Table of Contents

List of Figures

List of Tables

Chapter 3. Data and Methodology

1.1 Dataset Overview

This section includes a description of the storage of and framework used to interface with the data, insights on the value distributions and spatial variability of the input forcings from NLDAS-2, as well as a look at similar bulk properties of the target soil moisture states and governing processes within Noah-LSM. In this work, we define our valid domain to include all points falling within the conterminous United States, excluding those points within the NLDAS-2 domain falling with Canada and Mexico. We also exclude points that are classified as "water," "bedrock," or "other" in the STATSGO dataset, since they don't correspond to meaningful hydraulic properties, and have idle time series. What remains are 50,875 candidate grid cells within a 224x464 pixel domain.

1.1.1 Data Storage System

The data used in this project were acquired from the Goddard Earth Sciences Data and Information Services Center's Distributed Active Archive Center (GES DISC DAAC) in May of 2024. The DAAC archives the NLDAS-2 forcings and corresponding Noah-LSM model outputs as separate hourly files in a GRIB1 format, of which we downloaded the full 12-year time series from January 1, 2012

to December 31, 2023. This subset constitutes 210,384 files with a total size of just over 891.38 GB.

Since this project concerns developing 2-week time series of the forcings on a per-pixel basis, it would be inefficient to extract data from several hundred files for each sequence sample. Furthermore, it is widely recognized in deep learning that input/target pairs from heterogeneous datasets should be globally shuffled during the training process, as outlined by (Nguyen et al., 2022). This is because local subsets may have distribution characteristics that are distinct from the full dataset, so as the model trains on an unshuffled dataset, the loss gradients it experiences may encourage it to converge on a locally-optimal solution that does not generalize well to the overall task. Shuffling is especially salient for geoscience datasets like this one, which are highly spatially and temporally heterogeneous. With this in mind, the overhead from file I/O operations would be prohibitive for sporadically drawing samples from throughout the GRIB dataset during training or inference.

To address this problem while maintaining the spatiotemporal structure of the data, we develop a custom file standard using the HDF5 format – hereafter referred to as the TIMEGRID – and extract the full 12-year NLDAS and Noah-LSM record as a collection of them. The HDF5 format offers a system for memory-mapped data chunking in multiple dimensions, which means the data therein can be sparsely buffered and accessed on a per-chunk basis without loading the entire file into memory: a considerable advantage for thoroughly shuffling or accessing subsets of contiguous data within large files. In practice, each timegrid contains

3 years of data covering 1/6 of the spatial domain, and stores the 4-dimensional time-varying data (time, latitude, longitude, data type), 3-dimensional static data (latitude, longitude, data type), and timestamps alongside a string-serialized attribute dictionary. The attributes contain information on abbreviated and full data type names, ordering, units, and sources, which are sufficient to inform a variety of accessor methods with a wealth of downstream use cases.

1.1.2 Spatial Data Characteristics

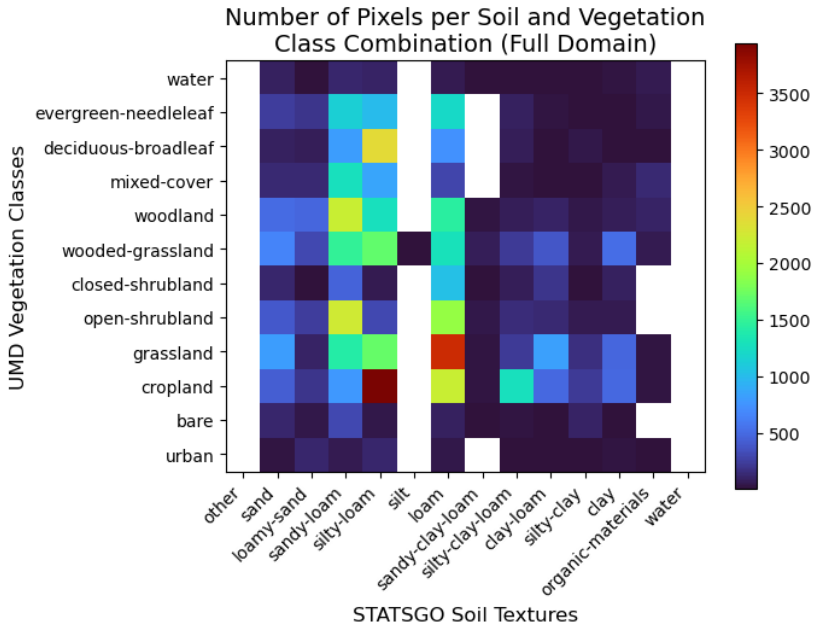

Figure 1.1: Full-domain combination matrix of vegetation and soil classes

One important aspect of the NLDAS-2/Noah-LSM datasets is the relationship between static and dynamic parameters, and the regional variance of both of these input types. Given any particular forcing time series, the subsequent

