Spatial Disaggregation and Harmonization of gSSURGO

Nathaniel Chaney, Jonathan Hempel, Nathan Odgers, Alex McBratney, Eric F. Wood
MOTIVATION: NEXT GENERATION LAND SURFACE MODELING

Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth’s terrestrial water

Eric F. Wood,1 Joshua K. Roundy,1 Tara J. Troy,1 L. P. H. van Beek,2 Marc F. P. Bierkens,2,3 Eleanor Blyth,4 Ad de Roo,5 Petra Döll,9 Mike Ek,7 James Famiglietti,8 David Gochis,9 Nick van de Giesen,10 Paul Houser,11 Peter R. Jaffé,1 Stefan Kollet,12 Bernhard Lehner,13 Dennis P. Lettenmaier,14 Christa Peters-Lidard,15 Murugesu Sivapalan,16 Justin Sheffield,1 Andrew Wade,17 and Paul Whitehead18

Received 6 October 2010; revised 21 January 2011; accepted 24 February 2011; published 6 May 2011.

Goal: ~100 meters global

Challenges:
• Model Structure
• Input Data
• Computation
Motivation: gSSURGO Tradeoffs

*Most Frequent Component per Map Unit*
SSURGO: COMPONENT INFO

Component Name: Cerini

- Rich database per component
- Uncertainty information
- Triangular Distribution
Motivation and Outline

**gSSURGO Tradeoffs**

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Detail</td>
<td>Survey Bias (Boundaries)</td>
</tr>
<tr>
<td>Rich Database</td>
<td>Incomplete</td>
</tr>
<tr>
<td>In Situ Observations</td>
<td>Variable Resolution</td>
</tr>
</tbody>
</table>

**Goal:** Address gSSURGO challenges

**Outline:**

A. Testbed: Northern Mississippi State
B. Methodology: DSMART
C. Application over CONUS (HPC)
D. Explore new dataset over CONUS

**Example:** Cerini

**Source:** [http://casoilresource.lawr.ucdavis.edu](http://casoilresource.lawr.ucdavis.edu)
Testbed: Northern Mississippi State

*Most Frequent Component per Map Unit*
Objective

Legacy Soil Data \rightarrow \text{Algorithm} \rightarrow \text{Corrected Product}

Soil Covariates
DIGITAL SOIL MAPPING

\[ S = f (S, C, O, R, P, A, N) + \varepsilon \]

- Soil Series (Posterior)
- Climate
- Organisms
- Relief
- Parent Material
- Age
- Lat, Lon
- Errors

Soil Series (Prior)

Legacy Soil Data (gSSURGO)

McBratney et al., 2003
## Soil Covariates: CONUS

<table>
<thead>
<tr>
<th>Relief</th>
<th>Dataset</th>
<th>Soil Covariate</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NED DEM</td>
<td>Topographic Index, Elevation, MRVBF, MRRTF, Curvature, Slope, Accumulation Area</td>
<td>30 meters</td>
</tr>
</tbody>
</table>

| Parent Material | USGS Aeroradiometric | Uranium, Thorium, Potassium | 4000 meters |

| Organisms | NLCD | Land Cover Type | 30 meters |

### Notes
- **Dataset**: NED DEM
- **Soil Covariate**: Topographic Index, Elevation, MRVBF, MRRTF, Curvature, Slope, Accumulation Area
- **Resolution**: 30 meters
- **Parent Material**: USGS Aeroradiometric
- **Soil Covariate**: Uranium, Thorium, Potassium
- **Resolution**: 4000 meters
- **Organisms**: NLCD
- **Soil Covariate**: Land Cover Type
- **Resolution**: 30 meters
Algorithm: DSMART

SOIL COVARIATES

- Elevation
- Gamma radiometric K
- Gamma radiometric Th
- MRVBF
- SAGA wetness index \((t = 10)\)
- Gamma radiometric U
- Landsat 5 TM NDVI
- SAGA modified catchment area \((t = 10)\)
- Valley depth
- Slope height
- MRR TF
- Mid slope position
- Landsat 5 TM Band 5
- Terrain ruggedness index
- Landsat 5 TM Band 1
- Landsat 5 TM Band 4
- Landsat 5 TM Band 7
- Landsat 5 TM Band 3
- Profile curvature
- Slope aspect
- Plan curvature
- Landsat 5 TM Band 2
- Slope gradient

