
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: EVT_PHYS (13 classes) and EVT_GP (80 classes). Qualitatively, for EVT_PHYS, model predictions align with ground truth. Trained models achieve 81% and 73% classification accuracy on EVT_PHYS validation sets in the USA and Canada, despite discussed limitations.

1 Introduction

High-quality environmental and climate data is essential for understanding and managing our planet. Nevertheless, comprehensive geospatial datasets that span all regions of the Earth remain rare. This lack of global coverage presents major challenges for applications such as environmental monitoring, disaster prediction, meteorological and climate analysis, and natural resource management.

Traditional geospatial data collection faces significant barriers: acquisition costs are often prohibitive, and integrating multiple datasets requires reconciling vastly different spatial resolutions and temporal sampling rates. While valuable datasets do exist, they are often confined to specific regions or capture

*These authors contributed equally to this work.

†A complete list of authors and their affiliations is available in the appendix.

only a limited subset of variables needed for comprehensive analysis. Developing reliable methods to extend data coverage into underrepresented areas significantly enhances our ability to monitor earth system changes, design novel applications in otherwise overlooked regions, and make informed predictions at global scales.

Representation learning models have the potential to fundamentally transform Earth observation capabilities, paralleling the transformative impact of large language models. By training on massive, heterogeneous geospatial datasets, these models learn rich embeddings that encode complex spatiotemporal dynamics and underlying geophysical processes across Earth systems. The key breakthrough lies in their generalization capabilities. They reduce the need for extensive labeled data for specific tasks, as their pre-training allows for more efficient fine-tuning. These learned representations can enable accurate predictions and interpolations of earth system properties in data-sparse regions.

In this paper, we leverage the embeddings generated by Google DeepMind’s AlphaEarth Foundations (AEF) model, to address the challenge of data scarcity. AEF generates high-resolution embeddings derived from satellite imagery sources, including Landsat 8 and 9 and Sentinel 1 and 2. A significant advantage of these embeddings is their inherent global coverage and consistent quality. This characteristic makes AEF embeddings particularly well-suited for synthesizing environmental data in otherwise data-deprived areas.

Our core contribution is the development and validation of a pipeline that utilizes the AEF embeddings to generate synthetic environmental datasets. Specifically, we provide a pipeline which trains machine learning models using the globally available AEF embeddings as input features and corresponding ground truth environmental data (e.g., vegetation indices or climatological parameters) as labels. After training, these models can be deployed to infer these same environmental features in areas lacking direct observations by leveraging the AEF embeddings as their input. The global nature of AEF ensures that this methodology is broadly applicable, offering a scalable solution for environmental data augmentation across the planet.

1.1 Related Works

Embedding Models The utility of learned vector representations or embeddings was first widely adopted in the field of natural language processing with models such as word2vec [23], GloVe [27], ELMo [28], and BERT [6]. In recent years, there has been a surge of research into the use of embeddings in geospatial applications [21] with works on models such as place2vec [34], tile2vec [16], hex2vec [33], space2vec [22], GeoVeX [7], Terramind [15], Prithvi [1], and SatCLIP [18]. Most recently, Google DeepMind introduced AlphaEarth Foundations, a multi-modal multi-source geospatial embedding model. This model is exposed to a larger breadth of high quality data sources than previous works and is explicitly trained to be performant across a diverse set of geospatial labeling tasks to make it general purpose [3]. As such, AEF is designed to be effectively used for regression, classification, and segmentation style tasks.

Classification and Segmentation Researchers have extensively applied machine learning to land cover classification and segmentation, resulting in a rich body of existing literature. Artificial Neural Networks and satellite data were used for land cover classification since the 1990s [13, 4]. Later on, decision-trees [8, 5], Random Forests [24, 9], and Support Vector Machines [14, 25] gained traction for land cover classification. Following the advent of deep neural networks in the early 2010s, a variety of such methods along with new sensors and satellite data were implemented for more accurate mapping and classification [31, 36]. In the early 2020s, vision transformers were extensively employed for classification and segmentation of remotely sensed data [30, 35]. A common theme amongst all these works is that they each utilize limited data sources with most limited to one or fewer instruments (e.g. visible, infrared, or SAR data). AEF provides an easy interface into a learned representation spanning many data sources and multiple types of satellite imagery including optical, thermal, and Synthetic Aperture Radar (SAR). As such, it can simplify the process of learning important labels and classifications.

1.2 AlphaEarth Foundations (AEF)

Representation learning has proven to be a powerful technique as evidenced in natural language processing and computer vision. The central technique is to utilize encoder models to turn raw

data into numerical embeddings. These embeddings are multi-dimensional vectors that capture the semantic essence and spatial relationships within data, thus translating high-dimensional data into lower-dimensional representations to improve data efficiency. Google DeepMind recently released the general-purpose geospatial model AlphaEarth Foundations (AEF) [3]. It learns a dense, global representation of the Earth's surface, capturing intricate patterns and interdependencies. Essentially, the AEF transforms complex Earth observation data into a more structured and dense latent representation, which although less semantically meaningful for humans, provides a structure more amenable for applying ML methods towards downstream tasks.

