Estimating Daily Surface Soil Moisture Using Recurrent Neural Networks and SMAP Data on a Global Grid Scale

Mohammad H. Erfani
NASA Goddard Institute for Space Studies
Center for Climate Systems Research, Columbia University

Background

In the field of hydrology, there has been a growing interest in the application of machine learning across various hydrological modeling tasks. This recent surge is particularly driven by the emergence of new data sources:

- The development of large datasets
- The increased availability of satellite-based data

Soil moisture is a key variable that controls various hydrologic processes, including infiltration, evapotranspiration, and subsurface flow.

$$\theta = \frac{volume\ of\ water}{total\ volume}$$

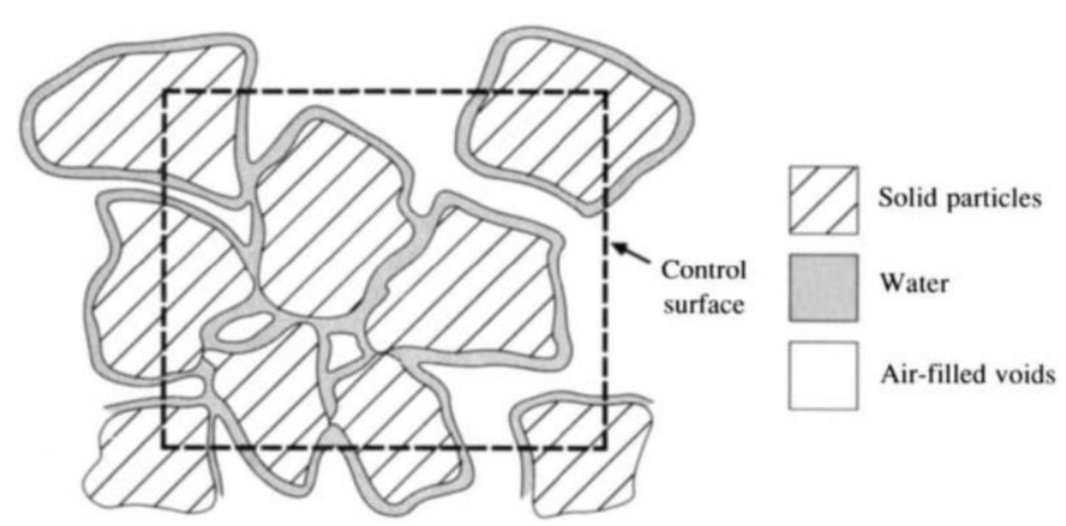


Fig 1: Cross section through an unsaturated porous medium

In this project, we aim to use climate forcing data from **ECMWF Reanalysis v5 (ERA5)**, **The Soil Moisture Active Passive (SMAP)**, as well as several static terrain attributes to train an LSTM model that can predict the soil moisture on a grid-scale for the region of interest.

Study Area

The Central Valley, also known as the Great Valley of California, covers about 20,000 square miles and is one of the more notable structural depressions in the world. The Valley is a vast agricultural region drained by the Sacramento and San Joaquin Rivers. The Valley averages about 50 miles in width and extends about 400 mi northwest from the Tehachapi Mountains to Redding.

