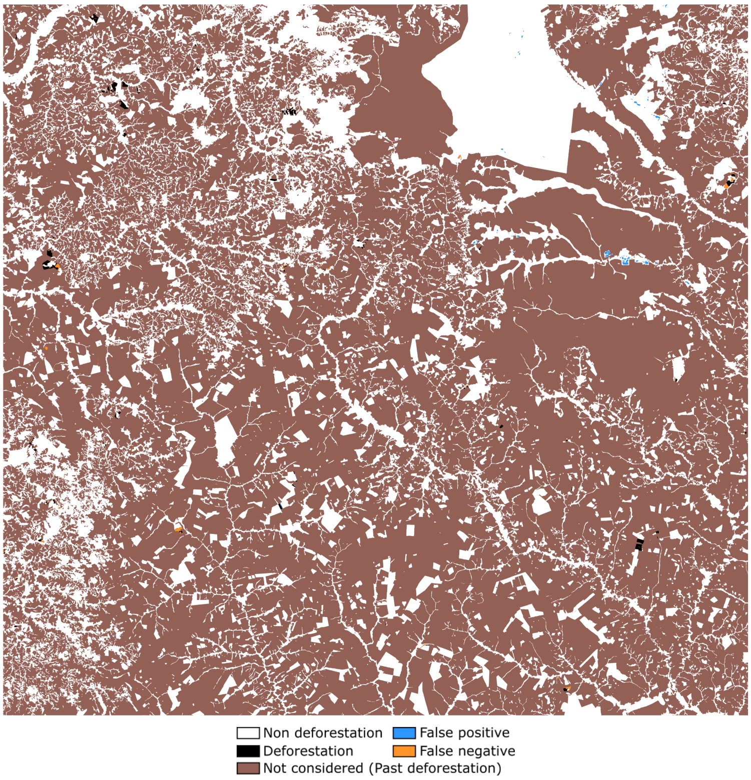


Figure 38: Classification results for MS site. Training in [2018, 2019] and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Mutual Information.

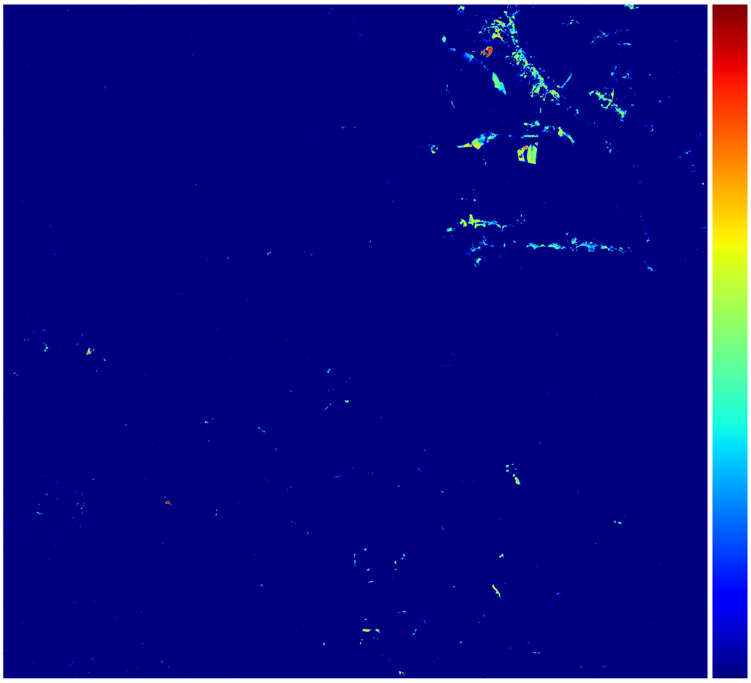


Figure 39: Uncertainty results for MS site. Training in [2018, 2019] and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Mutual Information.

5.3.4 Uncertainty Estimation Results in PI Site

Table 14 presents results for PI site when training and testing on the same date pair (2019/2020). Compared to the remaining sites, absolute performance was lower in the PI site. The highest F_1 was achieved by the best Single Run, with $F_1 = 77.9\%$. However, Single run presented high variability, with its worst result being 19.9% lower than the best result in terms of $F_{1_{low}}$. Likewise, high variability was obtained in EDL, although average metrics for EDL were higher than Single run, with F_1 and $F_{1_{low}}$ being 4% and 4.3% higher respectively.

Table 14: Results for PI site obtained by training and testing on a pair of images from 2019/2020 with $AA = 3\%$. Values are presented in percentages (%). First, second, and third best results in each column are highlighted in **red**, **blue** and **green** respectively

Method	Uncertainty Metric	F_1	$F_{1_{low}}$	$F_{1_{high}}$	$F_{1_{audit}}$
MCD	Predictive Entropy	**73.5**	83.5	30.6	86.0
	Predictive Variance		**87.3**	44.3	**90.2**
	Mutual Information		82.1	36.0	84.6
	Expected KL		80.8	35.3	83.2
Ensemble	Predictive Entropy	**77.5**	88.7	28.6	**90.5**
	Predictive Variance		**90.3**	**47.3**	**92.5**
	Mutual Information		85.1	**48.1**	87.6
	Expected KL		84.0	**49.6**	86.5
Single run	Entropy (Worst)	55.8	66.9	9.7	69.3
	Entropy (Mean)	68.9	77.2	25.9	79.5
	Entropy (Best)	**77.9**	**86.8**	35.5	88.6
EDL	Worst	53.0	62.0	13.2	63.7
	Mean	71.8	80.7	32.3	82.7
	Best	**77.5**	85.1	44.9	87.7

Without considering Single run and EDL due to their high variability, the best-performing method in terms of $F_{1_{low}}$ for $AA = 3\%$ was Ensemble with the Predictive Variance metric, closely followed by MCD with Predictive Variance. With $F_{1_{audit}} = 92.5\%$, its result still matches the PRODES accuracy from manual annotation, which was estimated as 87.1%. The auditor may increase the AA threshold for further accuracy improvements.

Table 15 presents results for the more realistic scenario of training with an earlier date pair (2018/2019) and testing with an upcoming date pair (2019/2020). As expected, absolute values are lower compared to the ideal scenario of training and testing on the same date pair, with a decrease of 14.3% for the best F_1 without considering Single run or EDL due to their high variability. The best-performing method in terms of $F_{1_{low}}$ was Ensemble with Predictive Variance, reaching $F_{1_{audit}} = 81.2\%$ which, different from the remaining study areas, is below the accuracy required by PRODES to replace manual annotation. This result is further discussed in Section 5.3.7.

Table 15: Results for PI site obtained by training on a pair of images from 2018/2019 and testing on a pair of images from 2019/2020 with $AA = 3\%$. Values are presented in percentages (%). First, second, and third best results in each column are highlighted in **red**, **blue** and green. respectively

Method	**Uncertainty Metric**	F_1	$F_{1_{low}}$	$F_{1_{high}}$	$F_{1_{audit}}$
MCD	Predictive Entropy	**61.7**	71.4	26.2	76.2
	Predictive Variance		**74.3**	**34.1**	**80.3**
	Mutual Information		69.8	27.3	74.6
	Expected KL		68.6	27.1	73.3
Ensemble	Predictive Entropy	**63.2**	**73.4**	25.3	**78.0**
	Predictive Variance		**75.8**	**33.1**	**81.2**
	Mutual Information		71.8	31.4	76.9
	Expected KL		70.3	31.9	75.3
Single run	Entropy (Worst)	51.5	59.6	17.8	64.5
	Entropy (Mean)	59.0	67.2	24.7	71.3
	Entropy (Best)	**63.1**	72.6	**32.2**	77.2
EDL	Worst	50.8	59.8	22.8	64.1
	Mean	55.9	64.4	21.9	67.5
	Best	58.7	69.8	20.8	72.8

Figures 40 and 41 present results for varying AA threshold values. In both cases, the auditor may select a larger AA value for improved outcomes. The desired PRODES accuracy (87.1% F1-score) would be attained using $AA = 8.2\%$ when training in an earlier date pair.

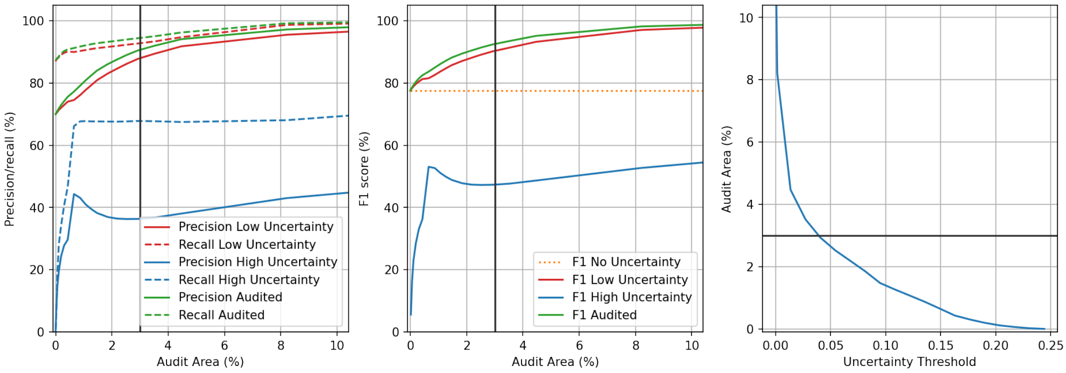


Figure 40: Classification metrics for multiple uncertainty threshold values in PI site. Training and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Variance. A sample AA threshold of 3% is highlighted in gray.

Figures 42 and 43 present classification results and the uncertainty map for the entire study area in the PI site. False positive error areas (Blue in the classification map) presented high uncertainty values. For example, a false

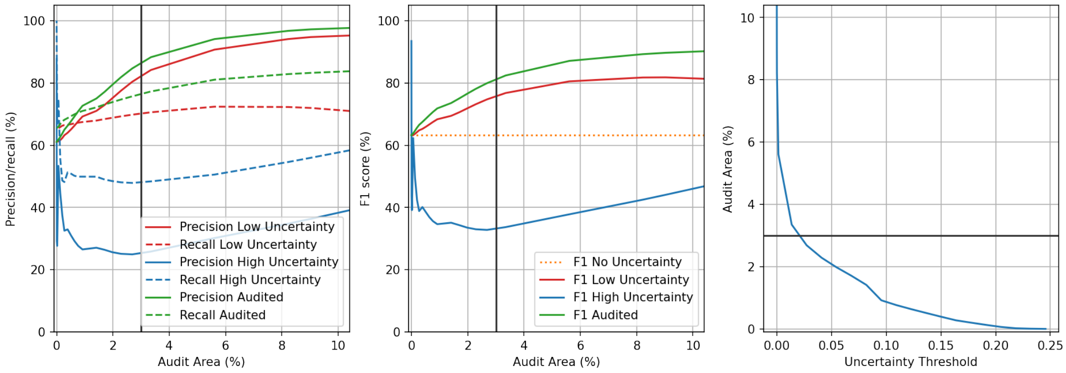


Figure 41: Classification metrics for multiple uncertainty threshold values in PI site. Training in [2018, 2019] and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Variance. A sample AA threshold of 3% is highlighted in gray.

positive error area in the lower center extreme presents high uncertainty values. However, different from previous sites, many of the false negative outcomes (Orange in the classification map) did not present high uncertainty.

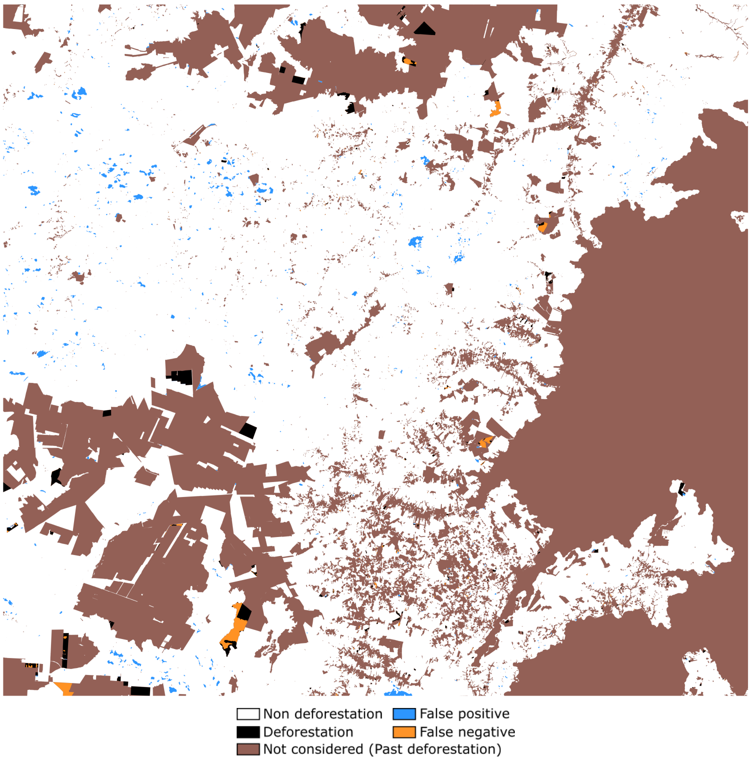


Figure 42: Classification results for PI site. Training in [2018, 2019] and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Variance.

5.3.5

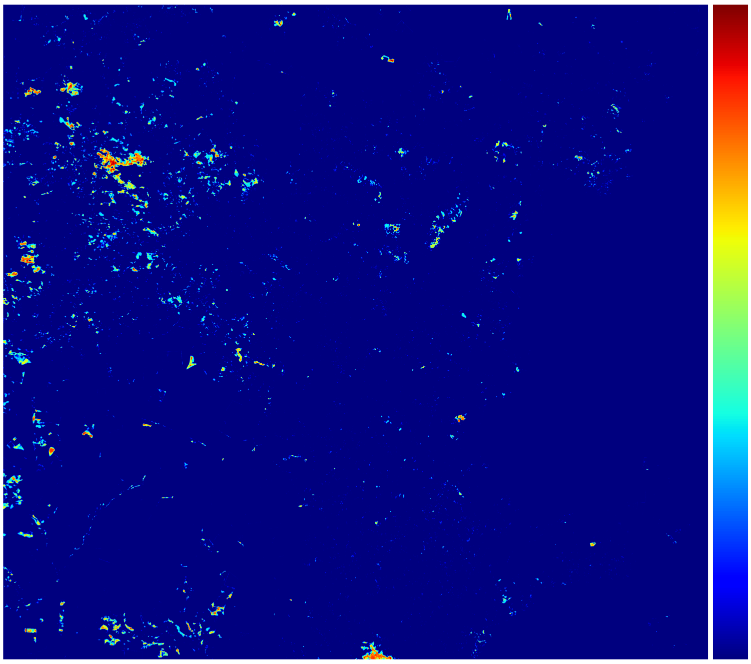


Figure 43: Uncertainty results for PI site. Training in [2018, 2019] and testing in [2019, 2020]. Uncertainty method: Ensemble. Uncertainty metric: Predictive Variance.

Computation time analysis

The proposed method is expected to be applied on large scale to entire Brazilian biomes such as the Amazon and Cerrado biomes. For such large-scale applications, computational complexity at training and inference is important when deciding which uncertainty estimation method to apply. Hence, this section presents a comparison of training and inference times for each uncertainty estimation method. Measurements are based on the more realistic case of training on an earlier date pair. Table 16 presents training times for the PA, MT, MS, and PI sites for the MCD, Ensemble, Single Run, and EDL uncertainty estimation methods.

Table 16: Training times for the assessed uncertainty estimation methods in the assessed study areas. Values are presented in minutes. Results training on a pair of images from an earlier date.

Method / Site	**PA**	**MT**	**MS**	**PI**
MCD	12.5	8.2	3.6	12.4
Ensemble	124.7	81.5	35.9	124.4
Single run	12.5	8.2	3.6	12.4
Evidential	17.7	9.5	5.0	15.6

Training times are presented as an average value over $n = 10$ repetitions for MCD, Single run, and Evidential methods. As expected, Ensemble resulted in nearly n times larger training times compared to the remaining methods, representing a computational disadvantage despite having presented the best accuracy outcomes. Nevertheless, it is essential to emphasize that, in contrast to inference, the training phase occurs only once. Therefore, the additional time invested in training ensembles may not pose a significant concern in practical operational applications. On the flip side, employing multiple trained models during inference could potentially lead to increased utilization of RAM or GPU memory, depending on the specific implementation. In general, the remaining single-training methods presented similar average training times in each study area, with EDL having slightly higher training times due to converging on a later epoch. The differences in training times among study areas are related to the number of overlapping training patches extracted from each site, which is correlated to the amount and size of deforestation polygons in the ground truth mask due to having selected only patches with a minimum percentage of the deforestation class during training. The duration of training may also depend on the complexity of the dataset at each site, which may influence the resulting early stopping epoch. For completeness, the amount of training and validation patches in each site are presented in Table 17. The lowest training times obtained in the MS site correspond with the lowest number of training patches.

Table 17: Number of training and validation patches for each study area. Results training on a pair of images from an earlier date.

N. of patches / Site	PA	MT	MS	PI
Training	3369	2962	1539	5495
Validation	1184	623	328	1422

Table 18: Inference times for the assessed uncertainty estimation methods in the assessed study areas. Values are presented in minutes.

Method / Site	PA	MT	MS	PI
MCD	11.0	11.0	24.2	29.8
Ensemble	11.1	11.3	23.6	23.4
Single Run	1.0	1.1	2.1	2.3
EDL	1.1	1.1	2.4	2.3

Table 18 presents inference times for the PA, MT, MS, and PI sites and all the assessed uncertainty estimation methods. As expected, the multiple inference approaches MCD and Ensemble took approximately n times more time to obtain inference compared to the single inference approaches Single Run and EDL, where n was the number of inferences, and it was equal to 10 in our experiments. This difference in execution time motivates further research in the EDL method, such as exploring variations as in [77, 83], with the objective of improving its performance so that it reaches the same accuracy outcomes as the best performing method in this study, namely the Ensemble approach, and reducing the computational complexity in an operational scenario. Due to the larger spatial extension, the Cerrado sites (MS and PI) presented larger inference times compared to the Amazon sites (PA and MT).

5.3.6 Summary of Uncertainty Estimation Results

This subsection presents a summary of the outcomes obtained in the previous subsections as well as providing advise on which uncertainty method to use for deforestation detection applications in an operational scenario.

