# Scalable Geospatial Data Generation Using AlphaEarth Foundations Model

**Luc Houriez**[*]
X, the Moonshot Factory, Bellwether
Stanford University
houriezl@google.com

**Sebastian Pilarski**[*]
X, the Moonshot Factory
Bellwether
sebpilarski@google.com

**Behzad Vahedi**[*]
X, the Moonshot Factory
Bellwether
vahedi@google.com

**Ali Ahmadalipour**
X, the Moonshot Factory
Bellwether
aliahma@google.com

**Teo Honda Scully**
X, the Moonshot Factory
Bellwether
teonnaise@google.com

**Nicholas Aflitto**
X, the Moonshot Factory
Bellwether
aflitto@google.com

**David Andre**[†]
X, the Moonshot Factory
Bellwether
davidandre@google.com

## Abstract

High-quality labeled geospatial datasets are essential for extracting insights and understanding our planet. Unfortunately, these datasets often do not span the entire globe and are limited to certain geographic regions where data was collected. Google DeepMind's recently released AlphaEarth Foundations (AEF) provides an information-dense global geospatial representation designed to serve as a useful input across a wide gamut of tasks. In this article we propose and evaluate a methodology which leverages AEF to extend geospatial labeled datasets beyond their initial geographic regions. We show that even basic models like random forests or logistic regression can be used to accomplish this task. We investigate a case study of extending LANDFIRE's Existing Vegetation Type (EVT) dataset beyond the USA into Canada at two levels of granularity: EVTPHYS (13 classes) and EVTGP (80 classes). Qualitatively, for EVTPHYS, model predictions align with ground truth. Trained models achieve 81% and 73% classification accuracy on EVTPHYS validation sets in the USA and Canada, despite discussed limitations.

## 1 Introduction

High-quality, global environmental datasets are critical for applications ranging from natural resource management to climate analysis, yet their creation is often hampered by prohibitive costs and inconsistent data sources. While valuable datasets for variables like vegetation type exist, they are often confined to specific regions, such as a single country. This geographic scarcity creates major gaps in our ability to monitor earth system changes and build globally applicable models.

Representation learning, a technique popularized in natural language processing [20, 22, 6], offers a path to overcoming these limitations. Its adoption in geospatial applications [18] has led to a rich ecosystem of models designed to encode spatial information [28, 13, 27, 19, 1, 12, 15]. Building on

[*]These authors contributed equally to this work.
[†]A complete list of authors and their affiliations is available in the appendix.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

this trend, Google DeepMind's AlphaEarth Foundations (AEF) provides a breakthrough: globally consistent, high-resolution embeddings derived from a fusion of satellite data including Landsat and Sentinel [3]. The resulting "Satellite Embedding" dataset [8] encodes a rich variety of geophysical properties, making it an ideal input feature for diverse downstream tasks.

However, raw embeddings seldom provide easily extractable insights. Insight extraction usually involves visualizations or transformations to simpler easier-to-understand representations. In geospatial settings, this is often done by creating data labels of important features, e.g., a road or a crop type. Unfortunately, for many applications, labeled data only exists for certain regions. This greatly limits access to model and data interpretability across large swaths of the world.

**Contribution** This paper's core contribution is a validated pipeline for leveraging AEF embeddings to generate synthetic environmental variables, effectively extending datasets from data-rich to data-poor regions. We demonstrate how to train a model to map AEF embeddings to known ground-truth labels in one area, and then apply this model to infer those same labels elsewhere. This task builds on a long history of remote sensing classification methods, from early artificial neural networks and decision trees [11, 5, 7] to modern deep learning architectures [25, 29, 24]. We showcase the efficacy of this approach with a case study extending a vegetation type dataset from the USA into Canada, demonstrating AEF can serve as a powerful basis for global data interpolation and synthesis.

## 2 Data and models

**EVT** LANDFIRE [23, 16] provides an ecological dataset called Existing Vegetation Type (EVT) which has historically been used for wildfire management efforts [4]. The EVT dataset consists of labels at various levels of classification granularity in the USA. LANDFIRE provides mappings across these different granularity levels, e.g., "Western Hemlock-Yellow-cedar Forest" at medium-level granularity (EVTGP: collapsed vegetation type) maps into "Conifer" at a lower granularity (EVTPHYS: physiognomy). This paper describes the training results for both EVTPHYS and EVTGP.

**Data selection** We train on data from Alaska and northern continental US (CONUS) above the 41.6 degrees latitude line as we believe it provides an effective balance between data quantity and regions that exhibit most similar ecological or environmental characteristics to the target Canadian regions. We use LANDFIRE's 2020 release for EVT in our study [17].

