UNCERTAINTY QUANTIFICATION FOR PROBABILISTIC MACHINE LEARNING IN EARTH OBSERVATION USING CONFORMAL PREDICTION.

Geethen Singh & Tamara B. Robinson *
Centre for Invasion Biology
Department of Botany and Zoology
Stellenbosch University
Stellenbosch, Matieland 7602, South Africa
{gsingh,trobins}@sun.ac.za

Glenn Moncrieff
Global science
The Nature Conservancy
Cape Town, South Africa
Centre for ecology, environment and conservation
Department of Statistical Sciences
University of Cape Town
Cape Town, Rondebosch 7701, South Africa
{glenn.moncrieff}@tnc.org

Zander S. Venter
Norwegian Institute of Nature research -NINA
The Nature Conservancy
Oslo, Sognsveien 0855, Norway
{zander.venter}@nina.no

Kerry-Cawse Nicholson
Carbon cycles and Ecosystems
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA
{Kerry-anne.cawse-nicholson}@jpl.nasa.gov

Jasper Slingsby
Department of Biological Sciences and
Centre for ecology, environment and conservation
Department of Statistical Sciences
University of Cape Town
Cape Town, Rondebosch 7701, South Africa
{jasper.slingsby}@uct.ac.za

ABSTRACT

Machine learning is applied to Earth Observation (EO) data to derive data sets that are used to characterise, comprehend and conserve natural resources, contributing to progress towards international accords. However, the derived datasets contain inherent uncertainty and need to be quantified reliably to avoid negative downstream consequences. In response to the increased need to report uncertainty, we bring attention to the promise of conformal prediction within the domain of EO. Conformal prediction is an Uncertainty Quantification (UQ) method that offers statistically valid and informative prediction regions while concurrently being computationally efficient, model-agnostic, distribution-free and can be applied in a post-hoc manner without requiring access to the underlying model and training dataset. We assessed the current state of uncertainty quantification in the EO domain and found that only 21% of the reviewed datasets incorporated a degree of uncertainty information, with unreliable methods prevalent. Next, we introduce Google Earth Engine native modules that can integrate into existing predictive modelling workflows and demonstrate the versatility, efficiency, and scalability of these tools by applying them to datasets spanning continental to global scales, regression, and classification tasks, featuring both traditional and deep learning-based workflows. We anticipate that the increased availability of easy-to-use implementations of conformal predictors, such as those provided here, will drive

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Funding acknowledgements go at the end of the paper.

wider adoption of rigorous uncertainty quantification in EO, thereby enhancing the reliability of uses such as operational monitoring and decision-making.

1 INTRODUCTION

Machine learning-derived datasets are widely used in ecology and conservation but may contain unquantified or unreliable uncertainty information. Moreover, the continued use of the underlying predictive models will inevitably expose them to data beyond the scope of their training data, compromising system performance (Quinonero-Candela et al., 2008; Sugiyama & Kawanabe, 2012). Data acquired by satellites, including reflectance spectra, backscatter or waveform data contain inherent uncertainty owing to measurement noise, randomness, unpredictability in a system, sensor anomalies (for example Landsat-8 thermal calibration issues, (Barsi et al., 2020)), imperfect preprocessing steps (for example, atmospheric correction, orthorectification and terrain corrections) and partial data acquisition (For example, due to the scan-line error in Landsat-7 or the acquisition footprint of Global Ecosystem Dynamics Investigation (GEDI) (Paasche et al., 2022; Wang et al., 2020)). These sources of uncertainty represent irreducible error and are denoted as aleatoric uncertainty (Gruber et al., 2023). In addition, uncertainties that arise from the lack of knowledge or understanding of a system, the selected modelling framework and the stochastic nature of model fitting are cumulatively referred to as epistemic uncertainty (Gruber et al., 2023). For instance, the lack of knowledge/data may be attributed to regions of the world that are critically under-sampled, which may not be well characterized by models trained on data from the global North, reducing the suitability of the derived data and end-user trust (Ludwig et al., 2023).

