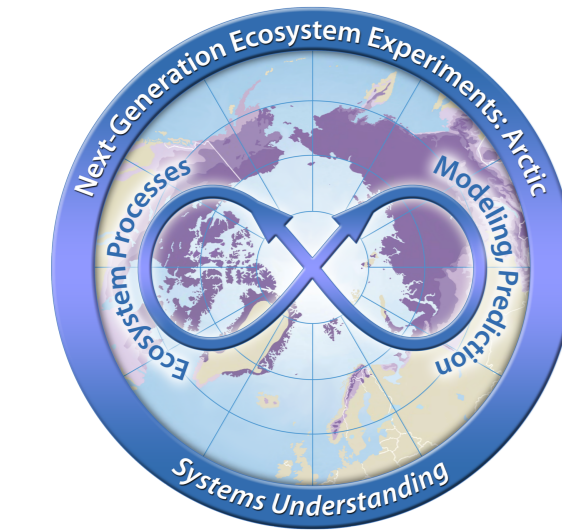# Landscape Characterization of Arctic Ecosystems Using Data Mining Algorithms and Large Geospatial Datasets

Zachary L. Langford[†‡] , Jitendra Kumar[‡], and Forrest M. Hoffman[‡]

[†]University of Tennessee - Knoxville and [‡]Oak Ridge National Laboratory (ORNL)

Introduction

- Accurate characterization is useful to understand the properties and organization of the landscape, optimal sampling network design, measurement and process up-scaling and to establish a landscape-based framework for multi-scale modeling of ecosystem processes.
- This study seeks to map land cover types in the Seward Peninsula of Alaska using large volumes of high-resolution satellite remote sensing datasets (Figure 2).
- We used data analytics algorithms applied to Phased Array type L-band Synthetic Aperture Radar (PALSAR), Interferometric Synthetic Aperture Radar (IfSAR), and Landsat 8 Operational Land Imager (OLI) datasets to develop high-resolution (~12 m) land cover maps (Table 1).
- PALSAR's L-band SAR yields detailed, all-weather, day-and-night observation.
- **We seek to evaluate the sensitivity between optical and PAL-SAR datasets for specific land cover types.**
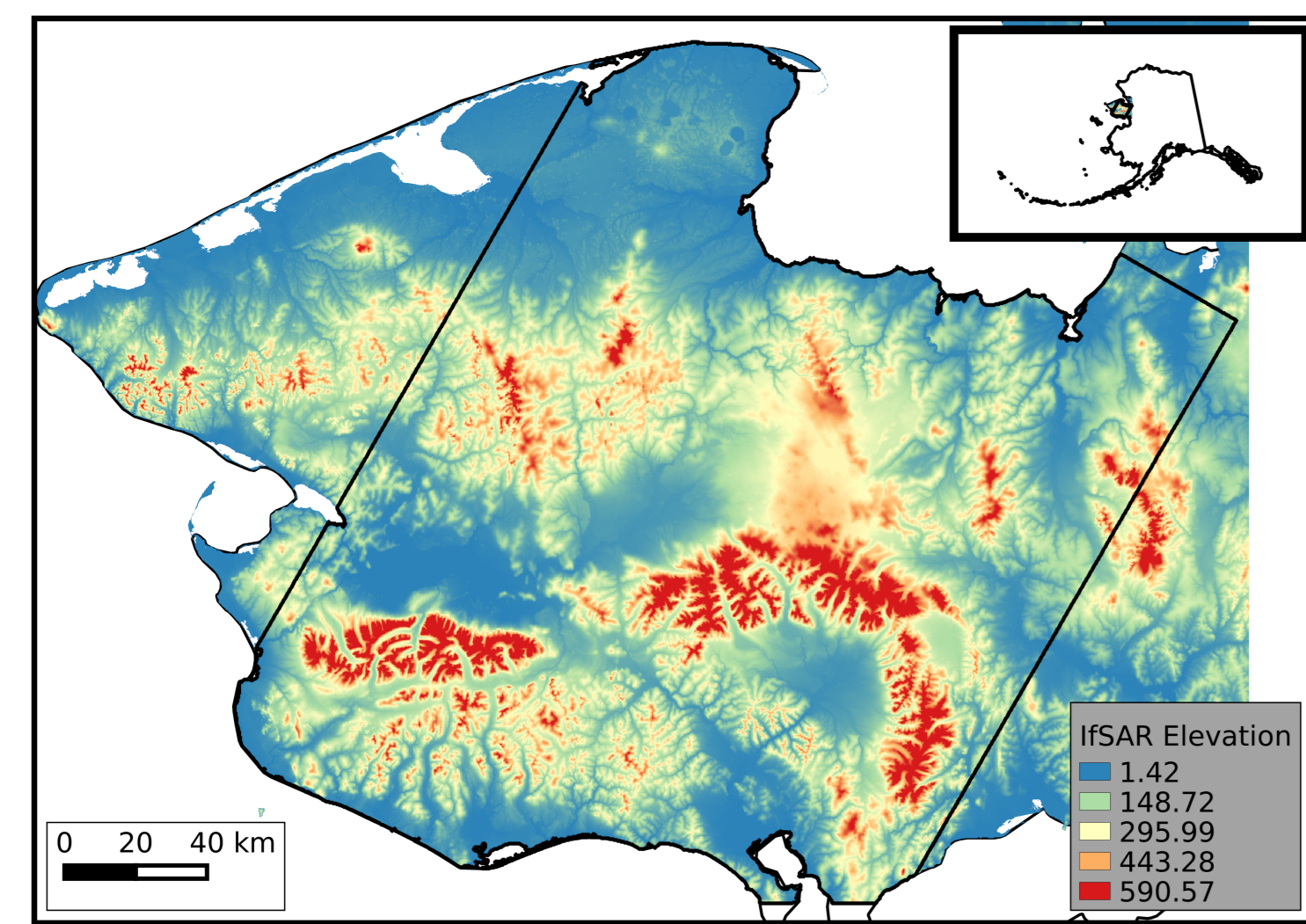
Figure 1: *Study area (black line) over the Seward Peninsula of Alaska. The background is Interferometric Synthetic Aperture Radar (IfSAR) elevation values in meters provided by Geographic Information Network of Alaska.*