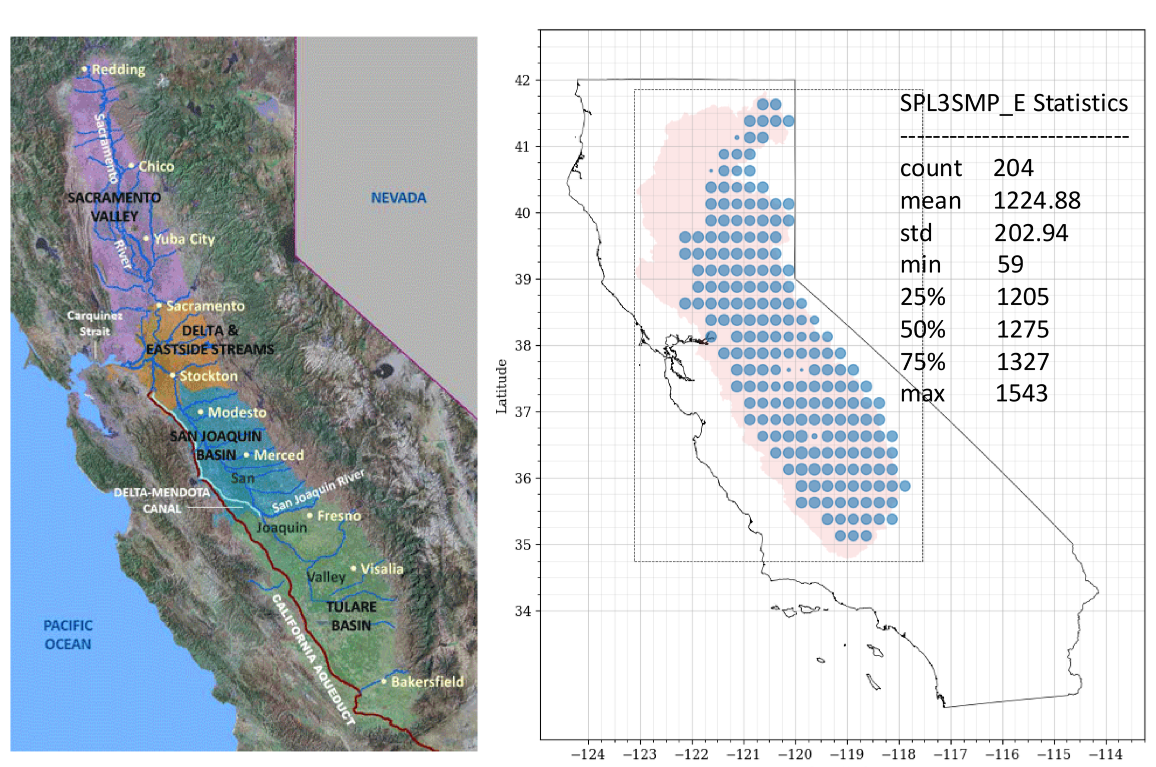


Fig 2: Study area and number of observations for each grid box (0.25° x 0.24°) based on SMAP data.

Dataset

Atmospheric Forcings

Atmospheric forcings were downloaded **ECMWF Reanalysis v5 (ERA5),** spanning from **1990-01-01** to **2022-12-31**. Input variables from atmospheric forcing include Daily Average Temperature, Minimum Temperature, Maximum Temperature, Precipitation, Sea Level Pressure, Solar Insolation, Outgoing Longwave Radiation, Wind Speed, Evaporation, and Relative Humidity. The selected variables are not a product of Land Surface models (LSM).

Terrain Attributes

Static attributes include Latitude, Longitude, NLDAS soil texture datasets, `NLDAS vegetation class datasets, National Land Cover Database (NLCD) 2016, and Digital Elevation Model (DEM).

SMAP Enhanced L3 Radiometer Global and Polar Grid Daily 9 km EASE-Grid Soil Moisture (SPL3SMP_E), Version 5

This enhanced Level-3 (L3) soil moisture product provides a composite of daily estimates of global land surface conditions retrieved by the SMAP radiometer. This product download from 2015-04-01 to 2022-12-31.