Consequently, communicating reliable uncertainty information can be beneficial for data creators and data users. Data users may rely on model predictions for decision-making. The ability to understand the confidence associated with predictions is important to prevent erroneous decisions and for risk management. UQ addresses this by encouraging analytical thinking around the data-generating process and by reducing the overreliance on low-confidence predictions. Data creators can use uncertainty information to identify systematic errors, bias, and instances where the model encounters difficulties. This may prompt targeted labelling efforts, the correction of incorrectly labelled data, or making changes to a model to efficiently improve its accuracy and reduce uncertainty. Despite the value that UQ holds, there is no current consensus on the best practices for UQ in EO. This has led to the use of inappropriate UQ methods such as ensemble methods, Bayesian methods, bagging, and quantile regression. To address this, we carried out the first systematic review of large-scale EO datasets to provide empirical evidence and support for the need for a UQ framework like conformal prediction. Conformal prediction is a mathematical framework, that provides uncertainty information with a coverage guarantee. This translates to the provision of uncertainty regions with a constrained error rate. For example, if a 95% confidence level is specified, the conformal predictor will provide a prediction region that contains the true value with a 95% probability (Angelopoulos & Bates, 2023; Molnar, 2023). This coverage guarantee (validity property) remains conspicuously absent from all other pixel-wise UQ methodologies, except under limited distribution assumptions (Vovk et al., 2005; Shafer & Vovk, 2008; Manokhin, 2022). Moreover, it has been shown to hold for satellite data despite spatial autocorrelation (Valle et al., 2023). An additional study explored the trade-offs between different confidence levels and the statistical efficiency of conformal prediction for classification. To encourage the wider adoption of conformal prediction in operational settings, we implemented conformal prediction in the freely available, cloud computing GEE platform (Python and JavaScript API) that is widely used for developing datasets across large extents. Thereafter, we demonstrate the versatility and scalability of the introduced tools with two case studies. The GitHub repository [1] provides the code used for the analyses, annotated notebooks, demos, and a Google Earth Engine application for users to visualize our results and create their own uncertainty maps.

[1] GitHub repo: `https://github.com/Geethen/GEEConformal`

2 METHODS

2.1 ASSESSING THE STATUS OF UNCERTAINTY QUANTIFICATION IN EARTH OBSERVATION.

To assess the status of UQ in EO, we examined all machine learning-derived datasets in the GEE and the GEE community catalogue (last accessed update: 2 November 2023) (Roy et al., 2023). These catalogues were selected because they contain commonly used datasets with national to global coverage. For the core GEE catalogue, "machine learning", "uncertainty" and "UQ" keywords were used to filter and select all machine learning-derived datasets that were reviewed. We examined each of the 241 resulting datasets in conjunction with their research papers to determine if i) machine learning was used to derive the dataset, ii) uncertainty was quantified for the dataset and, if it was quantified, iii) which UQ method was employed.

2.2 DEMONSTRATING THE UTILITY OF CONFORMAL PREDICTORS

The selected case studies look at UQ for canopy height estimation using GEDI for Africa (Dubayah et al., 2020)(Dubayah et al., 2020), and land cover classification ($>$110M instances) using the Dynamic World dataset (Brown et al., 2022). Google's Dynamic World is a near-real-time global (9-class) landcover dataset that is readily available in GEE (Brown et al., 2022). GEDI is a space-based laser altimeter with a full-waveform detector that captures the vertical structure and distribution of vegetation, a proxy for biomass and tree canopy height (Duncanson et al., 2022). These volumetric tree-stand variables are captured in 100 relative height (rh) bands. For instance, the selected rh98 response band corresponds to the height at which 98% of energy is returned to the detector from a 25x25 m footprint (Dubayah et al., 2020). This band was combined with the visible-NIR (RGB and Near-Infrared) bands of the first 2020 biannual composite from the NICFI PlanetScope dataset to estimate tree canopy height. This regression task used a quantile Light Gradient Boosting Machine (LGBM) model with a train, test and calibration split in the ratio 65:20:15.