Regarding accuracy, the best results were consistently obtained with Ensemble in both the Amazon and Cerrado biomes. Ensemble produced the best outcomes for the base metric F_1 and the uncertainty-related metrics $F_{1_{low}}$ and $F_{1_{audit}}$. Such superiority may be related to an enhanced generalization due to multiple training runs resulting in multiple locally optimal solutions instead of a single one within the parameter space (Figure 13). As presented in Subsection 5.3.5, Ensembles are computationally expensive, resulting in n times larger training and inference times. Such increased times may prove impractical in a real-world, large-scale operational context, primarily due to economic and time constraints. However, good results were achieved with a low amount of training and inference runs ($n = 10$), which may not represent such a significant overhead. Such computational costs may be overcome by further exploring more recent single-outcome approaches such as EDL in future works. Considering that the most critical factor for the operational implementation of a semi-automatic deforestation detection system is related to the high accuracy requirements currently achieved by manual annotators, the recommended uncertainty estimation method, according to the results in this work, is Ensemble. Regarding the uncertainty metric, the best choice varied across study areas. Out of 8 total experiments presented (training in a current and an earlier pair of dates for 4 study areas), 50% of the times, Predictive Entropy produced the best result, followed by Predictive Variance (37.5%) and MI (12.5%). Particularly, in the more operational scenario of training with an earlier date, two sites obtained the best result with predictive entropy, corresponding to the Amazon sites. The remaining Cerrado sites obtained the best results with MI and Predictive Variance. However, in both cases, the winning metric was closely followed by Predictive Entropy, with differences of 0.9% and 2.4% for MS and PI sites, respectively. Thus, across experiments and sites, the most consistent uncertainty metric was the Predictive Entropy, which is recommended in an operational scenario. In the following section, uncertainty interpretation is performed on the recommended uncertainty method: Ensembles with Predictive Entropy as an uncertainty metric.

5.3.7 Uncertainty Interpretation

The results shown in the previous section indicate that it is possible to increase the productivity of the auditing process by focusing on areas characterized by high uncertainty levels. In order to further analyze those results, a team of PRODES auditors examined the uncertainty maps and tried to associate them with specific forest characteristics or ongoing processes that might have caused the high uncertainty levels.

One of the hypotheses explored in this analysis is the potential of high uncertainty as an early warning sign for ongoing deforestation processes. This hypothesis pertains to regions experiencing prolonged deforestation over multiple years, becoming identified as deforested by PRODES only at a later stage when detection certainty is higher. Such areas may be difficult to classify and therefore may produce high uncertainty, which may in turn be correlated to early warning areas. To examine this hypothesis, the auditors adopted deforestation reports referring to the year after the detection date assumed in our previous experimental analysis. Henceforth, the year after the detection date is denoted as T_1, while T_0 and T_{-1} refer to the detection date and the prior year. The analysis was limited to the outcomes of the classifier trained on a previous pair of dates, i.e., T_{-2} and T_{-1}.

5.3.7.1 PA site

Figures 44 and 45 show image snippets covering parts of the PA site. The first rows of the figures display the input Landsat 8 images (Vegetation analysis composition with SWIR 1, NIR and red bands [110]), which are the same ones used for the PRODES report. The acquisition dates of such images are the so-called PRODES dates.

Deforested polygons for date T_0 are highlighted with red borders, while those for the future date T_1 are highlighted with yellow borders. In both cases, deforested polygons correspond to the PRODES ground truth reference. The black areas represent past deforestation (i.e., which occurred before T_0) and are irrelevant to the analysis. The second rows of the figures show the classification results for the detection date T_0. Specifically, they show a map depicting the predicted probabilities for the deforestation class on the left, followed by the prediction results on the center (white and black for non-deforestation and deforestation, blue and orange for false positive and true positive errors), and the uncertainty map on the right. Places of interest are indicated with uppercase letters within each snippet.

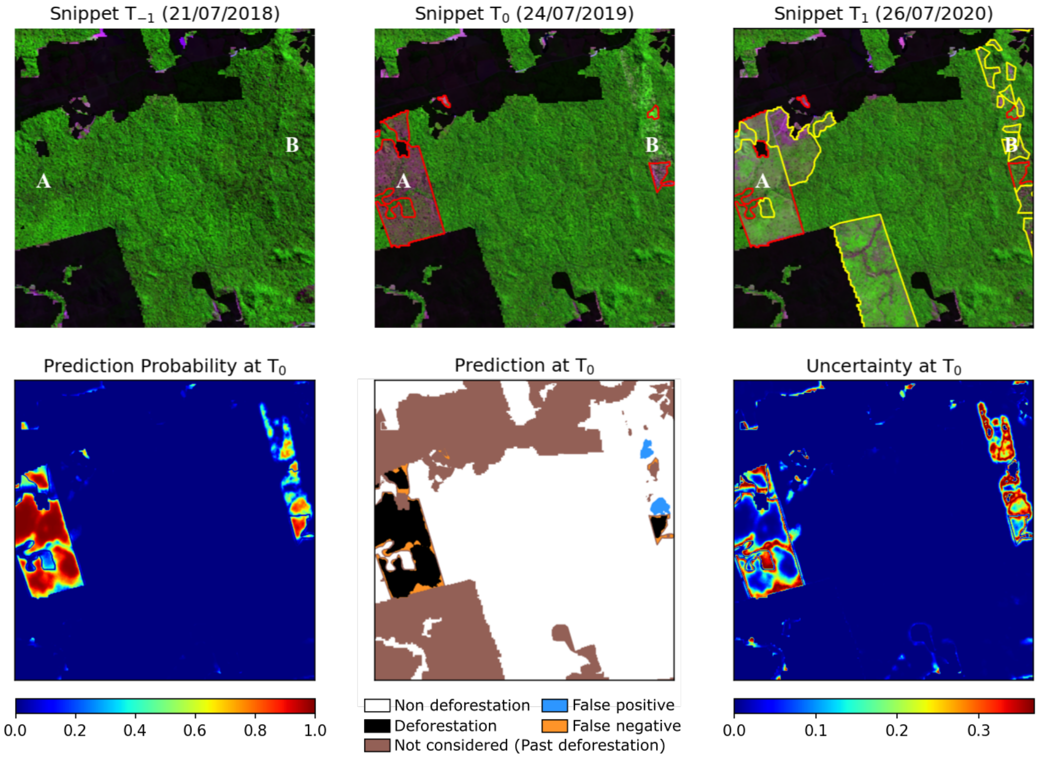


Figure 44: Qualitative results for PA site. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

As expected, misclassified pixels are typically found in areas with higher uncertainty values, whereas accurately classified pixels are associated with lower uncertainty values. Additionally, the borders of deforestation polygons exhibit high uncertainty. That observation is not unexpected, as at the polygon borders occur the transition between different land cover classes.

Place of interest A (Figure 44): On T_{-1} (2018), the region was covered by intact forest without visible signs of degradation. By T_0 (2019) most parts of the area suffered complete vegetation removal (smooth texture and reddish color), exposing the soil on the detection date. However, certain portions of the area kept clusters of preserved trees and riparian forests. The deforested regions underwent a complete removal of arboreal vegetation. In those regions, the classification was associated with low uncertainty. The automatic classification results closely resemble those reported by PRODES. Moreover, the areas encompassing remaining riparian forests with preserved canopies displayed low uncertainty and were not regarded as deforested in the PRODES report. However, some clusters of remaining trees exhibited high

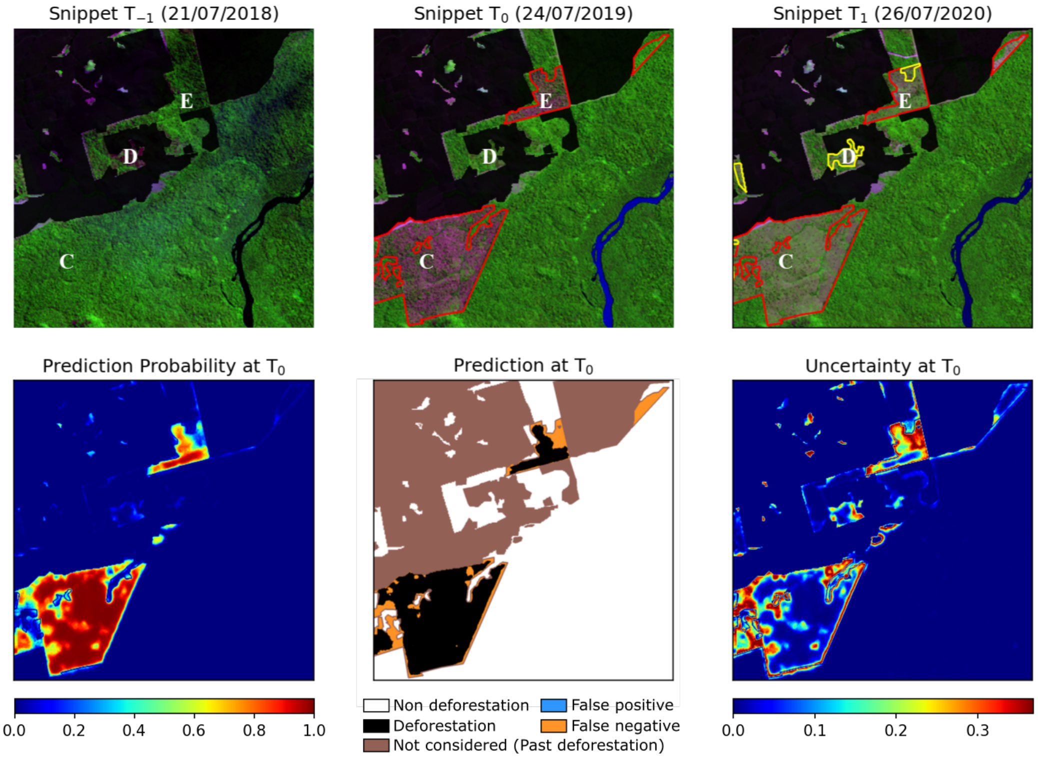


Figure 45: Qualitative results for PA site. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

uncertainty, indicating their degraded nature. In those cases, there were minor discrepancies between the classification results of this study and those found in the PRODES report.

Place of interest B (Figure 44): The image captured on T_{-1} (2018) shows no signs of degradation at place B. On the detection date, T_0 (2019), however, we can identify stains with advanced degradation characterized by a mix of reflectance typical of tree remnants and herbaceous (pasture), which may have given rise to the high observed uncertainty values. There were false positives in two areas near place B. We also verified that PRODES regarded most of the degraded areas on T_0, for which high uncertainty was observed, as deforestation in the following year (T_1). Once again, the high uncertainty was associated with an ongoing degradation process. The smooth textured polygon below place B indicates complete vegetation removal, with exposed soil on the detection date (T_0). Uncertainty was low at that location, and the detection results were consistent with the PRODES report.

Place of interest C (Figure 45): On the T_{-1} (2018) image, forest areas characterized by a dark green hue and rough texture can be observed, indicating an absence of apparent degradation. By the detection date T_0 (2019), vegetation was absent, revealing exposed soil over most of the polygon, represented by a smooth texture. The automatic classification exhibited high confidence at that location, as evidenced by the low uncertainty. Remarkably, the detection outcome was closely aligned with the findings reported by PRODES. Certain regions containing remnants of riparian forests, characterized by preserved canopies, displayed low uncertainty and were identified as deforestation by neither PRODES nor the network. In contrast, areas with high uncertainty within that location were degraded, encompassing clusters of remaining trees. Minor discrepancies emerged between the automatic classification and PRODES results in those areas.

Place of interest D (Figure 45): The image captured on T_0 (2019) contains degraded forest areas accompanied by adjacent pastures and signs of recent fire occurrence. However, no openings in the canopy can be observed in the image. Such a configuration may have led to the medium to high uncertainty values in some stretches. The PRODES report for the subsequent year T_1 (2020) classified that area as "deforestation due to progressive degradation." In this case, the high uncertainty served as an indication of ongoing deforestation processes.

Place of interest E (Figure 45): In the earlier image (T_{-1}), we observe a forest area with no signs of degradation. By the detection date T_0, a smooth texture indicates complete vegetation removal, accompanied by exposed soil. The automatic classification predictions in those regions exhibited low uncertainty and were closely aligned with the findings reported by PRODES. Nevertheless, some areas with clusters of preserved trees presented high uncertainty due to the presence of both tree individuals and exposed soil. The automatic classification did not detect deforestation in those small areas, contrary to PRODES. That is consistent with the hypothesis that areas characterized by high uncertainty corresponded to the error-prone classification results.

5.3.7.2 MT site

Figures 46 and 47 present qualitative results for the MT site. As in the PA figures, the first row presents Landsat 8 images for T_{-1}, T_0, and T_1, with black areas representing past deforestation areas. Deforestation polygons in T_0 and T_1, according to PRODES, are highlighted with red and yellow edges, respectively. The second row presents the automatic classification outcomes and uncertainty estimates for T_0. In the classification predictions (second row, second column), white and black represent correctly classified non-deforestation and deforestation classes, while blue and orange represent false positive and true positive errors.

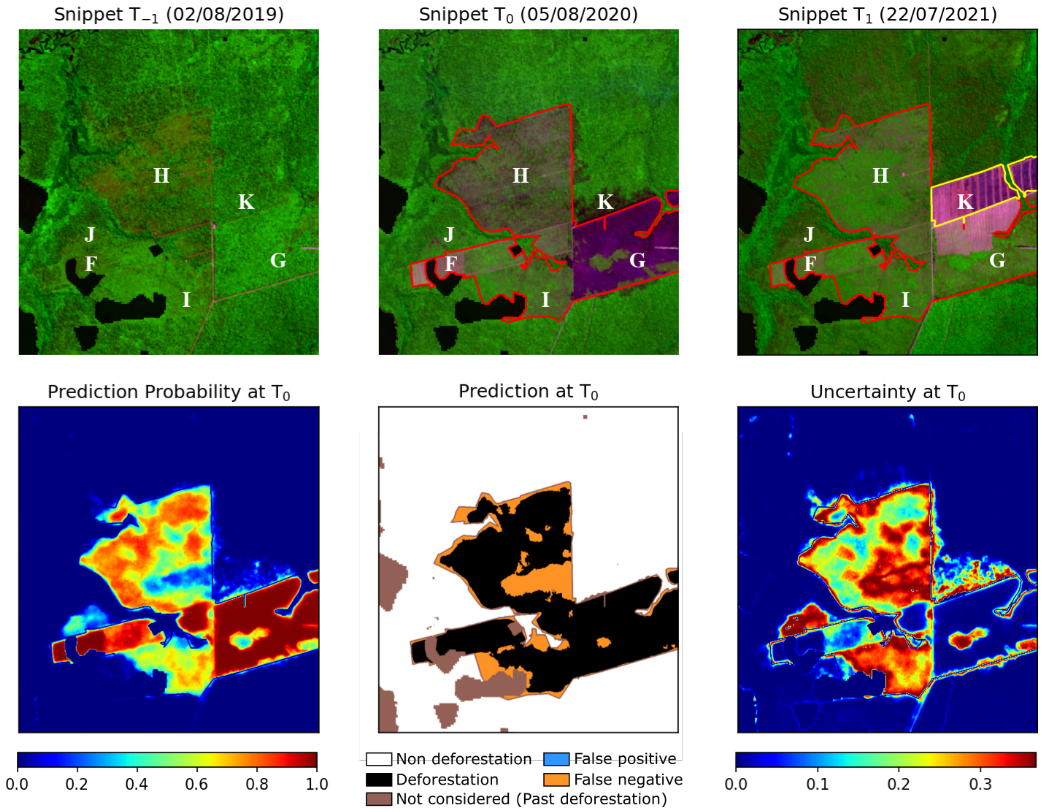

Snippet T$_{-1}$ (02/08/2019)
Snippet T$_0$ (05/08/2020)
Snippet T$_1$ (22/07/2021)
H
K
J
F
G
I
Prediction Probability at T$_0$
Prediction at T$_0$
Uncertainty at T$_0$
0.0
0.2
0.4
0.6
0.8
1.0
Non deforestation
False positive
Deforestation
False negative
Not considered (Past deforestation)
0.0
0.1
0.2
0.3

Figure 46: Qualitative results for MT site. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

Figure 46 shows a region with many false negatives. Accordingly, the respective uncertainty map shows high values for the misclassified locations and significantly lower values for correctly classified ones. On the other hand, the region shown in Figure 47 was almost perfectly classified, which is consistent with the low uncertainty values found. As expected, areas at the border of the deforestation polygons presented high uncertainty values.

Place of interest F (Figure 46): We identify slight signs of degradation on the T_{-1} image snippet (2019). In contrast, the vegetation was entirely cleared on T_0 (2020), with the soil becoming apparent. That must have led to the observed low uncertainty values. The detection outcome was closely consistent with the PRODES report for that date. Unlike what we observed in other spots, in this case the low-intensity degradation in the year preceding the detection did not contribute to heighten uncertainty.

Place of interest G (Figure 46): On the T_{-1} (2019) image, one can observe degraded forest associated with pastureland, rendering apparent canopy openings indiscernible. The lighter green color and smoother texture indicate degradation compared to non-degraded forests. The degradation persisted on T_0 (2020), with signs of fire intentionally set to facilitate complete clearing. The smooth texture indicates the loss of canopy cover. Consequently, PRODES classified this area as "deforestation due to progressive degradation." That type of deforestation occurs when selective wood removal and successive fires over several years lead to complete canopy opening, resulting in structural collapse, loss of the forest's ecological functions, and capacity for self-regeneration [111]. The uncertainty associated with that area was generally low, as the introduction of fire for complete clearing is easily detectable. Degradation without apparent canopy openings in the year preceding detection (2019) did not result in heightened uncertainty. For the most part, detection results were consistent with PRODES findings. However, some unburned areas within the polygon exhibited high uncertainty and were misclassified. Those particular areas hold significant interest for expert auditors because the canopy loss is not so obvious in the Landsat image used for the PRODES report.

Place of interest H (Figure 46): Covers areas of degraded forest with apparent canopy openings and exposed soil on T_{-1} (2019). In the following year, T_0 (2020), degradation reached its final stage, leading PRODES to classify the region as "deforestation due to progressive degradation." The automatic classification result showed high uncertainty, possibly because individual trees often occur in such cases, as observed in that polygon (indicated by a rough texture). The combination of degradation with exposed soil in the previous year and deforestation characterized by predominantly exposed soil, but with remaining individuals of trees on T_0 (2020), favored the increased uncertainty. The classification results diverged from PRODES in areas with a slightly denser cluster of trees, leading PRODES to detect more deforestation polygons in those regions.

Place of interest I (Figure 46): Like at place H, degraded forest areas with associated pasture can be observed on T_{-1} (2019). No apparent canopy openings are evident. The degradation is indicated by a lighter green color and smoother texture compared to the non-degraded forest. By T_0 (2020), degradation reached its final stage, leading PRODES to regard it as "deforestation due to progressive degradation." The observed high uncertainty values in those regions were most probably due to the characteristic association of exposed soil and pasture. Despite the absence of canopy, the degradation of the previous year, combined with the presence of exposed soil and pasture on T_0 (2020), increased uncertainty. Nevertheless, the automatic detection results were close to the corresponding PRODES classification.

Place of interest J (Figure 46): On T_{-1} (2019), the region showed degraded forest areas with associated pasture, making it impossible to identify canopy openings. No significant changes in the degradation process were noticed on T_0 (2020) and T_1 (2021), which caused PRODES not to regard it as deforestation. Only in 2022 was the area identified as "deforestation due to progressive degradation," which is consistent with the high uncertainty observed two years earlier.