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 [10]. This dataset is a global, annual dataset with a spatial resolution of 10 meters which is currently available from 2017 to 2024.

1.3 Problem Statement and Objective

Problem statement Raw or information dense data (e.g., satellite or embedding) seldom provides easily extractable insights. Insight extraction usually involves visualizations or transformations to simpler easier-to-understand representations. In geospatial settings, this often takes the form of specified geospatial 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.

Objective The main aim of this article is to investigate whether AEF could be leveraged to generate labels for previously unlabeled regions. Given that the AEF builds on high-quality data sources covering many inputs and modalities, it has the potential to serve as an interpolation base that provides labels with a high-degree of accuracy.

2 Extending Datasets to New Regions through AEF: A Case Study

Vegetation type is an important feature humans use when interpreting wildfire risk as it helps understand potential wildfire spread and ignition. Therefore, labeled vegetation information is extremely valuable to explain the outputs of machine-learned wildfire risk models. In this paper, we investigate extending an existing vegetation type dataset from the USA to Canada. This serves as a case study of AEF's potential for dataset extension.

2.1 Data

EVT LANDFIRE is a multi-agency program run by the US Forest Service, the Department of the Interior and US Geological Survey [29, 19]. LANDFIRE provides an ecological dataset called Existing Vegetation Type (EVT) which has historically been used for wildfire management efforts. Developing this dataset required coordinating multiple government agencies to build labeled datasets and training decision tree models to predict existing vegetation type. However, the dataset is limited to the United States (continental, Alaska, Hawaii and insular areas) and serves as a prime example for highly useful data which does not extend globally. In our motivating use case, extending the EVT data to Canada would enable us to make better sense of wildfire risk predictions in Canadian regions.

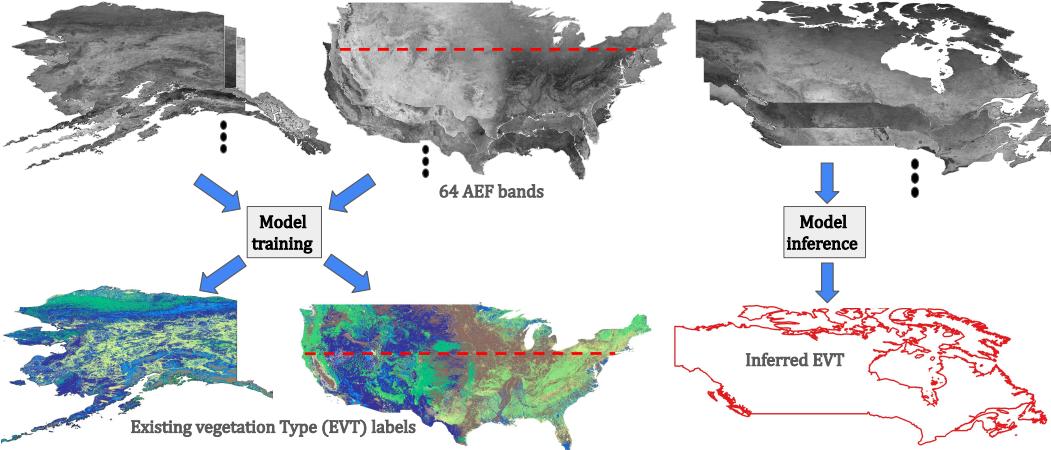


Figure 1: Schematic of model training and inference. The 64 bands of AEF data (input) and EVT data (target) from continental USA above the red dotted line and Alaska are used to train the model. Running inference on AEF data in Canada provides expected EVT in the previously unlabeled region.

EVT granularities The EVT dataset consists of labels at various levels of classification granularity. LANDFIRE provides label 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 on both EVTPHYS (low) and EVTGP (mid) level granularities.

For our model training we set AEF as input and EVT as target output as presented in Figure 1. Notably, AEF embeddings were evaluated with maximal classification targets of 40 classes. Hence EVTGP serves as an investigation into the ability of AEF to generalize to even more difficult classification tasks than originally optimized for.

Data selection Deliberate design decisions need to be made when extending a labeled dataset into another region like Canada. In our context, 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 [20].

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 in the training region to address class imbalance. This results in 80 classes spanning the selected continental US and Alaska regions. Pixels not belonging to one of these 80 are ignored in training via a mask. 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 [11] and train on AEF and EVT with 500m resolutions.

