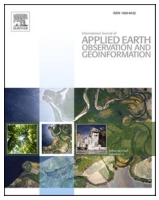
International Journal of Applied Earth Observation and Geoinformation 135 (2024) 104288

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



Local uncertainty maps for land-use/land-cover classification without remote sensing and modeling work using a class-conditional conformal approach

Denis Valle [a,*], Rodrigo Leite [b], Rafael Izbicki [c], Carlos Silva [a], Leo Haneda [a]

[a] School of Forest, Fisheries, and Geomatics Sciences, University of Florida, Gainesville, FL, USA
[b] NASA Postdoctoral Program Fellow, Goddard Space Flight Center, Greenbelt, MD, USA
[c] Department of Statistics, Federal University of Sao Carlos, Sao Paulo, Brazil

ARTICLE INFO

Keywords:
Conformal statistics
Classification uncertainty
Land-use land-cover
LULC
Image classification

ABSTRACT

Land use/land cover (LULC) is one of the most impactful global change phenomenon. As a result, considerable effort has been devoted to creating large-scale LULC products from remote sensing data, enabling the scientific community to use these products for a wide range of downstream applications. Unfortunately, uncertainty associated with these products is seldom quantified because most approaches are too computationally intensive. Furthermore, uncertainty maps developed for large regions might fail to perform adequately at the spatial scale in which they will be used and might need to be customized to suit the specific applications of end-users.

In this study, we describe the class-conditional conformal statistics method, an approach that quantifies uncertainty more uniformly for each class but that requires more calibration data than the conventional conformal method. Using the class-conditional method, we show that it is possible to create customized local uncertainty maps using local calibration data without requiring remote sensing and modeling work and that these local uncertainty maps outperform uncertainty maps calibrated based on global data. We use empirical data from Brazil (i.e., Dynamic World LULC product and Mapbiomas validation data) to demonstrate this methodology. The analysis of these data reveals substantial heterogeneity in observations of the same LULC class between Brazilian states, an indication that national-level data are not representative of the focal state, thus explaining why uncertainty maps calibrated using focal state-level data outperform maps calibrated using national-level data. Importantly, we develop straight-forward approaches to determine the spatial extent over which calibration data are still representative of the area of interest, ensuring that these data can be used to reliably quantify uncertainty. We illustrate the class-conformal methodology by creating uncertainty maps for a selected number of sites in Brazil. Finally, we show how these uncertainty maps can yield valuable insights for LULC map producers.

Our methodology paves the way for users to generate customized local uncertainty maps that are likely to be better than uncertainty maps calibrated based on global data while at the same time being more relevant for the specific applications of these users. A tutorial is provided to show how this methodology can be implemented without requiring remote sensing and modeling expertise to generate uncertainty maps.

1. Introduction

Land-use/land-cover (LULC) change is a pervasive phenomenon across the world and is the main driver of biodiversity and ecosystems integrity loss (Díaz et al. 2019; Tilman et al. 2017). As a result, regional, national, and global LULC maps have become increasingly important inputs for a wide range of downstream environmental science and ecological applications (Canibe et al. 2022; Jain 2020; Lyons et al. 2018; Stehman and Foody 2019). An integral part in the production of these LULC maps is the assessment of their accuracy/quality based on independent reference data, often in the form of error/confusion matrices and the associated user and producer accuracies (Foody 2002, 2012; Khatami et al. 2017; Stehman and Foody 2019). These accuracy metrics are useful to characterize the overall quality of the LULC map but unfortunately they fail to reveal how accuracy varies in space (Brown et al. 2009; Foody 2002; Stehman and Foody 2019). The spatial distribution

* Corresponding author.
E-mail address: drvalle@ufl.edu (D. Valle).

https://doi.org/10.1016/j.jag.2024.104288
Received 13 June 2024; Received in revised form 15 November 2024; Accepted 22 November 2024
Available online 10 December 2024

of accuracy is important because a LULC map might have great overall accuracy but low accuracy over the area of interest of a given map user.

Instead of mapping accuracy, multiple studies have focused on mapping uncertainty based on the class probabilities that are outputted by classification algorithms. One problem with class probabilities is that these probabilities, when generated by common classification algorithms (e.g., random forest and deep learning models), can be poorly calibrated (i.e., the estimated probabilities overestimate the likelihood that those class labels are actually correct) (Guo et al. 2017; Mukhoti et al. 2020; Niculescu-Mizil and Caruana 2005). Indeed, a classifier may be very certain (i.e., outputting a class probability close to 1 for a given class) despite being wrong and thus having low accuracy (Stehman and Foody 2019). Sampling uncertainty is another type of uncertainty and refers to the variability in predictions that arise due to the use of different datasets to train and tune the hyperparameters of the classification model. Sampling uncertainty is often quantified by bootstrapping the data used to train the model (Cheng et al. 2021; Hsiao and Cheng 2016; Lyons et al. 2018; Weber and Langille 2007) but we note that, in the case of stochastic model fitting algorithms, the uncertainty quantified by bootstrapping will include both sampling uncertainty and the variability inherent to these model fitting algorithms. Unfortunately, bootstrapping is often too computationally intensive to be implemented for large-scale LULC maps.

Conformal statistics has recently been proposed as a powerful approach to quantify uncertainty in LULC maps because it is simple to implement, it is not computationally intensive, and it works with any algorithm that outputs class probabilities (Valle et al. 2023). Indeed, the only required assumption is that observations are exchangeable (or the slightly stricter assumption that the observations are independent and identically distributed), a common assumption across the great majority of the machine learning methods (Shafer and Vovk 2008). Importantly, this approach to uncertainty combines class probabilities with information regarding the true LULC classes, thus combining elements of accuracy assessment and more standard uncertainty analysis based on class probabilities.

In this article, we investigate if conformal statistics can be used as an approach to create local uncertainty maps that outperform uncertainty maps calibrate with global data without requiring additional remote sensing and modeling work. By requiring less technical expertise and time, such an approach can allow the creation of local uncertainty maps by end-users, potentially leading to more reliable results and to customized uncertainty maps that are better tailored to the specific needs of each application. We start by describing the conformal statistics approach adopted in this article and how it can be used to generate uncertainty maps using local calibration data. Then, we use empirical data from Brazil (i.e., Dynamic World [DW] LULC product and Mapbiomas validation data) to demonstrate the benefits of this conformal approach. Finally, we illustrate the resulting uncertainty maps for a selected number of sites in Brazil and show how these uncertainty maps can also generate valuable insights for LULC map producers. We end this article by discussing remaining challenges and future research directions.

2. Methodology

2.1. Quantifying uncertainty using the class-conditional conformal approach

As introduced by Valle et al. (2023) in the context of LULC classification, conformal statistics is focused on generating predictive sets with a desired coverage C. A predictive set is a collection of LULC classes for a given pixel. For example, if there are four LULC classes in the landscape (e.g., “forest”, “water”, “urban”, and “agriculture”), the predictive set for a given pixel may consist of only a subset of these LULC classes (e.g., “forest” and “agriculture”). We refer to the frequency with which these predictive sets contain the true classes as empirical coverage. As a result, if the desired coverage C is equal to 95 %, then valid predictive sets should have empirical coverage close to 95 % (i.e., predictive sets should contain the true classes 95 % of the times for new pixels). In other words, the generated predictive sets need to satisfy the following relationship:

$$p\left(Y_{n+1} \in \Gamma_{n+1,C}\right) \geq C \tag{1}$$

where p() stands for probability, Y_{n+1} is the class label for a new pixel, and $\Gamma_{n+1,C}$ is the corresponding predictive set with coverage C. The size of the predictive sets can be used as a measure of LULC classification uncertainty. For example, a predictive set that contains only one class label (e.g., “forest”) indicates smaller classification uncertainty than a predictive set that contains multiple class labels (e.g., “forest”, “agriculture”, and “water”). However, note that predictive sets can also be empty, a situation that represents substantial classification uncertainty as none of the LULC class labels are probable. Importantly, conformal statistics does not rely on asymptotics, makes no assumption about the data generating mechanism (except that data are exchangeable), and focuses on uncertainty in class predictions rather than sampling uncertainty.

An undesirable feature of the conformal approach described in Valle et al. (2023) (onwards simply conventional conformal approach), however, is that the generated predictive sets might overcover certain classes while undercovering other classes. For example, if the desired coverage C is set to 95 %, it is possible that the empirical coverage of the predictive sets is higher for forested pixels and lower for agriculture pixels even if it is close to 95 % across all observations. For example, it could be that the predictive sets for forested pixels always contain the forest class label, resulting in empirical coverage of 100 % (i.e., overcoverage relative to the target C = 95 %). On the other hand, if the predictive sets for agriculture pixels only contain the agriculture class label half the time, then empirical coverage would be only 50 % (undercoverage relative to the target C = 95 %). The conformal approach would ideally avoid this problem by ensuring that the empirical coverage is at least C for each LULC class. This requirement can be described as:

$$p\left(Y_{n+1} \in \Gamma_{n+1,C} | Y_{n+1} = k\right) \geq C \tag{2}$$

where k is the class label. Eq. (2) states that for all pixels of class k, the corresponding predictive sets should contain class k with probability equal to or greater than C.

In this article, we introduce the class-conditional (or label-conditional) conformal approach, a method first proposed by Vovk (2012) that satisfies the requirement that the generated predictive sets have empirical coverage equal to or greater than C for each class. Because both the conventional and the class-conditional conformal approaches are split-conformal approaches, we start by describing the general procedure for split-conformal approaches. We then describe how local uncertainty maps can be created without additional remote sensing and modeling work to finally describe how the implementation of the class-conditional approach differs from that of the conventional conformal approach.