Place of interest K (Figure 46): Also, at this place, there is no sign of degradation on T_{-1} (2019). However, before T_0 (2020), the area was affected by fire from the deforested area to the south. The fire-induced degradation without complete deforestation is evident, as the forest still retained its canopy and potential for self-regeneration. That history led to high uncertainty at the detection date (2020). By T_1 (2021), complete deforestation occurred in the area.

Place of interest L (Figure 47): Selective geometric cutting is clearly visible on T_{-1} (2019), indicating selective logging conducted according to a management plan with prior planning, as the logging activity exhibited a regular pattern. On T_0 (2020), complete vegetation removal with exposed soil became evident. The associated low uncertainty values can be easily understood since complete deforestation with the removal of arboreal vegetation, indicated by the smooth texture and exposed soil, is easily detectable. The automatic detection results mostly agreed with PRODES observations.

Place of interest M (Figure 47): On T_0 (2020), a newly affected area with selective cutting became evident, highlighting logged areas. In that case, spots with medium to high uncertainty occurred here and there. It is worth noting that the forest in the same area was completely cleared in T_1 (2021).

Place of interest N (Figure 47): The region corresponds to non-degraded forest in all the assessed years (2019, 2020, and 2021). As expected, uncertainty was low in the area.

5.3.7.3
MS site

Similar to what was done for the Amazon sites, this section provides a visual analysis of the MS site, which is located in the Cerrado biome. The uncertainty results are interpreted across six points of interest, illustrated in

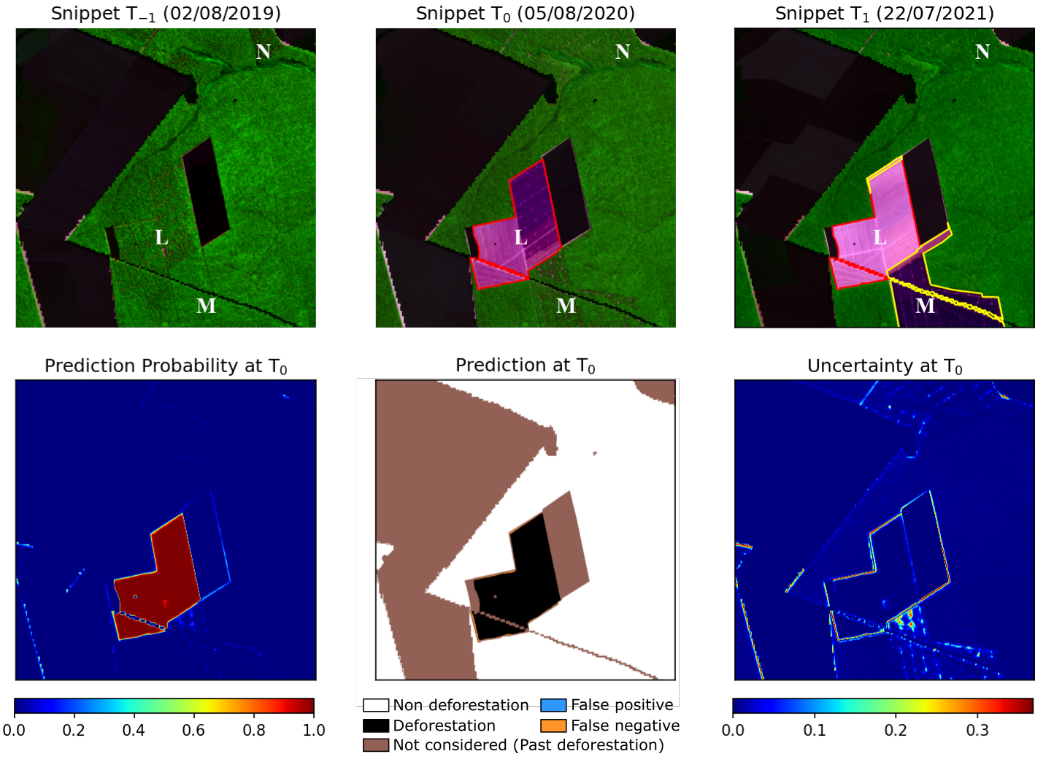


Figure 47: Qualitative results for MT site. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

Figures 48, 49, 50, 51, and 52. Following the established convention for the prior sites, the figures show input images from T_{-1} and T_0 dates in the first row. Each figure includes an optical image corresponding to one year ahead (T_1). The visual representations are derived from the RGB composition of Sentinel-2 satellite imagery. Deforested polygons (which correspond to the PRODES reference) from T_0 are outlined with red borders, while polygons associated with T_1 are showed with yellow borders. The lower row of the figures present the predicted probabilities for the deforestation class, the classification outcomes, and the corresponding uncertainty map.

In general, there was a stronger alignment between uncertainty and challenging-to-audit areas in the Amazon biome compared to the Cerrado biome. Given that deforestation polygons in MS were smaller and more isolated than those in the Amazon sites, each point of interest is presented in a distinct figure.

Place of interest O (Figure 48): Suppression of herbaceous vegetation, revealing exposed soil on the date of detection. The network generated a false

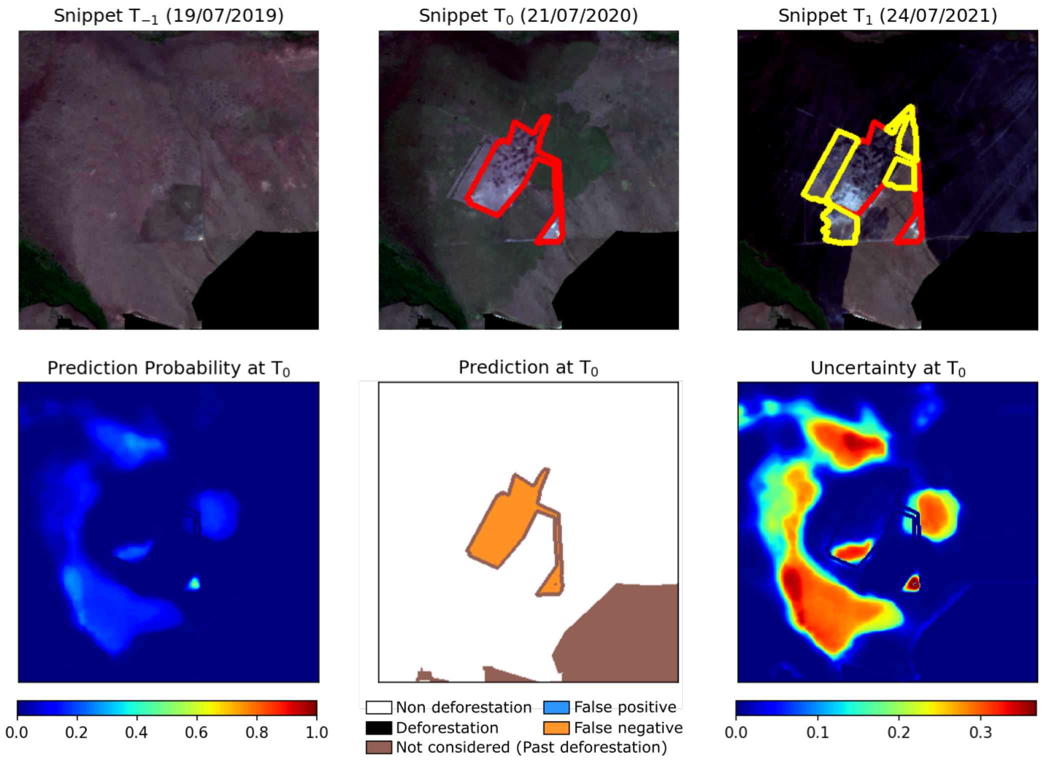


Figure 48: Qualitative results for MS site. Place of interest O. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

negative error for the entire polygon. While uncertainty was generally low in the error region, small areas of high uncertainty were notable in the southern portion. Additionally, some high-uncertainty sections in the polygon's vicinity aligned with subsequent deforestation in T_1.

Place of interest P (Figure 49): A visible fire scar was identified on vegetation primarily composed of herbaceous plants in the undergrowth, on the date of detection. Despite not being recognized as suppression in PRODES, the incident generated high uncertainty, and there was concordance between the classification and PRODES.

Place of interest Q (Figure 50): The suppression of arboreal vegetation, which is characterized by a closed canopy, occurred due to the introduction of agriculture on the date of detection. In the northwest portion of the polygon, despite deforestation taking place in T_0 (2020), it only involved partial vegetation removal, which the classification failed to detect, resulting in a false negative. Adequately, high uncertainty was observed in the northwest region. In the southern region, marked by low uncertainty, the classification accurately

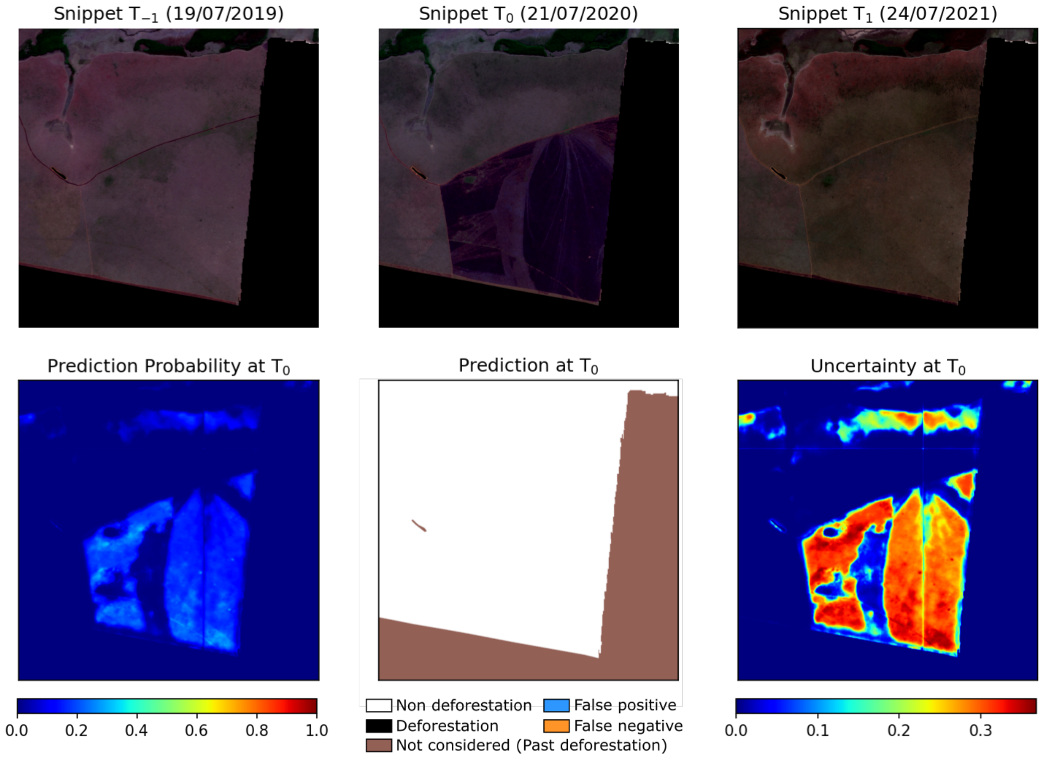


Figure 49: Qualitative results for MS site. Place of interest P. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

identified areas where vegetation was entirely removed (notable as exposed soil in the image).

Place of interest R (Figure 51): Vegetation suppression was observed in both the herbaceous (upper part of the polygon) and arboreal (lower part) components, marked by the emergence of exposed soil in the southwest section and the introduction of agriculture in the upper portion on the date of detection.

Place of interest S (Figure 52): The region, situated around a watercourse, features a blend of shrubbery and herbaceous vegetation. Herbaceous vegetation has a high seasonality throughout the year, being drier in the dry period and more vigorous in the wet period. This variation in vegetation vigor can also occur in the same months but in different years, as some years are drier than others. In 2019, the vegetation is noticeably drier along the watercourse in natural areas. Conversely, in 2020, the vegetation appears more vigorous, and the overall area is more humid, evident in the darker shade of vegetation. The substantial seasonality of natural herbs between 2019 and 2020

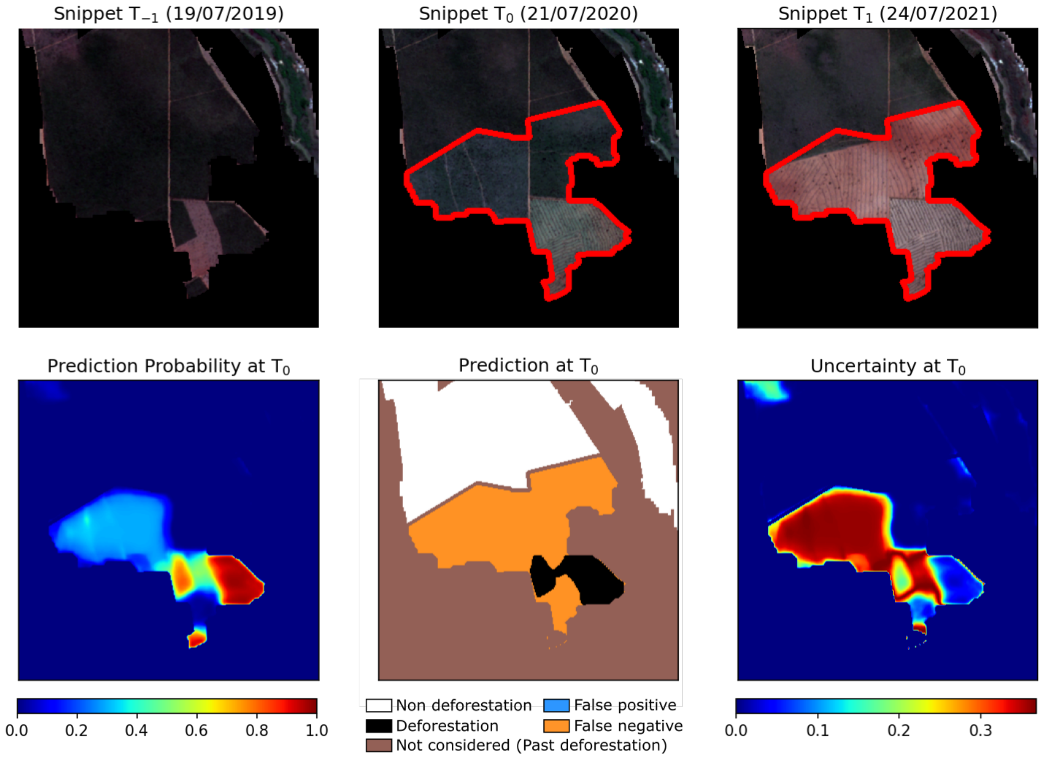


Figure 50: Qualitative results for MS site. Place of interest Q. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

contributed to elevated uncertainty. The classification process detected false positives, whereas PRODES did not detect deforestation in the area.

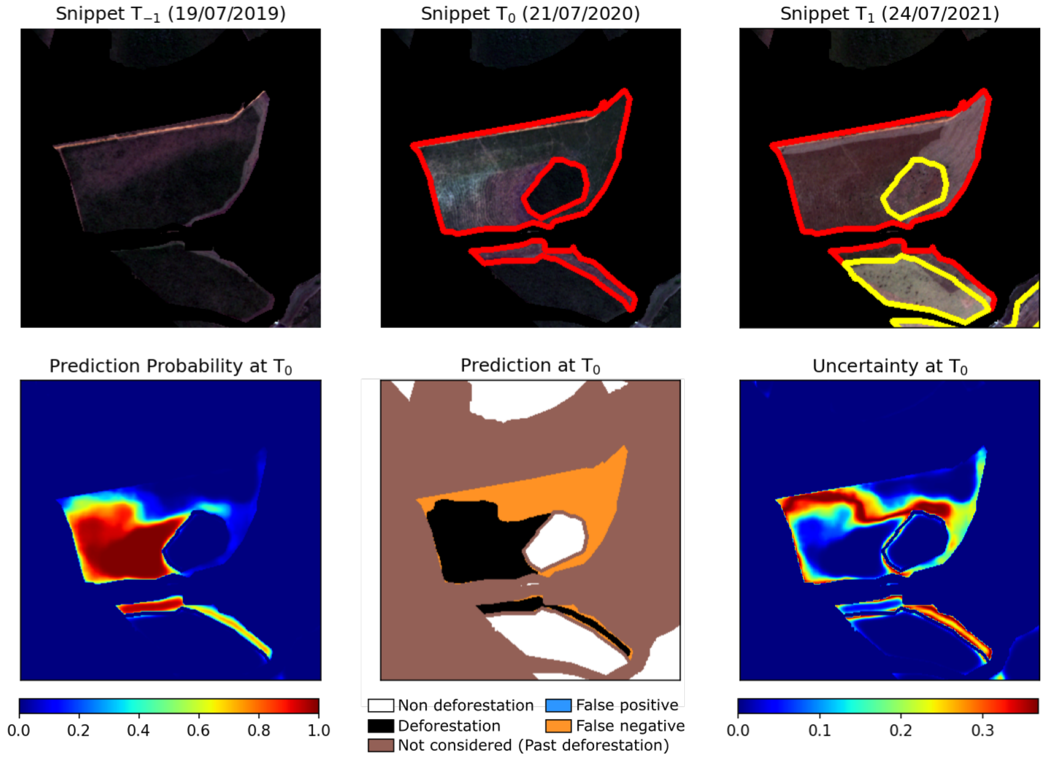


Figure 51: Qualitative results for MS site. Place of interest R. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

5.3.7.4 PI site

Classification performance in the PI site was lower compared to the other study areas. Even so, uncertainty still resulted in significant improvements in F_1. The results are elucidated in six areas of interest depicted in Figures 53, 54, 55, 56, 57, and 58. Similar to MS, where deforestation polygons were smaller and more isolated compared to the Amazon sites, each area of interest is shown in a separate figure. Like in the preceding sites, the first row displays the inference input images T_{-1} and T_0, along with the future date T_1. The second row presents the predicted probability for the deforestation class, followed by the classification outcomes and the uncertainty map.