Data splits We geographically tile our northern CONUS and Alaska data into tiles of size 64x64. We allocate 90% of these tiles for training and 10% for validation. The EVT dataset provides data for a southernmost 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. In all, 4.1 million pixels are used for testing (2.9 million in northern CONUS and 1.2 million in Alaska) and 34.6 million pixels are used for training and validation.

2.2 Models

Our proposed methodology to extend existing datasets is to train machine learning models to predict labels (e.g., EVT) from AEF inputs. We have developed a flexible and adaptable pipeline wherein model architectures can easily be swapped. This enables rapid experimentation with architectures designed to optimize for certain data characteristics or desired output granularity or resolution. While

we apply our methodology and pipeline to extend datasets into Canada from the USA in this article, they are flexible and can be applied to other regions. We evaluate four models for this case study:

Logistic regression We train 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 [12, 26].

Random forest We leverage scikit-learn’s RandomForestClassifier model to learn the AEF to EVT mapping at the pixel level as above [2, 26].

Gradient boosted tree We use the LGBMClassifier from the LightGBM library [17]. As for the linear and random forest models, individual pixels make up the inputs and outputs.

Segmentation Model We employ an encoder-decoder semantic segmentation pipeline. 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 opted for a fully-convolutional U-Net architecture using EfficientNet-B4 [32] pre-trained on advprop for the encoder (see Figure 5 in Appendix C for architecture details). We train without tile overlap to prevent data leakage to validation sets. During 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.

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.

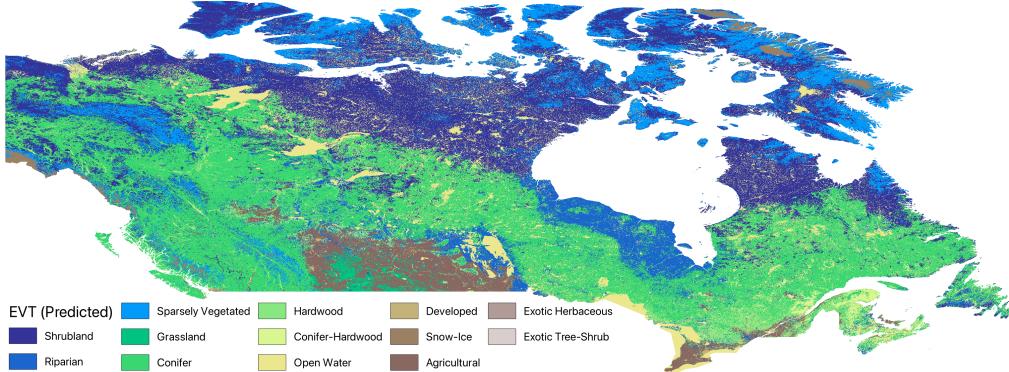
Extending datasets Once training is complete and validated to achieve acceptable metrics, the model is saved. The saved model can then be seamlessly deployed to perform inference in any desired region (including previously unlabeled ones) by running it on AEF inputs of that region. Models are appropriately tagged to prevent deployment to regions wherein the target features may fall into completely different distributions as this will likely not yield appropriate labels.

3 Results

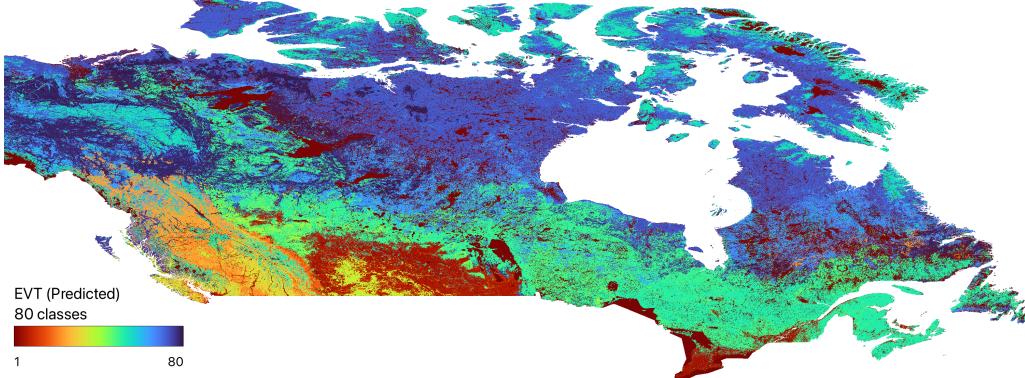
Qualitative assessment We present the segmentation model inference maps in Canada for EVT-PHYS and EVTGP (Figure 2). EVT-PHYS inference maps look very similar across all models (Figure 9) with differences discussed in Section 4. The inference maps seem to exhibit feasible vegetation type continuity. No obvious unexpected vegetation discontinuities or transitions are present. EVTGP inference maps generally agree outside northern Canada where significant differences exist. See Figure 10 for details.