Table 1: *Characteristics of the remotely sensed data used in this work.*

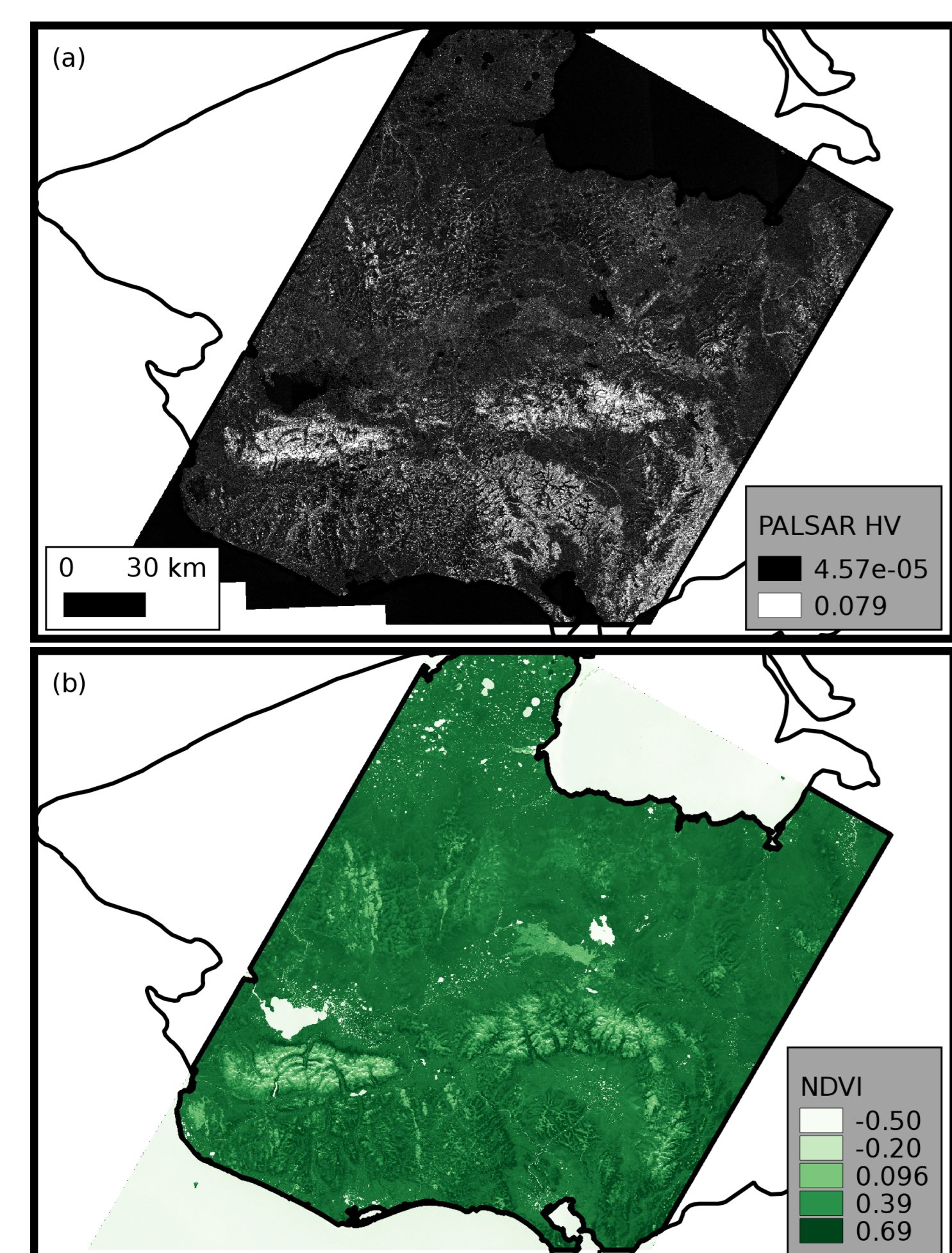| Platform	Sensor	Characteristics	Dates	Images
ALOS	PALSAR	L (HV)	07/28/08 - 09/07/08	37
Landsat 8	OLI	2,3,4,5,6, and 7	09/07/14	1
Airborne	IfSAR	DEM	Summer 2012	15

Figure 2: *(a) ALOS-1 PALSAR HV composite and (b) Landsat 8 OLI NDVI over the study region.*

Unsupervised Clustering

- Hoffman et al. (2008) developed a parallel version of the k-means algorithm to accelerate convergence, handle empty cluster cases, and obtain initial centroids through a scalable implementation of the Bradley and Fayyad (1998) method.
- Kumar et al. (2011) extended this to a fully distributed and highly scalable parallel version of the k-means algorithm for analysis of very large datasets, which was used in this study.
- ALOS-1 PALSAR, Landsat 8 OLI and IfSAR, consisting of ~270 M cells for the Seward Peninsula at 12 m resolution with 9 inputs, was analyzed using k-means clustering (Table 2).
- Three classifications were tested to assess the validity of using optical and SAR HV (horizontal transmitting, vertical receiving) datasets.
- We performed three classifications: (1) IfSAR and Landsat 8 OLI, (2) IfSAR, Landsat 8 OLI, and ALOS-1 PALSAR, and (3) IfSAR and ALOS-1 PALSAR (Figure 3).
- Clustering yields maps in which each cell is classified into one of k classes that is representative of land cover.

