

README document for the OzNet AOA experimental data collection

1. Collection description

This collection is structured to support reproducible research for "Spatial soil moisture prediction from in-situ data upscaled to Landsat footprint: Assessing area of applicability of machine learning models" (Yu et al., 2025). It provides all necessary input data, trained models, and soil moisture (SM) data extrapolated from 28 OzNet in-situ sites across a primary study area (100 km × 100 km) and an extended area (300 km × 300 km) in southeastern Australia (i.e., the Yanco agricultural region) during 2016-2021. The study period spans a cross-validation period (2016-2019) and an independent test period (2020-2021). The spatial resolution of SM prediction is 100 m and the temporal frequency is daily. A key focus is the characterisation of Area of Applicability (AOA) for Random Forests (RF) and eXtreme Gradient Boosting (XGB) models, delineating where predictions are statistically reliable. The collection includes multiple independent validation datasets from field campaigns, different in-situ networks, and SMAP L2 retrievals for further evaluations.

2. Data structure

Directory name	Data format / Explanation	Frequency	Resolution
0_OzNet_cleaned_data	OzNet_[sitename]_cleaned_data.csv	Daily	Point
1_ANUClimate	[var]/YYYY/ANUClimate_v2-0_[var]_daily_YYYYmm.nc	Daily	1 km
1_CMRSET_Landsat	[area]/CMRSET_Landsat_ET_monthly_100m_resampled_YYYYmm.tif	Monthly	100 m (resampled)
1_Downscaled_data	[area]/[var]/Downscaled_[var]_(ub)ESTARFM_daily_100m_YYYY(mm).tif	Daily	100 m
1_Static_data	[area]/[var]_100m_resampled.tif	Static	100 m (resampled)
2_Trained_models	No regular format. / This dir contains the trained models.	N/A	N/A
3_AOA	[area]/[model]/[cv_type]/AOA_[model]_YYYY(mm).nc	Daily	100 m
3_Spatial_SM_prediction	[area]/[model]/[cv_type]/Spatial_SM_prediction_[model]_daily_100m_YYYY(mm).nc	Daily	100 m
4_Independent_SM	No regular format. / This dir contains independent validation data.	Irregular	Point / interp
5_SMAP_L2_SM	SMAP_L2_SM_daily_10km_resampled_AusSubset_YYYY.nc	Daily	10 km (resampled)

3. Notes for data structure

- [sitename]: OzNet site names. The details can be found in [OzNet_study_sites.csv](#).
- [var]: Predictor variables used in training/validating the models.
- [area]: Either the primary study area (146.00-147.00 °E and 34.30-35.30 °S; 100 km × 100 km) or the extended area (144.50-147.50 °E and 33.00-36.00 °S; 300 km × 300 km).
- [model]: Trained models, including RF and XGB.
- [cv_type]: Cross-validation types, including spatial cross-validation and cross-cluster validation.

4. Experimental details and citation

Please refer to the [GitHub repository](#) for experimental details.

How to cite: Yu, Y., Malone, B. P., Renzullo, L. J., Burton, C. A., Tian, S., Searle, R. D., Bishop, T. F. A. and Walker, J. P., 2025. Spatial soil moisture prediction from in-situ data upscaled to Landsat footprint: Assessing area of applicability of machine learning models, *IEEE Transactions on Geoscience and Remote Sensing*, In Press.