Metrics We quantitatively evaluate the performance of each of our models on both EVT-PHYS and EVTGP classification granularities using three metrics: Accuracy (ACC), Jaccard Index (J), and F1 score. Both Jaccard Index and F1 are macro-averaged, meaning they are calculated as an unweighted average of the scores for each class. For all three metrics, a higher score indicates better performance. Table 1 and Table 2 present results across the training, validation, and test sets defined in Section 2.1.

EVT-PHYS Logistic regression, gradient boosted trees, and the segmentation model achieved similar performance across all metrics for both EVT-PHYS training and validation (near 0.79 accuracy, 0.52 Jaccard index, and 0.65 F1 score). The random forest model yielded slightly higher metrics in



(a) EVT_{PHYS} (13 classes)



(b) EVT_{GP} (80 classes)

Figure 2: Inference in Canada generated by the segmentation model.

Table 1: Accuracy (ACC), Jaccard (J), and F1 across data splits for EVT_{PHYS} (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

validation (0.81 accuracy, 0.55 Jaccard, 0.67 F1) and extremely high training values which suggest overfitting (0.97 accuracy, 0.95 Jaccard, 0.97 F1). Significant metric deterioration is observed for all models on the test split. All four models achieved very similar test metrics with random forest besting in all three: 0.73 accuracy, 0.43 Jaccard index, and 0.55 F1 score. Full results are shown in Table 1.

EVT_{GP} In the case of EVT_{GP}, gradient boosted trees and the segmentation model outperformed logistic regression substantially. Gradient boosted trees and segmentation achieve similar metrics for training, but the segmentation model much better generalized to validation and test sets. The segmentation model achieved 0.65 accuracy, 0.28 Jaccard, and 0.40 F1 scores for the training split and similar values for validation. On the test set performance dropped for all models, even more significantly than in EVT_{PHYS}; the segmentation model had 0.48 accuracy, 0.15 Jaccard, and 0.21 F1. The random forest model again experienced the overfitting phenomenon in training with 0.96 accuracy, 0.94 Jaccard, and 0.96 F1. It outperformed all other models significantly in validation with 0.71 accuracy, 0.39 Jaccard, and 0.53 F1, but performed similarly to segmentation on the test set. Jaccard index and F1 scores are not biased by majority class which, with many classes, likely explains their low values. Full results are shown in Table 2.

Table 2: 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

Table 3: 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 2.1).

	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

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 2 (see Data Splits in Section 2.1).

	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

Further test set investigation As metrics for all models are significantly lower on the test set, we investigate it in increased detail and add an additional independent set. First, we separate the original test set (Section 2.1) into two independent ones: the southern Canada 90km band and the western Canada 90km band. We also create a test set consisting of CONUS south of the 41.6 degrees latitude line as this was not included in training. Results for EVTPHYS and EVTGP are presented in Tables 3 and 4. Model performance varies drastically across these three regions. On EVTPHYS, all models achieve reasonably similar metric values when measured in Canada South and Canada West. A clear accuracy performance difference exists between Canada South and Canada West, with Canada West exhibiting 0.14 higher accuracy on average across all four models. Generally, Jaccard and F1 scores are also better for Canada West, but exceptions exist such as for gradient boosted trees. Overall, Canada West metrics are much closer to those for the validation set (we discuss possible reasons for this in Section 4: Inference Evaluation and Figure 3). Larger performance differences are observed in Southern CONUS, with random forest and segmentation achieving highest metrics (0.68 accuracy, 0.36 Jaccard, 0.48 F1). This performance difference is amplified for EVTGP, where random forest and segmentation clearly are best across all regions. Random forest achieves highest metric measures across nearly all test regions and metrics for both EVTPHYS and EVTGP.

4 Discussion

EVTPHYS vs EVTGP metrics All models trained to predict EVTGP (80 classes) achieve significantly lower metric-evaluated performances as compared to models trained for EVTPHYS (13 classes). This is expected as the number of classes and similarity between classes grows. Importantly, these metrics do not consider relative similarity between classes. Notably, even EVTPHYS contains similar classes, e.g., conifer, hardwood, and hardwood-conifer are 3 distinct classes. As can be seen in Figure 7b, the conifer-hardwood class is often misclassified as either conifer or hardwood. Such misclassifications are often much more tolerable in practice than the metrics would suggest.