Table 2: *The characteristics used in the k-means clustering algorithm for the classification of IfSAR, ALOS-1 PALSAR, and Landsat 8 Operational Land Imager (OLI), where TOA stands for top of atmosphere.*

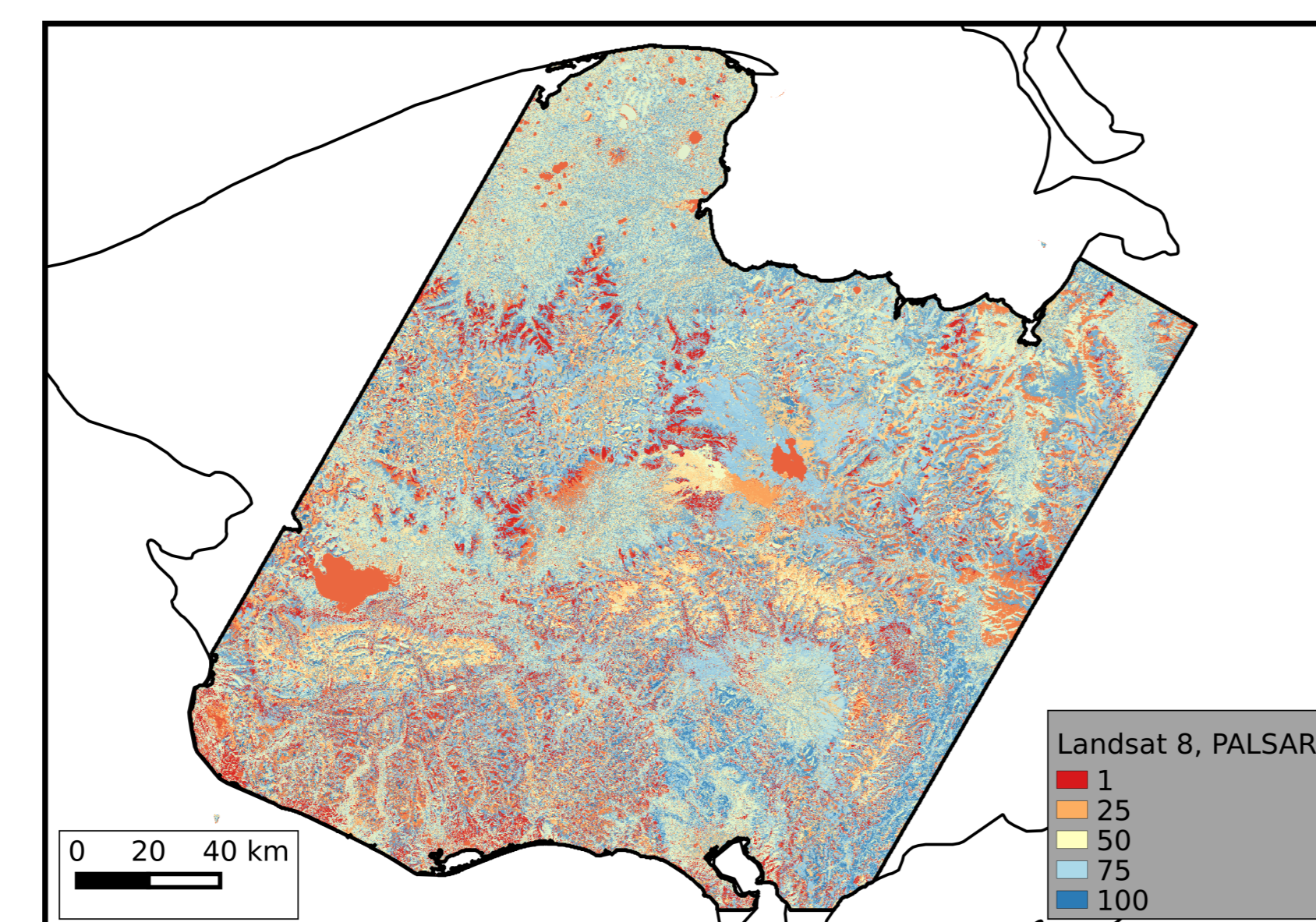| Variable	Mean	Standard Deviation	Platform	Resolution
Elevation	192.04	158.7	IfSAR	5 m
HV	0.02	0.05	ALOS-1 PALSAR	12 m
TOA Blue Band	0.1	0.01	Landsat 8 (OLI)	30 m
TOA Green Band	0.08	0.01	Landsat 8 (OLI)	30 m
TOA Red Band	0.07	0.02	Landsat 8 (OLI)	30 m
TOA NIR Band	0.2	0.08	Landsat 8 (OLI)	30 m
TOA SWIR 1	0.2	0.07	Landsat 8 (OLI)	30 m
TOA SWIR 2	0.1	0.04	Landsat 8 (OLI)	30 m
NDVI	0.5	0.3	Landsat 8 (OLI)	30 m

