ADAPTIVE CONFORMAL SETS UNDER DISTRIBUTION SHIFT: USING ENSEMBLE DISAGREEMENT AS AN EPISTEMIC NORMALIZER FOR EELGRASS SEGMENTATION UNDER TEMPORAL DRIFT

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Statistics

by

Dillon Murphy

March 2026

COMMITTEE MEMBERSHIP

TITLE:	Adaptive Conformal Sets Under Distribution Shift: Using Ensemble Disagreement as an Epistemic Normalizer for Eelgrass Segmentation under Temporal Drift
AUTHOR:	Dillon Murphy
DATE SUBMITTED:	March 2026

COMMITTEE CHAIR:	Kelly Bodwin, Ph.D. Associate Professor of Statistics and Data Science
COMMITTEE MEMBER:	Jonathon Ventura, Ph.D. Associate Professor of Computer Science and Software Engineering
COMMITTEE MEMBER:	Andrew Fricker, Ph.D. Associate Professor of Geography

ABSTRACT

Adaptive Conformal Sets Under Distribution Shift: Using Ensemble Disagreement as an Epistemic Normalizer for Eelgrass Segmentation under Temporal Drift

Dillon Murphy

Semantic segmentation of eelgrass (Zostera marina) from high-resolution drone-based imagery is essential for coastal habitat monitoring, restoration, and management, as these habitats see rapid changes due to climate change and human-induced influences. However, the reliability of generalizing these classification models not only relies on high-accuracy segmentation but also on rigorous uncertainty quantification that holds up when conditions change across years or locations. Conformal prediction (CP) converts a classifier's output into prediction sets with finite-sample marginal coverage; however, the standard score $s = 1 - p_y$ can under-cover in hard or out-of-distribution (OOD) regions under temporal drift. Inspired by heteroskedastic conformal regression, this study proposes normalizing conformal scores by an epistemic proxy, ensemble disagreement. These scores re-scale nonconformity by a data-driven estimation of local difficulty learned on the calibration set, so that the global quantile is conservative where OOD risk is high. This study evaluates (i) a parametric form $s' = (1 - p_y)/(1 + \lambda V)$, where V is ensemble variance of probabilities, and (ii) a nonparametric normalization $s' = s/\hat{\sigma}(V)$, where $\hat{\sigma}(V) = \mathrm{E}[s|V]$ is estimated through quantile binning and interpolation. On drone imagery of the Morro Bay estuary, California (2018-2022), the models were trained on 2018-2021, calibrated on 2021, and evaluated on 2022 on OOD points. Normalized scores recover much of the lost coverage seen by vanilla split conformal prediction and approximate target, increased the percentage of singletons, and reduced the spatial coverage variability. The study provides a simple-to-implement drop-in framework to regain coverage in difficult regions and out-of-distribution data without retraining or labels at test time.

Keywords: conformal prediction, adaptive sets, uncertainty, eelgrass, remote sensing.

ACKNOWLEDGMENTS

This page is not required, but if you have received funding for your research or assistance or guidance that you feel should be noted, it belongs on this page.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

1 Introduction 1

2 Background and Motivation 3

2.1 Uncertainty in Deep Learning 3

2.2 Deep Ensembles 4

2.3 Semantic Segmentation in Environmental Monitoring 4

2.4 Conformal Prediction Foundations 4

2.5 Adaptive or Shift-Aware CP 5

3 Methodology 7

3.1 Data 7

3.2 Model Training + Pipeline 7

3.3 Uncertainty Analysis 7

3.4 Conformalizing 7

3.4.1 Split CP for Classification 7

3.4.2 Variance-Aware Score Normalization 7

3.5 CP Evaluation 7

4 Results 8

4.1 Ensemble Results 8

4.2 In-Distribution Evaluation 8

4.3 Temporal OOD (2022) 8

4.4 Class Conditional Coverage 8

4.5 Spatial Robustness within 2022 8

4.6 Sensitivity 8
5 Discussion 9
6 Limitations 10
7 Conclusion 11
7.1 Computational Details 11
REFERENCES 12
APPENDICES
A Super Cool Fancy Function 14

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Eelgrass (*Zostera marina*) is a seagrass species of temperate waters that provides an anchor to coastal ecosystems by providing sediment stabilization, carbon sequestration, eliminating contamination, and providing habitats for protected species. Warming oceans, storm regimes, and local disturbance have quickly reshaped the intertidal morphology and habitat extent. As these threats increase, there is a need to monitor changes to determine appropriate environmental management. In recent years, deep networks have made this intertidal mapping practical in complex imagery, achieving strong pixel-wise accuracy in areas such as Morro Bay, California (Tallam et al., 2023). However, the reliability of general utilization of these classification models not only relies on high-accuracy segmentation, but also on rigorous uncertainty quantification under temporal and spatial distribution shift. Imagery collected over different years, at different tides or seasons, can erode the model's calibration, causing it to be over-confident. This study builds off these models to create a post hoc uncertainty framework.