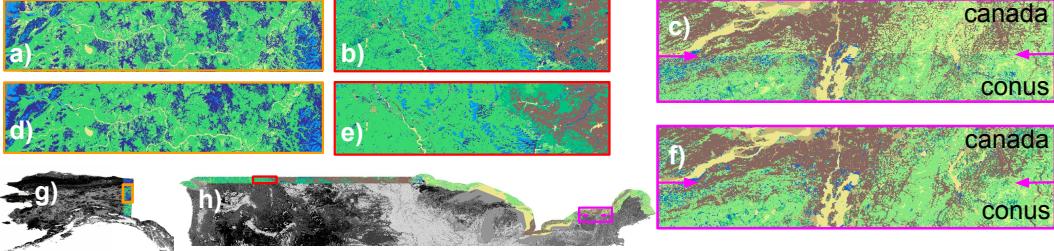


Figure 3: Ground truth EVTPHYS (a–c) compared to gradient boosted trees model inference (d–f) in Canada West (g) and South (h) test regions. Figures (c, f) additionally show land in CONUS across the border which is indicated by the magenta arrows. There, EVT values produced by LANDFIRE seem to exhibit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

Performance per class It is generally expected that model performance will not be the same across all classes. Figure 7 showcases the segmentation model performance across EVTPHYS classes. Performance does not directly correspond with class quantity in the training set (see Appendix C). 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.

Inference evaluation Most evaluated models experienced similar metric performance across training and validation sets (Tables 1 and 2). Inference on the test sets, which inherently have different underlying distributions from the training regions, results in significantly lower metrics. Performance across Southern CONUS is likely low due to significant ecological changes in latitudes not present in Alaska and northern CONUS. Notably, we observe that predicted class accuracy drops as a function of distance from labeled training regions (Table 5). As previously noted, models generally achieved higher performance metrics on Canada West than Canada South (Tables 3 and 4). Upon inspection of the ground truth labels, it can be seen in c) of Figure 3 that the EVT dataset exhibits an abrupt vegetation change across the CONUS/Canada border which is believed to be artificial, and which isn’t seen across the Alaska/Canada border. Across all investigated models, predicted outputs are continuous and do not exhibit such a discontinuity. This suggests that the ground truth label for significant portions of this band include some bias which could lead to lower model measured performance in testing than reality. Nevertheless, all models clearly capture the main EVTPHYS patterns as can be seen on the test tiles of Figure 3. Inference quality drops 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. Importantly, we expect that tolerance on accuracy may vary widely depending on the downstream tasks. EVTGP also experience performance discrepancies across test regions (Table 4) which can also be partially attributed to the observed EVT discontinuity.

Random forest performance Random forest, logistic regression, and gradient boosted trees all managed to achieve very good performance from singular pixel values. This is likely the result of inherent surrounding spatial information encoded into single pixels as a result of AEF being trained on 1.28km neighborhood tiles³. Random forest, surprisingly, outperformed the other models including the segmentation model. This may be due to the fact that EVT labels were generated by decision tree models [20]. These ground truth labels are not necessarily always correct (output of a predictive model) and they may introduce a exploitable structural bias for a particular model architecture.

Model discrepancies 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).

³It should be mentioned that this likely results in some minimal leakage of data splits along the tile borders, but should not be significant

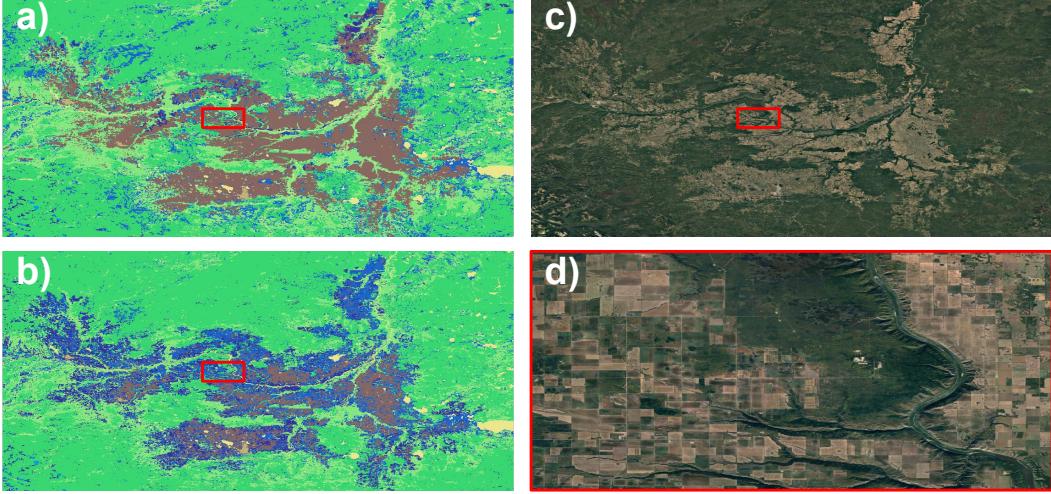


Figure 4: Segmentation model captures farmland (brown EVTPhys) better than gradient boosted tree near Peace River in AB, Canada. a) Segmentation model inference, b) Gradient boosted tree model inference, c) Satellite imagery, d) Satellite zoomed-in imagery.