Figure 3: *Unsupervised clustering (k=100) of Landsat 8 OLI and ALOS-1 PALSAR HV over the study region.*

Mapcurves & Label Stealing

- Clustering is an unsupervised classification technique, so clustered regions have no descriptive labels like **Moist Tundra**.
- Hargrove et al. (2006) developed a method for quantitatively comparing categorical maps that is 1) independent of differences in resolution, 2) independent of the number of categories in maps, and 3) independent of the directionality of comparison.
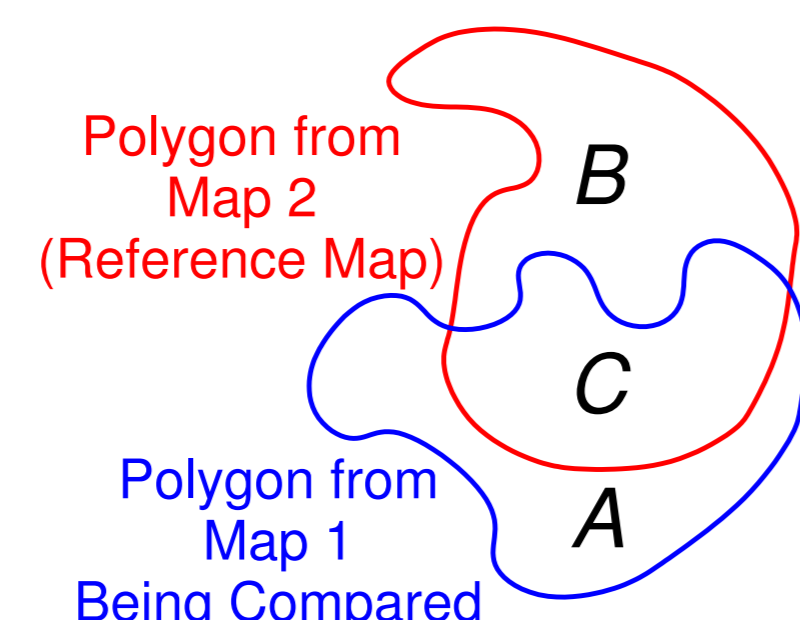
$$GOF = \sum_{polygons} \frac{C}{B+C} \times \frac{C}{A+C}$$

GOF provides "credit" for the area of overlap, but also "debit" for the area of non-overlap.

- Label stealing allows us to perform automated "supervision" to "steal" the best human-created descriptive labels to assign.
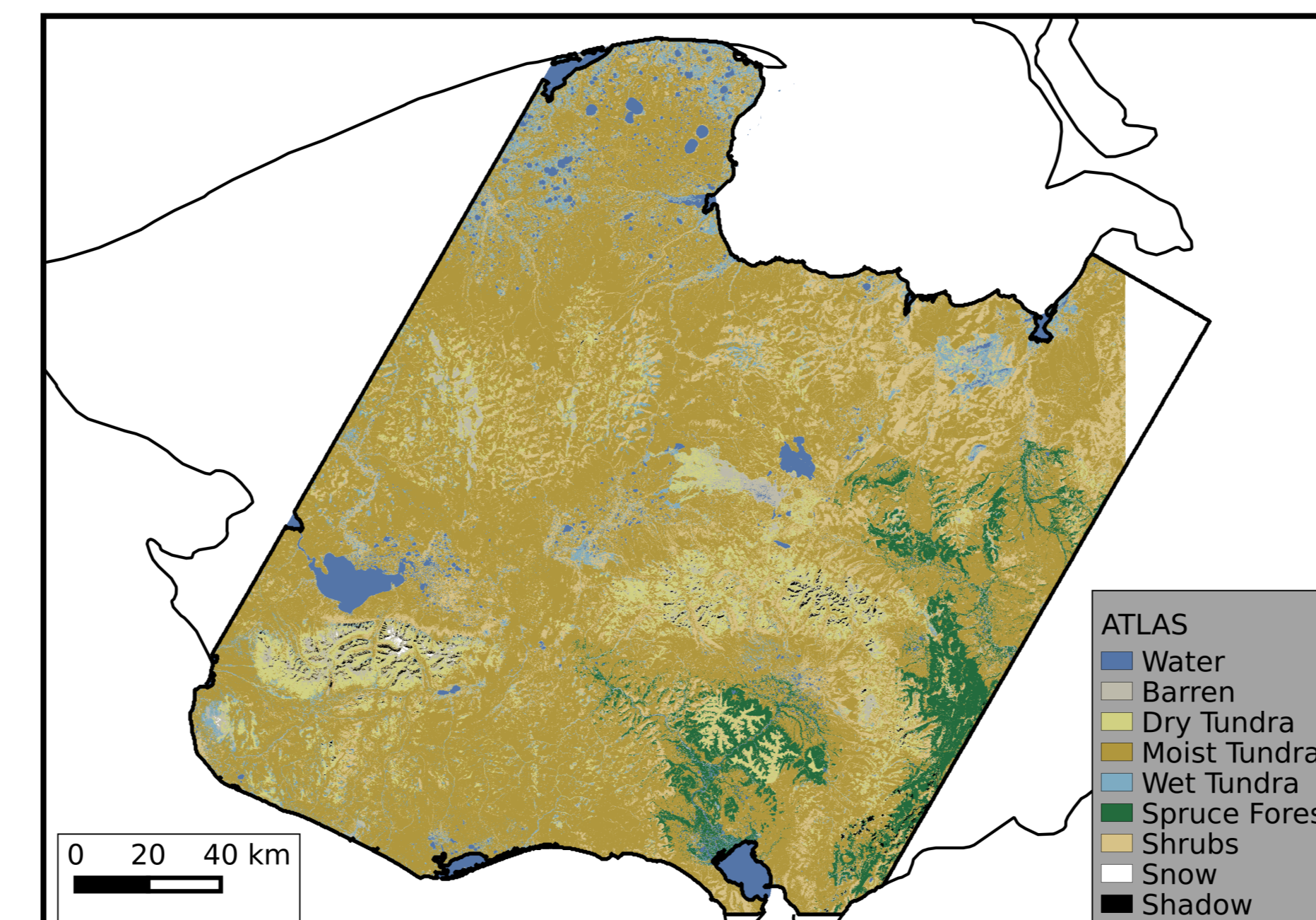- The following map (Figure 4) was used for "label stealing", the Arctic Transitions in the Land-Atmosphere System (ATLAS) (http://geobotanical.portal.gina.alaska.edu).