Place of interest T (Figure 53): Vegetation suppression, primarily comprising herbaceous plants in the undergrowth, was observed, along with the introduction of agriculture on the date of detection. While uncertainty was minimal across most of the error-prone area, an exception existed at the boundaries between the accurately classified portion and the error region in the northern

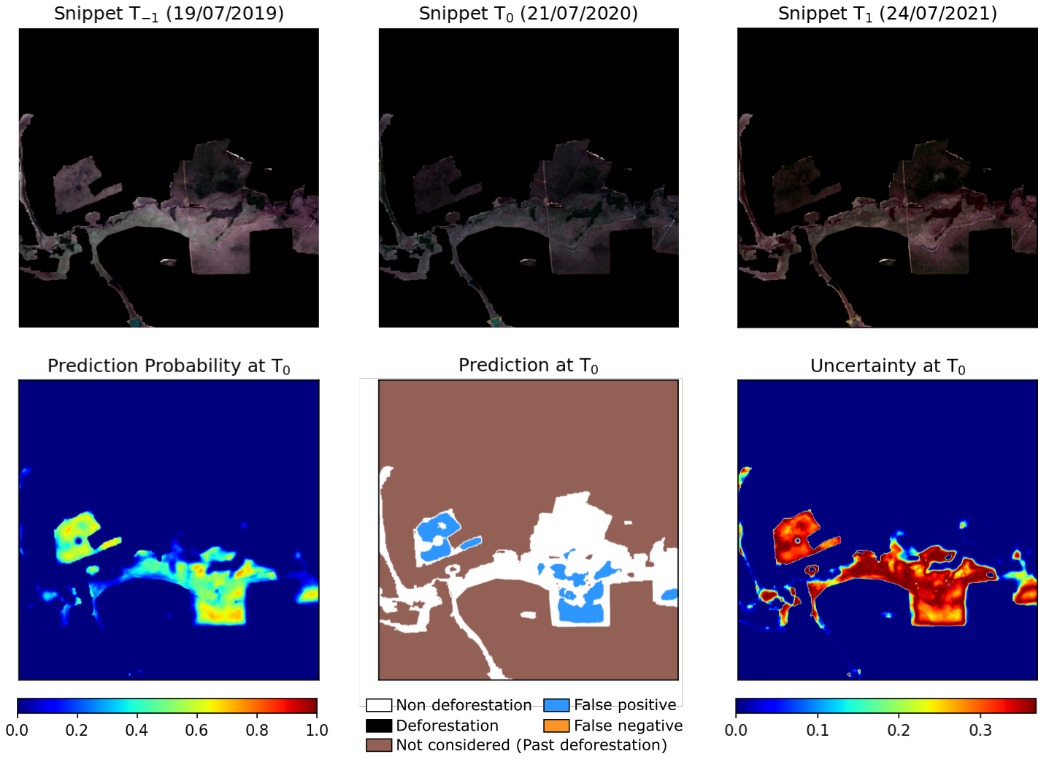


Figure 52: Qualitative results for MS site. Place of interest S. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

sector.

Place of interest U (Figure 54): There was suppression of vegetation with a predominance of herbaceous plants in the undergrowth, with the inclusion of agriculture on the date of detection. As a counter-example, uncertainty was not high in multiple error areas.

Place of interest V (Figure 55): Many phytophysiognomies in the Cerrado are predominantly herbaceous. Herbaceous vegetation becomes senescent in the dry period, and over the years there is an accumulation of combustible material. Fires have occurred naturally in the Cerrado for thousands of years but have intensified through human practices [112, 113]. The area of interest was degraded by fire. The image, however, does not show the burning but rather the subsequent degradation, with a lighter tone. This type of feature is not considered deforestation in PRODES, as burning can be considered as a form of natural management (elimination of dead herbaceous plants accumulated on the soil) and, after the fire, there is regrowth of the vegetation, which does not mean removal. False positives were detected in the classification.

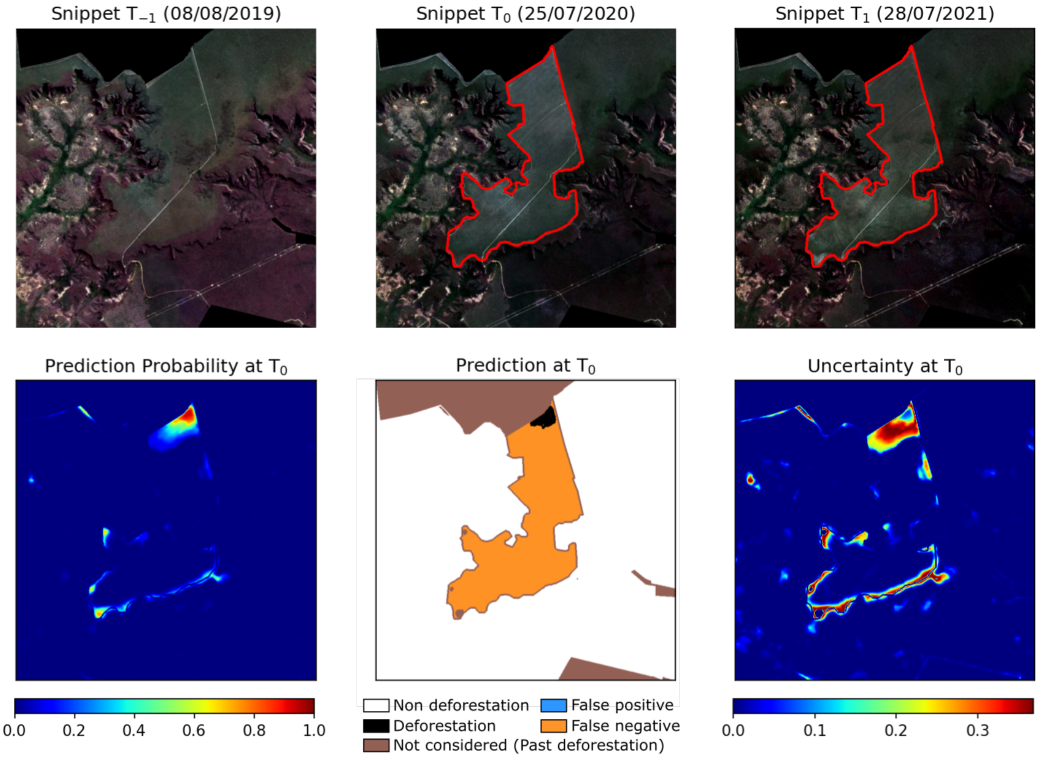


Figure 53: Qualitative results for PI site. Place of interest T. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

High uncertainty was obtained along a large percentage of the described phenomenon.

Place of interest W (Figure 56): The area is predominantly made up of degraded wooded savannah. Such vegetation shows seasonality throughout the year, being drier in the dry period and more vigorous in the wet period. This variation in vegetation vigor can also occur in the same months but in different years, as some years are drier than others. In the observed area, a certain level of degradation of herbaceous vegetation can also be seen. In T_{-1} (2019) the vegetation is more vigorous, while in T_0 (2020), it is drier. It is observed that more degraded places have an appearance very similar to exposed soil in 2020. In multiple areas, a fire occurred sometime between the dates of the 2019 and 2020 images. Such fire areas cannot be seen in the 2020 image anymore. However, those regions are not deforestation. The phenomenon above leaves several areas of high uncertainty and false positive detection.

Place of interest X (Figure 57): The area has a predominance of arboreal vegetation. Regions with high seasonality in 2020 (a drier year compared

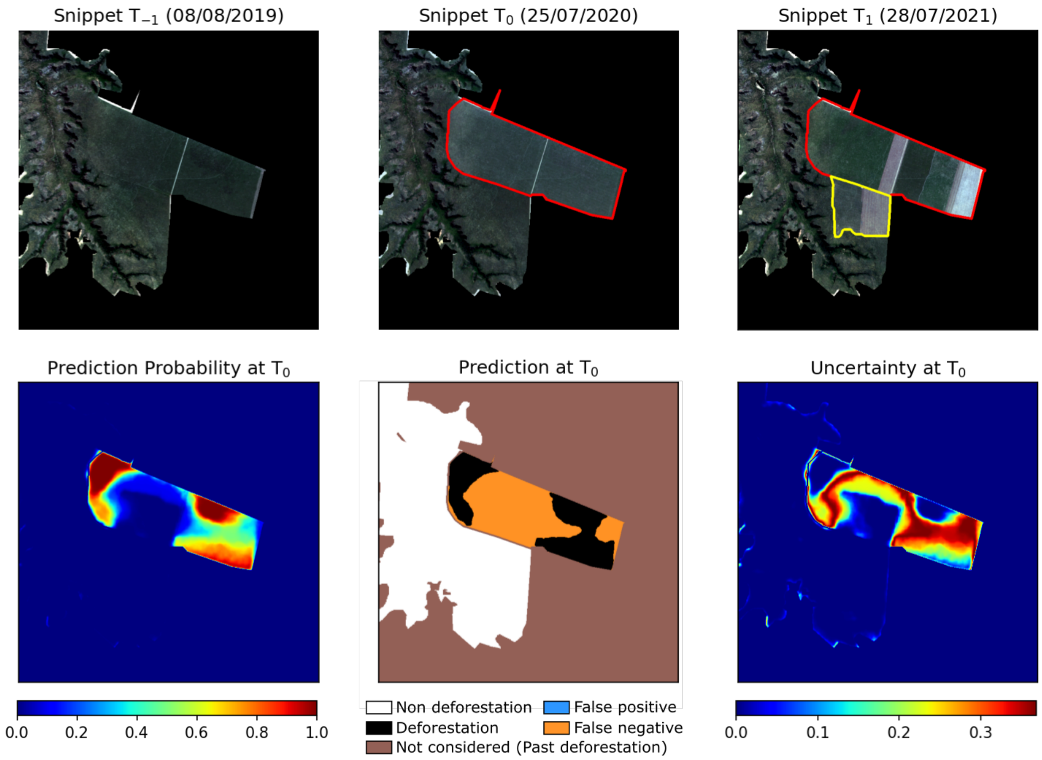


Figure 54: Qualitative results for PI site. Place of interest U. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

to 2019) to the south presented high uncertainty, and false positives were detected. Deforestation polygons featuring exposed soil at the time of detection in 2020 displayed low uncertainty and were classified similarly to PRODES.

Place of interest Y (Figure 58): The area had herbaceous shrub vegetation in 2019, which was suppressed in 2020. A region characterized by low uncertainty occurred in the northern section of the polygon. The area, exhibiting exposed soil in 2020, was accurately classified, mirroring the classification performed by PRODES. However, the polygon's most central and southern portions were highly uncertain and detected only by PRODES. Those error-prone regions featured less exposed soil, posing a challenge for detection. However, through visual inspection, one can discern the machinery's lines in the false-negative zones.

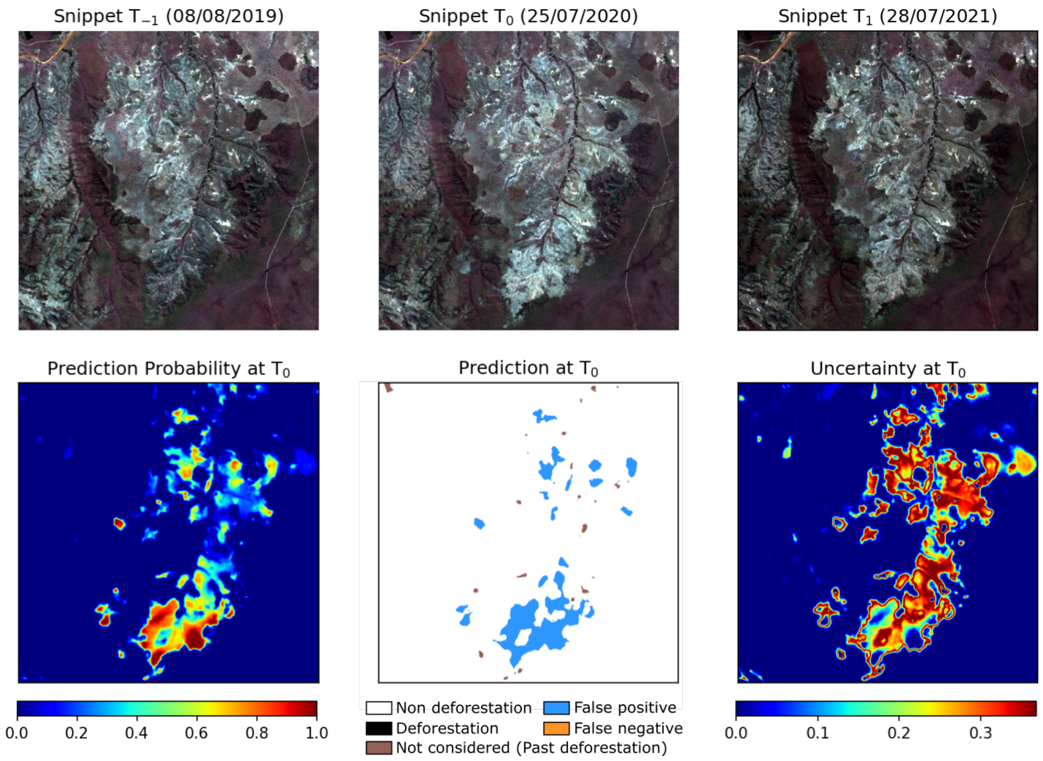


Figure 55: Qualitative results for PI site. Place of interest V. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

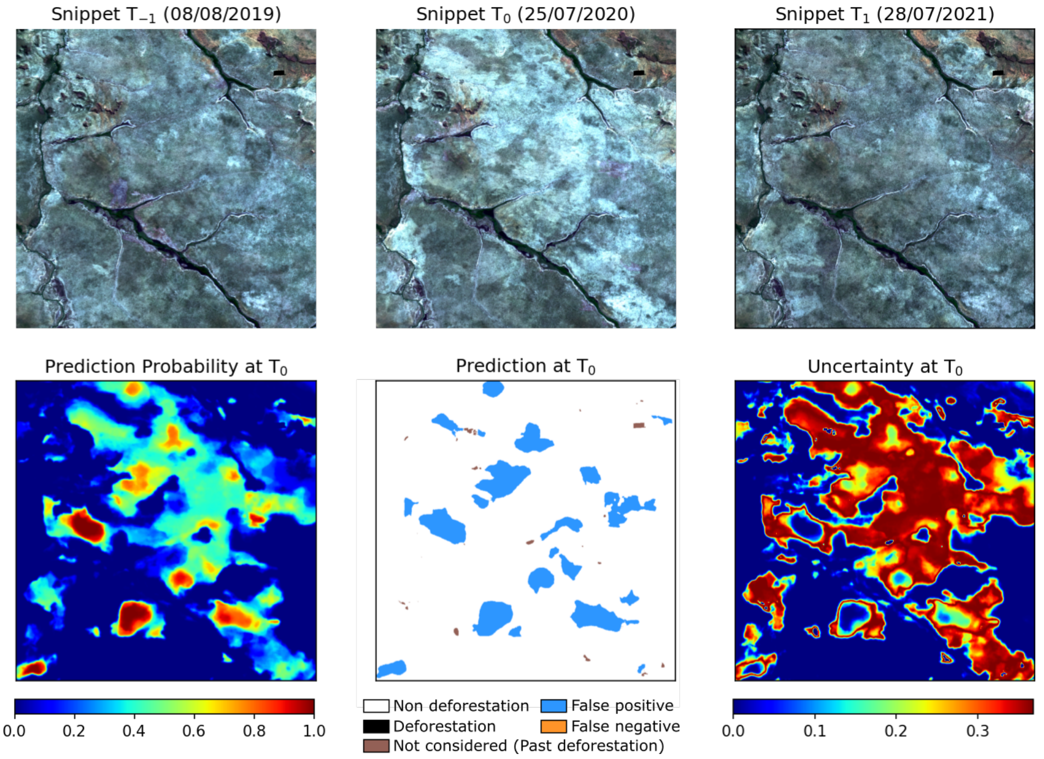


Figure 56: Qualitative results for PI site. Place of interest W. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

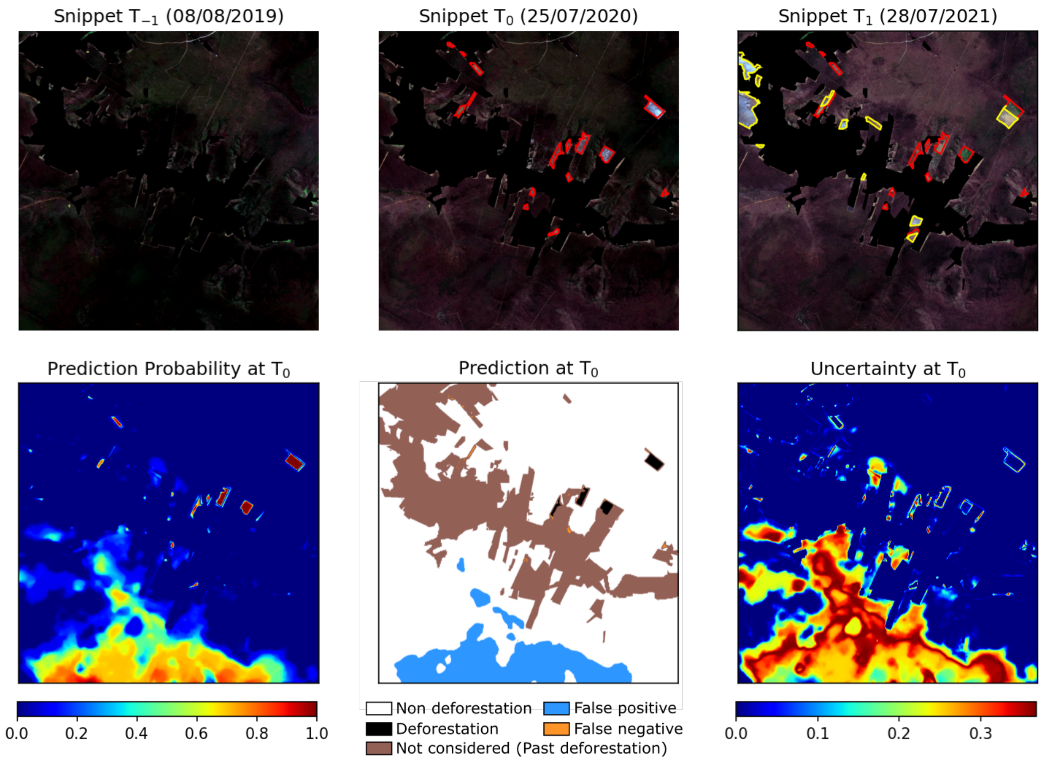

Snippet T_{-1} (08/08/2019)
Snippet T_0 (25/07/2020)
Snippet T_1 (28/07/2021)
Prediction Probability at T_0
Prediction at T_0
Uncertainty at T_0
0.0
0.2
0.4
0.6
0.8
1.0
Non deforestation
False positive
Deforestation
False negative
Not considered (Past deforestation)
0.0
0.1
0.2
0.3

Figure 57: Qualitative results for PI site. Place of interest X. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

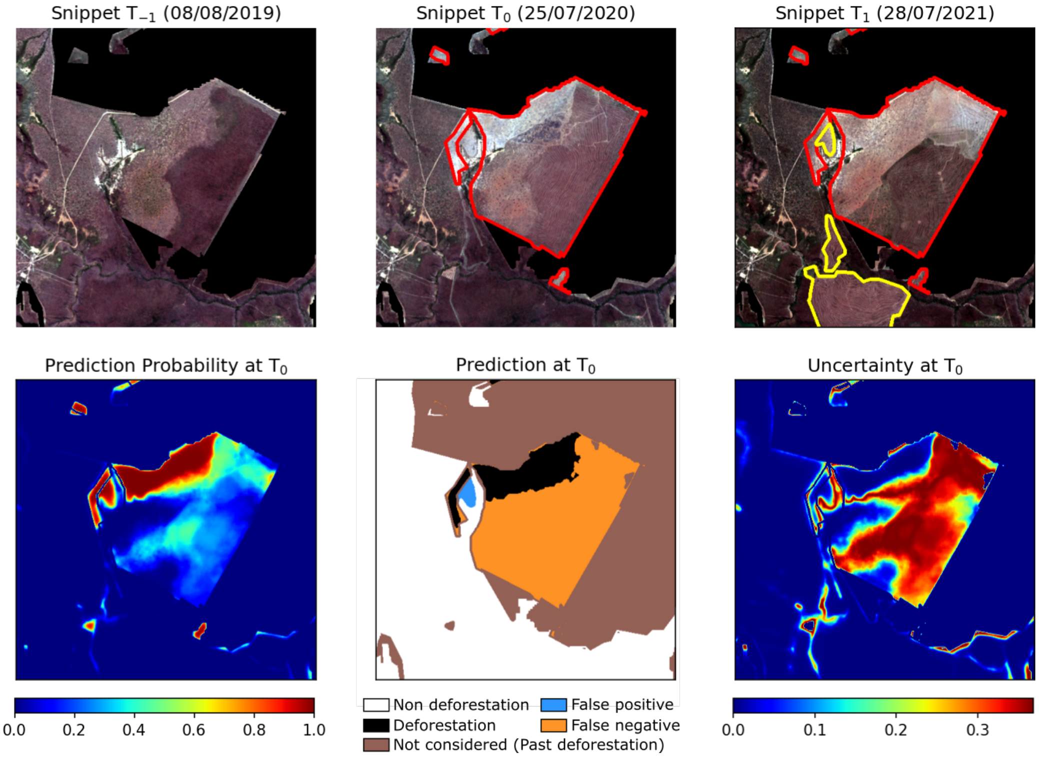


Figure 58: Qualitative results for PI site. Place of interest Y. The first row presents optical image snippets from T_{-1}, T_0, and T_1, at the PRODES dates. Deforested polygons for date T_0 are highlighted with red borders, and for the future date T_1, with yellow borders. The black areas represent past deforestation. The second row presents classification and uncertainty results for deforestation in T_0. Trained on an earlier pair of dates (T_{-2} and T_{-1}). Uncertainty method: Ensemble. Uncertainty metric: Predictive Entropy.