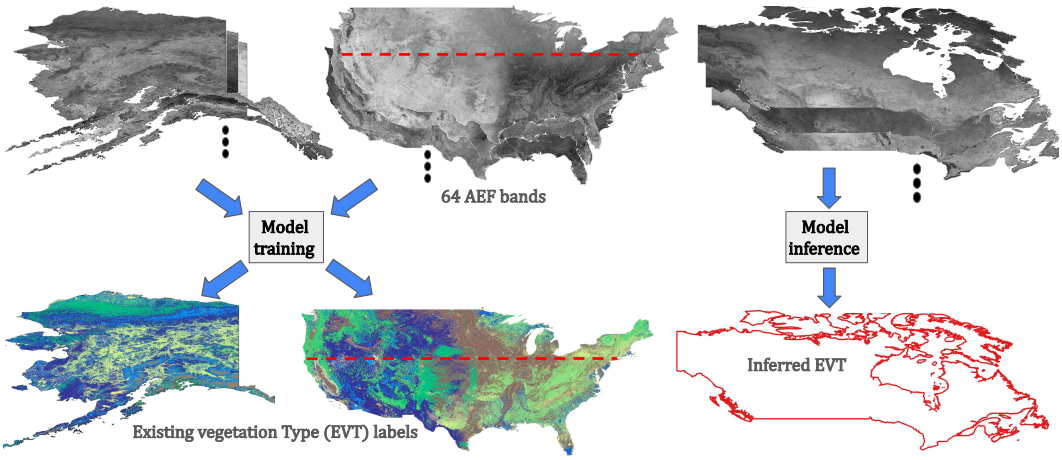



Figure 1: Schematic of model training and inference. The 64 bands of AEF data (input) and EVT data (target) from continental USA above the red dotted line and Alaska are used to train the model. Running inference on AEF data in Canada provides expected EVT in the previously unlabeled region.

**Data preprocessing** LANDFIRE's original EVTGP classifications consist of 194 unique classes, but we filter out classes comprising less than 0.1% of the dataset to address class imbalance. This results in 80 classes spanning the selected continental US and Alaska regions. We mask out pixels

not belonging to one of these 80 classes. This target tests AEF's ability to generalize to larger classification problems, as it was optimized on targets with 40 classes at most. For EVTPHYS we group all development related classes which yields 13 classes from an original 17. We download the AEF embeddings from Google Earth Engine [9] and train on AEF and EVT at 500m resolution.

**Data splits** We geographically tile our northern CONUS and Alaska data into tiles of size 64x64 pixels. We allocate 90% of these tiles for training and 10% for validation. The EVT dataset provides data for a Southern 90km band of Canada and a Western 90km wide band along the Alaska border in recent releases. We reserve this data for our final test set as it directly coincides with our desired target and allows us to evaluate the generalizability of our approach to an unseen region.

**Models** Our proposed methodology extends existing datasets by training machine learning models to predict labels (e.g., EVT) from AEF inputs. We use a flexible pipeline that allows for rapid experimentation by easily swapping model architectures. We evaluate four models for this case study: a **logistic regression** model which for a given AEF pixel determines a linear weighting of the 64 AEF band values to produce an EVT classification [10, 21]; a **RandomForestClassifier** from scikit-learn [2, 21]; an **LGBMClassifier** from the LightGBM library [14]; and a **segmentation model** with a U-Net architecture using EffecientNet-B4 [26] pre-trained on advprop for the encoder.

## 3 Results

Figure 2 shows the segmentation model's inference map of EVTPHYS in Canada which demonstrates good vegetation type continuity. EVTPHYS maps are qualitatively similar across all models, while the EVTGP maps show some differences in northern Canada (more details in Appendix Figures 8 and 9).

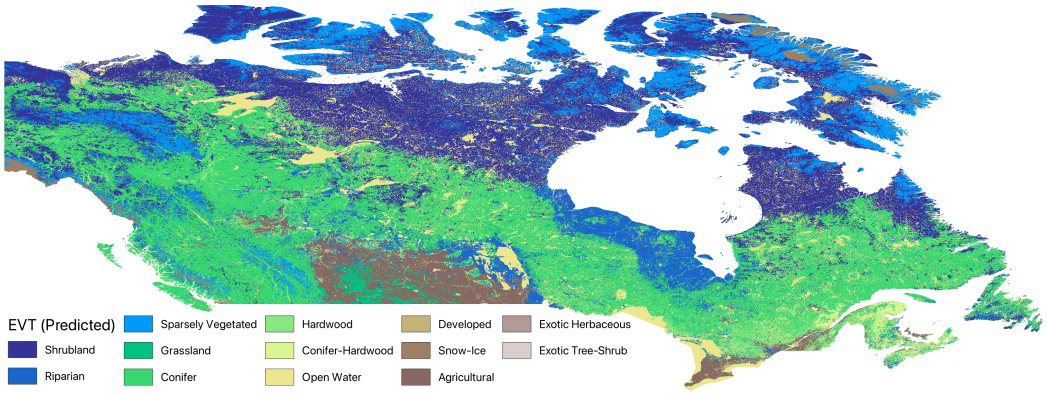



Figure 2: EVTPHYS(13 classes) Inference in Canada generated using the the segmentation model.