This behavior exhibited by the segmentation model doesn't mean it underperforms when comparing its outputs to satellite imagery, quite the contrary. One notable example is near Peace River in AB, Canada (Figure 4). 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). Figure 10 in Appendix C compares all inference maps for EVTGP. Certain discrepancies are visible between the segmentation model and gradient boosted tree, e.g., the darker blue class extends closer to the Canadian border in the gradient boosted tree model. While the examples in this paragraph are handpicked counterexamples and 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.

EVT Indeed, while we treat EVT as ground-truth in our experimentation, it is inherently noisy as the output of imperfect decision tree models trained on labeled field and satellite data. Misclassifications certainly exist which affect the true metric evaluation of trained models. AEF opens up new opportunities for potential EVT improvement. A pipeline like the one presented here could be used to train EVT models using AEF (and possibly other inputs) and LANDFIRE's raw labeled data (target). These approaches could possibly lead to better EVT labels across the USA and beyond. A caveat is that AEF is trained on Sentinel data, which limits it to 2017 and onwards.

5 Conclusion

This article presents a flexible framework and pipeline that leverages Google DeepMind's AEF model to extend labeled geospatial data to new regions. This pipeline trains machine learning model using AEF embeddings as input features and corresponding ground truth environmental data as targets. After training, this model can be deployed to predict labels in previously data-scarce (or missing) regions. We evaluate the approach by using it to extend LANDFIRE's existing vegetation type (EVT) dataset from the USA to Canada. At a lowest EVT granularity level, EVTPhys (13 classes), models achieved up to 69%, 83%, and 68% accuracy across Canada South, Canada West, and Southern CONUS test sets. At a higher granularity, EVTGP (80 classes), accuracy drops to 40%, 68%, and 48%, respectively. We discuss limitations of measured metrics. Visually and qualitatively, predictions, especially for EVTPhys, agree with ground truth classes. Given the AEF model is available globally, the work in this paper can be reapplied to other datasets or regions and environmental labels from data-rich to data-scarce regions where similar labels are expected. This opens up tremendous opportunities for improving interpretability of geospatial information across 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] Daniel L Civco. Artificial neural networks for land-cover classification and mapping. *International journal of geographical information science*, 7(2):173–186, 1993.
- [5] RS De Fries, M Hansen, JRG Townshend, and R Sohlberg. Global land cover classifications at 8 km spatial resolution: The use of training data derived from landsat imagery in decision tree classifiers. *International Journal of Remote Sensing*, 19(16):3141–3168, 1998.
- [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] Daniele Donghi and Anne Morvan. Geovex: Geospatial vectors with hexagonal convolutional autoencoders. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 3–13, 2023.
- [8] 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.
- [9] Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R Sveinsson. Random forests for land cover classification. *Pattern recognition letters*, 27(4):294–300, 2006.
- [10] Google and Google DeepMind. Google satellite embedding v1, 2025. Accessed: 2025-07-22.
- [11] Google Earth Engine Google DeepMind. Satellite Embedding V1, 2025.
- [12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
- [13] 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.
- [14] Chengquan Huang, LS Davis, and JRG Townshend. An assessment of support vector machines for land cover classification. *International Journal of remote sensing*, 23(4):725–749, 2002.
- [15] 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.
- [16] 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.
- [17] 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.

- [18] 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.
- [19] 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.
- [20] 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>.
- [21] 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.
- [22] 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.
- [23] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013.
- [24] Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005.
- [25] Mahesh Pal and Paul M Mather. Support vector machines for classification in remote sensing. *International journal of remote sensing*, 26(5):1007–1011, 2005.
- [26] 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.
- [27] 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.
- [28] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.
- [29] Matthew G Rollins. Landfire: a nationally consistent vegetation, wildland fire, and fuel assessment. *International Journal of Wildland Fire*, 18(3):235–249, 2009.
- [30] 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.
- [31] Grant J Scott, Matthew R England, William A Starns, 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.
- [32] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.
- [33] Szymon Woźniak and Piotr Szymbański. 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.
- [34] 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.

- [35] Jing Yao, Bing Zhang, Chenyu Li, Danfeng Hong, and Jocelyn Chanussot. Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.
- [36] 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 / Supplemental Material

Semantic segmentation model Figure 5 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.