6
Conclusions

In this study, a semiautomatic methodology based on uncertainty estimation to reduce the effort required for visual photo interpretation was proposed, aiming at maintaining (or even improving) the accuracy of deforestation reports in the Brazilian rainforest, as traditionally produced by human analysts since 1988 by the Brazilian National Institute of Space Research (INPE).

The proposed approach involves submitting only high-uncertainty predictions to a human expert for visual inspection and manual annotation while keeping the automatically generated low-uncertainty outcomes. One notable feature such a methodology is the flexibility to determine the trade-off between auditing effort and final accuracy by selecting the total area to be audited.

The experimental analysis was conducted on data from two sites of the Brazilian Legal Amazon, and two sites of the Brazilian Cerrado biome, where the methodology effectively distinguished reliable predictions from those more likely to be incorrect.

In the experiments we employed the proposed methodology considering that only 3% of the total imaged area would be audited. As a result, for most of the sites, the F1-score significantly improved from around 85%, produced by the fully automatic counterpart, to values exceeding 95%. An exception occurred on the PI site, where the absolute values were lower despite the significant improvements: for an audit area rate of 3%, the F1-score improved from 63.2% to 81.2%.

As mentioned before, the choice of a methodology based on human visual interpretation stems from the high accuracy requirements to guarantee overall accuracy as high as PRODES reports, considering [11] reported an overall accuracy of 98.8% for the 2022 report. However, regarding the $F1$-score for the deforestation class, within the Brazilian Legal Amazon, the PRODES accuracy reaches 87.1%. It should be noted that in this work, the deforestation polygons extracted from the PRODES database are considered ground truth, so the accuracy values calculated for the proposed method may not be completely precise. However, at least for the areas of the study sites considered in this work, the related errors should be negligible, as INPE's specialists reassessed the respective deforestation polygons. Anyway, those numbers indicate that, at least for the selected study areas that illustrate typical deforestation patterns in the region, the accuracy values obtained with the proposed method meet the high accuracy requirement of PRODES.

Various uncertainty estimation methods were also investigated based on single or multiple inferences for each input image. From the experiments it was concluded that multi-inference approaches, such as Monte-Carlo Dropout and particularly ensembles, attained significantly higher accuracy than alternatives based on a single result per input. Furthermore, among the tested uncertainty metrics, the predictive entropy and predictive variance yielded the best results in the experimental evaluation. Exceptionally, the MI metric yielded the best performance in the MS site when training on an earlier pair of dates.

Evidential Deep Learning (EDL) was also assessed as a more recent single-outcome approach, and although it had potential due to its low computational complexity, further studies need to be done for deforestation detection. Specifically, the effect of the inherent high class imbalance needs to be further explored [77]. Different from previous studies, the variability of multiple repetition runs for EDL was also assessed. The high variability among repetitions was similar to the single run option, suggesting that a single point estimate may not have been sufficient to approximate the underlying distribution of the predicted probabilities, which is an assumption in EDL [61, 109]. Such an assumption was originally made given a sufficiently large training set, which indicates that increasing the training data with images from additional dates and regions may improve EDL outcomes.

Moreover, we consulted field experts to interpret the meaning of uncertainty maps and found a correspondence between regions of high uncertainty and areas of interest for auditing experts. Notably, areas with high uncertainty may serve as early indicators of future deforestation.

Additionally, areas with non-degraded primary forest cover show low uncertainty, while forest degradation is usually indicated by high levels of uncertainty on the map. When clearcut deforestation occurs, characterized by a total removal of vegetation, there is low uncertainty. However, when deforestation due to progressive degradation with the presence of remaining tree individuals is identified, the maps indicate higher uncertainty. The indication of polygons with high uncertainty points the auditor to places where there will often be a need to consult complementary images of higher spatial resolution to verify whether there was an opening of the forest canopy. In qualitative terms, sites in the Amazon biome exhibited a stronger alignment between areas of high uncertainty and those challenging to annotate, in contrast to the Cerrado biome. This discrepancy may arise from increased year-round seasonality in vegetation and a notably greater diversity in deforestation types in the Cerrado biome.

The study used a fully convolutional network employed in previous works on automatic deforestation mapping. It is reasonable to expect that even greater levels of accuracy can be achieved with more suitable deep architectures tailored to the application. The question regarding the generality and scope of the analysis, which aimed to characterize the sites where the automatic model tends to produce results with high uncertainty, remains open for further investigation. Those issues warrant continued research for devising operational solutions for deforestation mapping.

During this doctorate research, a total of 8 scientific papers have been published, including four papers in research journals, and four in international conferences [114–121]. An additional work containing the results presented in this document has been submitted for a research journal [122], and is currently under review.

Future directions

As future directions, multiple research hypotheses may be derived from this work. First, additional uncertainty estimation methods may be considered. Particularly, evidential learning may be improved by addressing the high class imbalance, which is inherent in deforestation detection. The original EDL work in [18] did not specifically analyze class imbalance. Adapting more recent work addressing class imbalance in EDL is suggested [77]. Evidential learning approaches can be divided into model-based (as assessed in this work), and distance-based [83]. Hence, exploring and assessing distance-based evidential deep learning approaches is suggested. Furthermore, the combination of presented uncertainty methods may be explored. Particularly, the combination of Test-Time Augmentations (TTA) with the best performing approaches (MCD and ensembles) may bring additional improvements as in [37]. For multiple-outcome methods, the proposed uncertainty metric from [56] may be assessed, which measured the overlap between the top-2 predicted class distributions and showed promising results. Likewise, the Jensen-Shannon Distance may be explored as an alternative uncertainty metric, which improves the assessed Expected Kullback-Leibler divergence because it is symmetric and it always has a finite value [123].

Future works may examine the distinct significance of false negative and false positive errors in the context of deforestation detection. Notably, in this application, false positive errors carry considerably greater undesirability compared to false negative errors, given their potential implications for operational and legal costs. In the current study, both false positive and true positive errors were treated with equal concern. Future research could explore prioritizing lower false negative rates when identifying the most relevant areas for audit. In a similar way, the PRODES team embraces a conservative approach in their estimations. Specifically, they observe activities within a predefined spatial buffer (e.g., $15km$) surrounding previously identified deforestation polygons from earlier years. In this line, a possible approach might consider narrowing down areas of interest for auditing to regions within this specified spatial buffer.

Another future direction is related to active learning. In this work, samples of high uncertainty were selected for re-annotation by an expert auditor. Active learning schemes have been proposed for the same purpose, including two different stages. First, a predefined amount of high-relevance samples are pre-selected by selecting samples with the highest uncertainty. Finally, a smaller group of samples with the largest diversity are selected [89, 95]. Hence, a future direction would be to complement the proposed methodology with a diversity selection step, aiming at further reducing the auditing re-annotation effort. Such effort reduction might be reflected in smaller Alert Area (AA) percentage requirements.

Another direction is temperature scaling [15]. The authors show that most modern deep networks tend to produce overconfident probabilities, with the confidence (as in, the largest softmax value among classes) being close to 100% most of the time. Ideally, the confidence values should match the true likelihood of being correct. For example, if the average confidence in 100 inference samples is 80%, their corresponding accuracy should also match 80%, meaning the network should be calibrated. A calibrated network may produce more trustworthy uncertainty scores, particularly in the baseline confidence-based approach. Temperature scaling is a post-processing method for network calibration, which should be explored in future works. It divides the logits in the softmax function by a learned scalar parameter T, where T is learned from an additional validation set.

Considering that the best outcomes were obtained with deep ensembles, another future direction is to train a deep ensemble with a teacher-student knowledge distillation method as in [31]. This would allow to train an ensemble as a teacher and then train a single network as a student which learns the teacher's uncertainty in a supervised fashion, allowing it to estimate uncertainty in a single forward pass with similar outcomes to the teacher ensemble.

In order to improve absolute classification metrics, alternative base networks may be compared to the ResUnet-based architecture adopted in this work. Particularly, vision transformer-based [124] segmentation networks such as Swin Unet [107] and Dense Prediction Transformers (DPT) [125] may be worth exploring in conjunction with multiple uncertainty estimation techniques such as MCD, ensemble, and evidential learning. Transformers offer the advantage of paying attention to the entire image instead of a predefined neighboring kernel with a fixed size, as in convolutional networks. Examining the mentioned base networks, we can determine whether the suggested uncertainty-based methodology consistently yields the intended results independently of the selected base architectures.

Closely related to uncertainty estimation, out-of-distribution detection methods detect samples outside the training distribution as unknown outliers, which may also require re-annotation by an expert auditor. An example of samples outside the training distribution occurs when inference is made in geographical areas with modified spatial or temporal characteristics compared to the original training site (also known as *anomaly detection* or *domain shift detection*) [126]. Another example occurs when new classes different from the ones seen during training occur at inference (also known as *open set learning*) [115, 127]. Future works may combine out-of-distribution detection methods such as OpenPCS++ [115] with uncertainty estimation to produce more robust results regarding areas of interest for the auditing annotators in the context of deforestation detection.

Regarding the uncertainty interpretation analysis presented in this work, future works may extend such analysis by interpreting both the aleatoric and model uncertainty independently, and analyzing which sources of uncertainty belong to each of them. Aleatoric uncertainty may be estimated from Test-Time Augmentation (TTA).

In this work, the network was trained with an earlier image pair and tested on an unseen pair of dates. To improve classification accuracy, future works may add more training data by using multiple pairs of earlier images as input, taking advantage of the vast amount of ground truth information reported by PRODES since 1988.

In the Cerrado biome, there is a much larger variability in types of deforestation compared to the Amazon biome, with nearly ten types of deforestation [128]. Future works may cluster the training data into those multiple types of deforestation and balance the number of samples per cluster presented to the network during training to ensure that the network equally learns all types of deforestation. Such an approach was already used in the context of domain adaptation [129] and may improve the classification accuracy in the PI site, where the lowest-performing result was obtained.

Finally, future work may implement the proposed uncertainty estimation strategy in an operational setting, encompassing large spatial areas such as the entire Brazilian Amazon and Cerrado biomes instead of smaller study areas as the ones assessed in this work. In this work, it was assumed that the auditing expert would re-annotate the high-uncertainty samples with 100% accuracy. In further studies, human auditing experts may be included in the proposed semi-automatic methodology to validate its usefulness in an operational setting. It should be noted that the proposed methodology is currently being used in selected regions of interest for the new PRODES 2023 figures, with human auditing experts leveraging the proposed uncertainty estimates during their auditing procedure. The proposed uncertainty assessment methodology may be added to the currently operational monitoring system Brazil Data Cube (BDC) [130], as well as taking advantage of the open source time series analysis tool Satellite Image Time Series (SITS) [131], which accepts data from BDC as input.

References

1 KENDALL, A.; GAL, Y.. **What uncertainties do we need in bayesian deep learning for computer vision?** Advances in neural information processing systems, 30, 2017.

2 GAWLIKOWSKI, J.; TASSI, C. R. N.; ALI, M.; LEE, J.; HUMT, M.; FENG, J.; KRUSPE, A.; TRIEBEL, R.; JUNG, P.; ROSCHER, R. ; OTHERS. **A survey of uncertainty in deep neural networks.** Artificial Intelligence Review, p. 1–77, 2023.

3 FEITOSA, R. Q.. **Class Notes on Advanced Deep Learning for Image Analysis.** `http://www.ele.puc-rio.br/~raul/DL2CV/`, 2023. [Online; accessed 07-November-2023].

4 SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I. ; SALAKHUTDINOV, R.. **Dropout: A simple way to prevent neural networks from overfitting.** Journal of Machine Learning Research, 15:1929–1958, 2014.

5 THEVENOT, A.. **12 Main Dropout Methods: Mathematical and Visual Explanation for DNNs, CNNs, and RNNs.** `https://towardsdatascience.com/12-main-dropout-methods-mathematical-and-visual-explanation-58cdc2112293`, 2020. [Online; accessed 03-June-2020].

6 LEMOS, A. L. F.; SILVA, J. D. A.. **Desmatamento na amazônia legal: evolução, causas, monitoramento e possibilidades de mitigação através do fundo amazônia.** Floresta e Ambiente, 18(1):98–108, 2012.

7 MAURANO, L. E. P.; ALMEIDA, C. D. ; MEIRA, M. B.. **Monitoramento do desmatamento do cerrado brasileiro por satélite prodes cerrado.** Simpósio Brasileiro de Sensoriamento Remoto, 19:191–194, 2019.

8 INPE. **National institute for space research. general coordination of earth observation. monitoring program of the amazon and other biomes. deforestation - legal amazon -.** `http://terrabrasilis.dpi.inpe.br`, 2021.

9 PARENTE, L.; NOGUEIRA, S.; BAUMANN, L.; ALMEIDA, C.; MAURANO, L.; AFFONSO, A. G. ; FERREIRA, L.. **Quality assessment of the prodes cerrado deforestation data.** Remote Sensing Applications: Society and Environment, 21:100444, 2021.

10 LAURANCE, W. F.; ALBERNAZ, A. K.; SCHROTH, G.; FEARNSIDE, P. M.; BERGEN, S.; VENTICINQUE, E. M. ; DA COSTA, C.. **Predictors of deforestation in the brazilian amazon.** Journal of biogeography, 29(5-6):737–748, 2002.

11 INPE. **Avaliação da acurácia do mapeamento do PRODES 2022.** 2022.

12 ORTEGA, M. X.; FEITOSA, R. Q.; BERMUDEZ, J. D.; HAPP, P. N. ; DE ALMEIDA, C. A.. **Comparison of optical and sar data for deforestation mapping in the amazon rainforest with fully convolutional networks.** In: 2021 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM IGARSS, p. 3769–3772. IEEE, 2021.

13 ADARME, M. O.; COSTA, G. ; FEITOSA, R.. **Multi-attention ghostnet for deforestation detection in the amazon rainforest.** ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 3:657–664, 2022.

14 TORRES, D. L.; TURNES, J. N.; SOTO VEGA, P. J.; FEITOSA, R. Q.; SILVA, D. E.; MARCATO JUNIOR, J. ; ALMEIDA, C.. **Deforestation detection with fully convolutional networks in the amazon forest from landsat-8 and sentinel-2 images.** Remote Sensing, 13(24):5084, 2021.

15 GUO, C.; PLEISS, G.; SUN, Y. ; WEINBERGER, K. Q.. **On calibration of modern neural networks.** In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, p. 1321–1330. PMLR, 2017.

16 GAWLIKOWSKI, J.; TASSI, C. R. N.; ALI, M.; LEE, J.; HUMT, M.; FENG, J.; KRUSPE, A.; TRIEBEL, R.; JUNG, P.; ROSCHER, R. ; OTHERS. **A survey of uncertainty in deep neural networks.** arXiv preprint arXiv:2107.03342, 2021.

17 GAL, Y.; GHAHRAMANI, Z.. **Dropout as a bayesian approximation: Representing model uncertainty in deep learning.** In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, p. 1050–1059. PMLR, 2016.

18 SENSOY, M.; KAPLAN, L. ; KANDEMIR, M.. **Evidential deep learning to quantify classification uncertainty**. Advances in neural information processing systems, 31, 2018.

19 DE ANDRADE, R. B.; MOTA, G. L. A. ; DA COSTA, G. A. O. P.. **Deforestation detection in the amazon using deeplabv3+ semantic segmentation model variants**. Remote Sensing, 14(19):4694, 2022.

20 JHA, D.; SMEDSRUD, P. H.; RIEGLER, M. A.; JOHANSEN, D.; DE LANGE, T.; HALVORSEN, P. ; JOHANSEN, H. D.. **Resunet++: An advanced architecture for medical image segmentation**. In: 2019 IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA (ISM), p. 225–2255. IEEE, 2019.

21 AMOROCHO, J.; ESPILDORA, B.. **Entropy in the assessment of uncertainty in hydrologic systems and models**. Water Resources Research, 9(6):1511–1522, 1973.

22 WANG, D.; SHANG, Y.. **A new active labeling method for deep learning**. In: 2014 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN), p. 112–119. IEEE, 2014.

23 ZHAN, X.; WANG, Q.; HUANG, K.-H.; XIONG, H.; DOU, D. ; CHAN, A. B.. **A comparative survey of deep active learning**. arXiv preprint arXiv:2203.13450, 2022.

24 LAKSHMINARAYANAN, B.; PRITZEL, A. ; BLUNDELL, C.. **Simple and scalable predictive uncertainty estimation using deep ensembles**. Advances in neural information processing systems, 30, 2017.