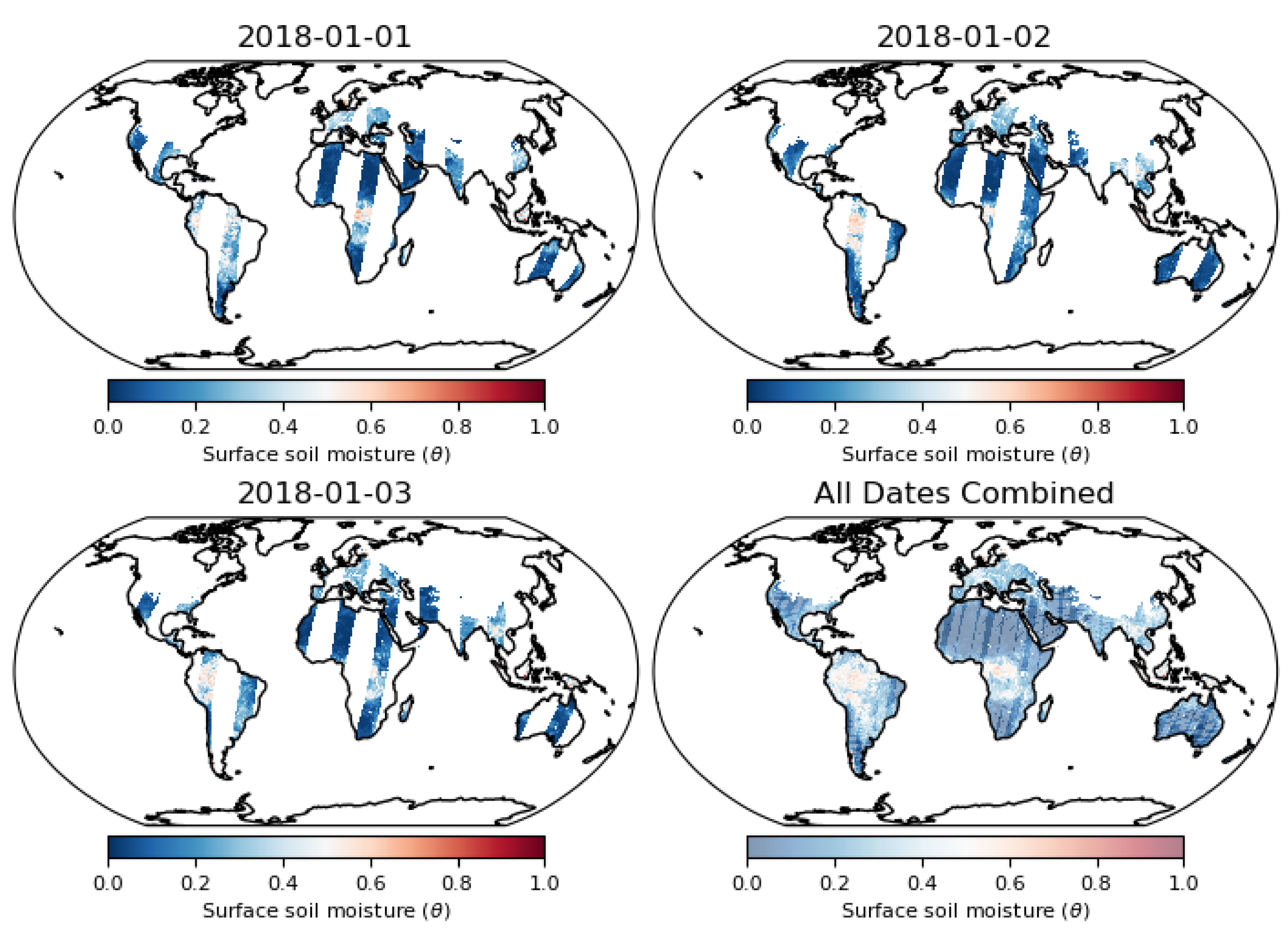


Fig 3: The spatial coverage of SMAP SPL3SMP_E products

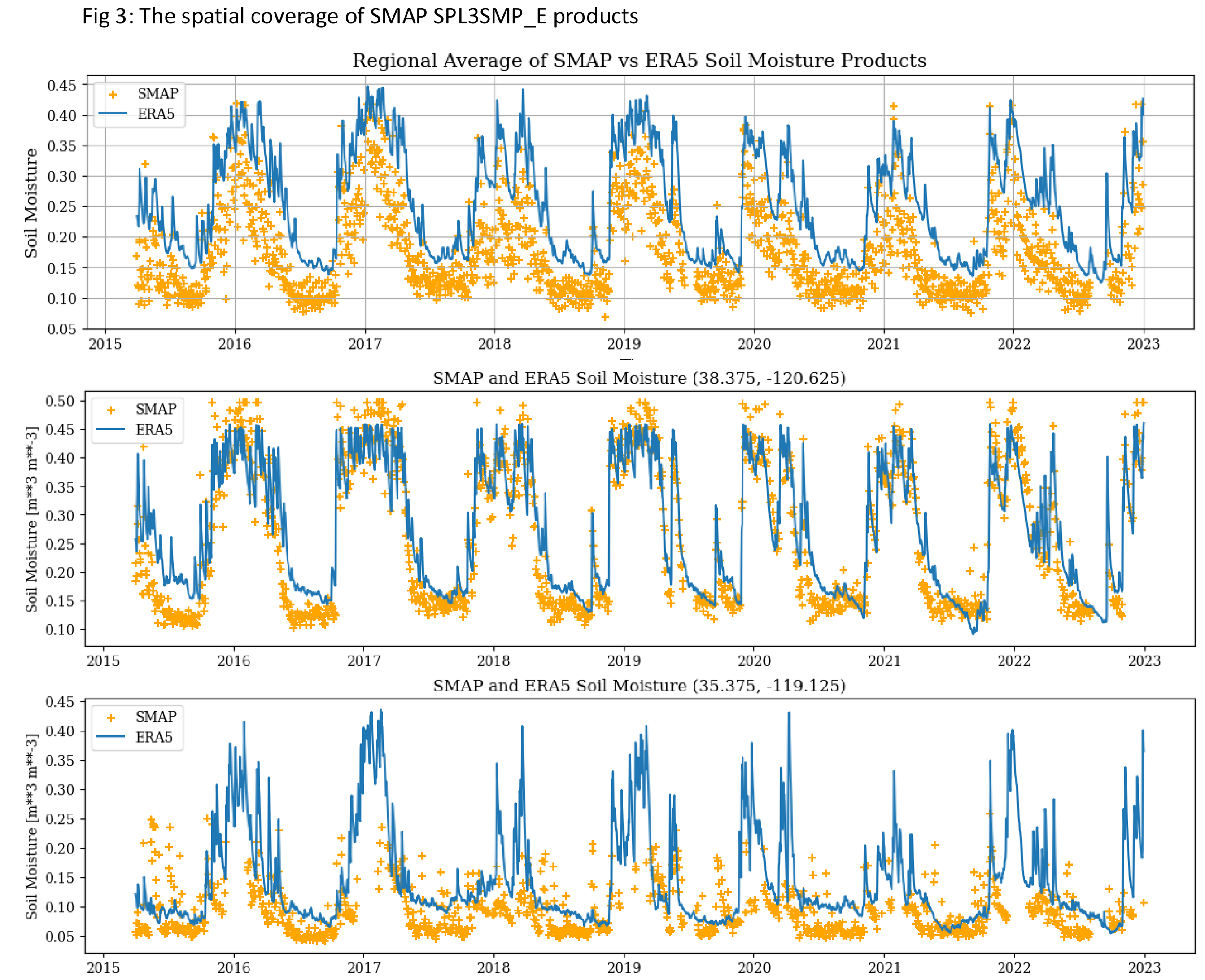
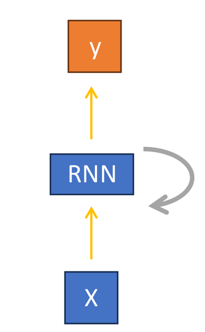


Fig 4: A comparison of SMAP SPL3SMP_E products vs ERA5 data for soil moisture

Methodology

We train LSTM to predict a single day of Surface Soil Moisture (SSM) from 60 days of precedent atmospheric forcing and terrain attributes. The shape (60, number of features), and since we use 16 input variables, the shape is (60, 16).

$$h_t = f_\delta(h_{t-1}, x_t)$$



Training model on ERA-5 data:

- Training set: 1990-01-01, 2009-12-31 (20 Years)
- Validation set: 2010-01-01, 2014-12-31 (5 Years)
- Test Set: 2015-01-01, 2022-12-31 (8 Years)

Fine-tune model on SMAP data:

- Training set: 2015-01-01, 2019-12-31 (5 Years)
- Test Set: 2020-01-01, 2022-12-31 (3 Years)