**Metrics** We used Accuracy, macro-averaged Jaccard Index, and F1 score to compare the performance of all models for both EVTPHYS and EVTGP (Tables 1 and 3). On the original training and validation sets, all models performed similarly, though the random forest model showed signs of overfitting. A significant performance drop was observed for all models on the test set.

Table 1: Accuracy (ACC), Jaccard (J), and F1 across data splits for EVTPHYS (13 classes).

| | **Training** | | | **Validation** | | | **Test** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** |
| **Logistic Regression** | 0.77 | 0.48 | 0.60 | 0.77 | 0.48 | 0.59 | 0.71 | 0.39 | 0.51 |
| **Random Forest** | 0.97 | 0.95 | 0.97 | 0.81 | 0.55 | 0.67 | 0.73 | 0.43 | 0.55 |
| **Gradient Boosted Trees** | 0.79 | 0.52 | 0.65 | 0.79 | 0.52 | 0.64 | 0.73 | 0.42 | 0.54 |
| **Segmentation Model** | 0.79 | 0.50 | 0.63 | 0.79 | 0.51 | 0.63 | 0.73 | 0.42 | 0.54 |

Table 2: Model performances for EVTPHYS (13 classes) across 3 distinct test regions. *Canada South* and *Canada West* combined comprise the test set in Table 1 (see Data Splits in Section 2).

| | Canada South | | | Canada West | | | Southern CONUS | | |
|---|---|---|---|---|---|---|---|---|---|
| | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** |
| **Logistic Regression** | 0.67 | 0.31 | 0.41 | 0.82 | 0.35 | 0.42 | 0.59 | 0.32 | 0.44 |
| **Random Forest** | 0.69 | 0.34 | 0.45 | 0.83 | 0.42 | 0.52 | 0.68 | 0.35 | 0.46 |
| **Gradient Boosted Trees** | 0.69 | 0.34 | 0.45 | 0.83 | 0.32 | 0.39 | 0.64 | 0.32 | 0.44 |
| **Segmentation Model** | 0.69 | 0.34 | 0.45 | 0.83 | 0.37 | 0.45 | 0.66 | 0.36 | 0.48 |

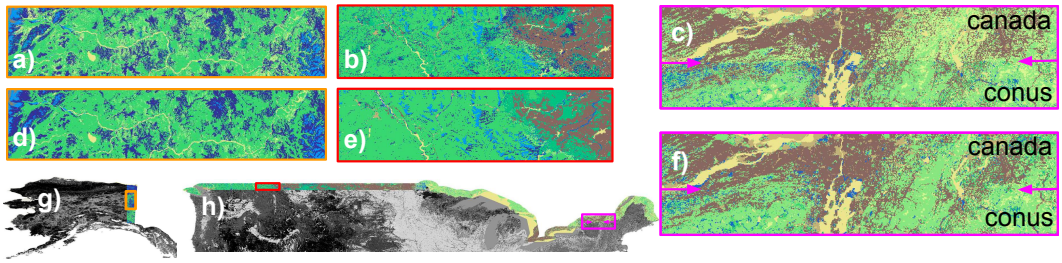



Figure 3: Ground truth EVTPHYS (a–c) compared to gradient boosted trees model inference (d–f) in Canada West (g) and South (h) test regions. Figures (c, f) additionally show land in CONUS across the border which is indicated by the magenta arrows. There, EVT values produced by LANDFIRE seem to exhibit an artificial discontinuity.

**Test set investigation** We investigate our EVTPHYS and EVTGP test sets in different regions (Southern and Western Canada) and use Southern CONUS as an additional test region (Tables 2 and 4). We found that performance varied drastically across these regions. All models performed best in Canada West, with metric values similar to those in the validation set. In Southern CONUS, a large performance difference between models was observed, with the random forest and segmentation models performing best for both EVTPHYS and EVTGP. Overall, the random forest model consistently achieved the highest metrics across most test regions and classification granularities.

**Inference evaluation** All EVTPHYS models successfully capture the main vegetation patterns (see Figure 3 as an example), though performance declines both qualitatively and quantitatively on the higher granularity EVTGP dataset. This is likely the result of increased class quantity and similarity, highlighting the tradeoff between granularity and accuracy. Notably, even EVTPHYS contains similar classes, e.g., conifer, hardwood, and hardwood-conifer are 3 distinct classes which are more often misclassified (see Appendix Figure 6b). Such misclassifications are sometimes more tolerable in practice than the metrics would suggest although tolerance may vary widely depending on the downstream tasks. The strong performance of non-spatial models like logistic regression, gradient boosted trees, and random forest is likely due to AEF encoding surrounding spatial information into single pixel values. There is a notable performance drop across models in regions with ecology that substantially differs from most of the training region (details in Appendix Table 5).