Uncertainty quantification is a requirement for flagging ambiguous areas. To answer this, split conformal prediction provides an appealing framework. It offers a wrapping of any probabilistic predictor to produce prediction sets with finite-sample marginal coverage under exchangeability, simply by using a calibration set and a quantile nonconformity score (Angelopoulos & Bates, 2022). In classification, a natural choice of noncormity score is $s = 1 - p_y$, yielding a prediction set of all labels whose scores do not exceed a global threshold learned on calibration. With stable data, this can be simple and effective; however, in reality, the distribution of scores can differ by difficulty. A threshold calibrated over one year can underperform when a distributional shift occurs, such as temporal drift.

To remedy this, it is necessary to make the conformal prediction adaptive to drift or local difficulty. First, work in adaptive or locally weighted CP in regression has shown that normalizing scores by a measure of local difficulty can make a single global quantile more appropriate across heterogeneous inputs, which improves approximate conditional behavior (Chernozhukov, Wüthrich, & Zhu, 2021; Dewolf, De Baets, & Waegeman, 2025; Lei, G'Sell, Rinaldo, Tibshirani, & Wasserman, 2018). Second, research on uncertainty in deep learning has demonstrated that diversified deep ensembles provide practical, shift-aware signals of difficulty that tend to grow on out-of-distribution inputs (Lakshminarayanan, Pritzel, & Blundell, 2017). Using these ideas, this study implements a simple strategy for rescaling the classification score by a calibration-learned difficulty proxy, so that the global threshold becomes conservative under drift, without the need for recalibration or labels.

This thesis develops this design for eelgrass segmentation of multi-year drone imagery. The proposed normalizers are deliberately simple to adopt into existing pipelines. This includes: 1. a linear normalization $s' = (1-p_y)/(1+\lambda V)$ where Vsummarizes ensemble disagreement and λ controls degree of conservativeness and 2. a learned normalizer $s' = s/\hat{\sigma}(V)$ where $\hat{\sigma}(V) = \mathrm{E}[s|V]$ is estimated on the calibration split via quantile binning and interpolation. Both approaches retain standard CP workflow and require no labels at test time or retraining.

Chapter 2

BACKGROUND AND MOTIVATION

Effective eelgrass monitoring requires models that both provide accurate predictions, as well as how certain they are for those predictions. This section reviews the uncertainty concepts used in deep learning.

2.1 Uncertainty in Deep Learning

Uncertainty in predictive modeling is commonly broken up into two components: *epistemic* and *aleotoric*. These two components collectively comprise the total uncertainty in a model's predictions. Aleatoric uncertainty represents the noise inherent in observations, which is often impossible to remove, even with the addition of more data (e.g. sensor noise or true label ambiguity). Even a perfect model cannot remove it. Epistemic uncertainty represents the uncertainty of the model itself, which can be reduced with additional or higher-quality data (Kendall & Gal, 2017).

For modern vision models, raw softmax probabilities are often miscalibrated, meaning the outputted confidence does not match empirical accuracy. This can be especially prevalent under shift, where models can be confident yet wrong. Post-hoc calibration, such as temperature scaling, rescales logits via $softmax(z/T)$ and improves in-distribution calibration without changing accuracy (Guo, Pleiss, Sun, & Weinberger, 2017). Under shift, however, calibration is not enough, as the structure of the errors can change. Thus, it requires an uncertainty representation that integrates decision rules that signal the difficulty of the prediction.

2.2 Deep Ensembles

Deep ensembles can address uncertainty by aggregating predictions from models trained independently with different initializations or architectures. For estimating epistemic uncertainty, member disagreements can be calculated by considering the variance in member probabilities or logits. Given members $\{f_k\}_{k=1}^{K}$ that produce $p^{(k)}(x)$, $V(x) = Var_k(p_y^{(k)}(x))$. Ensembles have not only been found to be more robust under shift, but taking epistemic uncertainty specifically into account can greatly increase a model's predictive uncertainty under dataset shift (Ovadia et al., 2019). Additionally, ensembles require no specialized inference and can easily integrate with existing training code, providing a difficulty proxy for uncertainty. This does come at the additional computational cost of training additional models. However, many models are often trained in pursuit of classification accuracy, which, when ensembled, can outperform individual models.

2.3 Semantic Segmentation in Environmental Monitoring

Semantic image segmentation is a deep learning neural network technique that assigns class labels to each pixel within an image, delineating objects. Recent work demonstrates that deep networks can accurately classify intertidal eelgrass from drone imagery at useful resolutions (Tallam et al., 2023). Environmental monitoring can pose unique challenges, as not only are annotations costly and sometimes ambiguous, but the habitat is constantly shifting over seasonal and annual time scales. Even at the same site, small variations such as time of day, tidal height, etc, can alter the conditions of the imagery.