Figure 4: *Arctic Transitions in the Land-Atmosphere System (ATLAS) land cover dataset over study area. This data was provided by Alaska Arctic Geoecological Atlas (http://geobotanical.portal.gina.alaska.edu).*

- Tables 3 list the percent error when comparing the area of land cover to ATLAS.
- Figure 5 compares the area (top) and GOF scores (bottom) of land cover classes to ATLAS.
- Figure 6 shows a zoomed in area of the Mapcurves method applied to ATLAS near the Council airport.
- Figure 7 show a zoomed in area of the Mapcurves method applied to ATLAS near Kougarok road.
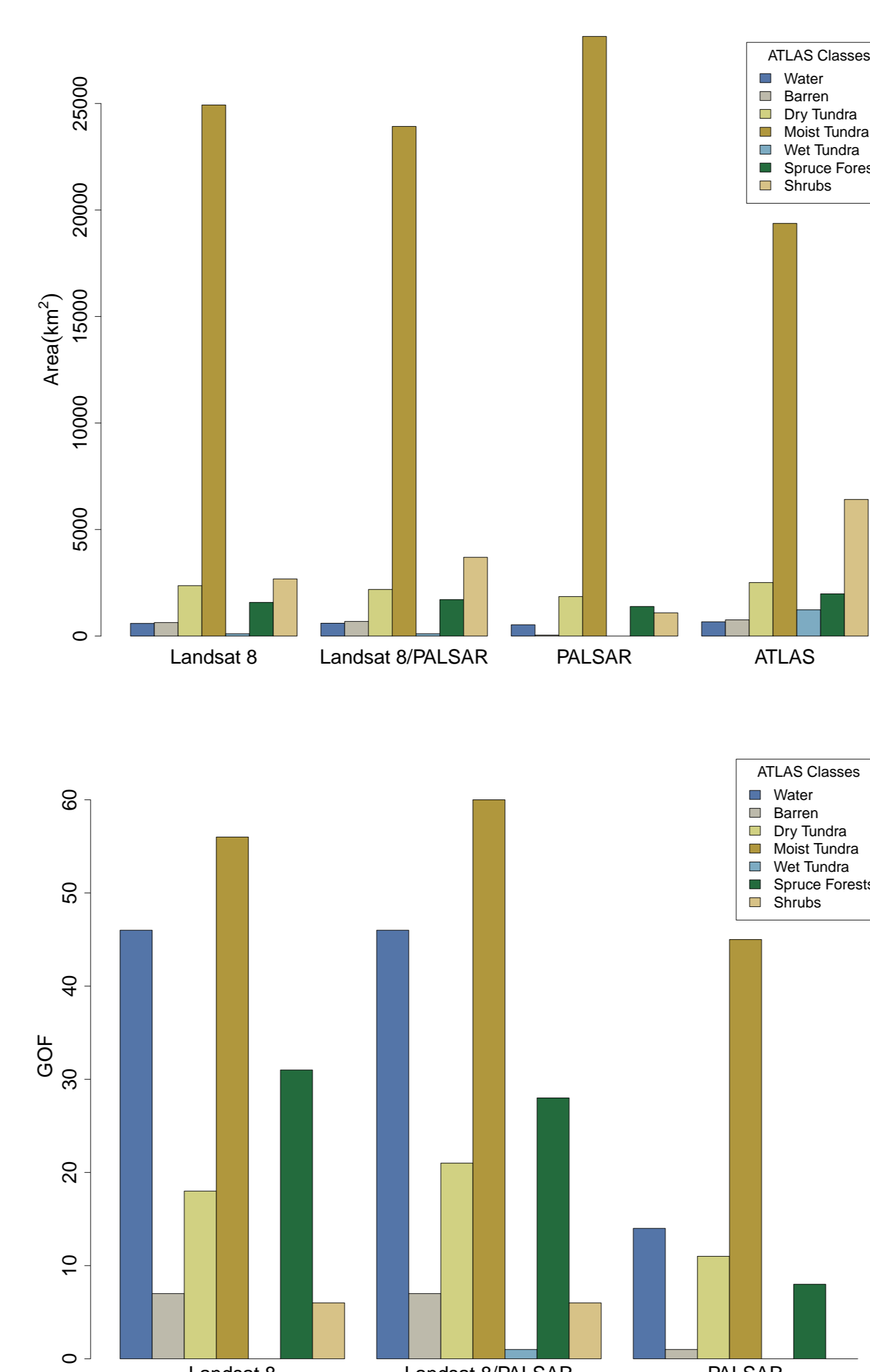
Figure 5: *(Top) Area (km^2) and GOF scores (Bottom) from Mapcurves and label stealing of the ATLAS land cover dataset for k=100.*