**Limitations** While we treat EVT as ground-truth in our experimentation, it is inherently noisy as the output of imperfect models trained on labeled field and satellite data. Misclassifications certainly exist which affect the true metric evaluation of trained models. The better performance of models in Canada West over Canada South could possibly be explained by a seemingly artificial vegetation discontinuity in the ground truth data near the CONUS/Canada border, which is not present in the model predictions (Figure 3). This suggests some bias and potential inaccuracies in the measured performance. Moreover, we hypothesis that random forest's superior performance, even over the segmentation model, could be due to a structural bias in the ground truth labels as the EVT dataset is the product of decision tree models.

## 4 Conclusion

Given the global availability of the AEF model, the framework presented in this paper can be applied to other datasets or regions, opening up new opportunities for creating and improving environmental labels. For example, this pipeline could be used to generate improved EVT labels by leveraging AEF and LANDFIRE's raw labels from its Public Reference Database [23]. However, a key limitation is that AEF is only available from 2017 onwards. This work demonstrates a powerful approach to expand geospatial data crucial to climate analysis efforts from data-rich to data-scarce regions.

## References


[1] Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, pages 1–8, 2025.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025.

[4] María Calviño-Cancela, María L Chas-Amil, Eduardo D García-Martínez, and Julia Touza. Wildfire risk associated with different vegetation types within and outside wildland-urban interfaces. *Forest Ecology and Management*, 372:1–9, 2016.

[5] Daniel L Civco. Artificial neural networks for land-cover classification and mapping. *International journal of geographical information science*, 7(2):173–186, 1993.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[7] Mark A Friedl and Carla E Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3):399–409, 1997.

[8] Google and Google DeepMind. Google satellite embedding v1, 2025. Accessed: 2025-07-22.

[9] Google Earth Engine Google DeepMind. Satellite Embedding V1, 2025.

[10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[11] Georgef Hepner, Thomas Logan, Niles Ritter, and Nevin Bryant. Artificial neural network classification using a minimal training set- comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing*, 56(4):469–473, 1990.

[12] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*, 2025.

[13] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

[14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[15] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4347–4355, 2025.

[16] Inga P La Puma. Landfire technical documentation. *Open-File Report 2023-1045. Washington DC: US Department of the Interior, US Geological Survey. 103 p.*, 2023.

[17] LANDFIRE. 2020 existing vegetation type layer, landfire 2.0.0, u.s. department of the interior, geological survey, and u.s. department of agriculture, accessed 16 july 2025 at. *http://www.landfire/viewer*.

[18] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

[19] Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. *arXiv preprint arXiv:2003.00824*, 2020.

[20] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[23] Matthew G Rollins. Landfire: a nationally consistent vegetation, wildland fire, and fuel assessment. *International Journal of Wildland Fire*, 18(3):235–249, 2009.

[24] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1422–1431, 2022.

[25] Grant J Scott, Matthew R England, William A Starms, Richard A Marcum, and Curt H Davis. Training deep convolutional neural networks for land–cover classification of high-resolution imagery. *IEEE Geoscience and Remote Sensing Letters*, 14(4):549–553, 2017.

[26] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.

[27] Szymon Woźniak and Piotr Szymański. Hex2vec: Context-aware embedding h3 hexagons with openstreetmap tags. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 61–71, 2021.

[28] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2017.

[29] Pengbin Zhang, Yinghai Ke, Zhenxin Zhang, Mingli Wang, Peng Li, and Shuangyue Zhang. Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery. *Sensors*, 18(11):3717, 2018.

## A Full Author List and Affiliations


- Luc Houriez; X, the Moonshot Factory; Bellwether; houriezl@google.com
  Stanford University, Mechanical Engineering Department; houriezl@stanford.com
- Sebastian Pilarski; X, the Moonshot Factory; Bellwether; sebpilarski@google.com
- Behzad Vahedi; X, the Moonshot Factory; Bellwether; vahedi@google.com
- Ali Ahmadalipour; X, the Moonshot Factory; Bellwether; aliahma@google.com
- Teo Honda Scully; X, the Moonshot Factory; Bellwether; teonnaise@google.com
- Nicholas Aflitto; X, the Moonshot Factory; Bellwether; aflitto@google.com
- David Andre; X, the Moonshot Factory; Bellwether; davidandre@google.com
- Caroline Jaffe; X, the Moonshot Factory; Bellwether; cjaffe@google.com
- Martha Wedner; X, the Moonshot Factory; Bellwether; wedner@google.com
- Rich Mazzola; X, the Moonshot Factory; Bellwether; richmazzola@google.com
- Josh Jeffery; X, the Moonshot Factory; Bellwether; joshuajeffery@google.com
- Ben Messinger; X, the Moonshot Factory; Bellwether; bmessinger@google.com
- Sage McGinley-Smith; X, the Moonshot Factory; Bellwether; sagems@google.com
- Sarah Russell; X, the Moonshot Factory; Bellwether; sarahrussell@google.com


## B Acknowledgments


The authors wish to acknowledge the valuable input from the Google DeepMind AlphaEarth Foundations team [3] with regards to experiment design and paper review.