Conformal prediction can be summarised in six steps. 1) A dataset is split into a train, calibration and test set. 2) The train set is used to fit a predictive model that is then used to estimate the class probabilities or regressed values within the calibration and test sets. 3) the predicted values for the calibration set are combined with the reference/expected values in the calibration set using a non-conformity function that provides a measure of conformity of a calibration instance with the train dataset. Simple, but popular nonconformity score functions include hinge loss for classification and the absolute residual for regression tasks. 4) The computed nonconformity scores are used to compute a quantile-based threshold corresponding to the user-defined confidence level. 5) Next, the computed threshold is used to derive prediction regions, a multi-class prediction set for classification and prediction intervals for regression, for the test set that is used in combination with the reference values to evaluate the quality of the calibrated conformal predictor. 6) Lastly, the calibrated predictor is applied to new instances to quantify uncertainty. For a more detailed description refer to (Angelopoulos & Bates, 2023; Molnar, 2023). A drawback of the mentioned scoring function for regression tasks is the lack of adaptability i.e., all prediction intervals have the same width, irrespective of their difficulty. Therefore, conformal quantile regression has been introduced to provide adaptability (Angelopoulos & Bates, 2023; Romano et al., 2019). For the classification task, we use the least ambiguous set-valued conformal classifier method (Sadinle et al., 2019). For the regression task, we use conformal quantile regression (Romano et al., 2019).

3 RESULTS AND DISCUSSION

Uncertainty is seldom considered in EO with only 18 of the 87 (21%) reviewed datasets in the GEE catalogues citing studies that quantified uncertainty (Figure 1). This may be attributed to a lack of consensus in methodologies, a lack of easy-to-use tools, and a lack of access to computational power required for some methods (Duncanson et al., 2022; Valle et al., 2023). For the 18 studies, quantile regression is commonly used for regression tasks while model ensembles are commonly used for classification tasks. Bootstrapping and design-based area estimates are commonly used to provide confidence intervals for accuracy scores and area estimation, respectively. However, they cannot provide uncertainty information at the pixel-level. In addition, an evaluation of the quality (validity and efficiency) of the quantified uncertainty was lacking, except for its partial consideration

in a single study (Lang et al., 2023). While ensemble and quantile regression methods are generally accessible and straightforward to implement, it is important to note that neither of these approaches offer valid coverage guarantees in a distribution-free manner, producing either over-conservative or overly optimistic prediction regions.

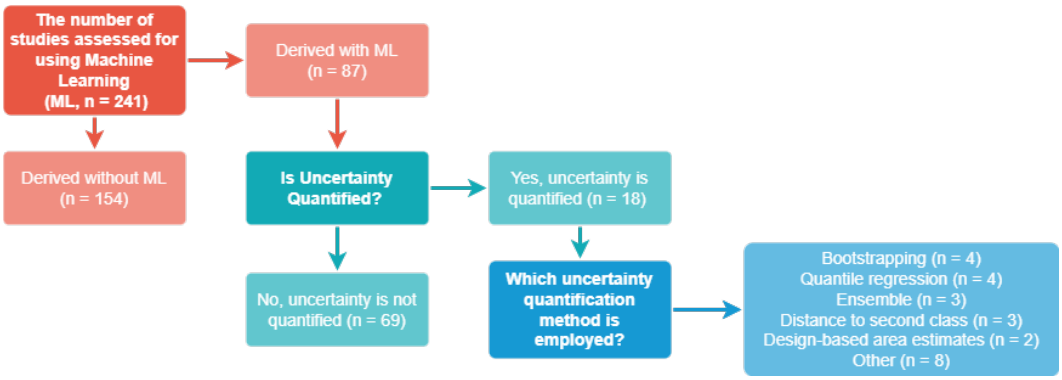

Figure 1: The classification of each reviewed dataset (n = 241) from both the GEE catalogue and the GEE community catalogue. Only the datasets that were derived using machine learning were considered from the main GEE catalogue. For the community catalogue, all datasets comprising the catalogue up to the 2 November 2023 update were considered. Five studies used more than one UQ method.