25 MEHRTASH, A.; WELLS, W. M.; TEMPANY, C. M.; ABOLMAESUMI, P. ; KAPUR, T.. **Confidence calibration and predictive uncertainty estimation for deep medical image segmentation**. IEEE transactions on medical imaging, 39(12):3868–3878, 2020.

26 OVADIA, Y.; FERTIG, E.; REN, J.; NADO, Z.; SCULLEY, D.; NOWOZIN, S.; DILLON, J.; LAKSHMINARAYANAN, B. ; SNOEK, J.. **Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift**. Advances in neural information processing systems, 32, 2019.

27 GUSTAFSSON, F. K.; DANELLJAN, M. ; SCHON, T. B.. **Evaluating scalable bayesian deep learning methods for robust computer**

vision. In: PROCEEDINGS OF THE IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS, p. 318–319, 2020.

28 RAHAMAN, R.; OTHERS. **Uncertainty quantification and deep ensembles**. Advances in Neural Information Processing Systems, 34:20063–20075, 2021.

29 FORT, S.; HU, H. ; LAKSHMINARAYANAN, B.. **Deep ensembles: A loss landscape perspective**. arXiv preprint arXiv:1912.02757, 2019.

30 LANDGRAF, S.; WURSTHORN, K.; HILLEMANN, M. ; ULRICH, M.. **Dudes: Deep uncertainty distillation using ensembles for semantic segmentation**. arXiv preprint arXiv:2303.09843, 2023.

31 HOLDER, C. J.; SHAFIQUE, M.. **Efficient uncertainty estimation in semantic segmentation via distillation**. In: PROCEEDINGS OF THE IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION, p. 3087–3094, 2021.

32 HOCHGESCHWENDER, N.; PLÖGER, P.; KIRCHNER, F.; VALDENEGRO-TORO, M. ; OTHERS. **Evaluating uncertainty estimation methods on 3d semantic segmentation of point clouds**. arXiv preprint arXiv:2007.01787, 2020.

33 HAAS, J.; RABUS, B.. **Uncertainty estimation for deep learning-based segmentation of roads in synthetic aperture radar imagery**. Remote Sensing, 13(8):1472, 2021.

34 MILANÉS-HERMOSILLA, D.; TRUJILLO CODORNIÚ, R.; LÓPEZ-BARACALDO, R.; SAGARÓ-ZAMORA, R.; DELISLE-RODRIGUEZ, D.; VILLAREJO-MAYOR, J. J. ; NÚÑEZ-ÁLVAREZ, J. R.. **Monte carlo dropout for uncertainty estimation and motor imagery classification**. Sensors, 21(21):7241, 2021.

35 LEIBIG, C.; ALLKEN, V.; AYHAN, M. S.; BERENS, P. ; WAHL, S.. **Leveraging uncertainty information from deep neural networks for disease detection**. Scientific reports, 7(1):17816, 2017.

36 VAN MOLLE, P.; VERBELEN, T.; DE BOOM, C.; VANKEIRSBILCK, B.; DE VYLDER, J.; DIRICX, B.; KIMPE, T.; SIMOENS, P. ; DHOEDT, B.. **Quantifying uncertainty of deep neural networks in skin lesion classification**. In: UNCERTAINTY FOR SAFE UTILIZATION OF

MACHINE LEARNING IN MEDICAL IMAGING AND CLINICAL IMAGE-BASED PROCEDURES: FIRST INTERNATIONAL WORKSHOP, UNSURE 2019, AND 8TH INTERNATIONAL WORKSHOP, CLIP 2019, HELD IN CONJUNCTION WITH MICCAI 2019, SHENZHEN, CHINA, OCTOBER 17, 2019, PROCEEDINGS 8, p. 52–61. Springer, 2019.

37 COMBALIA, M.; HUETO, F.; PUIG, S.; MALVEHY, J. ; VILAPLANA, V.. **Uncertainty estimation in deep neural networks for dermoscopic image classification.** In: PROCEEDINGS OF THE IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS, p. 744–745, 2020.

38 HERZOG, L.; MURINA, E.; DÜRR, O.; WEGENER, S. ; SICK, B.. **Integrating uncertainty in deep neural networks for mri based stroke analysis.** Medical image analysis, 65:101790, 2020.

39 PRINCE, E. W.; GHOSH, D.; GÖRG, C. ; HANKINSON, T. C.. **Uncertainty-aware deep learning classification of adamantinomatous craniopharyngioma from preoperative mri.** Diagnostics, 13(6):1132, 2023.

40 RONNEBERGER, O.; FISCHER, P. ; BROX, T.. **U-net: Convolutional networks for biomedical image segmentation.** In: INTERNATIONAL CONFERENCE ON MEDICAL IMAGE COMPUTING AND COMPUTER-ASSISTED INTERVENTION, p. 234–241. Springer, 2015.

41 DECHESNE, C.; LASSALLE, P. ; LEFÈVRE, S.. **Bayesian u-net: Estimating uncertainty in semantic segmentation of earth observation images.** Remote Sensing, 13(19):3836, 2021.

42 NGUYEN, D.; BARKOUSARAIE, A. S.; BOHARA, G.; BALAGOPAL, A.; MCBETH, R.; LIN, M.-H. ; JIANG, S.. **A comparison of monte carlo dropout and bootstrap aggregation on the performance and uncertainty estimation in radiation therapy dose prediction with deep learning neural networks.** Physics in Medicine & Biology, 66(5):054002, 2021.

43 KWON, Y.; WON, J.-H.; KIM, B. J. ; PAIK, M. C.. **Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation.** Computational Statistics & Data Analysis, 142:106816, 2020.

44 JOSHI, I.; KOTHARI, R.; UTKARSH, A.; KURMI, V. K.; DANTCHEVA, A.; ROY, S. D. ; KALRA, P. K.. **Explainable fingerprint roi segmentation using monte carlo dropout.** In: PROCEEDINGS OF THE IEEE/CVF WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION, p. 60–69, 2021.

45 HUANG, P.-Y.; HSU, W.-T.; CHIU, C.-Y.; WU, T.-F. ; SUN, M.. **Efficient uncertainty estimation for semantic segmentation in videos.** In: PROCEEDINGS OF THE EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), p. 520–535, 2018.

46 LI, W.; WANG, G.; FIDON, L.; OURSELIN, S.; CARDOSO, M. J. ; VERCAUTEREN, T.. **On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task.** In: INFORMATION PROCESSING IN MEDICAL IMAGING: 25TH INTERNATIONAL CONFERENCE, IPMI 2017, BOONE, NC, USA, JUNE 25-30, 2017, PROCEEDINGS 25, p. 348–360. Springer, 2017.

47 DEVRIES, T.; TAYLOR, G.. **Leveraging uncertainty estimates for predicting segmentation quality, 2018.** ArXiv abs, 2018.

48 EATON-ROSEN, Z.; BRAGMAN, F.; BISDAS, S.; OURSELIN, S. ; CARDOSO, M. J.. **Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions.** In: MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION–MICCAI 2018: 21ST INTERNATIONAL CONFERENCE, GRANADA, SPAIN, SEPTEMBER 16-20, 2018, PROCEEDINGS, PART I, p. 691–699. Springer, 2018.

49 BRAGMAN, F. J.; TANNO, R.; EATON-ROSEN, Z.; LI, W.; HAWKES, D. J.; OURSELIN, S.; ALEXANDER, D. C.; MCCLELLAND, J. R. ; CARDOSO, M. J.. **Uncertainty in multitask learning: joint representations for probabilistic mr-only radiotherapy planning.** In: MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION–MICCAI 2018: 21ST INTERNATIONAL CONFERENCE, GRANADA, SPAIN, SEPTEMBER 16-20, 2018, PROCEEDINGS, PART IV 11, p. 3–11. Springer, 2018.

50 NAIR, T.; PRECUP, D.; ARNOLD, D. L. ; ARBEL, T.. **Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation.** Medical image analysis, 59:101557, 2020.

51 JUNGO, A.; MEIER, R.; ERMIS, E.; HERRMANN, E. ; REYES, M.. **Uncertainty-driven sanity check: application to postoperative brain tumor cavity segmentation.** arXiv preprint arXiv:1806.03106, 2018.

52 SEDAI, S.; ANTONY, B.; MAHAPATRA, D. ; GARNAVI, R.. **Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using bayesian deep learning.** In: COMPUTATIONAL PATHOLOGY AND OPHTHALMIC MEDICAL IMAGE ANALYSIS: FIRST INTERNATIONAL WORKSHOP, COMPAY 2018, AND 5TH INTERNATIONAL WORKSHOP, OMIA 2018, HELD IN CONJUNCTION WITH MICCAI 2018, GRANADA, SPAIN, SEPTEMBER 16-20, 2018, PROCEEDINGS 5, p. 219–227. Springer, 2018.

53 JUNGO, A.; MEIER, R.; ERMIS, E.; BLATTI-MORENO, M.; HERRMANN, E.; WIEST, R. ; REYES, M.. **On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation.** In: MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION–MICCAI 2018: 21ST INTERNATIONAL CONFERENCE, GRANADA, SPAIN, SEPTEMBER 16-20, 2018, PROCEEDINGS, PART I, p. 682–690. Springer, 2018.

54 SOBERANIS-MUKUL, R. D.; NAVAB, N. ; ALBARQOUNI, S.. **An uncertainty-driven gcn refinement strategy for organ segmentation.** arXiv preprint arXiv:2012.03352, 2020.

55 SEEBÖCK, P.; ORLANDO, J. I.; SCHLEGL, T.; WALDSTEIN, S. M.; BOGUNOVIĆ, H.; KLIMSCHA, S.; LANGS, G. ; SCHMIDT-ERFURTH, U.. **Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct.** IEEE transactions on medical imaging, 39(1):87–98, 2019.

56 YU, L.; WANG, S.; LI, X.; FU, C.-W. ; HENG, P.-A.. **Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation.** In: MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION–MICCAI 2019: 22ND INTERNATIONAL CONFERENCE, SHENZHEN, CHINA, OCTOBER 13–17, 2019, PROCEEDINGS, PART II 22, p. 605–613. Springer, 2019.

57 CAO, X.; CHEN, H.; LI, Y.; PENG, Y.; WANG, S. ; CHENG, L.. **Uncertainty aware temporal-ensembling model for semi-supervised**

abus mass segmentation. IEEE transactions on medical imaging, 40(1):431–443, 2020.

58 HASAN, S. K.; LINTE, C. A.. **Calibration of cine mri segmentation probability for uncertainty estimation using a multi-task cross-task learning architecture.** In: MEDICAL IMAGING 2022: IMAGE-GUIDED PROCEDURES, ROBOTIC INTERVENTIONS, AND MODELING, volumen 12034, p. 174–179. SPIE, 2022.

59 WU, S.; HEITZLER, M. ; HURNI, L.. **A closer look at segmentation uncertainty of scanned historical maps.** The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 43:189–194, 2022.

60 WANG, G.; LI, W.; AERTSEN, M.; DEPREST, J.; OURSELIN, S. ; VERCAUTEREN, T.. **Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks.** Neurocomputing, 338:34–45, 2019.

61 MALININ, A.; GALES, M.. **Predictive uncertainty estimation via prior networks.** Advances in neural information processing systems, 31, 2018.

62 MALININ, A.; GALES, M.. **Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness.** Advances in Neural Information Processing Systems, 32, 2019.

63 WU, Q.; LI, H.; LI, L. ; YU, Z.. **Quantifying intrinsic uncertainty in classification via deep dirichlet mixture networks.** arXiv preprint arXiv:1906.04450, 2019.

64 CHARPENTIER, B.; ZÜGNER, D. ; GÜNNEMANN, S.. **Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts.** Advances in Neural Information Processing Systems, 33:1356–1367, 2020.

65 MOŻEJKO, M.; SUSIK, M. ; KARCZEWSKI, R.. **Inhibited softmax for uncertainty estimation in neural networks.** arXiv preprint arXiv:1810.01861, 2018.

66 NANDY, J.; HSU, W. ; LEE, M. L.. **Towards maximizing the representation gap between in-domain & out-of-distribution examples.** Advances in Neural Information Processing Systems, 33:9239–9250, 2020.

67 OALA, L.; HEISS, C.; MACDONALD, J.; MÄRZ, M.; SAMEK, W. ; KUTYNIOK, G.. **Interval neural networks: Uncertainty scores.** arXiv preprint arXiv:2003.11566, 2020.

68 HSU, Y.-C.; SHEN, Y.; JIN, H. ; KIRA, Z.. **Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data.** In: PROCEEDINGS OF THE IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 10951–10960, 2020.

69 RAGHU, M.; BLUMER, K.; SAYRES, R.; OBERMEYER, Z.; KLEINBERG, B.; MULLAINATHAN, S. ; KLEINBERG, J.. **Direct uncertainty prediction for medical second opinions.** In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, p. 5281–5290. PMLR, 2019.

70 RAMALHO, T.; MIRANDA, M.. **Density estimation in representation space to predict model uncertainty.** In: ENGINEERING DEPENDABLE AND SECURE MACHINE LEARNING SYSTEMS: THIRD INTERNATIONAL WORKSHOP, EDSMLS 2020, NEW YORK CITY, NY, USA, FEBRUARY 7, 2020, REVISED SELECTED PAPERS 3, p. 84–96. Springer, 2020.

71 OBERDIEK, P.; ROTTMANN, M. ; GOTTSCHALK, H.. **Classification uncertainty of deep neural networks based on gradient information.** In: ARTIFICIAL NEURAL NETWORKS IN PATTERN RECOGNITION: 8TH IAPR TC3 WORKSHOP, ANNPR 2018, SIENA, ITALY, SEPTEMBER 19–21, 2018, PROCEEDINGS 8, p. 113–125. Springer, 2018.

72 LEE, J.; ALREGIB, G.. **Gradients as a measure of uncertainty in neural networks.** In: 2020 IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP), p. 2416–2420. IEEE, 2020.

73 DEMPSTER, A. P.. **A generalization of bayesian inference.** Journal of the Royal Statistical Society: Series B (Methodological), 30(2):205–232, 1968.

74 BAO, W.; YU, Q. ; KONG, Y.. **Evidential deep learning for open set action recognition.** In: PROCEEDINGS OF THE IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION, p. 13349–13358, 2021.

75 GHESU, F. C.; GEORGESCU, B.; MANSOOR, A.; YOO, Y.; GIBSON, E.; VISHWANATH, R.; BALACHANDRAN, A.; BALTER, J. M.; CAO, Y.; SINGH, R. ; OTHERS. **Quantifying and leveraging predictive**

uncertainty for medical image assessment. Medical Image Analysis, 68:101855, 2021.

76 ZHAO, J.; LIU, X.; HE, S. ; SUN, S.. **Probabilistic inference of bayesian neural networks with generalized expectation propagation**. Neurocomputing, 412:392–398, 2020.

77 XIA, T.; HAN, J.; QENDRO, L.; DANG, T. ; MASCOLO, C.. **Hybrid-edl: Improving evidential deep learning for uncertainty quantification on imbalanced data**. In: WORKSHOP ON TRUSTWORTHY AND SOCIALLY RESPONSIBLE MACHINE LEARNING, NEURIPS 2022, 2022.

78 VAN AMERSFOORT, J.; SMITH, L.; TEH, Y. W. ; GAL, Y.. **Uncertainty estimation using a single deep deterministic neural network**. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, p. 9690–9700. PMLR, 2020.

79 DO NASCIMENTO, G. H.; EVALD, P. J. D. D. O. ; DREWS, P. L. J.. **Epistemic uncertainty estimation with evidential learning on semantic segmentation of underwater images**. In: 2022 LATIN AMERICAN ROBOTICS SYMPOSIUM (LARS), 2022 BRAZILIAN SYMPOSIUM ON ROBOTICS (SBR), AND 2022 WORKSHOP ON ROBOTICS IN EDUCATION (WRE), p. 1–6. IEEE, 2022.

80 HOLMQUIST, K.; KLASÉN, L. ; FELSBERG, M.. **Evidential deep learning for class-incremental semantic segmentation**. In: SCANDINAVIAN CONFERENCE ON IMAGE ANALYSIS, p. 32–48. Springer, 2023.

81 TONG, Z.; XU, P. ; DENOEUX, T.. **Evidential fully convolutional network for semantic segmentation**. Applied Intelligence, 51:6376–6399, 2021.

82 ZOU, K.; YUAN, X.; SHEN, X.; CHEN, Y.; WANG, M.; GOH, R. S. M.; LIU, Y. ; FU, H.. **Evidencecap: Towards trustworthy medical image segmentation via evidential identity cap**. arXiv preprint arXiv:2301.00349, 2023.

83 LI, H.; NAN, Y.; DEL SER, J. ; YANG, G.. **Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation**. Neural Computing and Applications, p. 1–15, 2022.

84 ZOU, K.; YUAN, X.; SHEN, X.; WANG, M. ; FU, H.. **Tbrats: Trusted brain tumor segmentation**. In: INTERNATIONAL CONFERENCE ON

MEDICAL IMAGE COMPUTING AND COMPUTER-ASSISTED INTERVENTION, p. 503–513. Springer, 2022.

85 GRANNAS, L.. **Real-time uncertainty estimation for semantic segmentation: Improving uncertainty estimates with temperature scaling and predicted dirichlet distributions**, 2020.

86 JOSHI, A. J.; PORIKLI, F. ; PAPANIKOLOPOULOS, N.. **Multi-class active learning for image classification**. In: 2009 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 2372–2379. IEEE, 2009.

87 WANG, K.; ZHANG, D.; LI, Y.; ZHANG, R. ; LIN, L.. **Cost-effective active learning for deep image classification**. IEEE Transactions on Circuits and Systems for Video Technology, 27(12):2591–2600, 2016.

88 HOULSBY, N.; HUSZÁR, F.; GHAHRAMANI, Z. ; LENGYEL, M.. **Bayesian active learning for classification and preference learning**. arXiv preprint arXiv:1112.5745, 2011.

89 ZHDANOV, F.. **Diverse mini-batch active learning**. arXiv preprint arXiv:1901.05954, 2019.

90 DI SCANDALEA, M. L.; PERONE, C. S.; BOUDREAU, M. ; COHEN-ADAD, J.. **Deep active learning for axon-myelin segmentation on histology data**. arXiv preprint arXiv:1907.05143, 2019.

91 GAL, Y.; ISLAM, R. ; GHAHRAMANI, Z.. **Deep bayesian active learning with image data**. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, p. 1183–1192. PMLR, 2017.

92 KIRSCH, A.; VAN AMERSFOORT, J. ; GAL, Y.. **Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning**. Advances in neural information processing systems, 32, 2019.

93 AGHDAM, H. H.; GONZALEZ-GARCIA, A.; WEIJER, J. V. D. ; LÓPEZ, A. M.. **Active learning for deep detection neural networks**. In: PROCEEDINGS OF THE IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION, p. 3672–3680, 2019.