Luc Houriez acknowledges the support of the Stanford Data Science Scholars, advising from Martin Fischer (Stanford University, Civil and Environmental Engineering Department), and Eric Darve (Stanford University, Institute for Computational and Mathematical Engineering).


## C Appendix

**AlphaEarth Foundations (AEF)** In order to train the AEF model, the Google DeepMind team leveraged diverse datasets to serve as training inputs and targets. The training inputs consist of Sentinel and Landsat images. The targets consist of various data types including topography (Copernicus DEM), land cover (NLCD), and climate (ERA5-Land). The AEF model uses a self-supervised autoencoder network to learn a representation that enables the reconstruction of individual target datasets from only the input data. It utilizes a novel Space Time Precision (STP) Encoder specifically designed to process long-range relationships across time and space. Learning the final embeddings was accomplished with three neural networks 1) a teacher network that processes complete, unaltered input imagery, 2) a student network that has the same architecture as the teacher network and attempts to produce the same embedding as the teacher network albeit from perturbed or incomplete input data, and 3) a text alignment network that takes text descriptions from wikipedia and produces an embedding. The combination of four loss functions – reconstruction loss, consistency loss, text contrastive loss, and batch uniformity loss across these three networks – produces the final embedding. These learned representations are publicly available as a dataset of 64-dimensional vectors for each year, called "Satellite Embedding" dataset [8]. This dataset is a global, annual dataset with a spatial resolution of 10 meters which is currently available from 2017 to 2024.

**Hardware** The models in this paper are trained on a virtual machine with 160 Intel Broadwell vCPUs and 3844 GB of memory. For the segmentation model, we use a single A100 GPU.

**Semantic segmentation model** Figure 4 presents the U-Net architecture used in the semantic segmentation model. It consists of EfficientNet-B4 and a default U-Net decoder. EfficientNet-B4 is a convolutional neural network architecture built from MBConv blocks. It learns by training the encoder and decoder in tandem on an AEF image input and corresponding EVT image target.

The encoder processes the AEF's representations, extracting and compressing the most relevant signals for the target geospatial feature layer. The decoder reconstructs the label image from this compressed representation. By continuously comparing the model's predicted output against available ground-truth label images during training, the model learns to accurately encode and decode the environmental signals. We train without tile overlap to prevent data leakage to validation sets. During

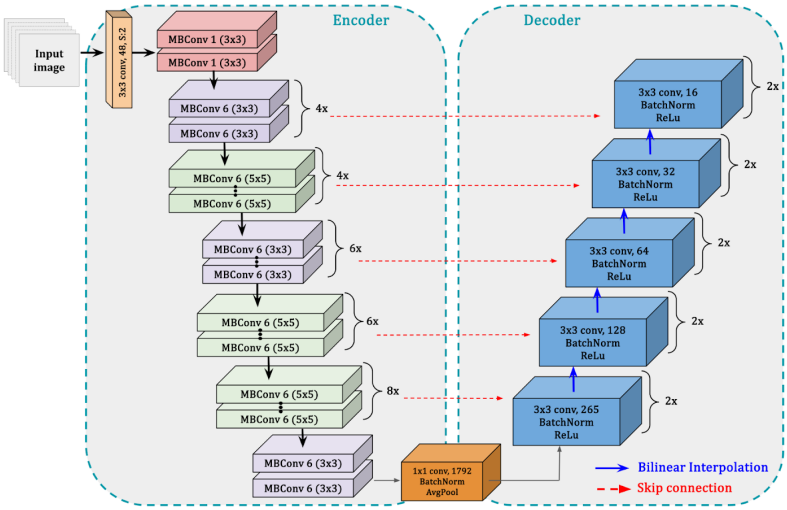



Figure 4: Model architecture. An encoder-decoder semantic segmentation network based on U-Net which uses EfficientNet-B4 as the encoder, and default U-Net decoder.

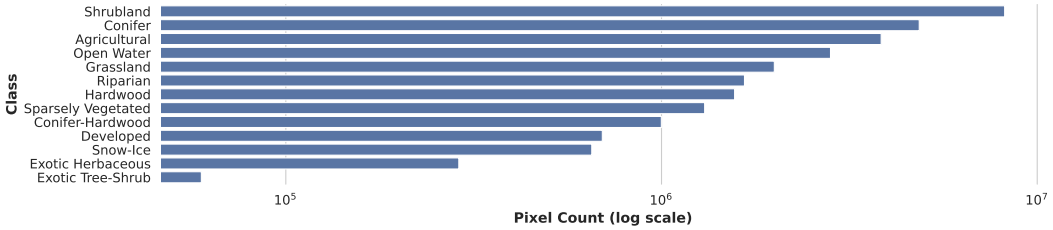
inference to unlabeled regions we set a 50% overlap on inference tiles to limit border artifacts. Final inference maps are generated by taking output probabilities taken from a final softmax layer and selecting the highest probability class after averaged smoothing across overlaps. We perform training data augmentation consisting of horizontal and vertical flips, random 90 degree rotations, and transposes each with 50% probability. During training we use Adam as our optimizer, with a learning rate scheduler which reduces on plateau. We train our segmentation model using cross-entropy loss. Training is limited to a maximum of 350 epochs with early stoppage (15 epochs) enabled.