As illustrated in Fig. 1, the first step in the split-conformal approach consists of dividing the ground-truth data into a training dataset and a calibration dataset and the second step consists of fitting the classification model to the training data. Then, in the third step, this classification model is used to predict the probability of each LULC class for the calibration data and the remaining pixels in the study area. In the fourth step, the LULC probabilities and the true LULC for the observations in the calibration dataset are used to determine the criterion that will enable the generation of predictive sets with the desired coverage C. Finally, in the fifth step, LULC probabilities calculated by the classification model in step 3 and the criterion derived in step 4 are used to create predictive sets for all pixels in the study area.

For the purposes of our goal of creating local uncertainty maps without modeling and remote sensing work, we assume that steps one

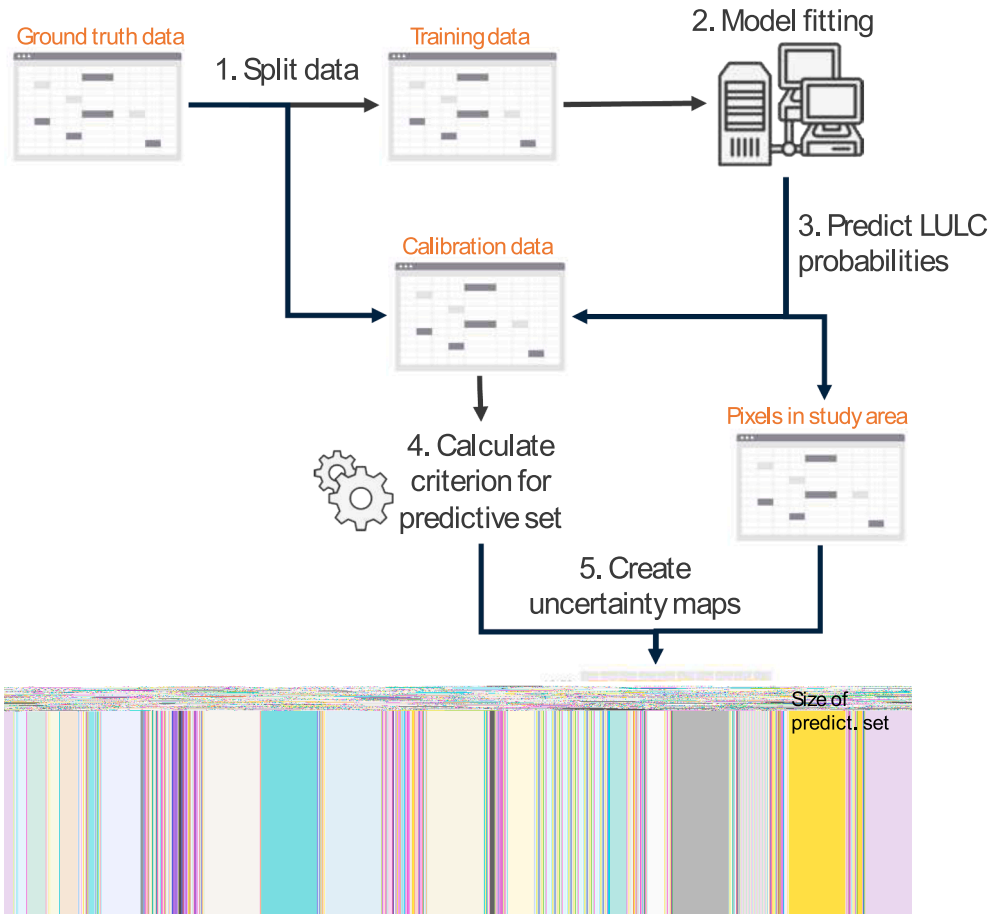


Fig. 1. General steps (numbered 1 through 5) for the split-conformal approach used to generate predictive sets and quantify uncertainty in LULC classification.

through three (see Fig. 1) have already been done. In other words, we assume that a classification model has already been fitted by large-scale LULC map producers and that maps containing the probability associated with LULC class k for each pixel i ($\widehat{p}_{ik}$) are available for the study region (e.g., as in Google's Dynamic World (DW) product; Brown et al. 2022; Venter et al. 2022). We further assume that we have local and independent ground-truth data in which the true LULC class is known for a set of pixels in the study region. This dataset will be used as the calibration dataset. Based on these two inputs (i.e., maps with the probabilities of each LULC class and a local calibration dataset), we show below how the class-conditional conformal approach can be used to create local uncertainty maps without requiring remote sensing and modeling work.

Let the score s_i be the class probability associated with the true class y_i^{true} (i.e., $s_i = \widehat{p}_{i,y_i^{true}}$) for each pixel i in this calibration dataset. For example, as illustrated in the first line of Table 1, say that the class probabilities for pixel i are equal to 0.8, 0.1, and 0.1 for LULC classes 1, 2, and 3, respectively. If this pixel is known to belong to LULC class 1, then the corresponding score will be the probability associated with LULC class 1 (i.e., $s_i = 0.8$). Therefore, assuming that the local calibration dataset contains n observations, we can calculate the score for each observation in this dataset (i.e., $s_1, \cdots, s_n$) as exemplified in Table 1. The main difference between the class-conditional and the conventional conformal approaches is that, in the class-conditional conformal approach, the criterion used to generate predictive sets (step 4 in Fig. 1) is a quantile for each LULC class k, $\widehat{q}_{1-C,k}$, instead of a single quantile over all LULC classes, $\widehat{q}_{1-C}$. For example, for the class-conditional conformal approach, if the desired coverage C is set to 90 %, then we need to calculate the 10 % quantile for the scores associated with each class. In Table 1, we assume that the quantiles calculated based on the subset of scores for each class are equal to 0.4, 0.8, and 0.1 for classes 1, 2, and 3, respectively (i.e., $\widehat{q}_{0.1,1} = 0.4, \widehat{q}_{0.1,2} = 0.8, \widehat{q}_{0.1,3} = 0.1$).

To create uncertainty maps, we need to generate predictive sets for each pixel in the image (step 5 in Fig. 1). We generate the predictive set for pixel i by including class k in the predictive set if the probability for this class is greater than $\widehat{q}_{1-C,k}$(i.e., if $\widehat{p}_{ik} > \widehat{q}_{1-C,k}$). In other words, $\widehat{q}_{1-C,k}$ is a probability threshold because it determines if class k has probability high enough to be included in the predictive set. For example, applying the calculated quantile for class 1 given in the last line of Table 1, we find that class 1 is part of the predictive set for pixels 1, 4, and 8 (Table 2). On the other hand, when using the quantile for class 2 given in Table 1, we find that class 2 is only part of the predictive set for pixel 7 (Table 2). This information can be summarized by determining the size of each predictive set. For example, Table 2 reveals that some pixels in our uncertainty map have just a single class in their

Table 1
Example of the calculation of the scores for hypothetical observations in the calibration dataset. Cells with bold numbers correspond to the probabilities associated with the true classes.

Observations	True class y_i^{true}	Class probabilities $\widehat{p}_{ik}$			Scores s_i for class 1	Scores s_i for class 2	Scores s_i for class 3
		1	2	3			
1	1	**0.80**	0.10	0.10	0.80		
2	1	**0.75**	0.15	0.10	0.75		
3	2	0.00	**0.85**	0.15		0.85	
4	2	0.05	**0.95**	0.00		0.95	
5	3	0.10	0.60	**0.30**			0.30
6	1	**0.25**	0.60	0.15	0.25		
…	…	**…**	…	…	…	…	…
Calculated quantiles ($\widehat{q}_{0.1,k}$)					0.40	0.80	0.10

Table 2
Example of generating 90 % predictive sets for different pixels in the study region. We assume that the quantiles for classes 1, 2, and 3 were calculated to be $\widehat{q}_{0.1,1} = 0.4$, $\widehat{q}_{0.1,2} = 0.8$, and $\widehat{q}_{0.1,3} = 0.1$, respectively, based on the local calibration data. Cells with bold numbers correspond to outcomes that satisfy the inequality $\widehat{p}_{ik} \geq \widehat{q}_{0.1,k}$.

Pixel	Class probabilities $\widehat{p}_{ik}$			90 % Predictive sets	Size of 90 % Predictive sets
	1 ($\widehat{q}_{0.1,1} = 0.4$)	2 ($\widehat{q}_{0.1,2} = 0.8$)	3 ($\widehat{q}_{0.1,3} = 0.1$)		
1	**0.85**	0.04	**0.11**	{1,3}	2
2	0.25	0.25	**0.50**	{3}	1
3	0.10	0.10	**0.80**	{3}	1
4	**0.75**	0.00	**0.25**	{1,3}	2
5	0.20	0.60	**0.20**	{3}	1
6	0.20	0.75	0.05	{}	0
7	0.15	**0.9**	0.05	{2}	1
8	**0.50**	0.45	0.05	{1}	1

predictive sets (e.g., pixels 2, 3, 5, 7, and 8), indicating low classification uncertainty. On the other hand, some pixels have 2 classes in their predictive sets (e.g., pixels 1 and 4), indicating larger classification uncertainty, and one pixel has an empty predictive set (i.e., pixel 6), indicating substantial uncertainty.

How does one determine the amount of calibration data that is required? The following expression, derived by Marques (2023), describes the distribution of empirical coverage $\widehat{C}^{(n,C)}$ as a function of the desired coverage C and size of the calibration dataset n:

$$\widehat{C}^{(n,C)} \sim Beta(\lceil C(n+1) \rceil, \lfloor (1-C)(n+1) \rfloor) \quad (3)$$

In this expression, the symbols $\lceil\ \rceil$ and $\lfloor\ \rfloor$ represent the ceiling and floor functions and Beta() refers to the Beta distribution. This expression reveals that empirical coverage is, on average, approximately equal to the desired coverage C (i.e., the mean of the beta distribution is $E\left[\widehat{C}^{(n,C)}\right] = \frac{\lceil C(n+1) \rceil}{\lceil C(n+1) \rceil + \lfloor (1-C)(n+1) \rfloor} \approx C$). Furthermore, the variance of the empirical coverage is given by $Var\left[\widehat{C}^{(n,C)}\right] = \frac{\lceil C(n+1) \rceil \lfloor (1-C)(n+1) \rfloor}{(\lceil C(n+1) \rceil + \lfloor (1-C)(n+1) \rfloor)^2 (\lceil C(n+1) \rceil + \lfloor (1-C)(n+1) \rfloor + 1)} \approx \frac{C(1-C)}{(n+2)}$, revealing that, as expected, it decreases as the size of the calibration data n increases. Eq. (3) is useful because it enables one to calculate the required calibration data sample size n once the desired coverage C and the range of acceptable values of $\widehat{C}^{(n,C)}$ have been defined. This calculation is straight-forward to implement in any software that has functions to evaluate the cumulative density function (CDF) of a beta distribution (e.g., R; R Core Team 2020). For example, if we want empirical coverage to be on average equal to 95 % and to be between 93 % and 97 % with probability of 0.99, then the number of observations n in the calibration dataset has to be equal to, or greater than, 785 pixels.