94 REN, P.; XIAO, Y.; CHANG, X.; HUANG, P.-Y.; LI, Z.; GUPTA, B. B.; CHEN, X. ; WANG, X.. **A survey of deep active learning**. ACM computing surveys (CSUR), 54(9):1–40, 2021.

95 YANG, L.; ZHANG, Y.; CHEN, J.; ZHANG, S. ; CHEN, D. Z.. **Suggestive annotation: A deep active learning framework for biomedical image segmentation.** In: INTERNATIONAL CONFERENCE ON MEDICAL IMAGE COMPUTING AND COMPUTER-ASSISTED INTERVENTION, p. 399–407. Springer, 2017.

96 ASH, J. T.; ZHANG, C.; KRISHNAMURTHY, A.; LANGFORD, J. ; AGARWAL, A.. **Deep batch active learning by diverse, uncertain gradient lower bounds.** arXiv preprint arXiv:1906.03671, 2019.

97 HEMMER, P.; KÜHL, N. ; SCHÖFFER, J.. **Deal: Deep evidential active learning for image classification.** Deep Learning Applications, Volume 3, p. 171–192, 2022.

98 MITTAL, S.; NIEMEIJER, J.; SCHÄFER, J. P. ; BROX, T.. **Best practices in active learning for semantic segmentation.** In: GERMAN CONFERENCE ON PATTERN RECOGNITION (GCPR), 2023.

99 SIDDIQUI, Y.; VALENTIN, J. ; NIESSNER, M.. **Viewal: Active learning with viewpoint entropy for semantic segmentation.** In: PROCEEDINGS OF THE IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 9433–9443, 2020.

100 BELUCH, W. H.; GENEWEIN, T.; NÜRNBERGER, A. ; KÖHLER, J. M.. **The power of ensembles for active learning in image classification.** In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 9368–9377, 2018.

101 MAKING, M.-O. D.. **Synthesis lectures on artificial intelligence and machine learning.**

102 LEE, S.; LEE, C.. **Revisiting spatial dropout for regularizing convolutional neural networks.** Multimedia Tools and Applications, 79(45):34195–34207, 2020.

103 MCCLURE, P.; RHO, N.; LEE, J. A.; KACZMARZYK, J. R.; ZHENG, C. Y.; GHOSH, S. S.; NIELSON, D. M.; THOMAS, A. G.; BANDETTINI, P. ; PEREIRA, F.. **Knowing what you know in brain segmentation using bayesian deep neural networks.** Frontiers in neuroinformatics, 13:67, 2019.

104 LEARNED-MILLER, E. G.. **Entropy and mutual information.** Department of Computer Science, University of Massachusetts, Amherst, p. 4, 2013.

105 KULLBACK, S.; LEIBLER, R. A.. **On information and sufficiency**. The annals of mathematical statistics, 22(1):79–86, 1951.

106 CHEN, L.-C.; ZHU, Y.; PAPANDREOU, G.; SCHROFF, F. ; ADAM, H.. **Encoder-decoder with atrous separable convolution for semantic image segmentation**. In: PROCEEDINGS OF THE EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), p. 801–818, 2018.

107 CAO, H.; WANG, Y.; CHEN, J.; JIANG, D.; ZHANG, X.; TIAN, Q. ; WANG, M.. **Swin-unet: Unet-like pure transformer for medical image segmentation**. In: EUROPEAN CONFERENCE ON COMPUTER VISION, p. 205–218. Springer, 2022.

108 ORTEGA ADARME, M.; DOBLAS PRIETO, J.; QUEIROZ FEITOSA, R. ; DE ALMEIDA, C. A.. **Improving deforestation detection on tropical rainforests using sentinel-1 data and convolutional neural networks**. Remote Sensing, 14(14):3290, 2022.

109 ULMER, D.; HARDMEIER, C. ; FRELLSEN, J.. **Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation**. Transactions on Machine Learning Research, 2023.

110 BUTLER, K.. **Band Combinations for Landsat 8**. `https://www.esri.com/arcgis-blog/products/product/imagery/band-combinations-for-landsat-8/`, 2013. [Online].

111 ALMEIDA, C.; MAURANO, L.; VALERIANO, D.; CÂMARA, G.; VINHAS, L.; MOTTA, M.; GOMES, A.; MONTEIRO, A.; SOUZA, A.; MESSIAS, C. ; OTHERS. **Metodologia utilizada nos sistemas prodes e deter-2ª edição (atualizada)**. Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2.

112 FRANKE, J.; BARRADAS, A. C. S.; BORGES, M. A.; COSTA, M. M.; DIAS, P. A.; HOFFMANN, A. A.; OROZCO FILHO, J. C.; MELCHIORI, A. E. ; SIEGERT, F.. **Fuel load mapping in the brazilian cerrado in support of integrated fire management**. Remote Sensing of Environment, 217:221–232, 2018.

113 MESSIAS, C. G.; FERREIRA, M. C.. **Modelo geoespacial para a identificação de áreas com perigo de propagação de queimadas no parque nacional da serra da canastra**. Revista do Departamento de Geografia, 38:154–168, 2019.

114 MARTINEZ, J. A. C.; LA ROSA, L. E. C.; FEITOSA, R. Q.; SANCHES, I. D. ; HAPP, P. N.. **Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences**. ISPRS Journal of Photogrammetry and Remote Sensing, 171:188–201, 2021.

115 MARTINEZ, J. A. C.; OLIVEIRA, H.; DOS SANTOS, J. A. ; FEITOSA, R. Q.. **Open set semantic segmentation for multitemporal crop recognition**. IEEE Geoscience and Remote Sensing Letters, 19:1–5, 2021.

116 CHAMORRO, J.; FEITOSA, R.; HAPP, P. ; BERMUDEZ, J.. **Towards lifelong crop recognition using fully convolutional recurrent networks and sar image sequences**. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 43:923–929, 2021.

117 MARTINEZ, J.; ADARME, M.; TURNES, J.; COSTA, G.; DE ALMEIDA, C. ; FEITOSA, R.. **a comparison of cloud removal methods for deforestation monitoring in amazon rainforest**. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 43:665–671, 2022.

118 ROGOZINSKI, M.; MARTINEZ, J. A. C. ; FEITOSA, R. Q.. **3d convolution for multidate crop recognition from multitemporal image sequences**. International Journal of Remote Sensing, 43(15-16):6056–6077, 2022.

119 ROGOZINSKI, M.; MARTINEZ, J. A. C.; HAPP, P. N. ; FEITOSA, R. Q.. **Exploring temporal context at multiple scales for crop mapping with fully convolutional recurrent nets and fully connected crfs**. In: 2021 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM IGARSS, p. 2305–2308. IEEE, 2021.

120 SANCHES, I.; FEITOSA, R.; MONTIBELLER, B.; DIAZ, P.; LUIZ, A.; SOARES, M.; PRUDENTE, V.; VIEIRA, D.; MAURANO, L.; HAPP, P.; CHAMORRO, J. ; OLDONI, L.. **First results of the lem benchmark database for agricultural applications**. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43:251–256, 2020.

121 CARVALHO, T.; CHAMORRO, J.; OLIVEIRA, H.; DOS SANTOS, J. ; FEITOSA, R.. **Outlier exposure for open set recognition from multitemporal image sequences**. IEEE Geoscience And Remote Sensing Letters, 2023.

122 CHAMORRO, J.; ; COSTA, G. ; FEITOSA, R.. **A semi-automatic methodology for deforestation mapping based on uncertainty estimation for automatic deforestation mapping in the brazilian amazon**. Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023.

123 ENDRES, D. M.; SCHINDELIN, J. E.. **A new metric for probability distributions**. IEEE Transactions on Information theory, 49(7):1858–1860, 2003.

124 DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. ; OTHERS. **An image is worth 16x16 words: Transformers for image recognition at scale**. arXiv preprint arXiv:2010.11929, 2020.

125 RANFTL, R.; BOCHKOVSKIY, A. ; KOLTUN, V.. **Vision transformers for dense prediction**. In: PROCEEDINGS OF THE IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION, p. 12179–12188, 2021.

126 BEVANDIĆ, P.; KREŠO, I.; ORŠIĆ, M. ; ŠEGVIĆ, S.. **Simultaneous semantic segmentation and outlier detection in presence of domain shift**. In: PATTERN RECOGNITION: 41ST DAGM GERMAN CONFERENCE, DAGM GCPR 2019, DORTMUND, GERMANY, SEPTEMBER 10–13, 2019, PROCEEDINGS 41, p. 33–47. Springer, 2019.

127 GENG, C.; HUANG, S.-J. ; CHEN, S.. **Recent advances in open set recognition: A survey**. IEEE transactions on pattern analysis and machine intelligence, 43(10):3614–3631, 2020.

128 BRITO, A.; VALERIANO, D. D. M.; FERRI, C.; SCOLASTRICI, A. ; SESTINI, M.. **Metodologia da detecção do desmatamento no bioma cerrado: Mapeamento de áreas antropizadas com imagens de média resolução espacial**. São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2018.

129 VEGA, P. J. S.; DA COSTA, G. A. O. P.; FEITOSA, R. Q.; ADARME, M. X. O.; DE ALMEIDA, C. A.; HEIPKE, C. ; ROTTENSTEINER, F.. **An unsupervised domain adaptation approach for change detection and its application to deforestation mapping in tropical biomes**. ISPRS Journal of Photogrammetry and Remote Sensing, 181:113–128, 2021.

130 FERREIRA, K. R.; QUEIROZ, G. R.; VINHAS, L.; MARUJO, R. F.; SIMOES, R. E.; PICOLI, M. C.; CAMARA, G.; CARTAXO, R.; GOMES, V. C.; SANTOS, L. A. ; OTHERS. **Earth observation data cubes for brazil: Requirements, methodology and products**. Remote Sensing, 12(24):4033, 2020.

131 SIMOES, R.; CAMARA, G.; QUEIROZ, G.; SOUZA, F.; ANDRADE, P. R.; SANTOS, L.; CARVALHO, A. ; FERREIRA, K.. **Satellite image time series analysis for big earth observation data**. Remote Sensing, 13(13):2428, 2021.

7
Appendix

In the following figures, the RGB composition for the Sentinel-2 input images at inference (T_{-1} and T_0) in the PA (Figures 59 and 60), MT (Figures 61 and 62), MS (Figures 63 and 64), and PI (Figures 65 and 66) sites are presented.

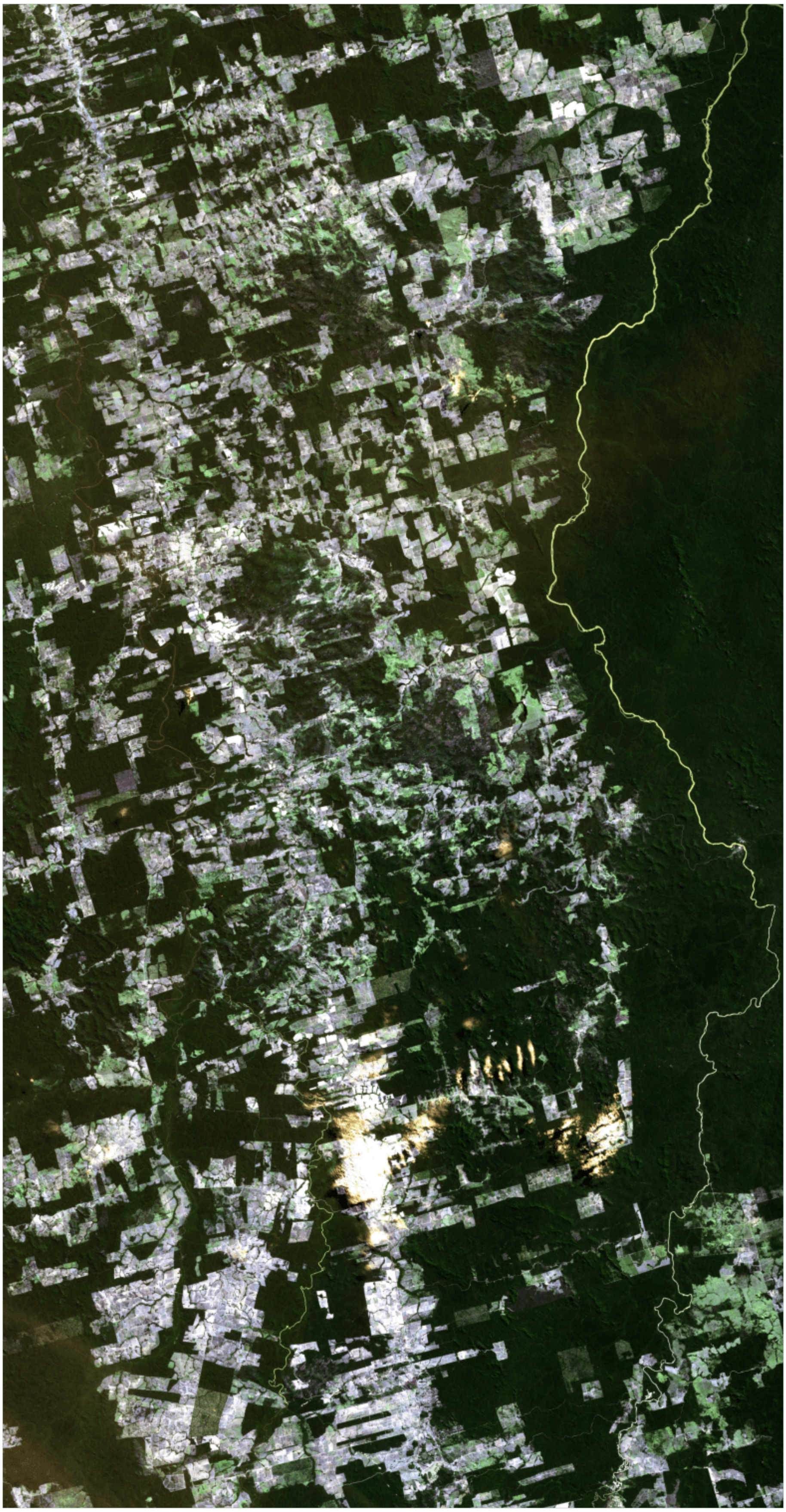


Figure 59: RGB composition for the S2 image in T_{-1}, for the PA site.

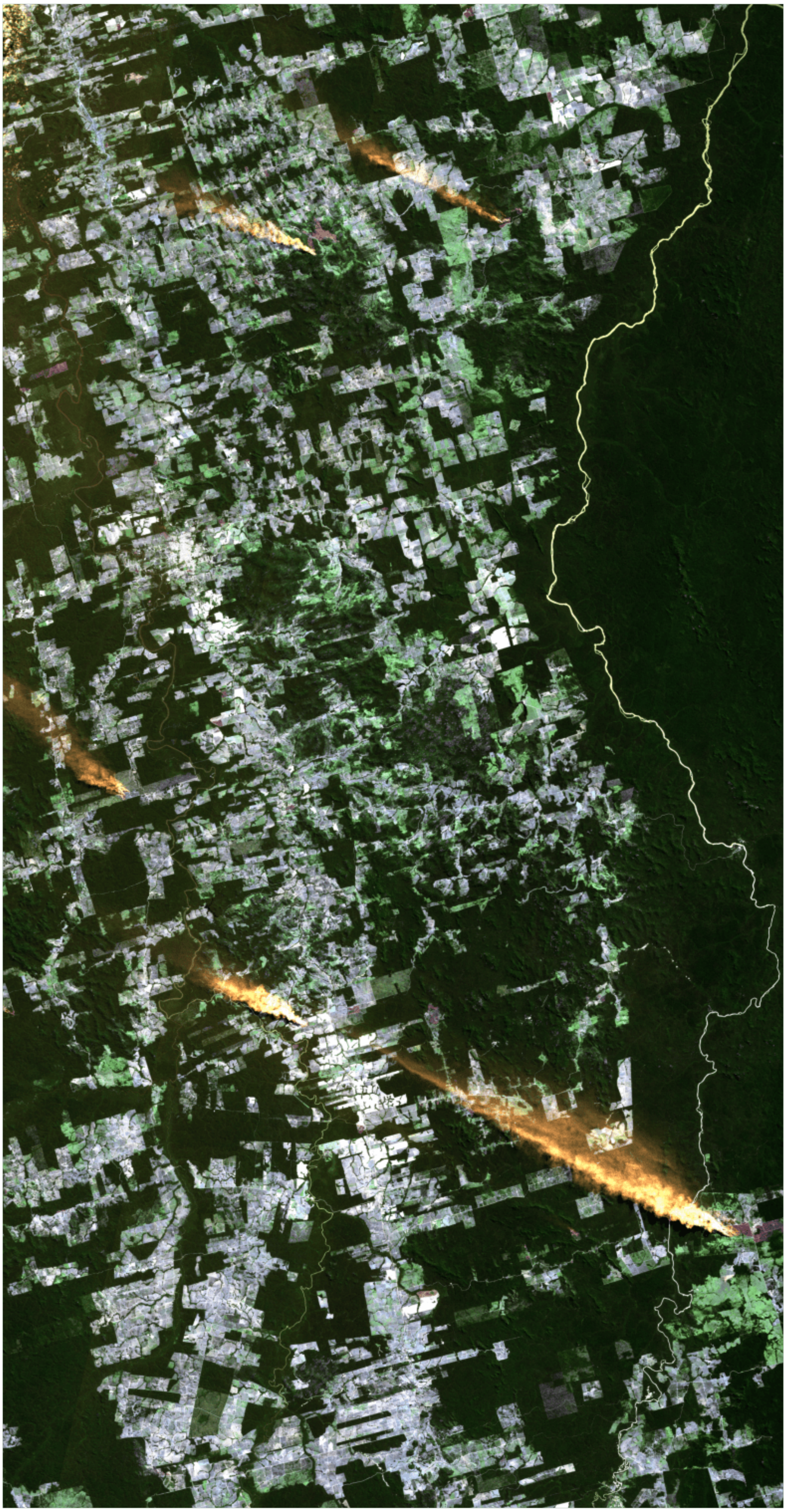


Figure 60: RGB composition for the S2 image in T_0, for the PA site.