**Class distributions** We present the class distributions for EVTPHYS (13 classes) and EVTGP (80 classes) in log scale in Figure 5. We provide names for the EVTPHYS classes. EVTGP class names are omitted due to quantity.

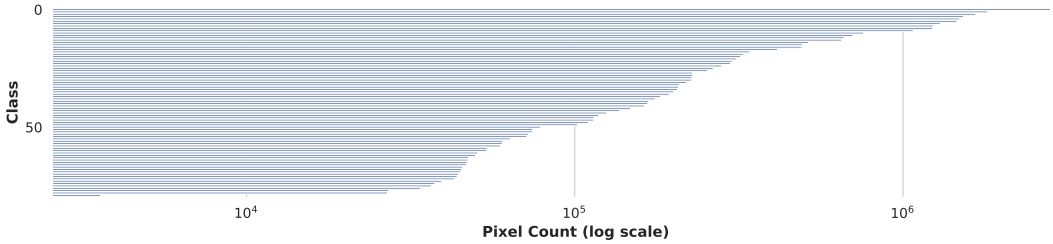
**Performance per class** It is generally expected that model performance will not be the same across all classes. Figure 6 showcases the segmentation model performance across EVTPHYS classes. Performance does not directly correspond with class quantity in the training set. Perhaps unsurprisingly, open water and snow-ice achieve highest precision values. Minority classes with similar more-common classes (e.g., exotic tree shrub, conifer-hardwood, exotic herbaceous) achieve lowest performance scores. Further grouping/clustering would likely significantly improve overall performance and better balance per-class metrics.

**Results for EVTGP** Results for EVTGP are presented in Table 3. A clear drop in performance can be observed for logistic regression. The segmentation model much better generalized to validation and test sets than gradient boosted trees. The random forest model experienced significant overfitting in training but outperformed all other models in validation. On the test set, performance dropped for all models, even more significantly than in EVTPHYS; which can also be partially attributed to the observed EVT discontinuity for Canada South. For EVTGP, the segmentation model performed similarly to random forest.

**Performance as a function of distance from the training region** In Table 5 we presented metric results for gradient boosted trees and segmentation model for EVTPHYS and EVTGP, respectively.

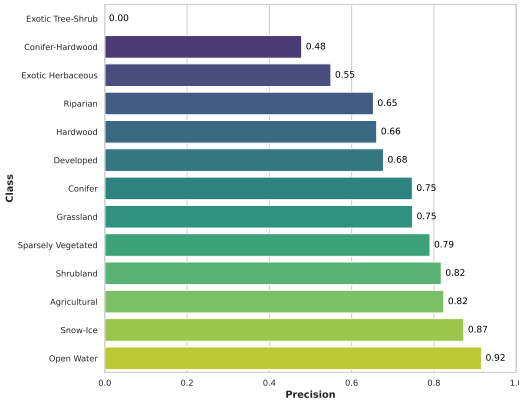



(a) Class distribution for EVTPHYS (13 classes)

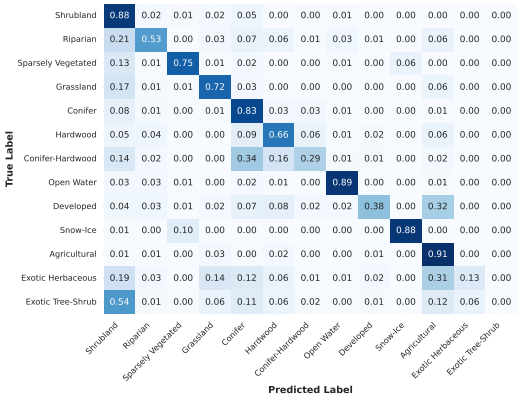



(b) Class distribution for EVTGP (80 classes)

Figure 5: Class distribution in the training data split for both EVTPHYS and EVTGP.




