Building Trust in Earth Science Findings through Data Traceability and Results Explainability

P. Olaya[1], D. Kennedy[1], R. Llamas[2], L. Valera[1], R. Vargas[2], J. Lofstead[3], and M. Taufer[1]

[1] *University of Tennessee, Knoxville;* [2] *University of Delaware;* [3] *Sandia National Labs.* ***DOI:*** *10.1109/TPDS.2022.3220539*

Science Question

Can we explain soil moisture findings predicted by machine learning (ML) models at different resolutions? Can we trace the data provenance in the predicting workflow?

Analysis

We develop a computational environment that enables traceability and explainability of high-resolution soil moisture predictions using SOMOSPIE (a Soil Moisture Spatial Inference Engine) and container technologies, starting from 27 km resolution satellite data from the ESA-CCI soil moisture database and terrain parameters.

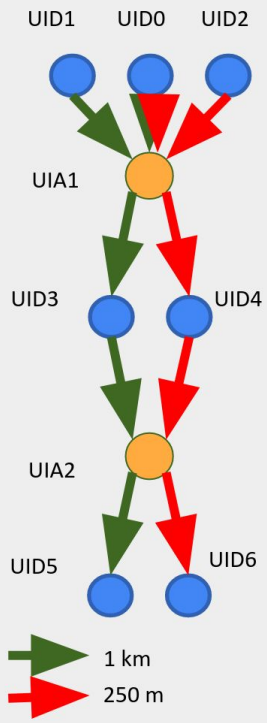
Results

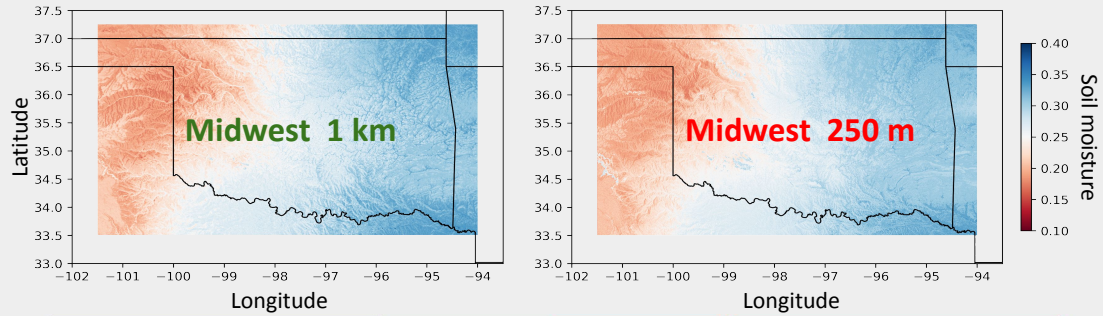
We use our environment for two use cases centered around the Midwest region in which we trace back differences in predictions due to input data and ML methods. Our environment generates automatic provenance of scientific findings with limited overhead in terms of time (10%) and storage space (5%).

Significance

Trustworthy high-resolution soil moisture is necessary for practical use in earth sciences, including precision forestry and agriculture, hydrology for landscape ecology, and regeneration dynamics.

Automatic data provenance of soil moisture predictions for the Midwest region at 1 km and 250 m





Input data

UID1	Terrain predict
Container_name	eval_1km.sif
Creation_time	2022-11-28T09:40:02-4
Command_line	noop
Record_trail	Out: [D1, eval_1km.sif]

UID0	Satellite train.
Container_name	train_27km.sif
Creation_time	2022-11-28T12:07:26-4
Command_line	noop
Record_trail	Out: [D0, train_27km.sif]

UID2	Terrain predict
Container_name	eval_250m.sif
Creation_time	2022-11-28T13:40:02-4
Command_line	noop
Record_trail	Out: [D2, eval_250m.sif]

Reusable artifact

UIA1	ML method
Container_name	knn.sif
Creation_time	2022-11-28T12:14:54-EDT
Command_line	noop
Record_trail	NULL

Intermediate data

UID3	Intermediate predictions
Container_name	predictions_oklahoma.sif
Creation_time	2022-11-28T12:48:57-4
Command_line	python3 knn.py Train/train.csv Eval/eval_1km.csv Predictions/predictions_oklahoma.csv
Record_trail	Out: [D3, predictions_oklahoma.sif] **App: [A1, knn.sif]** **In: [D1, eval_1km.sif]** **In: [D0, train_27km.sif]**

UID4	Intermediate predictions
Container_name	predictions_oklahoma.sif
Creation_time	2022-11-28T13:40:06-4
Command_line	python3 knn.py Train/train.csv Eval/eval_250m.csv Predictions/predictions_oklahoma.csv
Record_trail	Out: [D4, predictions_oklahoma.sif] **App: [A1, knn.sif]** **In: [D2, eval_250m.sif]** **In: [D0, train_27km.sif]**

Reusable artifact

UIA2	Visualization
Container_name	visualization.sif
Creation_time	2022-11-28T04:29:08-EDT
Command_line	noop
Record_trail	NULL

Output data

UID5	Soil moisture 1km
Container_name	output_oklahoma.sif
Creation_time	2022-11-28T04:29:25-4
Command_line	python3 visualization.py Predictions/predictions_oklahoma.csv Output/out_oklahoma.png 0.175 0.35
Record_trail	Out: [D5, output_oklahoma.sif] **App: [A2, visualization.sif]** **In: [D3, predictions_oklahoma.sif]**

UID6	Soil moisture 250m
Container_name	output_oklahoma.sif
Creation_time	2021-09-11T04:34:17-4
Command_line	python3 visualization.py Predictions/predictions_oklahoma.csv Output/out_oklahoma.png 0.175 0.35
Record_trail	Out: [D6, output_oklahoma.sif] **App: [A2, visualization.sif]** **In: [D4, predictions_oklahoma.sif]**

SOMOSPIE Webpage: https://globalcomputing.group/somospie/

SOMOSPIE Github Repository: https://github.com/TauferLab/SOMOSPIE

Computational Environment Github Repository: https://github.com/TauferLab/ContainerizedEnv