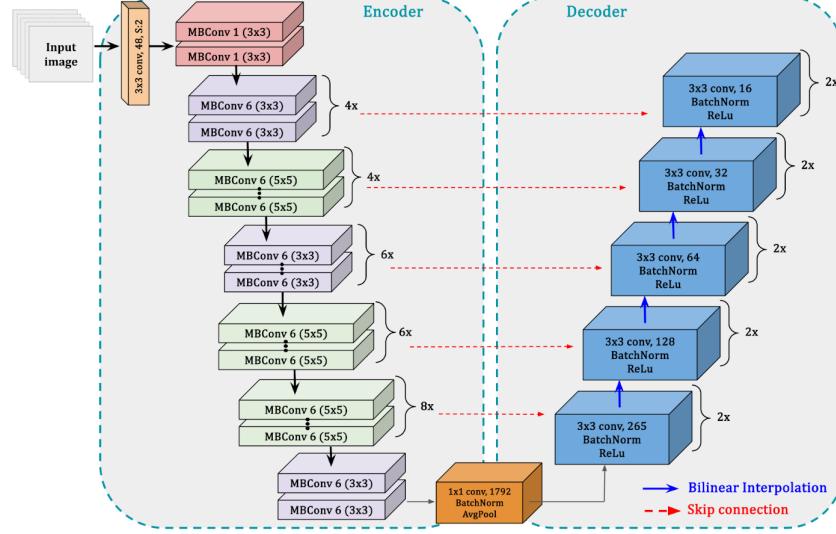


Figure 5: Model architecture. An encoder-decoder semantic segmentation network based on U-Net which uses EfficientNet-B4 as the encoder, and default U-Net decoder.

Class distributions We present the class distributions for EVTPHYS (13 classes) and EVTGP (80 classes) in log scale in Figure 6. We provide names for the EVTPHYS classes. EVTGP class names are omitted due to quantity.

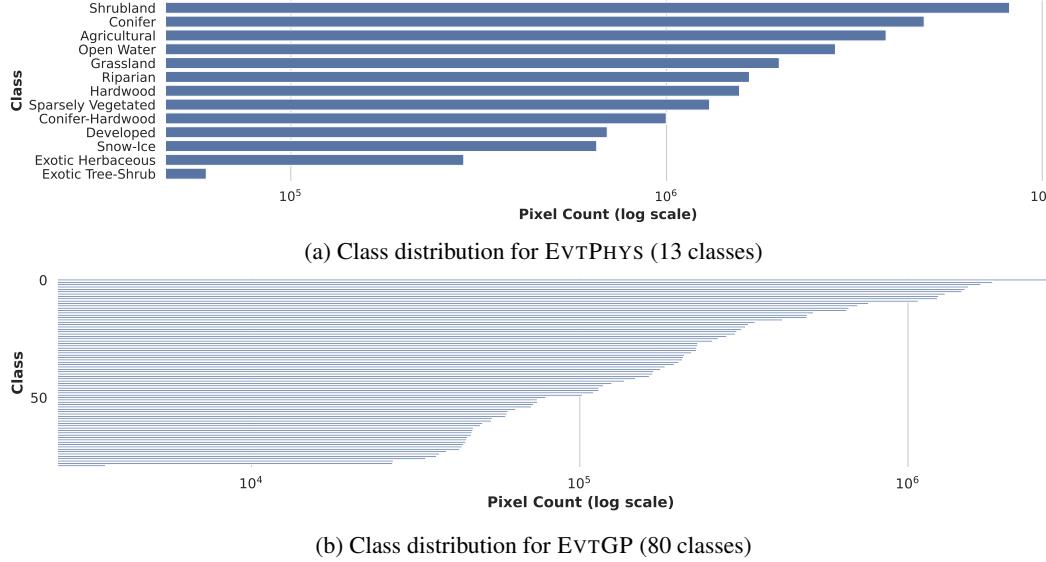


Figure 6: Class distribution in the training data split for both EVTPHYS and EVTGP.

Class performance by segmentation model Model performance is not consistent across all classes. In Figure 7 we present class-specific performance for the segmentation model. Note that precision drops to 0 for the exotic tree shrub as the model misclassifies all instances as just regular shrubland.

Random forest achieves 7% accuracy for this class. Prediction accuracy often drops for less-prevalent classes with similar larger classes.

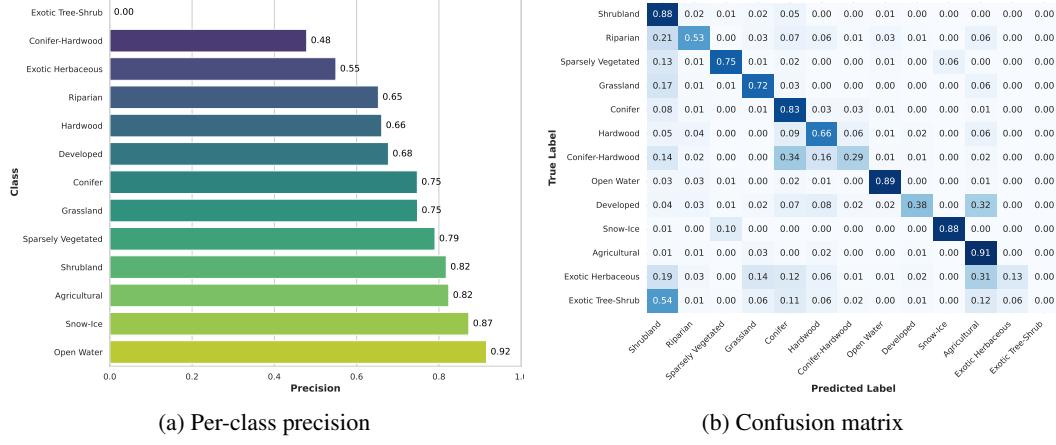


Figure 7: Per-class performance by segmentation model for EVTPHYS

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

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
Segmentation Model EVTGP (80 classes)	0.58	0.13	0.19	0.48	0.09	0.14	0.34	0.06	0.09

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.

