National Aeronautics and Space Administration

Groundwater data interpolation in California's Central Valley using multimodal data fusion and multivariate sequence-to-sequence transformation models

Principal Investigator: Kyongsik Yun (349); Co-Investigators: John Reager (329), Thomas Lu (349), Michael Turmon (398), Zhen Liu (329)

Program: FY21 R&TD Topics **Strategic Focus Area: Earth Science Data Analysis**

Objectives

The objective was to **create a multimodal data fusion and multivariate sequence-to-sequence transformation tool** to estimate groundwater, InSAR subsidence, and soil composition data in California's Central Valley. Given multiple time series data including groundwater storage, precipitation, and soil composition data, the state-of-the-art groundwater spatio-temporal data estimation models using machine learning showed a testing correlation coefficient below 0.8 between the estimated and ground-truth groundwater data. **Accurate interpolation and estimation of groundwater time series data has been challenging due to the different temporal and spatial resolution of multiple data sets.**

Background

California's Central Valley is responsible for $17 billion of annual agricultural output, producing 1/4 of the nation's food. However, land in the Central Valley is sinking at a rapid rate (as much as 20 cm per year) due to continued groundwater pumping. Land subsidence has a significant impact on infrastructure resilience and groundwater sustainability. It is important to understand subsidence and groundwater depletion in a consistent framework using improved models capable of simulating in-situ well observations and observed subsidence. Currently, **groundwater well data is sparse and sampled irregularly**, compromising our understanding of groundwater changes. Moreover, groundwater pumping data is a major missing piece of the puzzle. **Limited data availability and spatial/temporal uncertainty** in the available data have hampered understanding the complex dynamics of groundwater and subsidence.

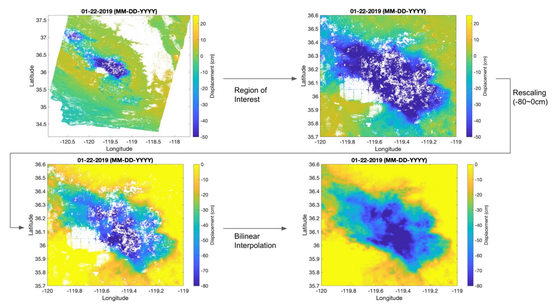
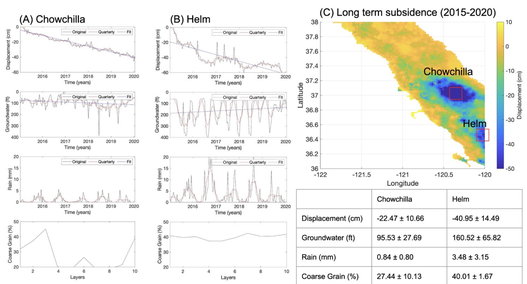


Figure 1. InSAR-based land displacement data interpolation process. We first identified the region of interest that exhibited significant displacement. We then rescaled the displacement visualization to -80~0 cm to focus on the negative displacement. Finally, we performed bilinear interpolation to fill in the missing data.



	Chowchilla	Helm
Displacement (cm)	-22.47 ± 10.66	-40.95 ± 14.49
Groundwater (ft)	95.53 ± 27.69	160.52 ± 65.82
Rain (mm)	0.84 ± 0.80	3.48 ± 3.15
Coarse Grain (%)	27.44 ± 10.13	40.01 ± 1.67

Figure 2. Two representative regions of the Central Valley with significant subsidence with different characteristics. (A) Chowchilla has been shown to maintain monotonically decreasing land displacements, less fluctuating groundwater depth, relatively low precipitation, and high fine-grain ratio across the middle soil layers. (B) Helm, on the other hand, exhibited fluctuating land displacements, relatively large seasonal changes in groundwater depth, high precipitation, and a higher overall coarse-grain ratio across all soil layers. (C) A displacement map including Chowchilla and Helm (2015-2020).

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

www.nasa.gov

Approach and Results

We first integrated multimodal data including InSAR, groundwater, precipitation, and soil composition by interpolating data with the same spatial and temporal resolutions (every 2 weeks on a 1kmX1km grid) (Figure 1). We then identified regions with different temporal dynamics of land displacement, groundwater depth, and precipitation (Figure 2).

We fed the integrated data into the deep neural network of a gated recurrent unit (GRU)-based sequence-to-sequence generation model (Figure 3). We found that the combination of InSAR, groundwater depth, and precipitation data had predictive power for soil composition using deep neural networks (R=0.84, NNSE=0.83). A random forest model was tested as baseline (Figure 4). We also achieved significant accuracy with only 40% of the training data, suggesting that the model can be generalized to other regions for indirect estimation of soil composition.

For uncertainty quantification, we compared the model regression performance between the proposed neural network and Random Forest. Neural network models showed generally lower aleatoric uncertainty than Random Forest, except for 0-10% and 70-80% coarse grain percentages (Figure 5). Central Valley's ground truth high coarse grain data (>70%) is limited and therefore experiences high epistemic uncertainty in high coarse grain estimation. Epistemic uncertainty can be reduced by adding training samples from other regions (e.g., North China Plains that have similar subsidence levels).