Results

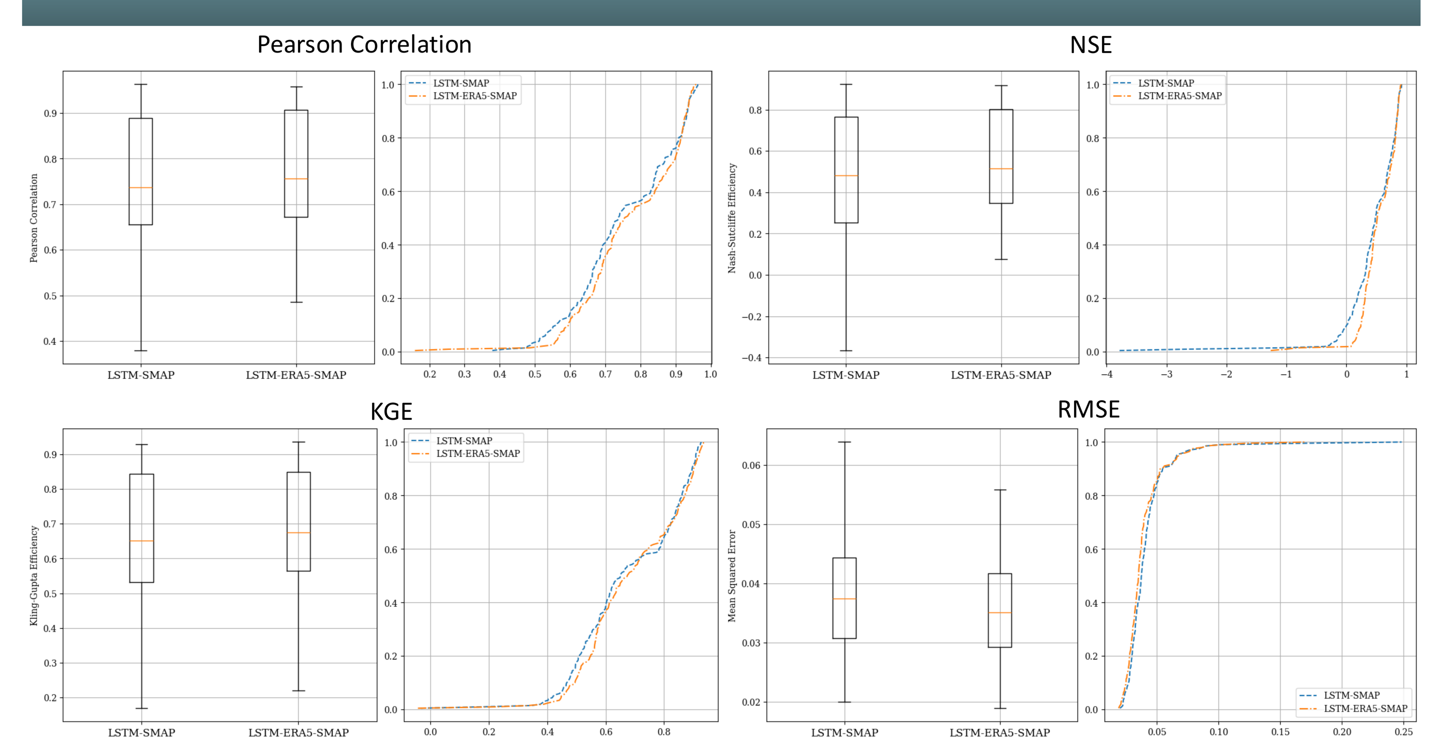


Fig 5: A comparison of two training strategies—utilizing transfer learning vs without it