Note that, in the class conditional method, the class-specific quantile for LULC class k is calculated based on the pixels in the calibration data that belong to class k (e.g., see Table 1). On the other hand, the conventional conformal approach calculates a single quantile based on all the calibration data. This is an important distinction because it reveals that the overall amount of calibration data required by the class conditional method is much larger than that required for the conventional conformal approach. For example, based on our results derived from Eq. (3), we would need 785 pixels for each LULC class k if we were using the class-conditional conformal approach whereas we would only need 785 pixels overall if we were using the conventional conformal approach. To help readers implement the class-conditional conformal methodology, we created an R tutorial that provides all the relevant code and explains each of the steps required to run this procedure and to calculate the required amount of calibration data (Appendix 2).

2.2. Empirical data

The creation of uncertainty maps without requiring remote sensing or modeling work is based on two main inputs: a) LULC class probabilities for all pixels in the area of interest; and b) independent ground-truth data to be used as calibration data. We rely on the 2018 LULC product provided by Google's Dynamic World (DW) product (Brown et al. 2022; Venter et al. 2022) because it is one of the few large-scale LULC maps that provide class-specific probabilities. In relation to ground-truth data, we rely on the data used by Mapbiomas to validate their annual LULC classification products for Brazil (freely available at https://mapbiomas.org/pontos-de-validacao). These data were created by visually inspecting satellite imagery for each year between 1985 and 2018. Pixels were selected for inspection based on stratified random sampling and each pixel was evaluated by 3 independent analysts (Souza et al. 2020). For our purposes, we only used pixels from 2018 for which the 3 analysts agreed on the LULC class (representing about 71 % of the original data) to avoid introducing additional uncertainty associated with inconsistent reference class labels.

One of the challenges associated with using the DW product together with the Mapbiomas dataset consists of the fact that these products rely on different LULC classification schemes. To harmonize the LULC classes in these products, we performed an exploratory analysis in which we assessed the DW probabilities for the different Mapbiomas classes. This revealed that the DW classes "water", "built", "crops", "trees" and "grass", consistently had high probabilities in areas classified as "water", "urban", "temporary crop", "forest", and "pasture"+"cropland", respectively, in the Mapbiomas map, indicating good agreement. On the other hand, the Mapbiomas classes "savanna", "perennial crop", "wetland" did not match well with any single DW class or combination of DW classes. As a result, these Mapbiomas classes were lumped into a single "other" class. Similarly, the DW probabilities associated with the "shrub_and_scrub", "flooded_vegetation", "snow_and_ice", and "bare" were summed and became the probability associated with the "other" class. A more detailed description of this LULC class harmonization process is provided in Appendix 1. Finally, we restricted our analyses to the Brazilian states that have at least 2,000 observations in the ground-truth Mapbiomas data.

We illustrate how a close examination of high uncertainty pixels can generate novel insights for LULC map producers by overlaying these pixels with a high spatial-resolution (i.e., 5 x 5 m pixels) image from the microsatellite VENμS (a product of the partnership between the Israel Space Agency and the CNES (French Space Agency). This image was collected in 2018 (the same year as the LULC data) from a region north of Manaus (the capital city of Amazonas state in Brazil). More information about this satellite and how to access its images are available at the Theia Land Data Centre (https://www.theia-land.fr/en/product/venus/#toc0).

2.3. Comparison of the class-conditional vs. conventional conformal approaches using state-level empirical data

The goal of this analysis is to determine if the class-conditional conformal approach can indeed generate predictive sets that have the desired coverage for each class, thus resulting in more uniform uncertainty quantification when compared to the conventional conformal approach. We randomly sampled 1,000 observations as calibration data and 500 additional observations as validation data for each state. Based on this calibration data, we used both conformal approaches to create predictive sets for the observations in the validation dataset. This allowed us to compare the empirical coverage of the predictive sets generated by the conventional vs. class-conditional conformal approaches. Note that there is no overlap between the validation and calibration datasets and that both of these datasets arise from the same focal state.

To ensure that our results are robust, for each state we perform 50

simulations, where different calibration and validation datasets were drawn in each simulation. Finally, we set C to 90 % and we only calculate empirical coverage for LULC classes that had at least 20 observations in the validation dataset. Requiring at least 20 observation per class is important to generate robust empirical coverage results, avoiding coverage from fluctuating considerably due to small sample sizes.

2.4. Comparison of uncertainty maps calibrated using state-level vs. national-level data and the class-conditional approach

The goal of this analysis is to determine if, and by how much, uncertainty maps calibrated using local data are better than maps created using global data. For the purposes of this analysis, we assume that local data consist of state-level data, mimicking the situation where a state agency collects these data to be able to create a customized state-level uncertainty map. On the other hand, we assume that global uncertainty maps are created using data from across Brazil, mimicking the situation in which a map producer (e.g., Mapbiomas) creates a national uncertainty map that is subsequently used by the state agency.

We follow a very similar methodology to the one described in section 2.3 but we only rely on the class-conditional conformal approach. The state-level calibration data consisted of 1,000 randomly sampled observations from the focal Brazilian state whereas the national calibration dataset had 10x more observations, consisting of 10,000 observations randomly sampled from across Brazil. The validation data consisted of 500 additional randomly selected observations for the focal state. We used these calibration datasets to create predictive sets for the validation observations. Note that there is no overlap between the validation and calibration datasets. Furthermore, to ensure that our results are robust, for each state we performed 50 simulations, where different calibration and validation datasets are drawn in each simulation, and we set C to 90 %. Finally, we only calculate empirical coverage for the LULC classes that had at least 20 observations in the validation dataset to ensure that the empirical coverage results are robust.

2.5. Determining the spatial extent over which calibration data are still representative of the area of interest

It is important to note that we define local calibration data as data from locations close to the study site (i.e., it does not necessarily mean just data from within the study site). Determining how close these data need to be from the location of interest will depend on how the exchangeability assumption is impacted by spatial extent. In other words, how small does the spatial extent have to be for the observations to still be representative of the area for which predictions are desired? Assuming that calibration data are available for a much larger area than just the study region, we can answer this question by creating buffer areas of different sizes around the study region to then determine how empirical coverage for validation observations changes as spatial extent is increased. As spatial extent is increased around the study region, the amount of available calibration data increases, decreasing the theoretical variance in empirical coverage (Marques 2023), but data are likely to be less representative for the study region (potentially resulting in "biased" predictive sets; sets that have mean empirical coverage different from the desired level). Understanding this "bias-variance" tradeoff is critical to determining the ideal spatial extent of the calibration data.

We illustrate the role of spatial extent on empirical coverage by selecting the state of Mato Grosso do Sul (MS) as our area of interest. More specifically, we assess how well the class conformal approach works when based on calibration data gathered from increasingly larger regions. To create these regions, we gather calibration data over increasingly larger buffer areas around the MS state (i.e., buffer distance was increased from 0 to 2400 km in 400 km increments). To quantify empirical coverage, we create a validation dataset with observations that were not part of the calibration data. The validation data contained 50 randomly selected observations of each LULC class within the area of interest. Finally, we performed this analysis 100 times, each time randomly selecting different sets of calibration and validation data. We set the target coverage to 90 % (i.e., C = 90 %). All of our analyses and figures were done in R (R Core Team 2020).

3. Results

3.1. Comparison of the class-conditional vs. conventional conformal approaches using state-level empirical data

Recall that we compare the class-conditional conformal approach to the conventional conformal approach using data from each state. This comparison reveals that the conventional conformal approach generates predictive sets that have the desired overall empirical coverage but this approach often under or over covers certain LULC classes (red boxes in Fig. 2). On the other hand, we find that the class-conditional conformal approach ensures more uniform performance of the generated predictive sets, yielding overall and per class empirical coverage close to 90 % (black boxes in Fig. 2).

3.2. Comparison of uncertainty maps calibrated using state-level and national-level data and the class-conditional approach

In this section, we compare the uncertainty maps calibrated using state-level and national-level data. We find that the proposed approach to quantify uncertainty using state-level data worked well. More specifically, the 90 % predictive sets created using state-level data generally encompassed the true class for the validation observations 90 % of the time for each class and state (black boxes in Fig. 3). These results were consistent across all 8 Brazilian states selected for analysis (i.e., the states with at least 2,000 MapBiomas observations). On the other hand, uncertainty maps calibrated based on national-level data did not perform as well, with occasional over or under coverage, regardless if empirical coverage was assessed for each class or across all classes (red boxes in Fig. 3). These results confirm that uncertainty maps created using local calibration data (i.e., state-level data) can result in improved uncertainty quantification when compared to uncertainty maps created using global calibration data (i.e., national-level data).

Although it is intuitive that uncertainty maps calibrated with local data will generate better results than uncertainty maps calibrated to global data, it is critical to quantify how much better the uncertainty analysis based on local data is because this will determine if it is worthwhile to create a local uncertainty map or if one should rely on readily available uncertainty maps calibrated with global data. Indeed, while for some states empirical coverage results were not substantially worse (e.g., Mato Grosso state in Fig. 3), for other states (e.g., Minas Gerais and Bahia) the performance of the predictive sets deteriorated substantially. Critically, the uncertainty map calibrated using the national-level data often resulted in worse uncertainty quantification despite relying on a dataset that was 10x larger than the state-level data.