Figure 6: *Zoomed in of (a) Landsat 8 OLI true color (b) ATLAS land cover classes, (c) Landsat 8 OLI "label stealing", (d) ALOS-1 PAL-SAR HV "label stealing", and (e) Landsat 8 OLI and ALOS-1 PAL-SAR HV "label stealing" near the Council airport.*
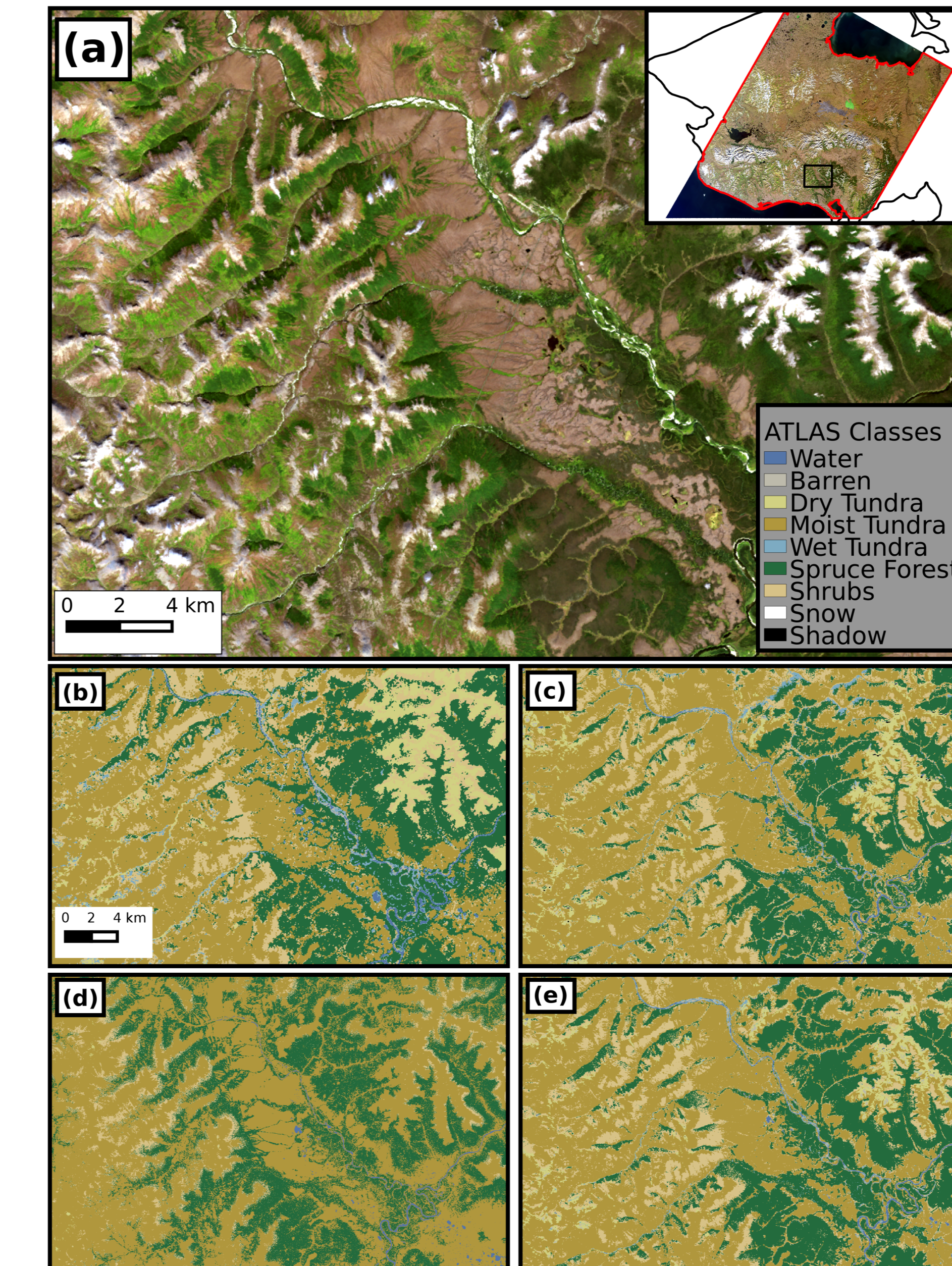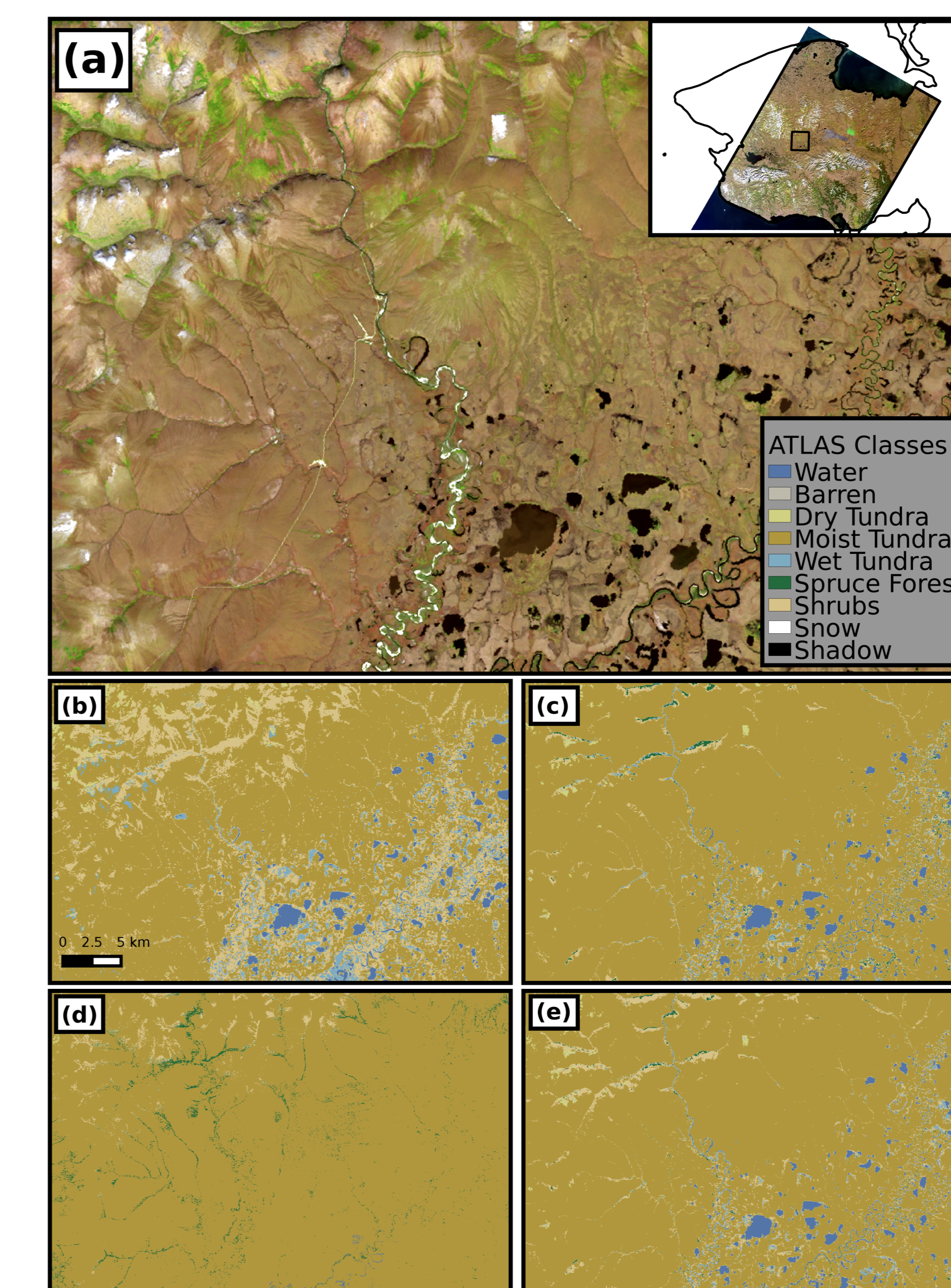
Figure 7: *Zoomed in of (a) Landsat 8 OLI true color (b) ATLAS land cover classes, (c) Landsat 8 OLI "label stealing", (d) ALOS-1 PAL-SAR HV "label stealing", and (e) Landsat 8 OLI and ALOS-1 PAL-SAR HV "label stealing" near Kougarok road.*