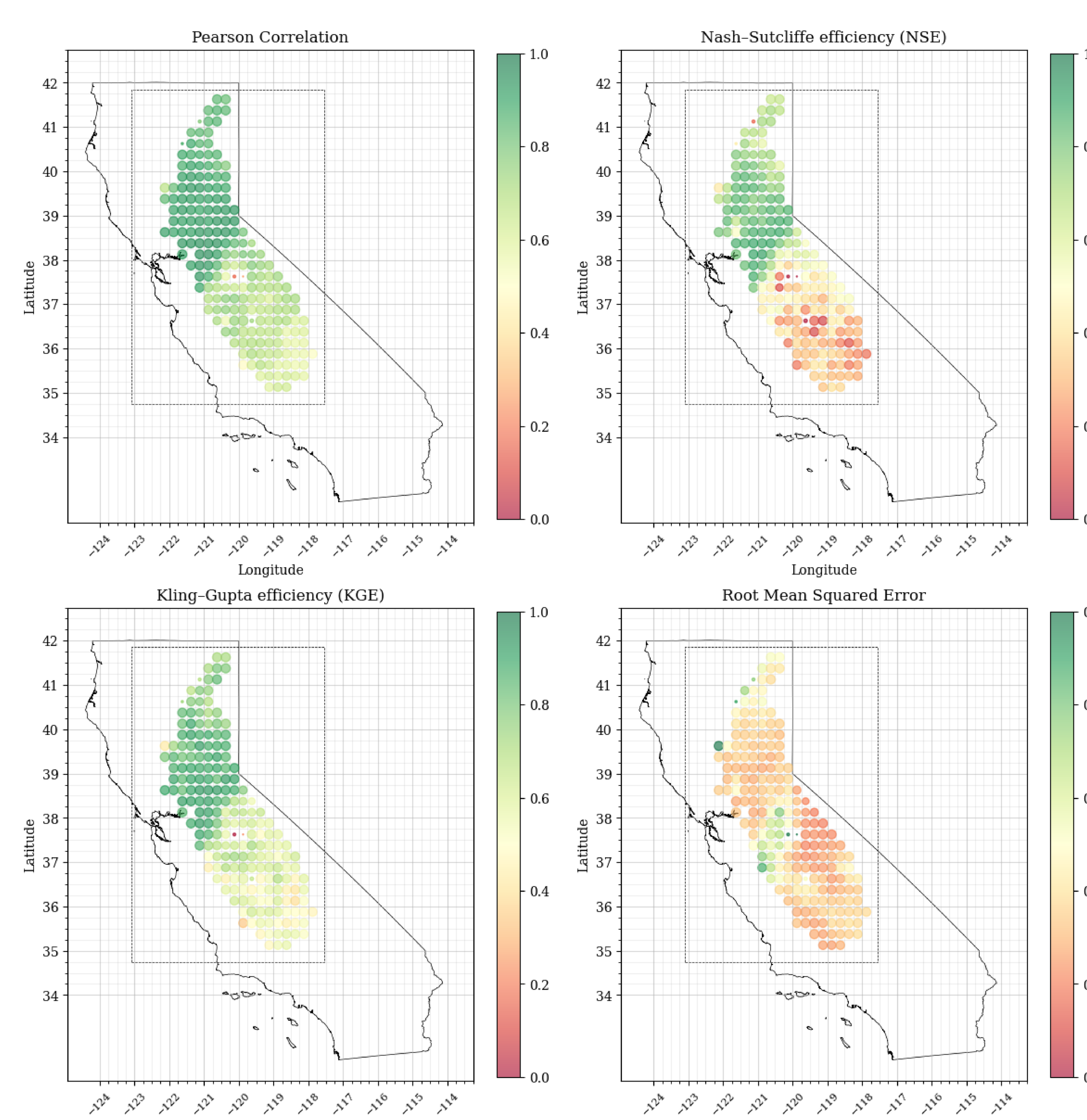


Fig 6: Illustration of the performance metrics for each grid box within the study area.

