Supplement of Earth Syst. Sci. Data, 15, 3761–3790, 2023
https://doi.org/10.5194/essd-15-3761-2023-supplement

*Supplement of*

# Developing a spatially explicit global oil and gas infrastructure database for characterizing methane emission sources at high resolution

**Mark Omara et al.**

*Correspondence to:* Mark Omara (momara@edf.org) and Ritesh Gautam (rgautam@edf.org)

# Supplementary Information

## S1: Further assessment of OGIM data coverage and spatial distribution: example for the Permian Basin using machine-learning-derived oil and gas datasets

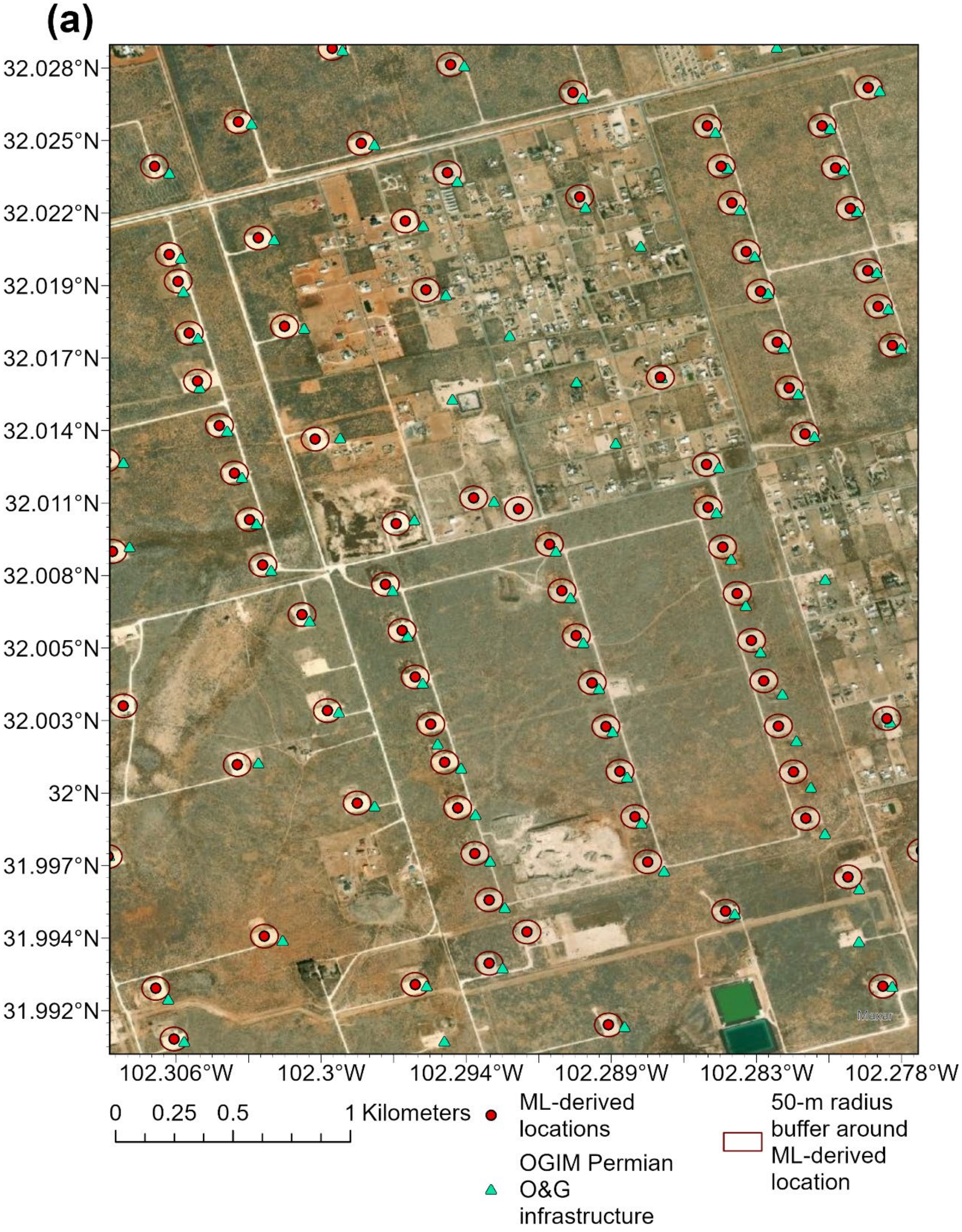
Here, we use a machine-learning derived dataset of oil and gas infrastructure in the Permian Basin, developed by training machine learning (ML) models to automatically detect and classify locations of oil and gas infrastructure in AirBus SPOT imagery (1.5 m pixel resolution) for 2019. Further details of the model development can be found in Lyon et al. (2020). The ML-dataset included over 190,000 locations in the Permian, which we filter to 35,107 locations with reported model raw confidence of >95%, indicating high confidence in the likelihood of the model detection being an oil and gas facility.