(a) Per-class precision




(b) Confusion matrix

Figure 6: Per-class performance by segmentation model for EVTPHYS

Table 3: Accuracy (ACC), Jaccard (J), and F1 across data splits for EVTGP (80 classes).

| | Training | | | Validation | | | Test | | |
|---|---|---|---|---|---|---|---|---|---|
| | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** |
| **Logistic Regression** | 0.60 | 0.28 | 0.40 | 0.60 | 0.27 | 0.39 | 0.42 | 0.11 | 0.17 |
| **Random Forest** | 0.96 | 0.94 | 0.96 | 0.71 | 0.39 | 0.53 | 0.48 | 0.16 | 0.23 |
| **Gradient Boosted Trees** | 0.63 | 0.29 | 0.43 | 0.62 | 0.28 | 0.41 | 0.44 | 0.12 | 0.17 |
| **Segmentation Model** | 0.65 | 0.28 | 0.40 | 0.66 | 0.29 | 0.41 | 0.48 | 0.15 | 0.21 |

All metrics decrease as the geographic distance from the training set (CONUS north of 41.6 degrees of latitude) increases.

**Southern CONUS test area** We compare random forest model inference to the ground truth in the Southern CONUS test area. Visual agreement seems relatively well achieved, with some notable discrepancies in Texas and New Mexico where vegetation types differ significantly from the training region (Alaska and northern CONUS).

Table 4: Model performances for EVTGP (80 classes) across 3 distinct test regions. *Canada South* and *Canada West* combined comprise the test set in Table 3 (see Data Splits in Section 2).

| | **Canada South** | | | **Canada West** | | | **Southern CONUS** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** |
| **Logistic Regression** | 0.33 | 0.07 | 0.11 | 0.64 | 0.21 | 0.29 | 0.38 | 0.10 | 0.16 |
| **Random Forest** | 0.40 | 0.11 | 0.16 | 0.68 | 0.27 | 0.37 | 0.48 | 0.14 | 0.21 |
| **Gradient Boosted Trees** | 0.36 | 0.07 | 0.10 | 0.63 | 0.08 | 0.11 | 0.30 | 0.06 | 0.11 |
| **Segmentation Model** | 0.40 | 0.10 | 0.15 | 0.68 | 0.23 | 0.31 | 0.46 | 0.10 | 0.16 |

Table 5: Test results (Accuracy, Jaccard and F1 scores) for different models across distinct latitude bands within the CONUS region.

| | **Lat. 41.6 to 38.6** | | | **Lat. 38.6 to 35.6** | | | **Lat. 35.6 to 33.6** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** | **ACC** | **J** | **F1** |
| **Gradient Boosted Trees EVTPHYS (13 classes)** | 0.76 | 0.42 | 0.53 | 0.69 | 0.34 | 0.45 | 0.55 | 0.26 | 0.37 |
| **Segmentation Model EVTGP (80 classes)** | 0.58 | 0.13 | 0.19 | 0.48 | 0.09 | 0.14 | 0.34 | 0.06 | 0.09 |

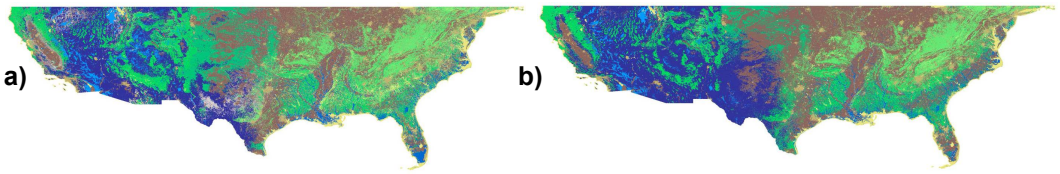



Figure 7: Southern CONUS test area (below latitude 41.6). EVTGP ground truth (a) versus inference using random forest (b)

**Comparing model inferences** We compare model inferences of EVTGP and EVTPHYSfor all models in Figures 8 and 9. Overall, good consistency is observed. Inference from logistic regression looks quite different for EVTGP.

On a zoomed-out prediction map, the majority of predicted pixel classes look consistent between the four evaluated models. One notable observation is that the segmentation model inference results in regions of more consistent vegetation than the other models (thresholding). Other model outputs appear more noisy, with neighboring pixels less likely to belong to the same vegetation class and less clearly defined class boundaries (more akin to LANDFIRE EVT data). This behavior exhibited by the segmentation model doesn't necessarily mean it underperforms when comparing its outputs to satellite imagery, quite the contrary. One notable example is near Peace River in AB, Canada (Figure 10). In this region, the segmentation model correctly identifies (verified by satellite imagery) sections of agricultural land (in brown). Logistic regression, random forests, and gradient boosted trees (pictured) all incorrectly label large sections of the region as shrubland (blue). While this example is not necessarily representative across the entire datasets, it does point to the possibility that the metrics are not capturing true generalization or performance of respective models.