3.3. Why are the results worse for the uncertainty map calibrated with national-level data when compared to the map calibrated with state-level data?

Recall that the threshold probabilities $\widehat{q}_{1-C,k}$ are used to determine which LULC classes belong to the predictive set. We set C to 90 % and find that there is considerable variability between states regarding the threshold probabilities for each LULC class (Fig. 4). For example, the threshold for the "built" class is equal to 0.69 for Amazonas (AM) but is equal to 0.05 for MatoGrosso do Sul (MS). These results suggest that the "built" class is well predicted in Amazonas because 90 % of the areas identified by MapBiomas as belonging to this class tend to have a high

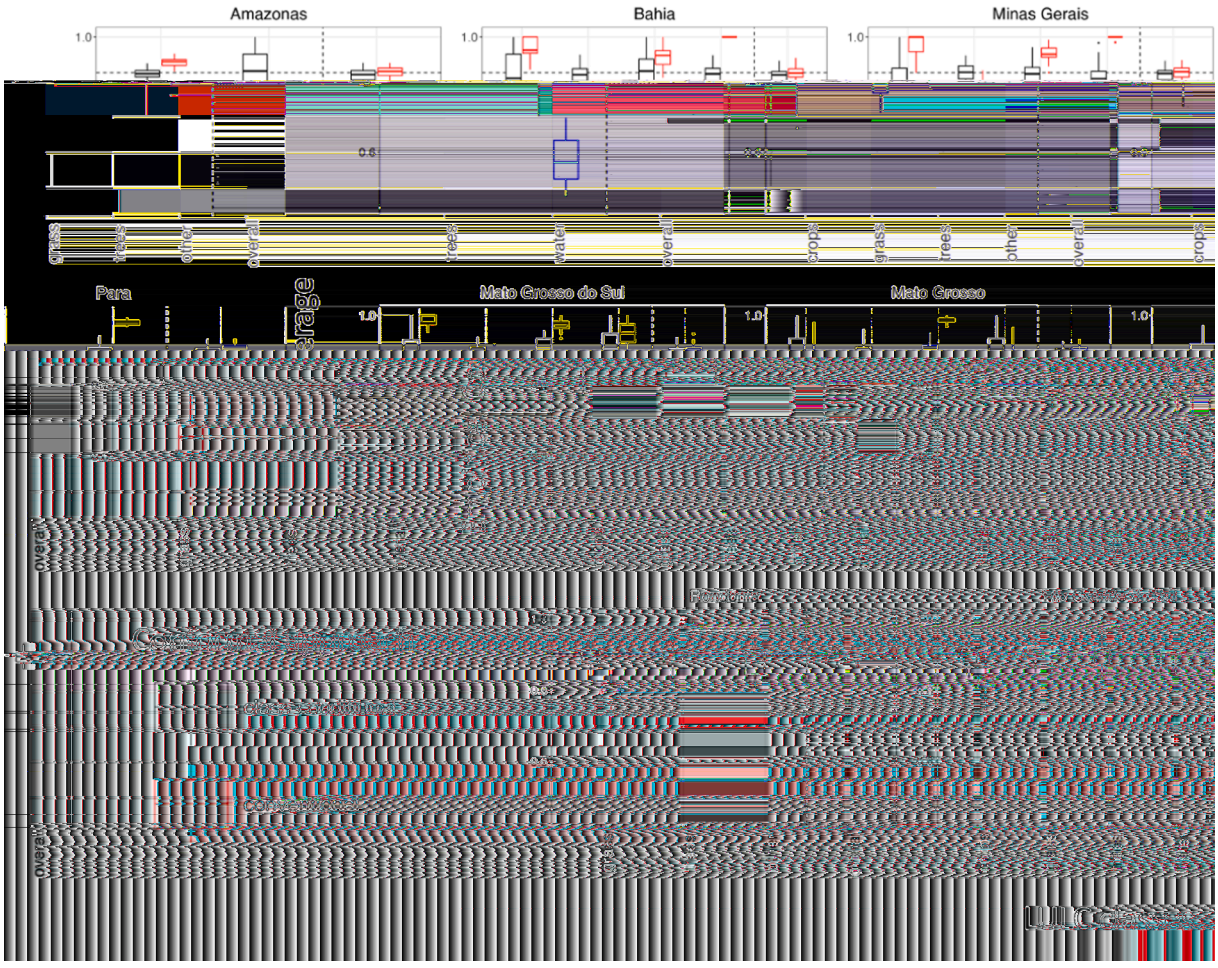


Fig. 2. Uncertainty quantification based on the class-conditional conformal approach (black boxes) avoids under or over-covering specific classes, differently from the conventional conformal approach (red boxes). Each panel shows the results for a distinct Brazilian state. Empirical coverage results are based on 50 validation datasets per state and we only report results for land use/land cover (LULC) classes that had at least 20 observations in the validation dataset. Horizontal dashed line shows the desired coverage level (C = 90 %) whereas the vertical dashed line separates the class-specific empirical coverage from the overall empirical coverage results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

DW "built" probability (>0.69). On the other hand, 90 % of the areas identified to be "built" by MapBiomas in MatoGrosso do Sul have a relatively low DW "built" probability (>0.05). The fact that the same LULC class is assigned very different threshold probabilities depending on the state suggests that there is considerable variability between states that is unaccounted for by the DW algorithm.

Notice that, had we used a global dataset to quantify uncertainty, we would have a single threshold probability for each LULC class, ignoring the inherent heterogeneity between states shown in Fig. 4. In short, the reason that national-level calibration data result in predictive sets with worse performance is because these data are often not representative of the focal state regardless of the number of observations, violating the exchangeability assumption underlying the conformal approach. The violation of this assumption ultimately results in a deterioration of the empirical coverage results of the generated predictive sets.

Given these results, a natural follow-up question is how small does the spatial extent have to be for the observations to still be representative of the area for which predictions are desired? We illustrate how this question can be answered by assuming that the area of interest is the state of Mato Grosso do Sul (MS) and performing a validation exercise to determine how empirical coverage changes as calibration data are collected over increasingly larger buffer areas around the area of interest.

As the spatial extent is increased, the size of the calibration dataset increases and the theoretical variance of the empirical coverage decreases. For example, if C = 90 % and sample size of the calibration data increases from 10, to 100, to 1000, the theoretical variance of the empirical coverage (determined by Eq. (3) decreases from 0.0075, to 0.0009, to 0.0001, respectively. However, empirical coverage results can often be increasingly different from the target coverage level as we include calibration data from locations that are farther and farther away from the area of interest, as illustrated in Fig. 5. Importantly, the maximum spatial extent for the calibration data to still be representative of the area of interest varies according to LULC class. For example, empirical coverage for trees tend to deteriorate if calibration data are gathered from regions farther than 400 km from the study region (Fig. 5c). On the other hand, empirical coverage levels for grass starts to differ from the desired coverage of 90 % only if calibration data arise from areas farther than 1,200 km (Fig. 5b).

Note that the variability in Fig. 5 does not decrease substantially with sample size as expected based on the theoretical variance derived from Eq. (3). The reason for this is because Eq. (3) provides the theoretical distribution of empirical coverage when conformal statistics is used on new training and calibration datasets with exchangeable observations. In contrast, the results displayed in Fig. 5 are based on calibration and validation data that are repeatedly sampled from the same original dataset and some of these data are clearly not exchangeable (as evidenced in Fig. 5 by empirical coverage being substantially different from the desired coverage).