Source: Microsoft Research

Source: Odgers et al., 2014

Train with legacy soil data
Enhanced DSMART: Random Forest

Forest output probability: \( p(c|\mathbf{v}) = \frac{1}{T} \sum_{t=1}^{T} p_t(c|\mathbf{v}) \)

Soil Covariates

Component
Enhanced DSMART: Result

gSSURGO

DSMART

Corrected Product

Soil Covariates
Detailed Info: Probabilities

- Decision Tree Leaf - Component Histogram
  A. Each grid cell (soil covariates) falls on a leaf
- Implication → Quantify component uncertainty
PROBABILITY RANKED

Component  Probability

Soil 1  Soil 2  Soil 3  Soil 4
<table>
<thead>
<tr>
<th>Component</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil 1</td>
<td></td>
</tr>
<tr>
<td>Soil 2</td>
<td></td>
</tr>
<tr>
<td>Soil 3</td>
<td></td>
</tr>
<tr>
<td>Soil 4</td>
<td></td>
</tr>
</tbody>
</table>
Soil 1
Soil 2
Soil 3
Soil 4

Component
Probability

PROBABILITY RANKED
PROBABILITY RANKED

Component  

Probability

<table>
<thead>
<tr>
<th>Component</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
PROBABILITY RANKED

<table>
<thead>
<tr>
<th>Component</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil 1</td>
<td>1</td>
</tr>
<tr>
<td>Soil 2</td>
<td>2</td>
</tr>
<tr>
<td>Soil 3</td>
<td>5</td>
</tr>
<tr>
<td>Soil 4</td>
<td>20</td>
</tr>
</tbody>
</table>

**Goal:** Obtain similar spatial detail over CONUS
Application over CONUS

**CONUS 30 meters \(\rightarrow\) \(~9\) billion grid cells**

**Feasible Approach: Moving window**
- Split up domain into overlapping blocks
- Run DSMART on each block
- Small region \(\rightarrow\) small sample size \(\rightarrow\) fast random forest
- \(~25,000\) blocks \(\rightarrow\) **500,000 core hours**
High Performance Computing: Blue Waters

<table>
<thead>
<tr>
<th>Machine Stats</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Cores</td>
<td>600,000</td>
</tr>
<tr>
<td>Memory</td>
<td>1.5 petabytes</td>
</tr>
<tr>
<td>Short Term Storage</td>
<td>25 petabytes</td>
</tr>
<tr>
<td>Long Term Storage</td>
<td>500 petabytes</td>
</tr>
</tbody>
</table>

Source: NCSA

500,000 hours (57 years)

5 hours
DSMART: Montana

gSSURGO

dSSURGO
DSMART: Texas

gSSURGO

dSSURGO
DSMART: Mississippi

gSSURGO

→

dSSURGO
DSMART: Washington

gSSURGO

dSSURGO
DSMART: New York

gSSURGO

dSSURGO
DSMART: California
DSMART: Southern California
DSMART: Greater LA Area
Angeles National Forest
Conclusions and Next Steps

- **dSSURGO** - CONUS at 30 meters
  - 50 most probable components (and probabilities)
  - \( \sim \)2 terabyte dataset (freely accessible)
  - [stream.princeton.edu/dSSURGO](stream.princeton.edu/dSSURGO)

- **Next Steps**
  - Applications (e.g. Hydrologic Modeling)
  - Validation (Need your help!)
Conclusions and Next Steps

- **dSSURGO** - CONUS at 30 meters
  - 50 most probable components (and probabilities)
  - ~2 terabyte dataset (freely accessible)
  - stream.princeton.edu/dSSURGO

- Next Steps
  - Applications (e.g. Hydrologic Modeling)
  - Validation (Need your help!)

Questions?