2.4 Conformal Prediction Foundations

Conformal prediction provides a model and distribution-free technique to convert probability scores into prediction sets with finite-sample marginal coverage under exchangeability (Angelopoulos & Bates, 2022). For a classifier with probabilities $p(x)$, a simple *nonconformity*

score is $s = 1 - p_y$, which is small when the model is confident in the true class. In split conformal prediction, scores are computed on a calibration set C of size n and the finite-sample-corrected quantile is selected $\hat{q}_\alpha = \text{Quantile}_{\lceil (n+1)(1-\alpha) \rceil / n}(\{s_i\})$, then the set value prediction is outputed

$$C_\alpha(x) = \{y : s \leq \hat{q}_\alpha\}$$

This procedure guarantees $P\{Y \in C_\alpha(X)\} \geq 1 - \alpha$ for new points exchangeable with C, regardless of the underlying model. Two limitations of conformal prediction motivate this thesis. The first limitation is that the guarantee is marginal. This means there is no guarantee of conditional coverage, that the per class coverage will be $\geq$ target, and that coverage is not guaranteed across differing inputs, such as varying difficulty. Second, conformal prediction has mathematical validity due to the concept of symmetry (data could have been seen in any order) as long as the data is exchangeable. Coverage guarantees are lost as distribution drifts, and the model can become overly confident, leading to under-coverage. To address these problems, we need to use an adaptive variant of conformal prediction, which is more conservative for harder inputs.

2.5 Adaptive or Shift-Aware CP

There is a substantial amount of literature that aims to make CP adaptive to heterogeneity or robust during dataset drift.

In regression, locally-weighted CP normalizes residuals by an estimate of local spread to equalize scores for a global quantile under heterogeneity (Lei et al., 2018). Distributional conformal prediction (DCP) transforms the score using an estimate of the conditional cumulative distribution function. If the distribution of the score is stable conditional on covariates, then a single global quantile should be appropriate across those covariates (Chernozhukov et al., 2021).

The two ideas loosely connect to this variance score normalization. Let $S = s(X, Y)$ denote the standard noncomformity score and let $V = V(X)$ be a one-dimensional proxy from the deep ensemble (member-probability variance after temperature scaling). We can use this proxy to stabilize the score with respect to *V*, rather than estimating an ultra-high-dimensional $F_{S|X}$ as in DCP. Additionally, we do not estimate a full conditional CDF transform and instead normalize under a multiplicative difficulty assumption

$$S = a(V)U, a(V) > 0, \text{U independent of V}.$$

Then, $S' = S/a(V) \approx U$, so $S'|V$ becomes approximately invariant, and locally tuned to difficulty. Normalization targets the actual failure under shift, that the global threshold will become too lenient for the hard OOD inputs. On the calibration set, we estimate $\hat{\sigma} \approx E[S|V = v]$ via quantile binning and smooth interpolation, and then conformalize the normalized score $S' = S/\hat{\sigma}(V)$. Thus, if the conditional distribution $S|V = v$ is roughly the same across years:

$$S \mid V = v(calibration) \approx S \mid V = v(test),$$

then we expect the normalized scores to be approximately stable conditional on V. Then the global quantile will be appropriate across the shifted years. Intuitively, pixels with large member disagreements (hard pixels) would have larger scores. Dividing by $\hat{\sigma}(V)$ will equalize the scores, so that hard and easy pixels are all scored equally, being precise on easy inputs and conservative on difficult ones. By using a difficulty signal (epistemic uncertainty) that is specifically "shift-aware," we expect that the score becomes shift-aware as well.

Chapter 3

METHODOLOGY

3.1 Data

3.2 Model Training + Pipeline

3.3 Uncertainty Analysis

3.4 Conformalizing

3.4.1 Split CP for Classification

3.4.2 Variance-Aware Score Normalization

3.4.2.1 Parametric linear shrink

3.4.2.2 Nonparametric normalization

3.5 CP Evaluation

Set Composition, Spatial Equality, and Conditional Coverage

Chapter 4

RESULTS

4.1 Ensemble Results

4.2 In-Distribution Evaluation

4.3 Temporal OOD (2022)

global coverage and set composition

4.4 Class Conditional Coverage

4.5 Spatial Robustness within 2022

4.6 Sensitivity

Chapter 5

DISCUSSION

Chapter 6

LIMITATIONS

Chapter 7

CONCLUSION

7.1 Computational Details

REFERENCES

Angelopoulos, A. N., & Bates, S. (2022). *A gentle introduction to conformal prediction and distribution-free uncertainty quantification*. Retrieved from https://arxiv.org/abs/2107.07511

Chernozhukov, V., Wüthrich, K., & Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, *118*(48), e2107794118. https://doi.org/10.1073/pnas.2107794118

Dewolf, N., De Baets, B., & Waegeman, W. (2025). Conditional validity of heteroskedastic conformal regression. *Information and Inference: A Journal of the IMA*, *14*(2), iaaf013. https://doi.org/10.1093/imaiai/iaaf013

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (pp. 1321–1330). PMLR. Retrieved from https://proceedings.mlr.press/v70/guo17a.html

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6405–6416. Red Hook, NY, USA: Curran Associates Inc.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*,

113(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., … Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Proceedings of the 33rd international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc.

Tallam, K., Nguyen, N., Ventura, J., Fricker, A., Calhoun, S., O'Leary, J., … Walter, R. K. (2023). Application of deep learning for classification of intertidal eelgrass from drone-acquired imagery. *Remote Sensing*, *15*(9). https://doi.org/10.3390/rs15092321

APPENDICES

Appendix A

SUPER COOL FANCY FUNCTION