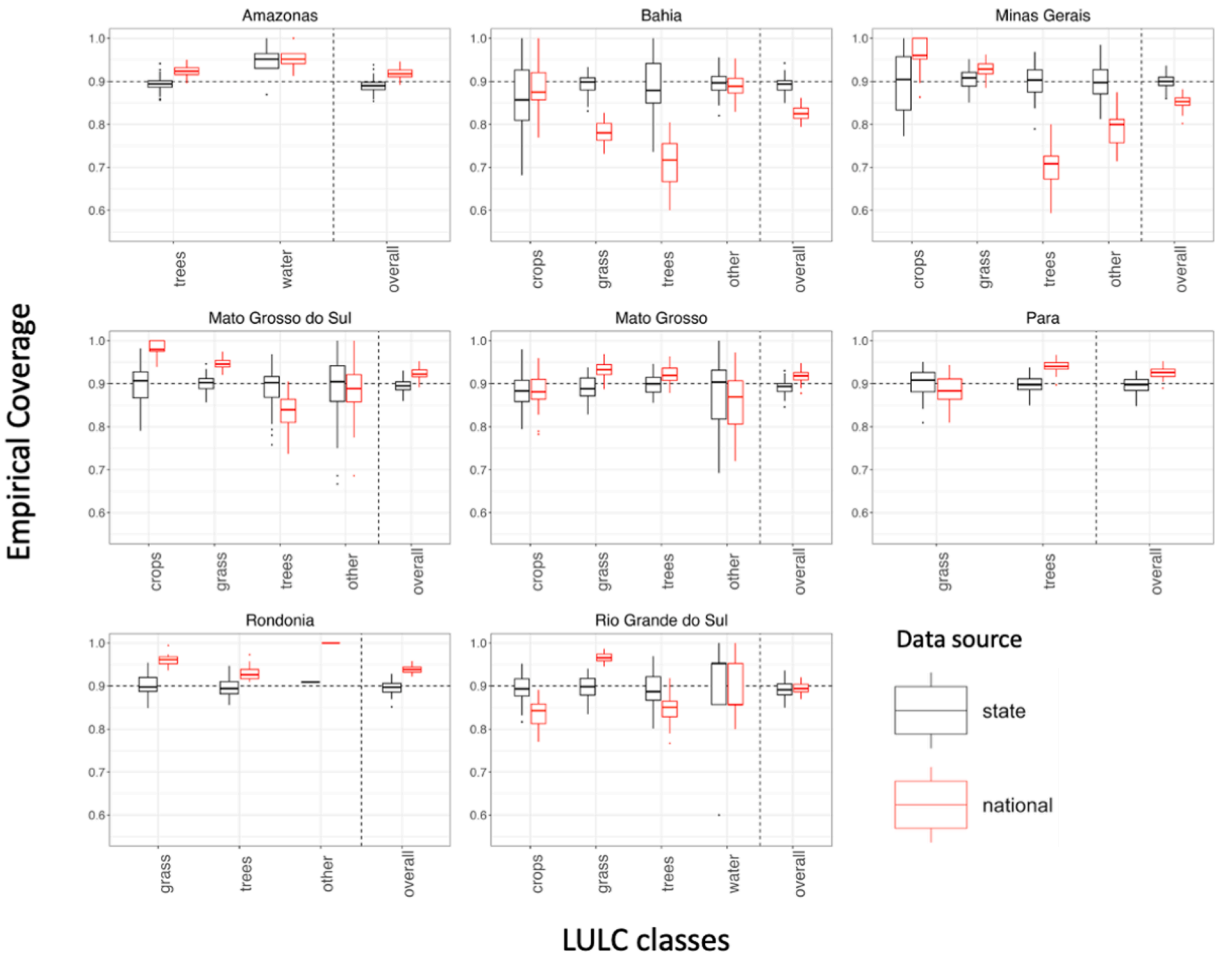


Fig. 3. Uncertainty maps calibrated using state-level data (black boxes) outperform uncertainty maps calibrated using national-level data (red boxes). Each panel shows the results for a given focal state. Empirical coverage results are based on 50 validation datasets per state and we only report results for LULC classes that had at least 20 observations in the validation dataset. Horizontal dashed line shows the desired coverage level (C = 90 %) whereas the vertical dashed line separates the class-specific from the overall empirical coverage results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

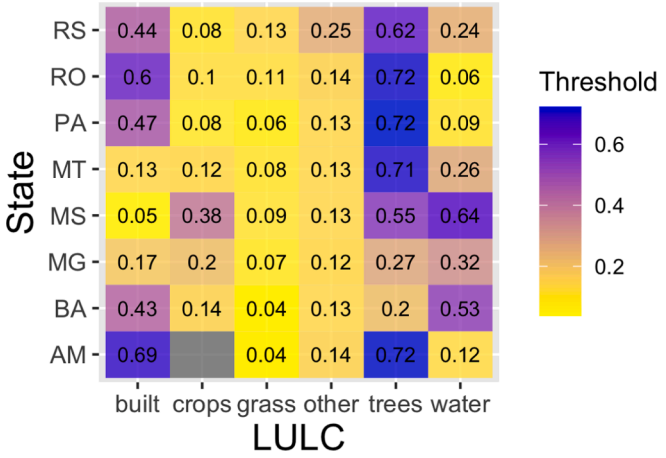


Fig. 4. Threshold probabilities $\widehat{q}_{1-C,k}$ for each state and LULC class, where C was set to 90%. Note that there is no information for crops in Amazonas state (AM) because there was no ground-truth data in this state for this particular LULC class. The state acronyms RS, RO, PA, MT, MS, MG, BA, and AM stand for Rio Grande do Sul, Rondonia, Para, Mato Grosso, Mato Grosso do Sul, Minas Gerais, Bahia, and Amazonas, respectively.

3.4. Illustration of the class-conditional conformal results

Based on the 90 % predictive sets created for the Mapbiomas dataset, we find that empty predictive sets were generally rare, always being less than 5 % of the observations for any given state and LULC class. After removing these empty sets, our results show that the LULC class which consistently had the largest predictive sets across all states was “crops”, suggesting that this is the most uncertain LULC class (Table 3). Furthermore, the predictive sets for “crop” pixels often included “grass” and “other”, indicating that it is often hard to distinguish between these LULC classes. These results agree with the very low threshold probabilities for “crops” and “grass” shown in Fig. 4, which suggest that these classes are hard to classify. Finally, the LULC class with smallest predictive sets on average was “trees” followed by “built”.

To illustrate the uncertainty maps generated by the class-conditional conformal approach, we selected areas in three different states (Rio Grande do Sul [RS], Amazonas [AM], and Mato Grosso [MT]) and show both the DW LULC classification map (Fig. 6a-c) and the corresponding uncertainty maps (Fig. 6d-l). Left to right panels in Fig. 6 show a gradient of increasing overall classification uncertainty. Interestingly, we find that the uncertainty maps can be relatively different from the classification maps because uncertainty depends on the individual class probabilities at each pixel. As a result, there can be considerable spatial heterogeneity in the amount of uncertainty within areas that seem very homogeneous based on the DW LULC product. For example, there are

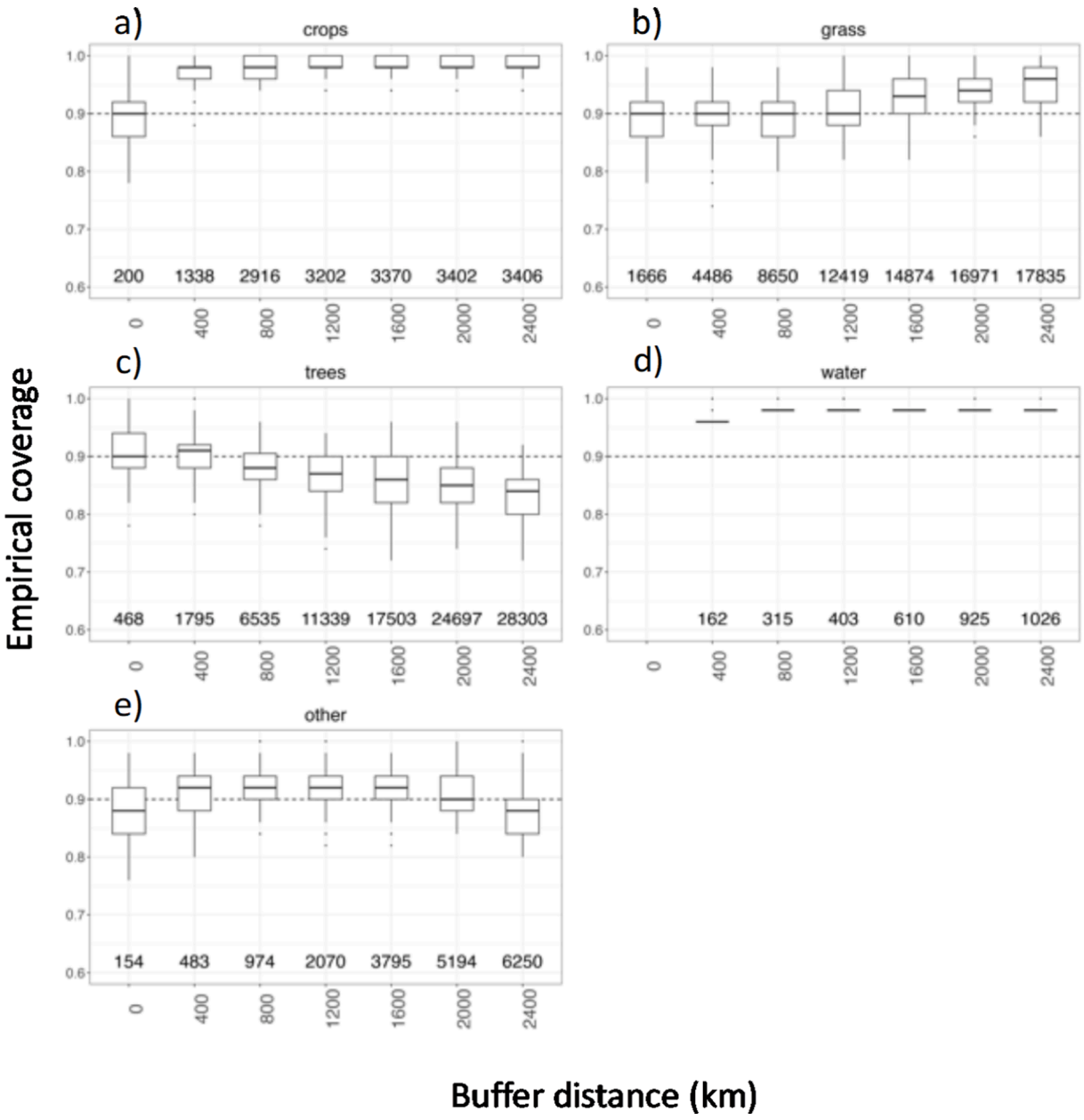


Fig. 5. Empirical coverage results for each LULC class for increasingly larger calibration datasets gathered over increasingly larger regions (defined based on the buffer distance used to create these regions) around the area of interest (i.e., the state of Mato Grosso do Sul [MS]). These results are based on 50 validation observations per LULC class in MS. Note that the "built" class was not included here and some combinations of buffer distance and size of calibration data are missing because there were not enough observations in the Mapbiomas dataset. Numbers at the bottom of each panel refer to the sample size of the calibration data.

Table 3
Mean size of the 90% predictive sets for each LULC class and each state based on the Mapbiomas dataset. Empty predictive sets were excluded from this calculation.