We have demonstrated the utility of conformal prediction to quantify uncertainty in a robust and scalable manner relevant to classification and regression tasks in EO. For classification, we represent pixel-wise uncertainty as the number of classes included in a pixels' prediction set. Highly uncertain predictions can either be represented with an empty set (length equal to zero) or a large multi-label set (length closer to the candidate number of Dynamic World classes (9)). A multi-label set suggests that the model is finding it challenging to distinguish between several possible class labels at the desired confidence level. Although such a prediction is not incorrect, it is inconclusive, and human intervention would be required to derive the true label. Empty set predictions are examples where the model could not assign any label, typically meaning that the example was very different from the training data. Conversely, the most confident predictions are shown with set lengths of one. For instance, inland water and forest cover predictions are among the most reliably mapped land cover classes (Figure 2A-C), whereas object boundaries typical of mixed landcover pixels, transition and seasonal areas are associated with larger prediction sets and higher uncertainty (Figure 2B).

For the canopy height regression task (test set RMSE = 3.30m), we represent uncertainty as the difference between the upper and lower prediction bound, referred to as the prediction interval. The prediction interval contains the actual canopy height, as based on GEDI, with a 95% probability (empirical marginal coverage, 95.15% ± 0.07). Prediction intervals with a greater width are indicative of higher uncertainty. The average prediction interval width is 9.28m ± 0.03m. Generally, taller canopy height (Figure 2D) corresponds to higher uncertainty (Figure 2E), but when one looks at water systems and pans, some instances deviate from this generalisation. For example, the Sua salt pan in Botswana and the Namibian Etosha pan (Figure 2E, red boxes). Moreover, a similar deviation can be seen for diagonal image artefacts in central Africa (Figure 2E, feint lines), comprising aleatoric uncertainty due to seamlines. By knowing such shortcomings, it becomes possible to enhance system performance through active-learning paradigms (Boulent et al., 2023; Ren et al., 2021). In this way, conformal prediction may expedite the transition towards collaborative human-AI systems. Such a paradigm shift is anticipated to engender enhanced trust, increased adoption rates, and overall improved operational efficiency (Dvijotham et al., 2023; Kamar, 2016; Wang et al., 2020).

4 Conclusion

The use of EO-derived datasets in data-driven decision-making has made a substantial contribution to the characterisation, comprehension, and conservation of planet Earth. Nevertheless, our examination of national to global scale datasets involved in these contributions highlights the lack of

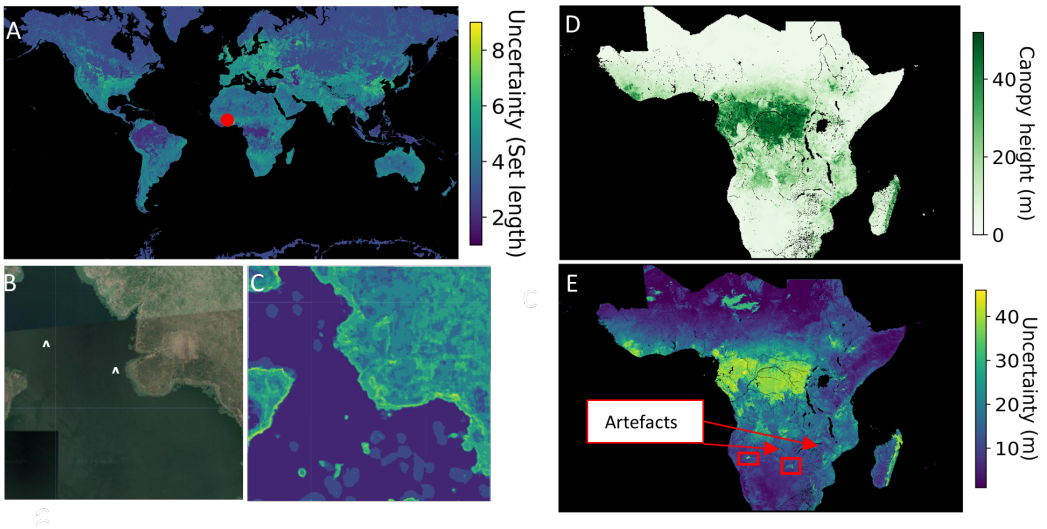

Figure 2: The quantified uncertainty for a) a 2020 first non-null global land cover image from Dynamic World, b-c) A high-resolution Google Earth Image (red point in A) with corresponding c) length of the prediction sets. The d) tree canopy height as estimated from GEDI and NICFI PlanetScope data, e) associated prediction intervals with a 95% confidence level, highlighting diagonal linear artefacts and large prediction intervals for pans (red boxes).