Results

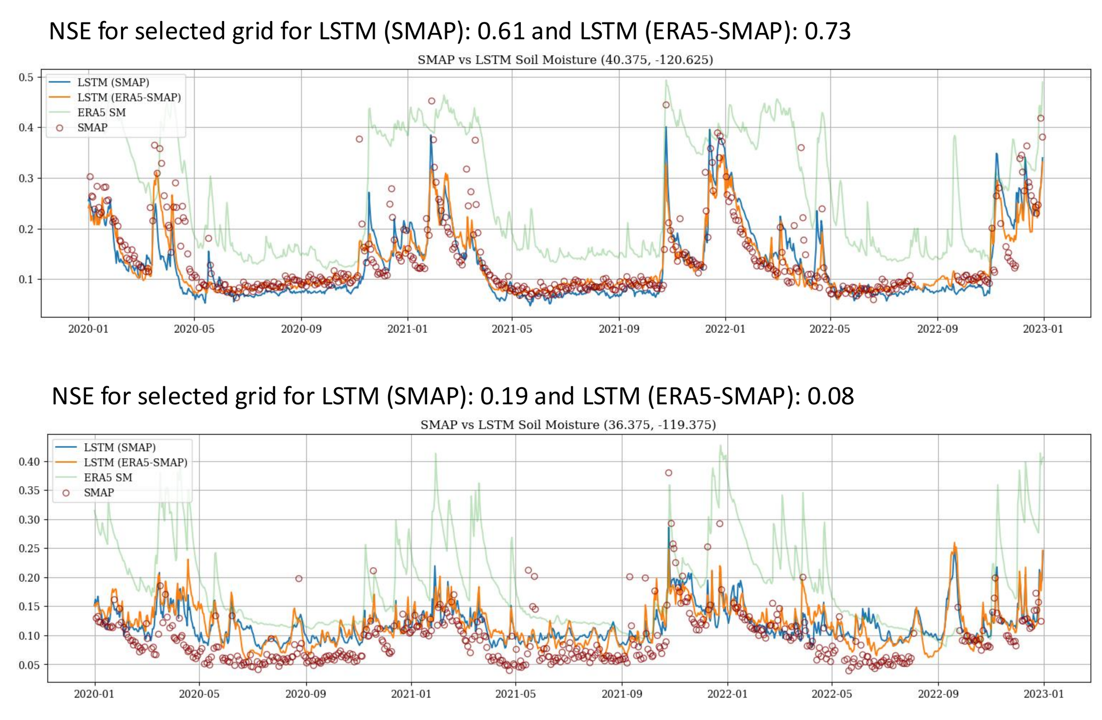


Fig 7: A comparison of models output, ERA5 SSM outputs and SMAP observations

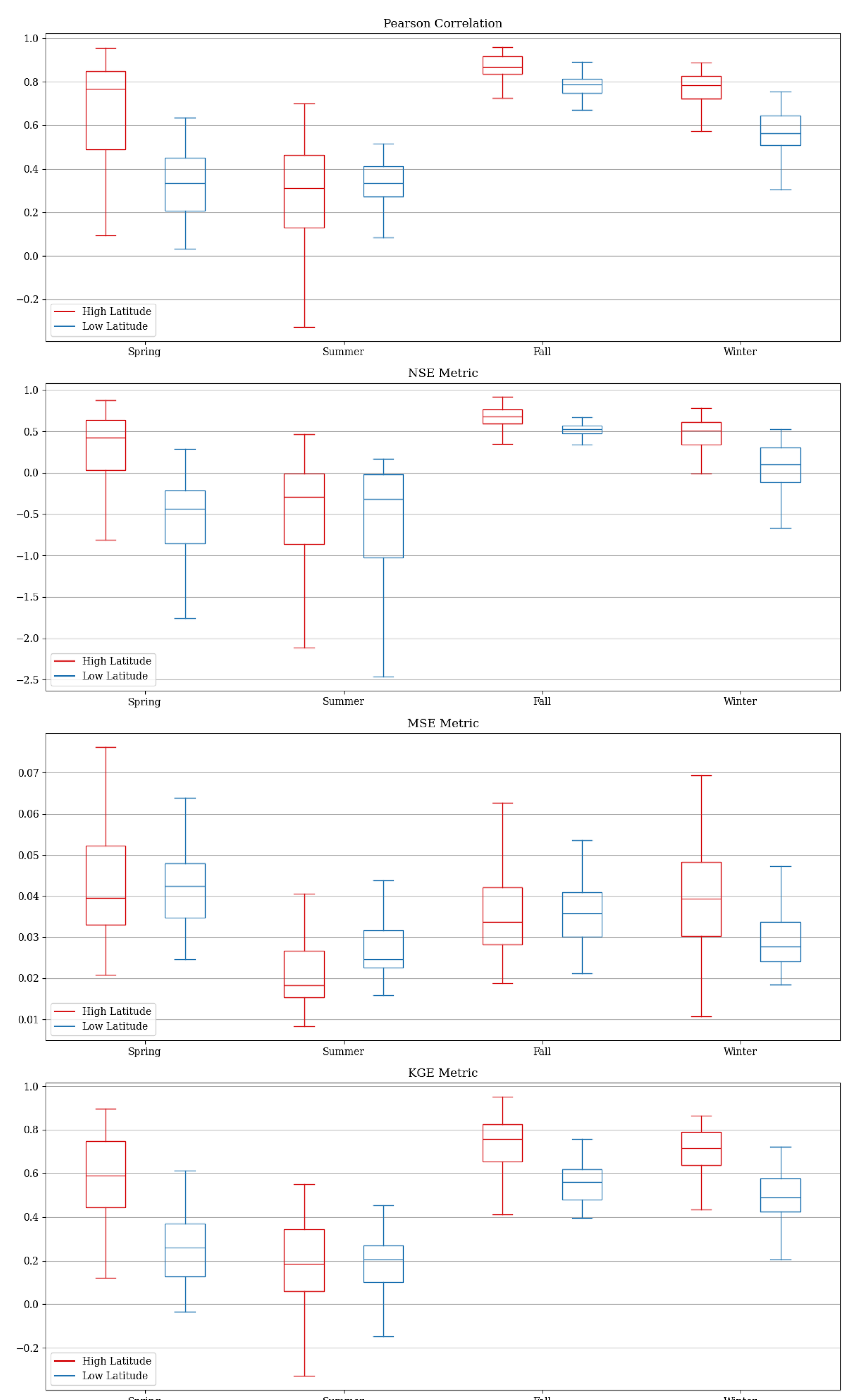


Fig 8: A comparison of model performance across different seasons.

Acknowledgments

Salient Predictions partially funded this research to study the Great Valley of California.