State	built	crops	grass	other	trees	water
AM	1.40	NA	1.91	2.05	1.57	2.17
BA	2.00	2.74	2.52	2.51	2.09	1.94
MG	2.03	2.74	2.24	2.22	2.01	2.09
MS	1.75	2.58	2.21	1.80	1.53	1.73
MT	2.29	2.74	2.21	1.75	1.85	1.96
PA	1.57	2.61	2.37	1.94	1.85	2.16
RO	1.80	2.79	2.36	1.77	1.33	1.89
RS	1.11	1.92	1.41	1.50	1.05	1.07
Average	1.74	2.59	2.15	1.94	1.66	1.88

large blocks of forest in Rio Grande do Sul (RS) and Mato Gross (MT) that seem homogeneous in the LULC map (Fig. 6a and 6c) but that nonetheless have varying levels of uncertainty (Fig. 6d and 6f). Similar to the results in Singh et al. (2024), these uncertainty maps reveal that areas with higher uncertainty tend to be those at the edge of LULC classes. For example, there is greater uncertainty at the edge of the areas classified as "built" and "water" in MT (Fig. 6f). This is likely due to the fact that it is challenging to classify transition regions such as areas with some vegetation but also exposed soil (e.g., *peri*-urban areas) and areas with trees and water (e.g., varzea ecosystems).

We can also compare uncertainty maps calibrated with national-level data (Fig. 6d-f) with uncertainty maps based on state-level data (Fig. 6g-i). This comparison reveals that uncertainty maps based on state-level data can sometimes be substantially different from uncertainty maps based on national-level data (e.g., Fig. 6d,e vs. Fig. 6g,h). Importantly, if

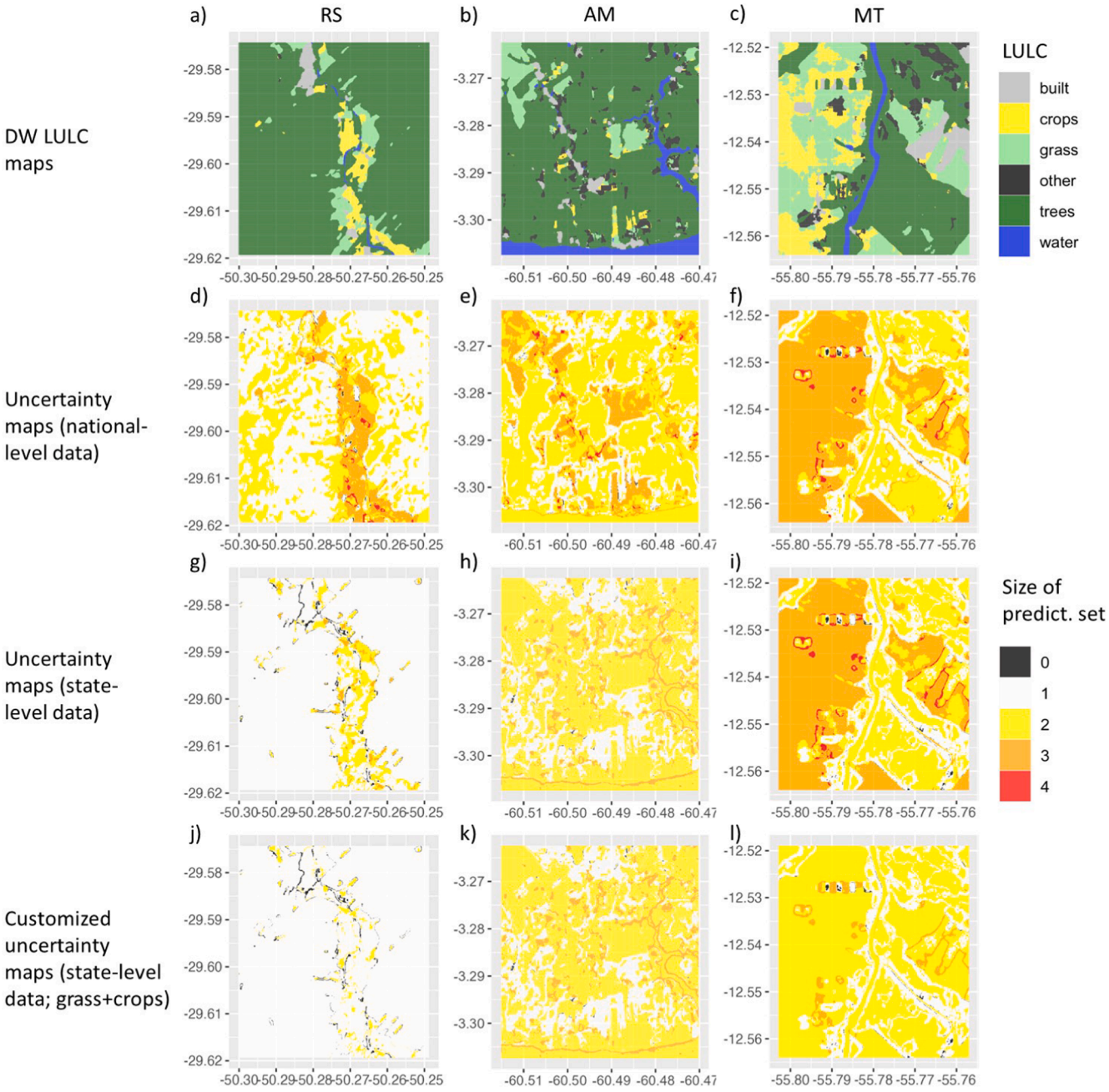


Fig. 6. LULC classification maps from Dynamic World (DW) (panels a-c) together with the corresponding uncertainty maps generated using the class-conditional conformal approach and C = 90 % (d-l). Each panel shows a selected area within a state (RS, AM, and MT stand for Rio Grande do Sul, Amazonas, and Mato Grosso states, respectively). Predictive sets with a single LULC class indicate low uncertainty whereas predictive sets with increasing number of classes indicate increasing uncertainty. Note that empty predictive sets are also indicative of high uncertainty. Uncertainty maps were calibrated using both national-level data (panels d-f) and state-level data (state data) (panels g-l). The customized local uncertainty maps (panels j-l) illustrate how combining the grass and crops LULC classes in the predictive sets can result in a substantial reduction in uncertainty. Note that, because there was no calibration data on crops for AM, we set the threshold probability for this class to 1 (i.e., no predictive set contained the "crops" class).

the end-user of the uncertainty map is not worried about distinguishing between crops and grass, the user can further customize their uncertainty map so that pixels with predictive sets that contain these two LULC classes are not deemed to have greater uncertainty than pixels that only contain one of these LULC classes. This customization can lead to a substantial reduction in uncertainty (Fig. 6g,i vs. Fig. 6j,l) because these two LULC classes are often found together in predictive sets. The resulting customized uncertainty map is useful in highlighting only the pixels that have more discrepant LULC classes.

Finally, we illustrate below how quantifying uncertainty using our methodology can also be useful for map producers. Fig. 7a shows the location of pixels with high uncertainty (i.e., pixels with empty predictive sets; yellow circles) over a background high spatial-resolution (5 x 5 m pixel) true color composite image from the VENμS microsatellite for an area that is predominantly forested in Amazonas state. One of the striking features of Fig. 7a is that several of the pixels with empty predictive sets are systematically arranged along two lines from north to south. A close-up look (Fig. 7c) reveals that these lines occur over large blocks of forest. While these areas are correctly classified as belonging to the "trees" class in the DW map (data not shown), the corresponding

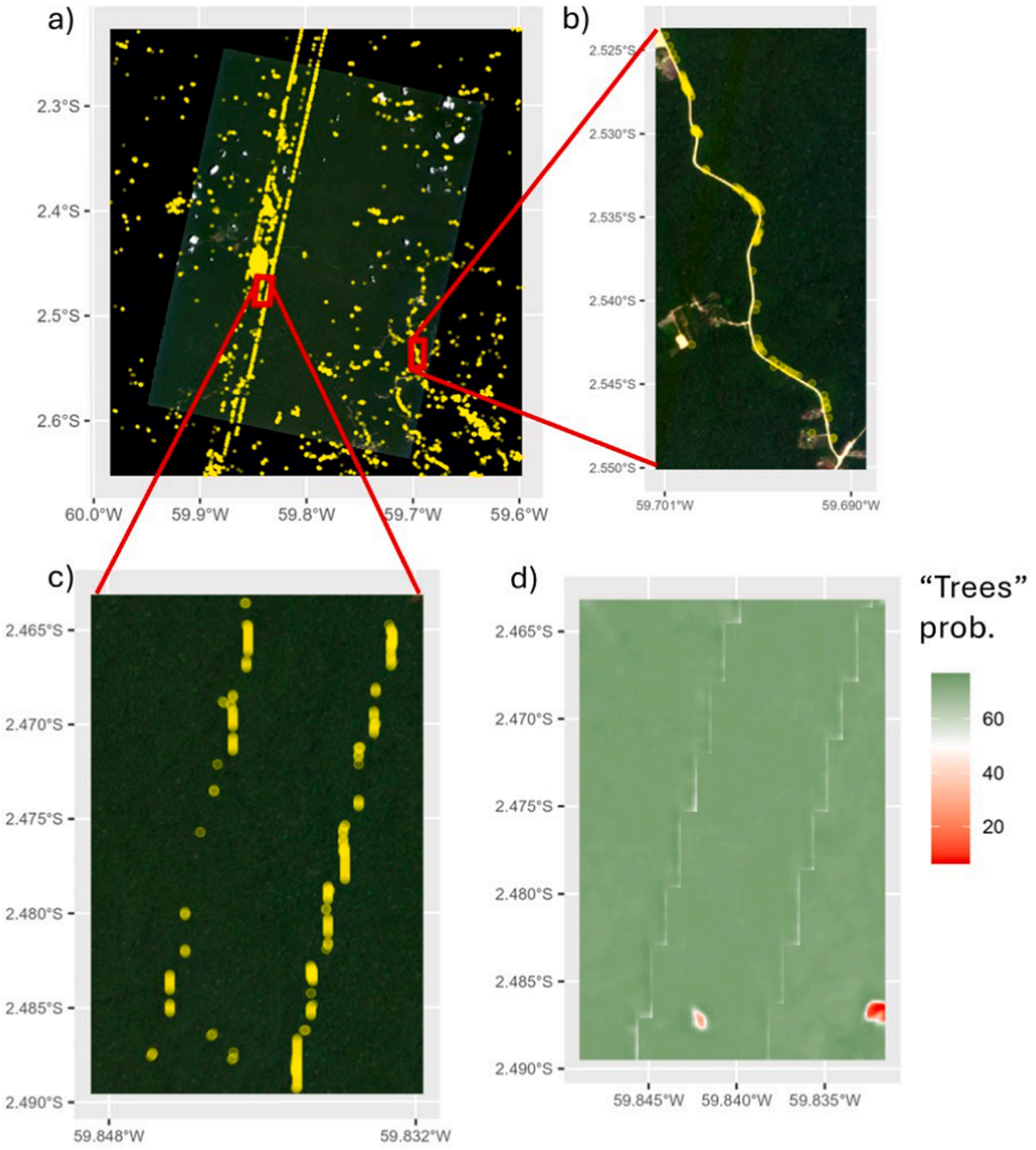


Fig. 7. Pixels with high uncertainty (i.e., pixels with empty predictive set; yellow circles) reveal important features for map producers. Panel a shows two north to south lines of high uncertainty pixels across a true color composite image from the VENμS microsatellite. Panel c shows a subregion, illustrating how these lines with high uncertainty pixels occur in largely forested landscapes. The corresponding "trees" class probabilities for this subregion is depicted in panel d. Panel b shows another subregion in which high uncertainty pixel tend to spatially coincide with the road network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

