Cyberinfrastructure Tools for Precision Agriculture in the 21st Century

Michela Taufer
The University of Tennessee Knoxville
Contributors and collaborators

- U. Delaware: Ricardo Llamas, Mario Guevara, and Rodrigo Vargas
- UTK: Danny Rorabaugh, Kae Suarez, Leobardo Valera, Ria Patel, and David Icove
- ORNL: Jimmy Landmesser
- UIUC: Craig Willis and Victoria Stodden
Sponsors and supporters

- NSF OAC 1854312 CIF21 DIBBs: PD: Cyberinfrastructure Tools for Precision Agriculture in the 21st Century (PIs: Taufer and Vargas)
- NSF OAC 1941443 EAGER: Reproducibility and Cyberinfrastructure for Computational and Data-Enabled Science (PIs: Stodden and Taufer)
- IBM Shared University Research (SUR) Award
- NSF XSEDE JetStream: Allocations EAR180011 and TRA180041
  - Many thanks to Jeremy Fischer, IU
Multiscale computational modeling

https://ajw-group.mit.edu/multiscale-modeling-clays

M Stan, Material Today, 12, 2009, 20-28
Multiscale data modeling (MSDM)

Scientific scales

Software ecosystem

Length

Time

km

m

cm

sec

hour

day
“Only a small fraction of real-world ML systems is composed of the ML code” D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips () Hidden Technical Debt in Machine Learning Systems
“Only a small fraction of real-world ML systems is composed of the ML code” D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips () Hidden Technical Debt in Machine Learning Systems
Feature extraction

Protein-ligand docking

**Linear regression:** map ligands into 3-D point representation

Feature extraction

Linear regression: map ligands into 3-D point representation

Numerical analyses: map secondary structures into eigenvalues

Feature extraction

Linear regression: map ligands into 3-D point representation

Numerical analyses: map secondary structures into eigenvalues

Deep learning: map both secondary and ternary structures into tensors

Protein-ligand docking

Protein folding

Protein engineering

Hidden (forgotten?) software ecosystem

“Only a small fraction of real-world ML systems is composed of the ML code” D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips () Hidden Technical Debt in Machine Learning Systems
Data collection at the edge

Point Field Measurements
Data collection at the edge

Point Field Measurements

Remote Sensor Measurements
Challenges in MMDM

• Design and implement robust and sustainable software ecosystems
• Combine analytics and computing across heterogenous platforms (i.e., HPC, Cloud, and edge computing)
• Build trust in results through reproducibility, replicability, and transparency (RRT)
Relevance of soil moisture data

- Satellite-borne remote sensing technology
  - Infrared to radio
  - Active and passive
Workflows for precision agriculture

Soil moisture leveraged for:
• Environmental sciences
• Precision agriculture

Data Generation
- Fine-grained, complete data
  - Fine-grained Soil Moisture
  - Landscape Surface DSM
  - Weather Data NOAA

Data Analytics
- Fine-grained, complete data
  - Analytics representations + algorithms
  - Data prediction

Data Feedback
- Coarse-grained, incomplete data
  - Soil Moisture ESA-CCI
  - Landscape Surface DSM
  - Weather Data NOAA

Application
- Computation

Multiscale data
- Weather Data NOAA
- Landscape Surface DSM
- Soil Moisture

Application
- Fine-grained, complete data
- Analytics representations + algorithms
- Data prediction

Data Feedback
- Coarse-grained, incomplete data
- Soil Moisture ESA-CCI
- Landscape Surface DSM
- Weather Data NOAA

Data Generation
- Weather Data NOAA
- Landscape Surface DSM
- Soil Moisture

Soil moisture leveraged for:
• Environmental sciences
• Precision agriculture
Design and implement a software ecosystem for precision agriculture

Collaborators: Rodrigo Varga’s Group (UD)
Platform: NSF XSEDE Jetstream
NSF OAC 1854312 CIF21 DIBBs: PD: Cyberinfrastructure Tools for Precision Agriculture in the 21st Century
Data analytics for soil moisture

Soil moisture leveraged for:
- Environmental sciences
- Precision agriculture

Data generation

Satellite and sensors

Data Analytics

Fine-grained, complete data

Analytics representations + algorithms

Soil Moisture ESA-CCI

Landscape Surface DSM

Weather Data NOAA

Data prediction

Coarse-grained, incomplete data

Data Feed Back

Application

Soil moisture leveraged for:
- Environmental sciences
- Precision agriculture

Computation

Data Feedback
Challenge 1: incomplete soil moisture data (I)

Satellites collect raster data across the surface of the Earth

Visualization example of the ESA-Climate Change Initiative Soil Moisture database with a coarse pixel size of 27x27km

(Liu et al. 2011 HESS, Liu et al. 2012 RSE)
Challenge 1: incomplete soil moisture data (II)

Causes of missing data:
- snow/ice cover
- frozen surface
- dense vegetation
- extremely dry surface

Dec. 2000 Average Soil Moisture (m$^3$/m$^3$)

Challenge 2: coarse-grained soil moisture data (I)

Original Resolution
27 km × 27 km

Desired Resolution
1 km × 1 km

Challenge 2: coarse-grained soil moisture data (II)

Original product ESA CCI (m$^3$ m$^{-3}$, mean 2013)

Integration of multiscale data: from satellites ...
... to terrain, climate, and weather data

Region of interest

Satellite data

Terrain parameters

Global Historical Climatology Network (GHCN) and other local data (field measurements)

Example of terrain parameters: water wetness index

Shaw et al., 2016 GRL, Moore 2012, Geomorphology.
SOMOSPIE: SOil MOisture SPatial Inference Engine