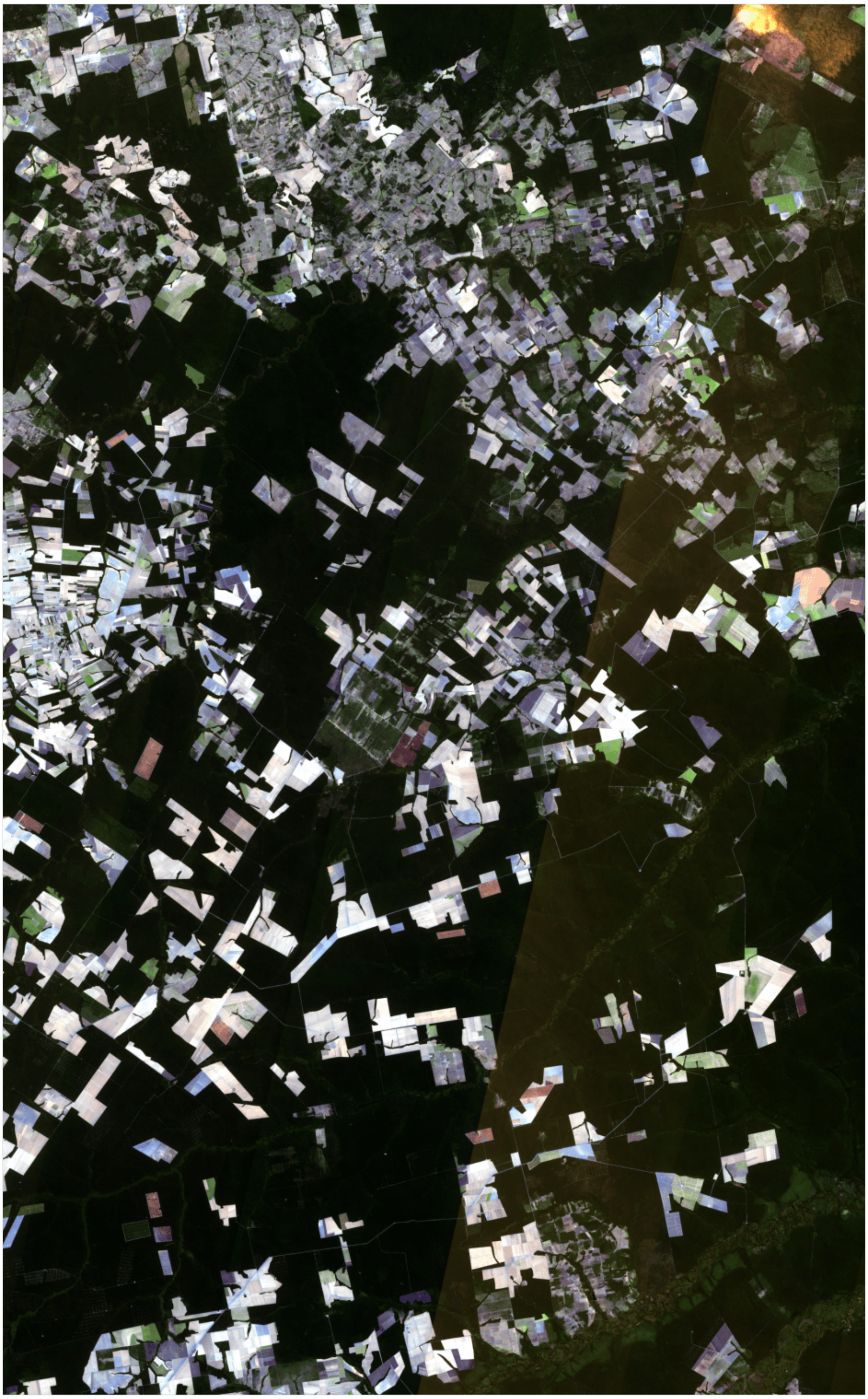


Figure 61: RGB composition for the S2 image in T_{-1}, for the MT site.

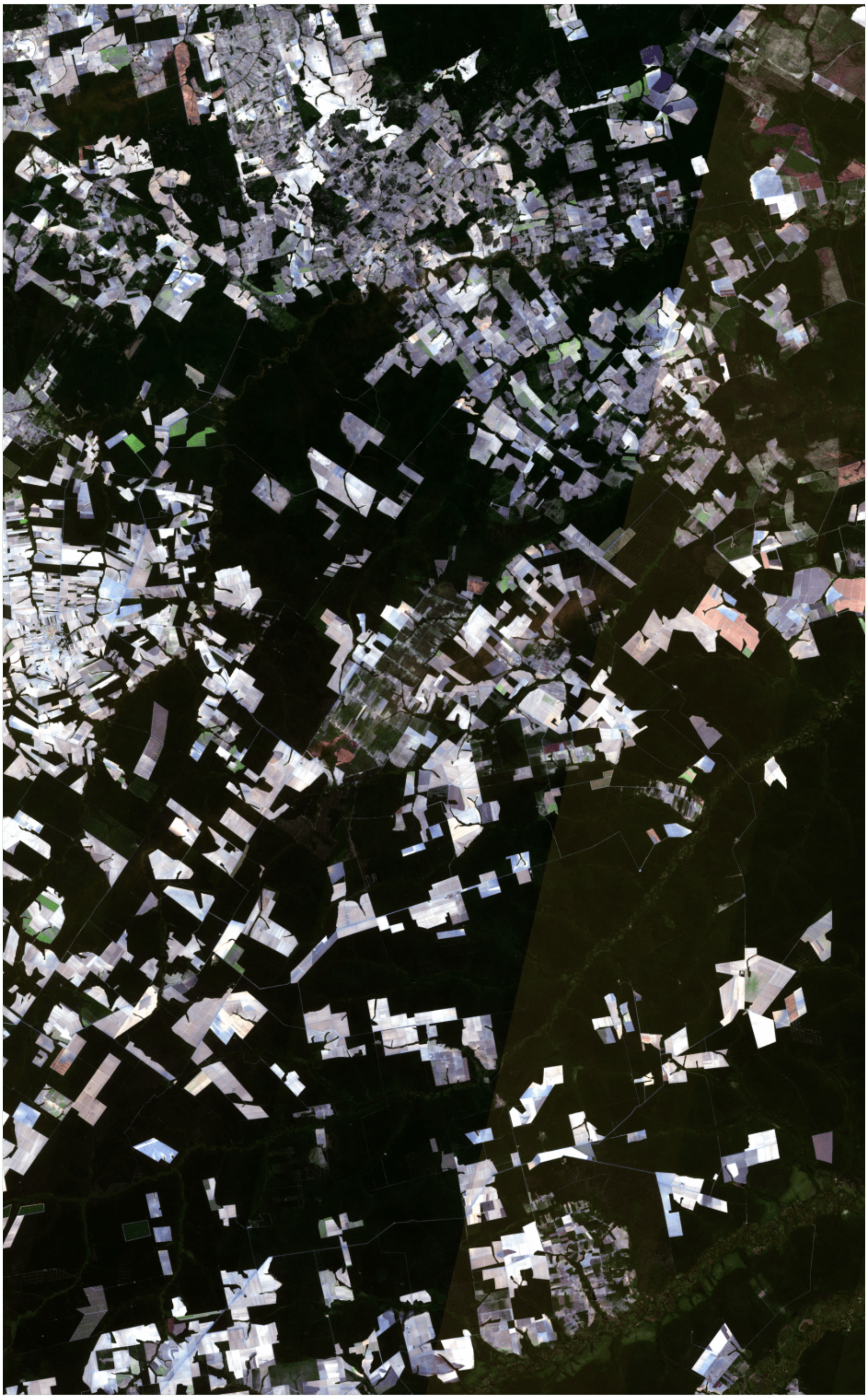


Figure 62: RGB composition for the S2 image in T_0, for the MT site.

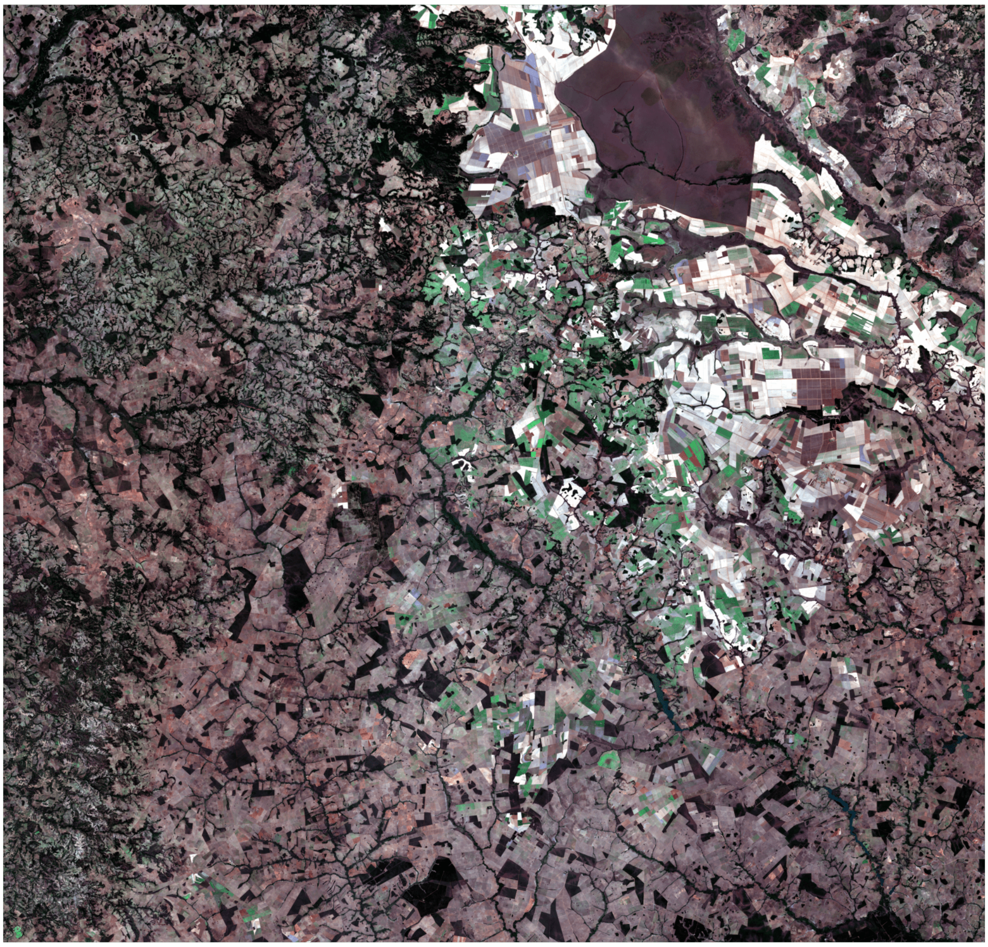


Figure 63: RGB composition for the S2 image in T_{-1}, for the MS site.

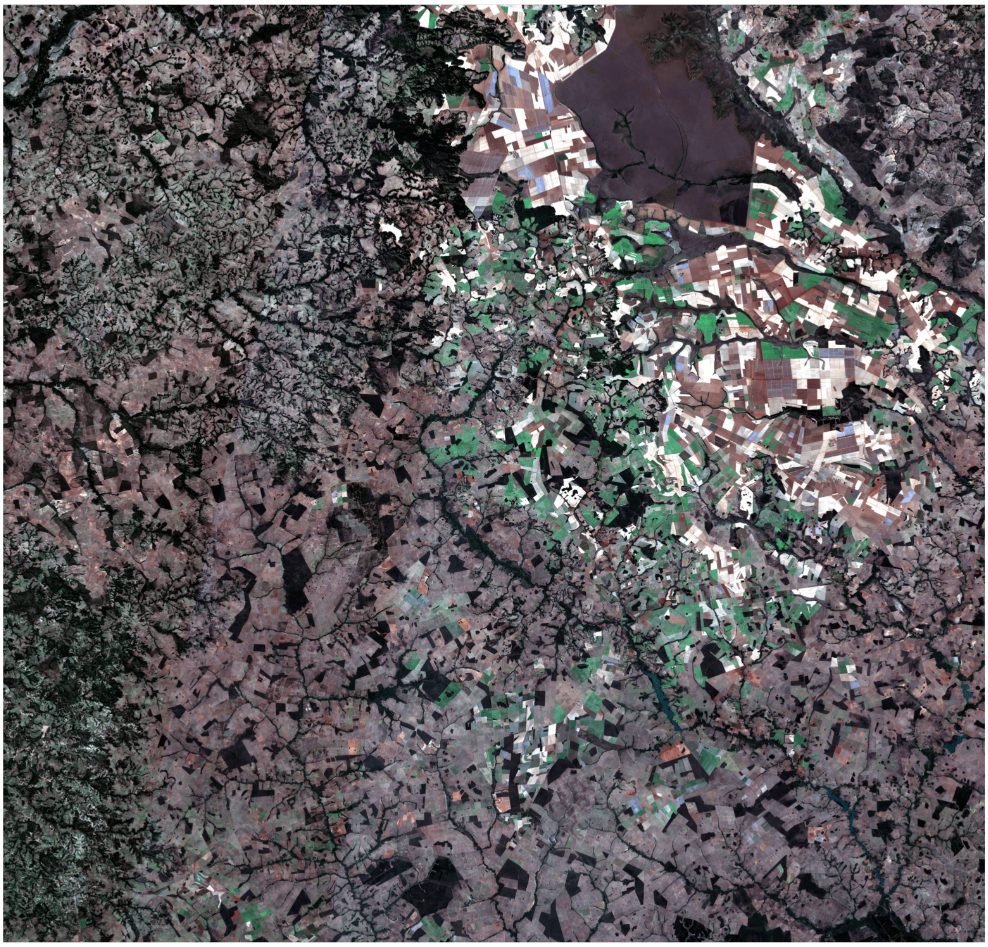


Figure 64: RGB composition for the S2 image in T_0, for the MS site.

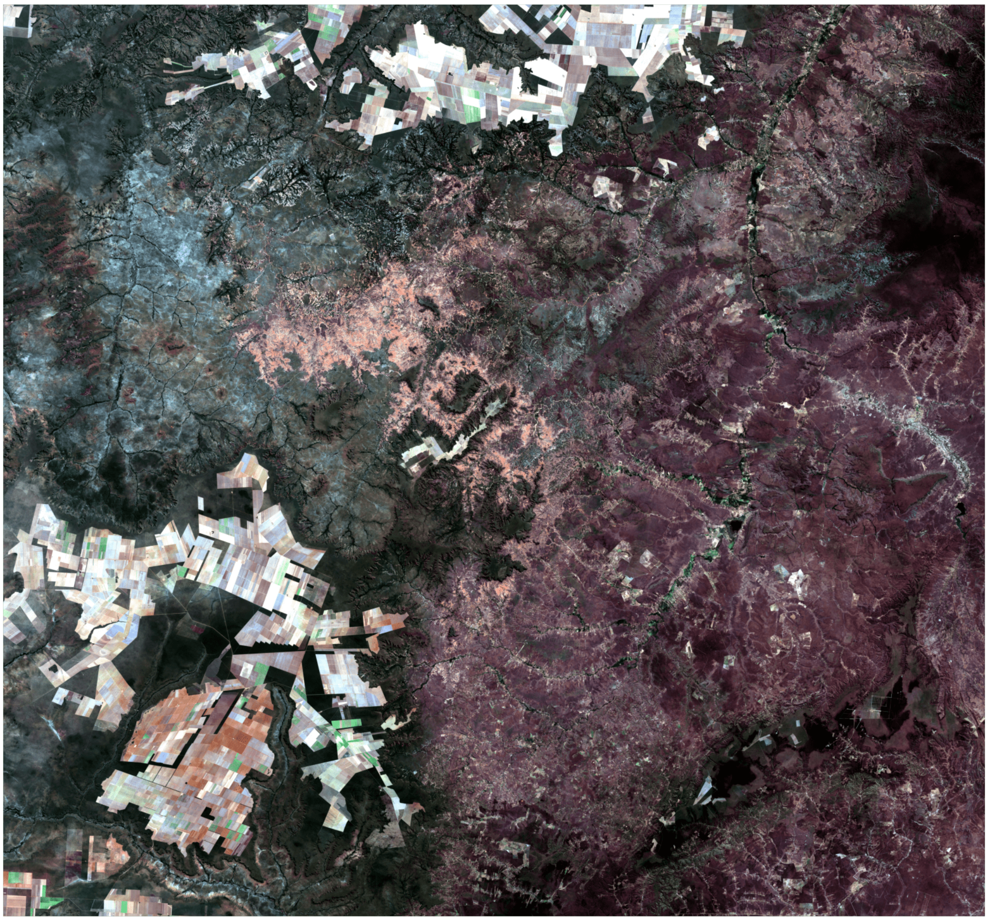


Figure 65: RGB composition for the S2 image in T_{-1}, for the PI site.

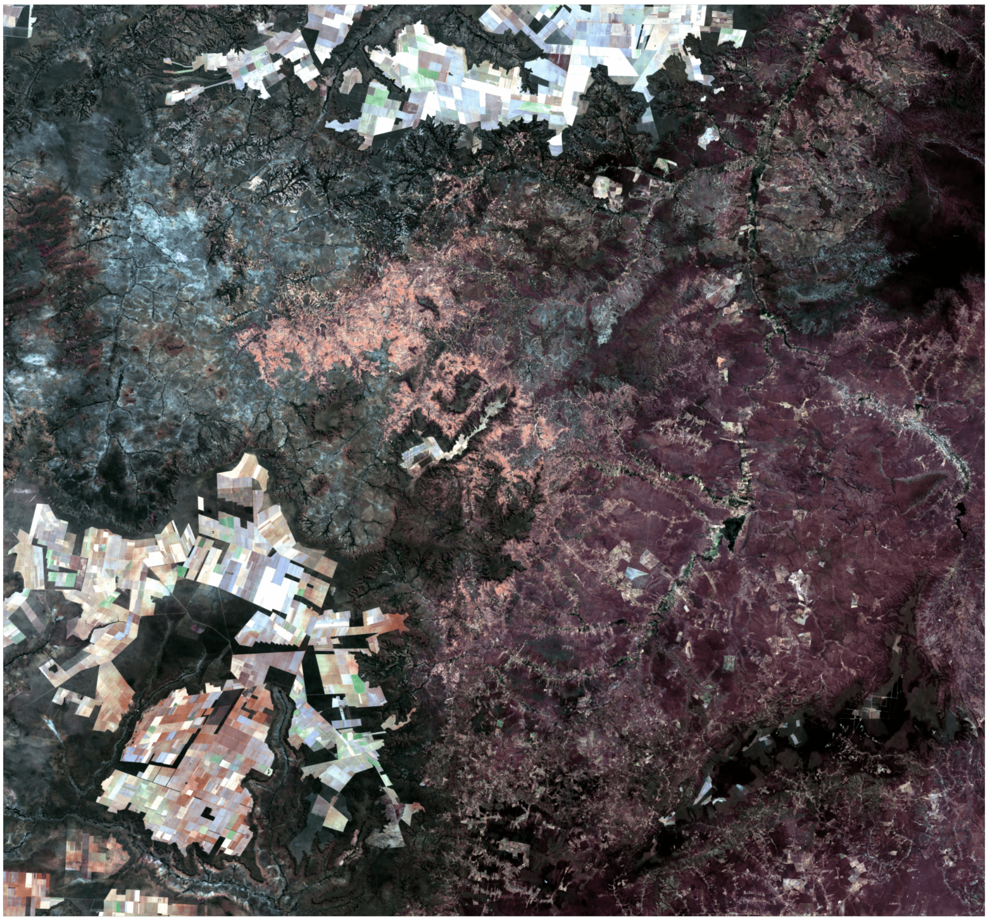


Figure 66: RGB composition for the S2 image in T_0, for the PI site.