map depicting the "trees" class probabilities clearly show lower values in the same north to south line patterns (Fig. 7d). We hypothesize that there might be artifacts in the underlying remote sensing data used to create the DW LULC map (or artifacts introduced during the processing of these images), ultimately resulting in the DW classifier algorithm predicting uncharacteristically low "trees" class probabilities (i.e., values below the threshold probability of 0.72 for the class "trees" in Amazonas state; see Fig. 4 for threshold probability values).

Pixels with empty predictive sets also spatially coincide with part of the rural road network (Fig. 7b). We believe the main reason for this is because DW does not contain a road class and the road's spectral signature does not match with any other DW class. As a consequence,

our conformal approach indicates that none of the DW LULC class probabilities are high enough to justify including the corresponding LULC class label to the predictive set, ultimately resulting in empty predictive sets. These results suggest that the inclusion of a rural road class to the DW LULC product may be beneficial to improve its overall quality.

4. Discussion

In this article, we have described a modified conformal approach that ensures more uniform per class uncertainty quantification but that requires more calibration data when compared to the conventional conformal approach. Using this method, we have demonstrated how customized local uncertainty maps can be created without requiring remote sensing and modeling work. This is possible because steps 1 to 3 of the conformal approach (illustrated in Fig. 1) can be skipped if class probabilities from a LULC map are readily available and if local calibration data are available. Critically, by relying on local data, we show that these customized local uncertainty maps can have better performance than uncertainty maps calibrated using global data. Finally, we show the importance of ensuring that local calibration data are representative of the study region and demonstrate an approach to determine if the spatial extent used to collect the local calibration data is appropriate.

To enable the creation of local uncertainty maps, it will be critical that large-scale LULC map producers (e.g., annual national LULC maps produced by Mapbiomas (mapbiomas.org; Souza et al. 2020) and global LULC maps produced by Buchhorn et al. (2020) and Potapov et al. (2022)) make available the LULC class probabilities generated by their classification algorithms (Singh et al. 2024). Another requirement for the proposed approach is the availability of considerable amount of high-quality local data. Collecting such a large dataset on the field can be daunting but other potential sources of local data also exist. For example, LULC map producers themselves might make their validation data available (e.g., the MapBiomas validation dataset), allowing users to just subset data for their area of interest. These local data can also be derived from visual interpretation of high-resolution satellite or unmanned aerial vehicle (UAV) imagery, data sources that are becoming increasingly more common. Alternatively, crowdsourcing can be another way to obtain abundant local LULC data (Hadi et al. 2022). Regardless of the source, local calibration data need to be representative of the landscape of interest to ensure accurate uncertainty quantification. Importantly, because local data should include observations that reflect the inherent variability of each LULC class (e.g., oligotrophic and eutrophic lakes; abandoned and well-maintained pastures), more heterogenous LULC classes may require a greater number of observations. Finally, it is also important for these data to follow a similar classification scheme as the adopted LULC product to avoid introducing additional errors due to the need to harmonize different classification schemes.

The amount of available calibration data is a key factor determining which conformal approach to use. Although the class-conditional approach tends to generate more uniform per class uncertainty quantification, empirical coverage for a given LULC class can vary substantially if there are too few observations in the calibration data for that specific LULC class. Importantly, in some cases certain LULC classes might be important but relatively rare in the study area. In this situation, calibration data from outside the study area might need to be collected. However, as shown in this study, one has to be careful with the spatial extent over which the calibration data are collected. If the spatial extent is too large, the calibration data might cease to be representative of the study area, potentially resulting in predictive sets with decreased performance. If calibration data over a much larger region are readily available (as in this study), then one approach to determine the appropriate spatial extent of the calibration data is to perform an analysis similar to the one we used. For example, the user might have determined that 200 observations are required for calibration (based on Eq. (3) but only 100 observations are available in the area of interest. In this case, one could increase the spatial extent until 200 observations are available and then check if the resulting predictive sets have approximately the desired coverage. An important shortcoming of this approach is that it might require too much data for the area of interest given that some of these data will need to be set aside for validation purposes. If this is a concern, one option is to revert back to the conventional conformal approach, in which case only 50 observations (for example) within the area of study would be required for validation purposes (instead of 50 observations per LULC class when using the class-conditional conformal approach). Another option is to compare the threshold probabilities $\widehat{q}_{1-C,k}$ for the calibration data gathered over increasingly larger areas. In this case, very different threshold values $\widehat{q}_{1-C,k}$ would suggest that the data from the area of interest are not exchangeable with those from the buffer area. We illustrate these alternative approaches of determining spatial extent in Appendix 3.

Local uncertainty maps can be used for several purposes. For example, a user might decide that uncertainty is too high for their area of interest and that a local LULC map will need to be created (or a different LULC map be used) instead of relying on the global DW map. Critically, this decision might be different from the decision taken based solely on the global accuracy information provided by map producers. Furthermore, if deciding on creating a local LULC map, a user might want to make sure that areas with high uncertainty are sampled because these areas are harder to classify (e.g., because they are transition areas or belong to a class that is not present in the current classification). For example, we demonstrate that pixels with high uncertainty (i.e., empty pixels) tend to spatially coincide with the road network in a region in the Amazonas state (Fig. 7b), suggesting that the inclusion of a "rural road" LULC class would be important. In addition, when attempting to understand how LULC influences different ecological and environmental processes (e.g., wildlife movement and occupancy, water pollution, deforestation and wildfire risk), users of uncertainty maps may discard observations associated with pixels for which LULC is too uncertain, this way ensuring a more robust analysis. On the other hand, when modeling the drivers of LULC change, users can test the robustness of their analysis by exploring the effect of different assumptions regarding how pixels with greater uncertainty are handled. Finally, local uncertainty maps allow the reporting of the minimum and maximum area that a given LULC class may cover, thus explicitly acknowledging uncertainty in LULC classification results. This is likely to be particularly important for professionals that have to assess temporal trends in LULC within the areas that they manage (e.g., protected areas or river basins).

Our results also show how uncertainty maps can help improve the generation of large-scale LULC maps. More specifically, uncertainty maps can identify LULC classes that are not well represented in the current LULC classification scheme (e.g., rural roads) and can help identify potential artifacts in the data or algorithms used to create these LULC maps (e.g., Fig. 7). In addition, despite the focus on local uncertainty maps, the proposed methodology can also enhance large-scale uncertainty maps. More specifically, similar to the increasingly common approach of using separate models to make predictions for different regions when creating large-scale biomass maps (e.g., Duncanson et al. 2022), one can systematically divide the region of interest into smaller subregions and quantify uncertainty for each of these subregions just using local data, where "local" is defined based on the approach that we have described to characterize the "bias-variance" tradeoff. This approach to generating large-scale uncertainty maps is likely to yield a much better characterization of uncertainty when compared to using conformal approaches that rely on global probability thresholds. Having said that, we note that an important limitation of this study is that both conventional and class-conditional conformal approaches described here assume data are independent, an assumption that is unlikely to hold due to spatial correlation. Recent research has focused on extending

these conformal approaches to accommodate for spatial and/or temporal dependencies by assuming local exchangeability instead of relying on the assumption that all observations are exchangeable (e.g., Barber et al. 2023; Mao et al. 2022). These newer methodologies will likely be critical to further improve uncertainty quantification in LULC maps.

Because the creation of LULC maps is not trivial, requiring considerable remote sensing and modeling expertise, there has been a trend of large-scale (regional, national, and global) LULC products being systematically produced by specialized groups and organizations, freeing map users to focus on a wide range of downstream applications. While these large-scale products may capture well overall LULC patterns, they might have important shortcomings when used for specific locations. For this reason, accurately quantifying the uncertainty associated with these products at the scale in which these products will be used is paramount. The benefit of creating customized local uncertainty maps is that, as we have shown, it can often improve the quantification of uncertainty. Furthermore, the creation of local uncertainty maps by users opens the opportunity of these maps being customized to their needs. Ultimately, the approach described in this article paves the way for users to generate local customized uncertainty maps that are likely to be much more relevant for their specific uses without requiring extensive remote sensing and modeling skills.

CRediT authorship contribution statement

Denis Valle: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Rodrigo Leite:** Writing – review & editing, Data curation. **Rafael Izbicki:** Writing – review & editing, Methodology. **Carlos Silva:** Writing – review & editing. **Leo Haneda:** Writing – review & editing.