reporting uncertainty and the lack of UQ methods that provide pixel-wise uncertainty information accompanied by coverage guarantees. We believe that UQ through the inclusion of conformal prediction into AI systems stands to significantly increase the role of EO data in operational monitoring systems, policy formulation, and regulatory reporting, accelerating progress towards the realisation of international planetary objectives and targets.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16:494–591, 2023. ISSN 1935-8237.

Julia A Barsi, Brian L Markham, Matthew Montanaro, Simon J Hook, Nina G Raqueno, Jeffrey A Miller, and Rasa Willette. Landsat-8 tirs thermal radiometric calibration status. volume 11501, pp. 70–84. SPIE, 2020.

Justine Boulent, Bertrand Charry, Malcolm McHugh Kennedy, Emily Tissier, Raina Fan, Marianne Marcoux, Cortney A Watt, and Antoine Gagné-Turcotte. Scaling whale monitoring using deep learning: A human-in-the-loop solution for analyzing aerial datasets. *Frontiers in Marine Science*, 10:1099479, 2023. ISSN 2296-7745.

Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, and Simon Ilyushchenko. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9:1–17, 2022. ISSN 2052-4463.

Ralph Dubayah, James Bryan Blair, Scott Goetz, Lola Fatoyinbo, Matthew Hansen, Sean Healey, Michelle Hofton, George Hurtt, James Kellner, and Scott Luthcke. The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography. *Science of remote sensing*, 1:100002, 2020. ISSN 2666-0172.

Laura Duncanson, James R Kellner, John Armston, Ralph Dubayah, David M Minor, Steven Hancock, Sean P Healey, Paul L Patterson, Svetlana Saarela, and Suzanne Marselis. Aboveground biomass density models for nasa's global ecosystem dynamics investigation (gedi) lidar mission. *Remote Sensing of Environment*, 270:112845, 2022. ISSN 0034-4257.

Krishnamurthy Dvijotham, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, Robert Stanforth, Nick Pawlowski, Patricia Strachan, Zahra Ahmed, Shekoofeh Azizi, and Yoram Bachrach. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, 29:1814–1820, 2023. ISSN 1078-8956.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*, 2023.

Ece Kamar. Directions in hybrid intelligence: Complementing ai systems with human intelligence. pp. 4070–4073, 2016.

Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the earth. *Nature Ecology 'I&' Evolution*, pp. 1–12, 2023. ISSN 2397-334X.

Marvin Ludwig, Alvaro Moreno-Martinez, Norbert Hölzel, Edzer Pebesma, and Hanna Meyer. Assessing and improving the transferability of current global spatial prediction models. *Global Ecology and Biogeography*, 32:356–368, 2023. ISSN 1466-822X.

Valery Manokhin. Machine learning for probabilistic prediction. 2022.

Christoph Molnar. *Introduction to Conformal Prediction with Python: A Short Guide to Quantifying Uncertainty of Machine Learning Models*. First edition, 2023.

Hendrik Paasche, Matthias Gross, Jakob Lüttgau, David S Greenberg, and Tobias Weigel. To the brave scientists: Aren't we strong enough to stand (and profit from) uncertainty in earth system measurement and modelling? *Geoscience Data Journal*, 9:393–399, 2022. ISSN 2049-6060.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54:1–40, 2021. ISSN 0360-0300.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Samapriya Roy, Kurt Schwehr, Valerie Pasquarella, Erin Trochim, and Tyson Swetnam. samapriya/awesome-gee-community-datasets: Community catalog, 10 2023. URL `https://doi.org/10.5281/zenodo.8435453`.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 2008. ISSN 1532-4435.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Denis Valle, Rafael Izbicki, and Rodrigo Vieira Leite. Quantifying uncertainty in land-use land-cover classification using conformal statistics. *Remote Sensing of Environment*, 295:113682, 2023. ISSN 0034-4257.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. pp. 1–6, 2020.