Table 3: *Percent error using **Mapcurves** between ATLAS and unsupervised clustering.*

		Moist Tundra	Shrubs	Dry Tundra	Spruce Forest
DEM, PALSAR	k=100	44.13 %	79.44 %	33.88 %	26.27 %
DEM, PALSAR, Landsat	k=100	25.18 %	48.51 %	4.05 %	12.10 %
DEM, Landsat	k=100	24.66 %	50.38 %	7.93 %	10.97 %

Conclusions & Future Work

- We used a k-means clustering algorithm on Landsat 8 OLI and ALOS-1 PALSAR over the Seward Peninsula of Alaska to create high-resolution (~12m) land cover maps.
- The Mapcurves approach was applied to steal land cover or vegetation type labels for unsupervised classifications.
- We compared three different clustering combinations (Table 3) for k=100.
- The combined Landsat 8 OLI and ALOS-1 PALSAR had the best combination with lowest percent error (~20 %) for the main land cover types.
- Label stealing of the ATLAS dataset indicates that ALOS-1 PAL-SAR could be used alone for identifying certain land cover types (Table 3).
- We also intend to look into:
  - Combination of multiple polarimetric signals (i.e., HH, HV, VV, VH) and k clustering values for shrub identification.
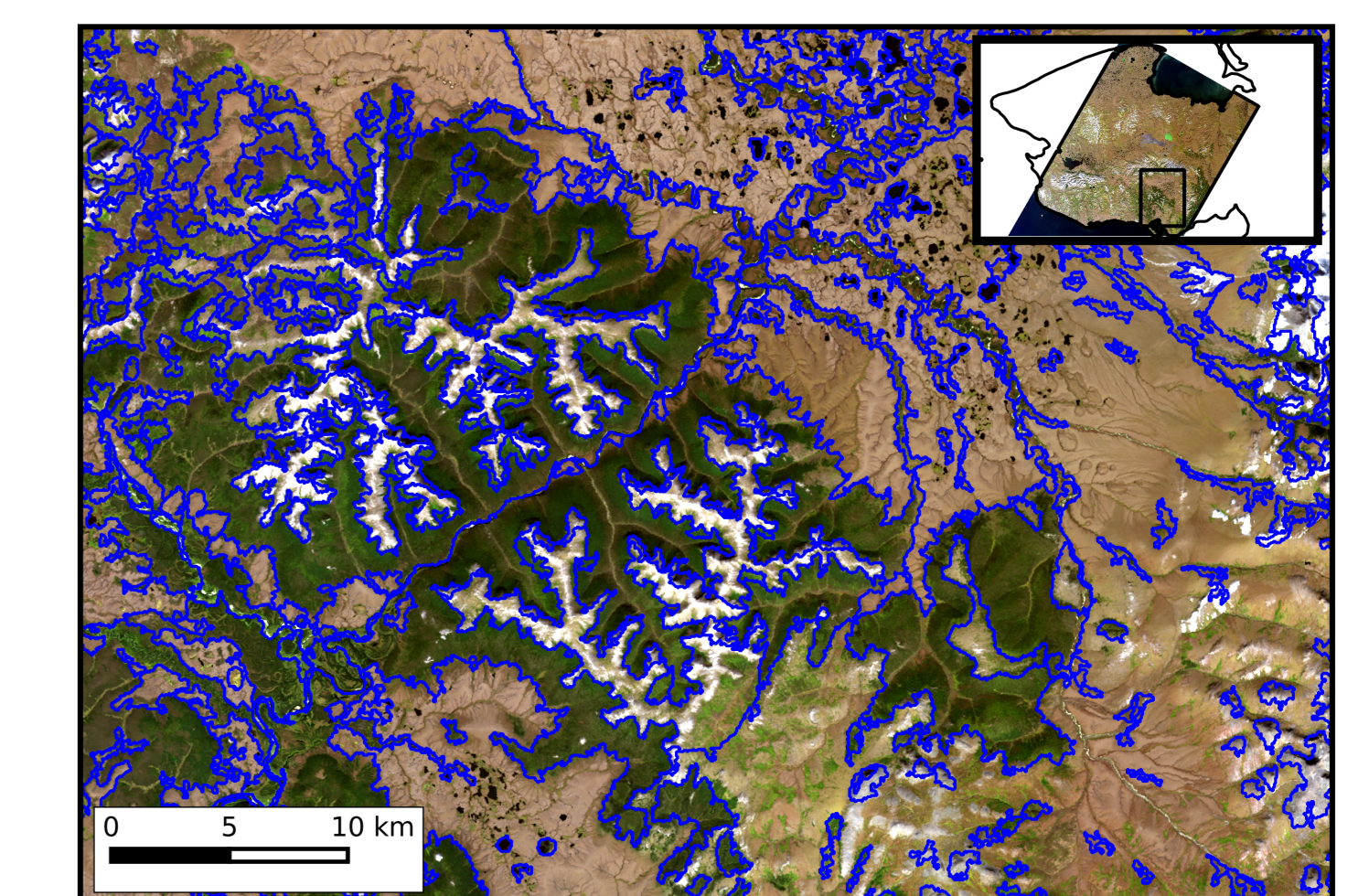 - Comparing an object-based image analysis algorithm (Figure 8) to unsupervised clustering.

Figure 8: *Image segmentation (blue lines) applied for an area near Council.*

References

Forrest M. Hoffman, William W. Hargrove, Richard T. Mills, Salil Mahajan, David J. Erickson, and Robert J. Oglesby. Multivariate Spatio-Temporal Clustering (MSTC) as a data mining tool for environmental applications. In Miquel Sànchez-Marrè, Javier Béjar, Joaquim Comas, Andrea E. Rizzoli, and Giorgio Guariso, editors, *Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on Environmental Modelling and Software Society (iEMSs 2008)*, pages 1774–1781, July 2008. ISBN 978-84-7653-074-0.

Paul S. Bradley and Usama M. Fayyad. Refining initial points for k-means clustering. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 91–99, San Francisco, CA, USA, July 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8.

Jitendra Kumar, Richard Tran Mills, Forrest M. Hoffman, and William W. Hargrove. Parallel k-means clustering for quantitative ecoregion delineation using large data sets. In Mitsuhisa Sato, Satoshi Matsuoka, Peter M. Sloot, G. Dick van Albada, and Jack Dongarra, editors, *Proceedings of the International Conference on Computational Science (ICCS 2011)*, volume 4 of *Procedia Comput. Sci.*, pages 1602–1611, Amsterdam, June 2011. Elsevier. doi: 10.1016/j.procs.2011.04.173.

William W. Hargrove, Forrest M. Hoffman, and Paul F. Hessburg. Mapcurves: A quantitative method for comparing categorical maps. *J. Geograph. Syst.*, 8(2):187–208, July 2006. doi: 10.1007/s10109-006-0025-x.

Contact Information

Zachary L. Langford
Oak Ridge National Laboratory
University of Tennessee
zach@climatemodeling.org

Get a copy of this poster!