The ML-derived locations are reported for the centroid of the facility footprint based on the satellite imagery. We note that even with this filter of high-confidence detections, it is possible that there are a small undetermined fraction of non-oil and gas infrastructure locations with footprints that are similar in appearance to oil and gas infrastructure footprints in satellite imagery. Nevertheless, this ML-derived dataset represents a very large and independent dataset of oil and gas infrastructure locations in the Permian Basin and provides a unique opportunity for a comprehensive comparison with the OGIM data to further assess OGIM data coverage and spatial accuracy.

For each of the 35,107 high-confidence ML-derived locations, we use a *k*-dimensional binary tree algorithm to search for its nearest neighbor in the OGIM dataset and compute the distance (in meters) between the ML-derived location and the OGIM nearest neighbor. Setting a distance threshold of 100-m as an approximate dimension (length/width) of the typical oil and gas facility in the Permian, we find 33,620 OGIM locations are within 100-m of the ML-derived locations, suggesting comprehensive coverage (96%) with high spatial accuracy at the set distance threshold. This coverage increases to 97% at 250-m threshold and 99% at 500-m threshold.

We note that the Permian is a highly dynamic basin in terms of oil and gas activity and had ~300 new well pad development per month in 2019 (Lyon et al. 2020). Additionally, public data reporting in this region can have reporting lags of more than three to six months. Thus, while it is possible to track monthly trends in new oil and gas development using machine learning approaches (depending on satellite imagery refresh rate), the public data reporting lag and update frequencies could help explain the <100% coverage assessed herein. A map showing an example of the locations of the ML-derived dataset and the OGIM dataset is shown in Figure S1 (a), while Figure S1 (b) shows a histogram of the computed distances between the ML-derived locations and the OGIM locations.

The above assessment provides an independent check on the OGIM data coverage and spatial accuracy in a dynamic oil and gas basin with dense oil and gas infrastructure. Future data verification work should leverage similar approaches to further characterize data coverage and spatial accuracy in other regions.


(a)
32.028°N
32.025°N
32.022°N
32.019°N
32.017°N
32.014°N
32.011°N
32.008°N
32.005°N
32.003°N
32°N
31.997°N
31.994°N
31.992°N
102.306°W
102.3°W
102.294°W
102.289°W
102.283°W
102.278°W
Maxar
0 0.25 0.5 1 Kilometers
● ML-derived locations
▲ OGIM Permian O&G infrastructure
□ 50-m radius buffer around ML-derived location

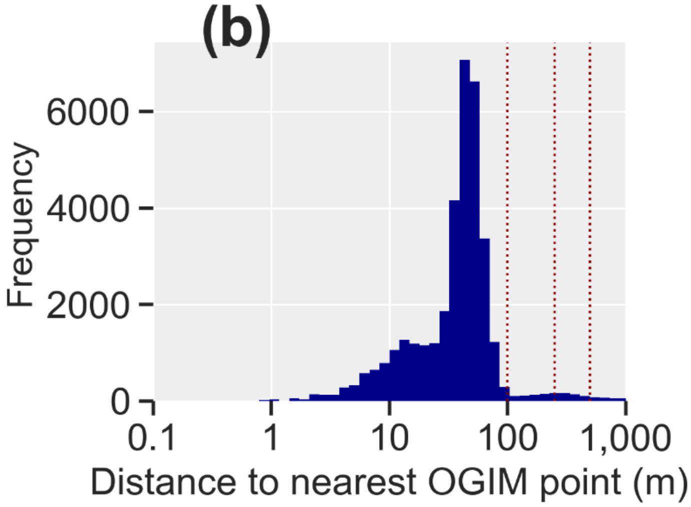


**Figure S1**. **(a)** Visualization of a machine-learning derived dataset of point locations of oil and gas infrastructure in the Permian Basin (Lyon et al. 2020) and the OGIM dataset in the same region. The red points show the filtered high-confidence ML-derived dataset, the red polygons show a 50-m radius buffer around each ML-derived point, and the cyan triangles show the OGIM oil and gas infrastructure points. **(b)** Histogram of the distance between an ML-derived dataset and its nearest neighbor in the OGIM dataset. The dashed dark-red lines show the distance thresholds of 100-m, 250-m, and 500-m. (ESRI basemap imagery, © Environmental System Research Institute)

## References

Lyon, D. R., Hmiel, B., Gautam, R., Omara, M., Roberts, K. A., Barkley, Z. R., Davis, K. J., Miles, N. L., Monteiro, V. C., Richardson, S. J., Conley, S., Smith, M. L., Jacob, D. J., Shen, L., Varon, D. J., Deng, A., Rudelis, X., Sharma, N., Story, K. T., Brandt, A. R., Kang, M., Kort, E. A., Marchese, A. J., and Hamburg, S. P.: Concurrent variation in oil and gas methane emissions and oil price during the COVID-19 pandemic, Atmos. Chem. Phys., 21, 6605–6626, https://doi.org/10.5194/acp-21-6605-2021, 2021.