Data collection

 Feature extraction

 Predictions

 Analysis tools

SOMOSPIE: SOil MOisture SPatial Inference Engine

Data collection
Feature extraction
Predictions
Analysis tools

Region selection: format of regions of interest

- **NEON**, "Mid Atlantic"
- **CEC**, "8.5.1"
- **BOX**, "-77_-75_37_40"
- **STATE**, "Delaware"

Algorithmic solutions: ML-based software suite

**Random Forest**
- Compute weighted mean of 500 prediction trees

**KKNN:**
- Use **local data**
- Compute k and distance kernel using cross validation automatically
- Compute weighted means with the kernel (**many values**)  

**Surrogate based model (SBM):**
- Use **all sampled data**
- Use regression to generate one single polynomial model (**single polynomial model**)
Algorithmic solutions: ML-based software suite

**Random Forest**
- Compute weighted mean of 500 prediction trees

**KKNN:**
- Use local data
- Compute k and distance kernel using cross validation automatically
- Compute weighted means with the kernel (*many values*)

**Surrogate based model (SBM):**
- Use all sampled data
- Use regression to generate one single polynomial model (*single polynomial model*)

**HYPPO (Hybrid Piecewise Polynomial Modeling):**
- Use local data
- Determine local polynomial degree using cross validation
- Use regression to generate local polynomial model (*many polynomial models*)

Computational solutions: Jupyter + XSEDE Jetstream
Computational solutions: Jupyter + XSEDE Jetstream

**SOMOSPIE**

**SOil MOisture SPatial Inference Engine**

**Introduction**

Soil moisture is a critical variable that links climate dynamics with water and food security. It regulates land-atmosphere interactions (e.g., via evapotranspiration—the loss of water from evaporation and plant transpiration to the atmosphere), and it is directly linked with plant productivity and survival. Information on soil moisture is important to design appropriate irrigation strategies to increase crop yield, and long-term soil moisture coupled with climate information provides insights into trends and potential agricultural thresholds and risks. Thus, information on soil moisture is a key factor to inform and enable precision agriculture.

The current availability in soil moisture data over large areas comes from remote sensing (i.e., satellites with radar sensors) which provide daily, nearly global coverage of soil moisture. However, satellite soil moisture datasets have a major shortcoming in that they are limited to coarse spatial resolution (generally no finer than tens of kilometers).

There do exist at higher resolution other geographic datasets (e.g., climatic, geological, and topographic) that are intimately related to soil moisture values. SOMOSPIE is meant to be a general-purpose tool for using such datasets to downscale (i.e., increase resolution) satellite-based soil moisture products. This Jupyter Notebook is a result of a collaboration between computer scientists of the Global Computing Laboratory at the University of Tennessee, Knoxville and soil scientists at the University of Delaware (funded by NSF awards #1724843 and #1854312).
Computational solutions: Jupyter + XSEDE Jetstream
Use case I: from 27x27km to 1x1km

Fine-grained modeling of Mid-Atlantic region in April 2017:

- Terrain parameters: Elevation, Slope, and Wetness Index

Level III Ecoregions of the Continental United States (CEVLv3)

Original satellite data (27x27km)

Random Forest fine-grained predictions 1x1km

Use case I: from 27x27km to 1x1km

Fine-grained modeling of Mid-Atlantic region in April 2017:
- Terrain parameters: Elevation, Slope, and Wetness Index

Use case II: Local scale predictions - 1x1m resolution

Use case II: Local scale predictions - 1x1m resolution

Combine computing and analytics: integration of soil moisture predictions into controlled (or prescribed) burn

Collaborator: David Icove’s group (UTK)
Planform: Tellico cluster (IBM Power9 system) – supported by 2019 IBM Shared University Research (SUR) Award
Soil moisture data for simulating controlled burn

Soil moisture leveraged for:
- Controlled or prescribed burn

Data Generation
- Multiscale data
  - Weather Data (NOAA)
  - Landscape
  - Surface DSM
  - Soil Moisture (ESA-CCI)
  - Weather Data (NOAA)

Data Analytics
- Fine-grained, complete data
  - Fine-grained Soil Moisture
  - Landscape Surface DSM
  - Weather Data NOAA
- Coarse-grained, incomplete data

Data Feedback
- Weather Data (NOAA)
- Landscape
- Surface DSM

Application

Computation

Fine-grained, complete data

Coarse-grained, incomplete data

Data Feedback
Elephant in the room: the soil moisture

- Simulation of the 2016 Gatlinburg wildfire
- Software:
  - Fire Dynamics Simulator (FDS) - large-eddy simulation (LES) for low-speed flows
- Platform:
  - IBM Power9 cluster at UTK
- Simulation specs:
  - 120m × 120m × 100m domain
  - 5 frames/sec temp. resolution

Firestarting area

Soil moisture is missing in FDS
FDS simulations

Soil moisture layer - 1x1m
FDS simulations
FDS simulations

Soil moisture layer - 1x1m
FDS simulations

Soil moisture layer - 1x1m
Build trust in results through reproducibility, replicability, and transparency

Collaborator: Victoria Stodden’s Group (UIUC)
Platform: NSF XSEDE Jetstream
NSF OAC 1941443 EAGER: Reproducibility and Cyberinfrastructure for Computational and Data-Enabled Science
Leveraging other NSF projects: Whole Tale

- Building an **open platform for computational reproducibility**
  - Create and publish **executable research objects** ("Tales")
- Simplify process of creating & verifying reproducible computational artifacts for scientific discovery

This material is based upon work supported by the National Science Foundation under Grant No. OAC-1541450
Capturing metadata
Plug into SOMOSPIE GitHub
Enable replicability of results
SOMOSPIE version 1.2

leobardovalera released this 13 hours ago - 2 commits to master since this release

Update README.md

Assets 3

- SOMOSPIE.tgz
- Source code (zip)
- Source code (tar.gz)