Declaration of competing interest

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

Acknowledgement

DV is grateful for the financial support provided by the US Department of Agriculture National Institute of Food and Agriculture McIntire–Stennis project 1005163 and US National Science Foundation award 2040819. RI is grateful for the financial support of FAPESP (grants 2019/11321-9 and 2023/07068-1) and CNPq (grants 309607/2020-5 and 422705/2021-7). RVL was supported by an appointment to the NASA Postdoctoral Program at the Goddard Space Flight Center, administered by Oak Ridge Associated Universities under contract with NASA. CS was funded through NASA's grants (ICESat-2, 80NSSC23K0941), Carbon Monitoring System (CMS, grant 80NSSC23K1257), and Commercial Smallsat Data Scientific Analysis (CSDSA, grant 80NSSC24K0055).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2024.104288.

Data availability

The data that we use is already publicly available on the internet

References

Barber, R.F., Candes, E.J., Ramdas, A., Tibshirani, R.J., 2023. Conformal prediction beyond exchangeability. The Annals of Statistics 51, 816–845.

Brown, C.F., Brumby, S.P., Guzder-Williams, B., Birch, T., Hyde, S.B., Mazzariello, J., Czerwinski, W., Pasquarella, V.J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., & Tait, A.M. (2022). Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data, 9*.

Brown, K.M., Foody, G.M., Atkinson, P.M., 2009. Estimating per-pixel thematic uncertainty in remote sensing classifications. Int. J. Remote Sens. 30, 209–229.

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.E., Linlin, L., Tarko, A., 2020. Copernicus Global Land Service: Land Cover 100m: Version 3 Globe 2015–2019: Product User Manual. Geneve, Switzerland.

Canibe, M., Titeux, N., Dominguez, J., Regos, A., 2022. Assessing the uncertainty arising from standard land-cover mapping procedures when modelling species distributions. Divers. Distrib. 28, 636–648.

Cheng, K.-S., Ling, J.-Y., Lin, T.-W., Liu, Y.-T., Shen, Y.-C., Kono, Y., 2021. Quantifying uncertainty in land-use/land-cover classification accuracy: a stochastic simulation approach. Front. Environ. Sci. 9. https://doi.org/10.3389/fenvs.2021.628214.

Díaz, S., Settele, J., Brondízio, E.S., Ngo, H.T., Agard, J., Arneth, A., Balvanera, P., Brauman, K.A., Butchart, S.H.M., Chan, K.M.A., Garibaldi, L.A., Ichii, K., Liu, J., Subramanian, S.M., Midgley, G.F., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., Polasky, S., Purvis, A., Razzaque, J., Reyers, B., Chowdhury, R.R., Shin, Y.-J., Visseren-Hamakers, I., Willis, K.J., Zayas, C.N., 2019. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science 366, 1327. https://doi.org/10.1126/science.aax3100.

Duncanson, L., Kellner, J.R., Armston, J., Dubayah, R., Minor, D.M., Hancock, S., Healey, S.P., Patterson, P.L., Saarela, S., Marselis, S., Silva, C.E., Bruening, J., Goetz, S.J., Tang, H., Hofton, M., Blair, B., Luthcke, S., Fatoyinbo, L., Abernethy, K., Alonso, A., Andersen, H.-E., Aplin, P., Baker, T.R., Barbier, N., Bastin, J.F., Biber, P., Boeckx, P., Bogaert, J., Boschetti, L., Boucher, P.B., Boyd, D.S., Burslem, D.F.R.P., Calvo-Rodriguez, S., Chave, J., Chazdon, R.L., Clark, D.B., Clark, D.A., Cohen, W.B., Coomes, D.A., Corona, P., Cushman, K.C., Cutler, M.E.J., Dalling, J.W., Dalponte, M., Dash, J., de-Miguel, S., Deng, S., Ellis, P.W., Erasmus, B., Fekety, P.A., Fernandez-Landa, A., Ferraz, A., Fischer, R., Fisher, A.G., García-Abril, A., Gobakken, T., Hacker, J.M., Heurich, M., Hill, R.A., Hopkinson, C., Huang, H., Hubbell, S.P., Hudak, A.T., Huth, A., Imbach, B., Jeffery, K.J., Katoh, M., Kearsley, E., Kenfack, D., Kljun, N., Knapp, N., Král, K., Krůček, M., Labrière, N., Lewis, S.L., Longo, M., Lucas, R.M., Main, R., Manzanera, J.A., Martínez, R.V., Mathieu, R., Memiaghe, H., Meyer, V., Mendoza, A.M., Monerris, A., Montesano, P., Morsdorf, F., Næsset, E., Naidoo, L., Nilus, R., O'Brien, M., Orwig, D.A., Papathanassiou, K., Parker, G., Philipson, C., Phillips, O.L., Pisek, J., Poulsen, J.R., Pretzsch, H., Rüdiger, C., Saatchi, S., Sanchez-Azofeifa, A., Sanchez-Lopez, N., Scholes, R., Silva, C.A., Simard, M., Skidmore, A., Stereńczak, K., Tanase, M., Torresan, C., Valbuena, R., Verbeeck, H., Vrska, T., Wessels, K., White, J.C., White, L.J.T., Zahabu, E., & Zgraggen, C. (2022). Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. *Remote Sensing of Environment, 270*, 112845, https://doi.org/10.1016/j.rse.2021.112845.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80, 185–201.

Foody, G.M., 2012. Latent Class Modeling for Site- and Non-Site-Specific Classification Accuracy Assessment Without Ground Data. IEEE Trans. Geosci. Remote Sens. 50, 2827–2838.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On Calibration of Modern Neural Networks, 34th International Conference on Machine Learning. Australia, Sydney.

Hadi, Yowargana, P., Zulkarnain, M.T., Mohamad, F., Goib, B.K., Hultera, P., Sturn, T., Karner, M., Durauer, M., See, L., Fritz, S.A., Hendriatna, A., Nursafingi, A., Melati, D. N., Prasetya, F.V.A.S., Carolita, I., Kiswanto, Firdaus, M.I., Rosidi, M., & Kraxner, F. A national-scale land cover reference dataset from local crowdsourcing initiatives in Indonesia Sci. Data 9 2022.

Hsiao, L.-H., Cheng, K.-S., 2016. Assessing uncertainty in LULC classification accuracy by using bootstrap resampling. Remote Sens. (Basel) 8. https://doi.org/10.3390/rs8090705.

Jain, M., 2020. The benefits and pitfalls of using satellite data for causal inference. Rev. Environ. Econ. Policy 14, 157–169.

Khatami, R., Mountrakis, G., Stehman, S.V., 2017. Mapping per-pixel predicted accuracy of classified remote sensing images. Remote Sens. Environ. 191, 156–167.

Lyons, M.B., Keith, D.A., Phinn, S.R., Mason, T.J., Elith, J., 2018. A comparison of resampling methods for remote sensing classification and accuracy assessment. Remote Sens. Environ. 208, 145–153.

Mao, H., Martin, R., Reich, B.J., 2022. Valid model-free spatial prediction. J. Am. Stat. Assoc. 1–11.

Marques, P.C. (2023). On the universal distribution of the coverage in split conformal prediction. In, *eprint arXiv:2303.02770*.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P.H.S., Dokania, P.K., 2020. Calibrating Deep Neural Networks using Focal Loss, *34th Conference on Neural Information Processing Systems*. Vancouver, Canada.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities With Supervised Learning, *Proceedings of the 22nd International Conference on Machine Learning*. Bonn, Germany.

Potapov, P., Hansen, M.C., Pickens, A., Hernandez-Serna, A., Tyukavina, A., Turubanova, S.A., Zalles, V., Li, X., Khan, A., Stolle, F., Harris, N., Song, X.-P., Baggett, A., Kommareddy, I., Kommareddy, A., 2022. The global 2000–2020 land cover and land use change dataset derived from the ladsat archive: first results. Front. Remote Sens. 3, 856903.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria; R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Shafer, G., Vovk, V., 2008. A tutorial on conformal prediction. J. Mach. Learn. Res. 9, 371–421.

Singh, G., Moncrieff, G., Venter, Z.S., Cawse-Nicholson, K., Slingsby, J., Robinson, T.B., 2024. Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction. Sci. Rep. 14.

Souza, C.M., Jr.; , Shimbo, J., Rosa, M.R., Parente, L.L., Alencar, A.A., Rudorff, B.F.T., Hasenack, H., Matsumoto, M., Ferreira, L.G., Souza-Filho, P.W.M., de Oliveira, S.W., Rocha, W.F., Fonseca, A.V., Marques, C.B., Diniz, C.G., Costa, D., Monteiro, D., Rosa, E.R., Vélez-Martin, E., Weber, E.J., Lenti, F.E.B., Paternost, F.F., Pareyn, F.G.C., Siqueira, J.V., Viera, J.L., Neto, L.C.F., Saraiva, M.M., Sales, M.H., Salgado, M.P.G., Vasconcelos, R., Galano, S., Mesquita, V.V., & Azevedo, T. (2020). Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine *Remote Sensing, 12*, https://doi.org/10.3390/rs12172735.

Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. Remote Sens. Environ. 231.

Tilman, D., Clark, M., Williams, D.R., Kimmel, K., Polasky, S., Packer, C., 2017. Future threats to biodiversity and pathways to their prevention. Nature 546, 73–81.

Valle, D., Izbicki, R., Leite, R., 2023. Quantifying uncertainty in land-use land-cover classification using conformal statistics. Remote Sens. Environ. 295.

Venter, Z.S., Barton, D.N., Chakraborty, T., Simensen, T., Singh, G., 2022. Global 10 m land use land cover datasets: a comparison of dynamic world, world cover and Esri land cover. Remote Sens. 14.

Vovk, V., 2012. Conditional validity of inductive conformal predictors. JMLR: Workshop Conf. Proc. 475–490.

Weber, K.T., Langille, J., 2007. Improving classification accuracy assessments with statistical bootstrap resampling techniques. GIsci. Remote Sens. 44, 237–250.