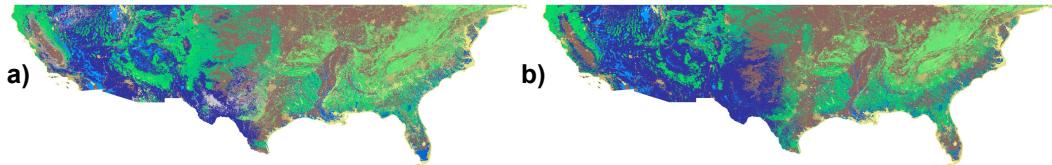


Figure 8: Southern CONUS test area (below latitude 41.6). EVTGP ground truth (a) versus inference using random forest (b)

Comparing model inferences We compare model inferences for all models for EVTGP and EVTPHYS. Overall, good consistency is observed. Inference from logistic regression looks quite different for EVTGP.

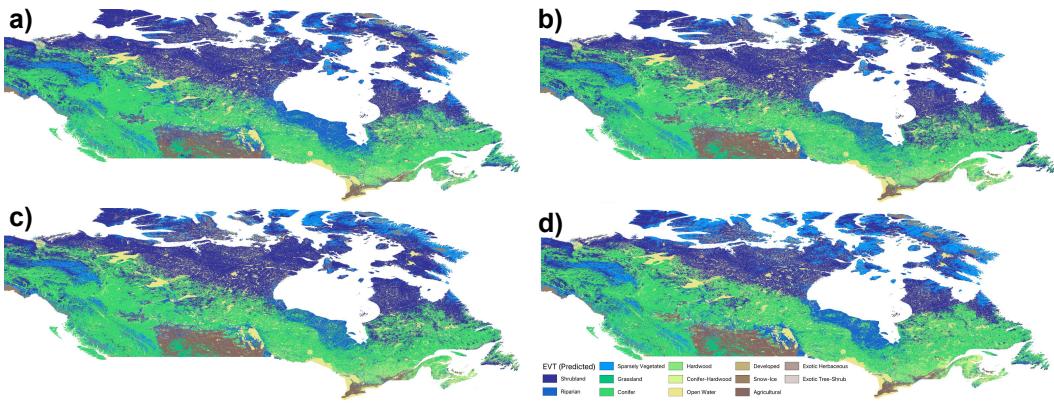


Figure 9: EVT_{PHYS} (13 classes) inference in Canada using: (a) logistic regression, (b) random forest, (c) gradient boosted trees and (d) segmentation model.

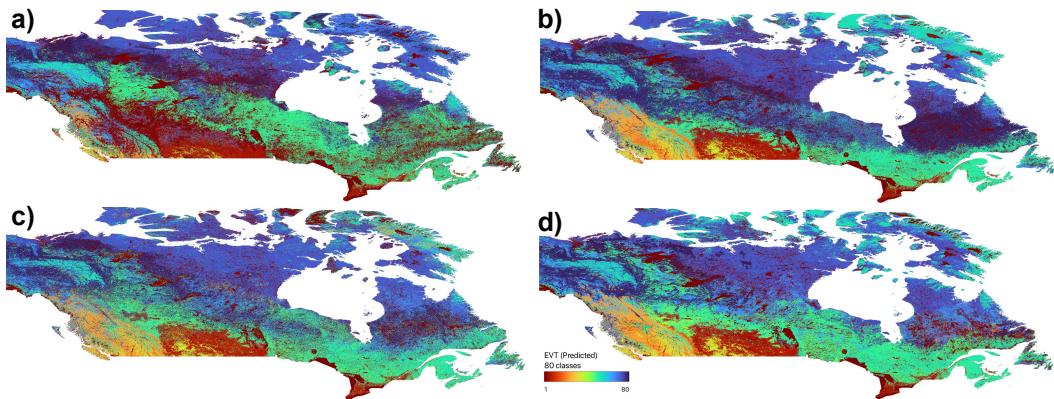


Figure 10: EVT_{GP} (80 classes) inference in Canada using: (a) logistic regression, (b) random forest, (c) gradient boosted trees and (d) segmentation model.