Our results indicate that soil composition can be estimated using InSAR, groundwater depth and precipitation data. In-situ measurements of soil composition can be expensive and time consuming and may be impractical in some areas. The generalizability of the model sheds light on high spatial resolution soil composition estimation utilizing existing measurements.

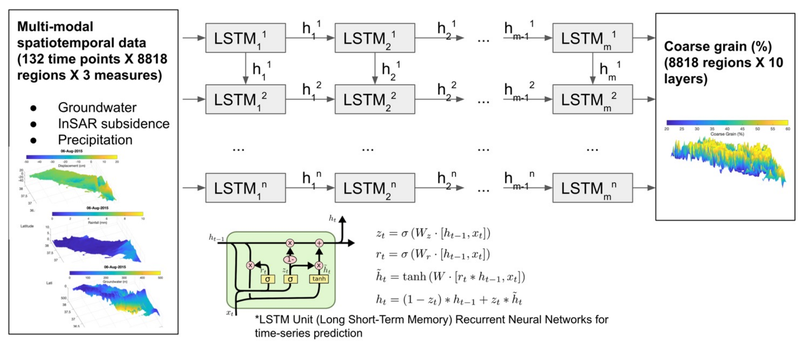


Figure 3. Model architecture. We used LSTM to estimate multidimensional time series. The system has multiple sigmoid and tanh activation functions to selectively pass information to the next layer based on the training process of the final output estimate. The system is inspired by the human memory process, which transfers some perceptual information to long-term memory based on attention and importance calculated by a value function. Input data include groundwater, InSAR subsidence, and precipitation data, covering 8818 different locations, and 132 biweekly time points (5 years). Here, the coarse grain percent of 10 layers was estimated. Coarse-grained soil is defined as containing no more than 50% fine grains (i.e., silt and clay, or particles smaller than 0.075 mm).

PI/Task Mgr Contact
Email: kyongsik.yun@jpl.nasa.gov

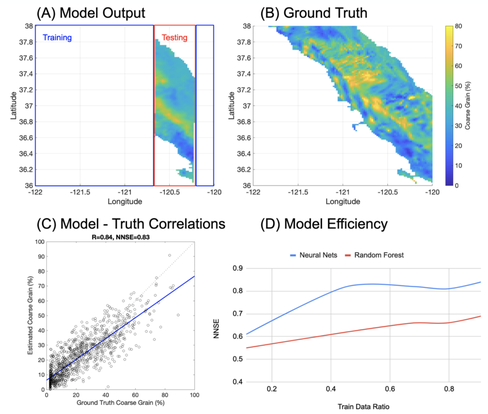


Figure 4. Deep neural network model of soil composition estimation. (A) Example training and testing areas randomly selected for validation. (B) Ground truth coarse-grained ratio of Central Valley in soil layer 1. (C) Correlation plot between ground-truth coarse-grain and the estimated coarse-grain ratios (correlation coefficient R=0.84, normalized Nash-Sutcliffe model efficiency NNSE=0.83). (D) NNSE of neural networks and random forest models over various training data ratios (0.1-0.9).

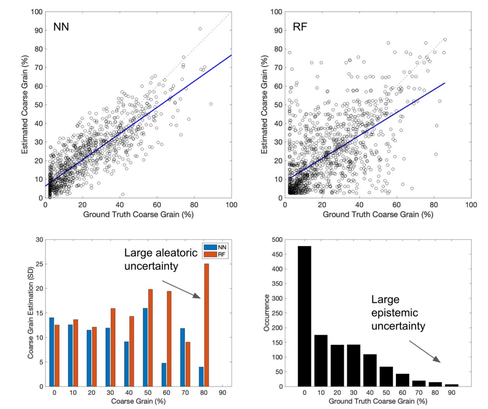


Figure 5. Uncertainty quantification. We compared the model regression performance between the proposed neural network (correlation coefficient r=0.84) and Random Forest (r=0.62). Neural network models showed generally lower aleatoric uncertainty than Random Forest, except for 0-10% and 70-80% coarse grain percentages. Central Valley's ground truth high coarse grain data (>70%) is limited and therefore experiences high epistemic uncertainty in high coarse grain estimation.

Significance/Benefits to JPL and NASA

Groundwater and subsidence prediction and management are critical to understanding how Earth is changing, a JPL Quest. Our project will help drive JPL to innovate novel data processing technologies to estimate future groundwater availability and observe Earth's response to natural and human-induced events (e.g., precipitation, agriculture groundwater usage). Our proposed work is directly relevant to the Strategic Theme of "Monitoring Freshwater Availability", identified in the 2018 JPL Strategic Implementation Plan's Earth Science and Applications Directorate. Also, this work will produce a high-resolution groundwater depletion data set for the Central Valley, as well as a data analysis framework that can be applied for study in other global aquifers.

Publications

Kyongsik Yun, Kyra Kim, Anshuman Pradhan, John Reager, Zhen Liu, Michael Turmon, Alexander Huyen, Thomas Lu, Venkat Chandrasekaran, Andrew Stuart, "Filling the gap: Estimation of soil composition using InSAR, groundwater depth, and precipitation data in California's Central Valley", submitted to *AGU 2021*

Clearance Number:
RPC/JPL Task Number: R21132