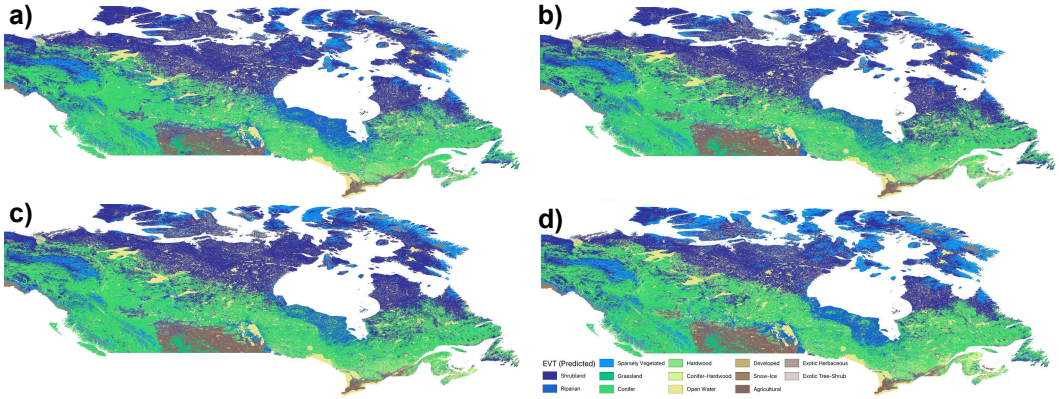



Figure 8: EVTPHYS (13 classes) inference in Canada using: (a) logistic regression, (b) random forest, (c) gradient boosted trees and (d) segmentation model.

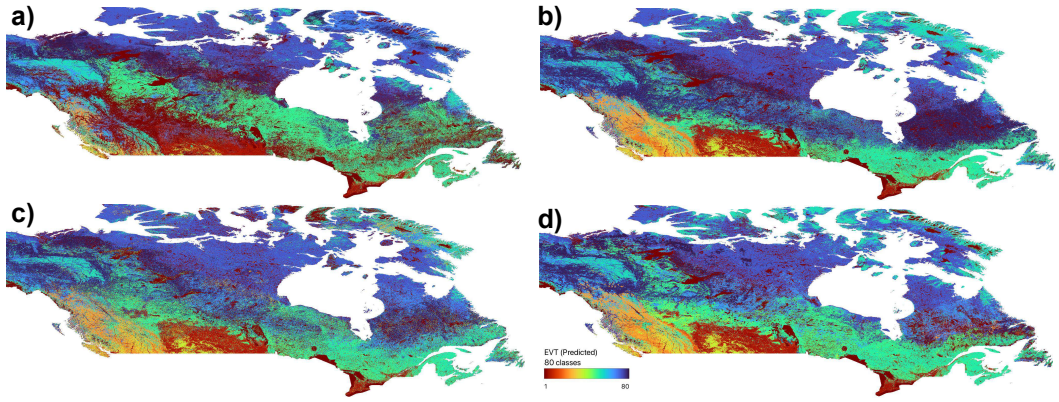



Figure 9: EVTGP (80 classes) inference in Canada using: (a) logistic regression, (b) random forest, (c) gradient boosted trees and (d) segmentation model.

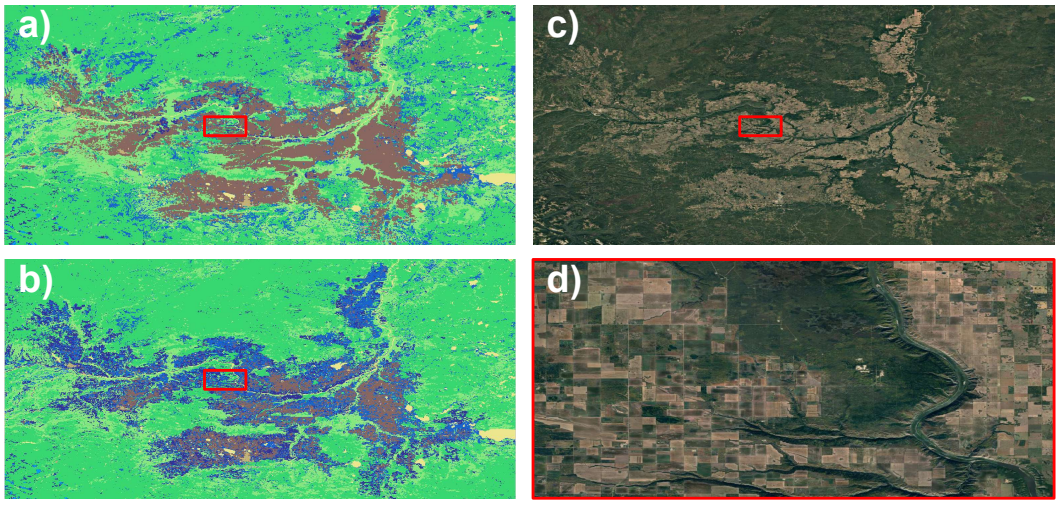



Figure 10: Segmentation model captures agriculture land (brown EVTPHYS) better than gradient boosted tree near Peace River in AB, Canada. a) Segmentation model inference, b) Gradient boosted tree model inference, c) Satellite imagery, d) Satellite zoomed-in imagery.