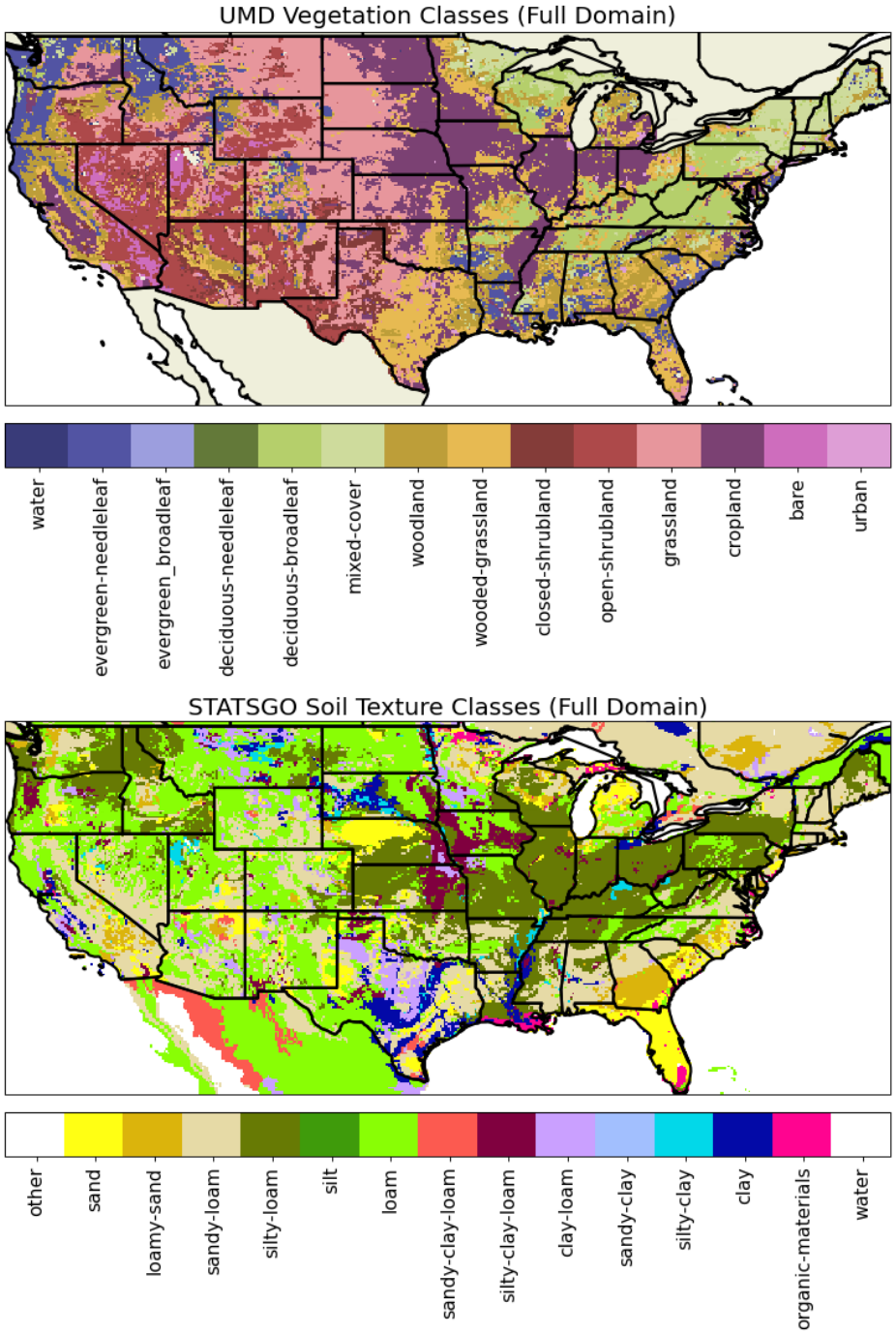

Figure 1.2: Spatial distribution of vegetation and soil classes

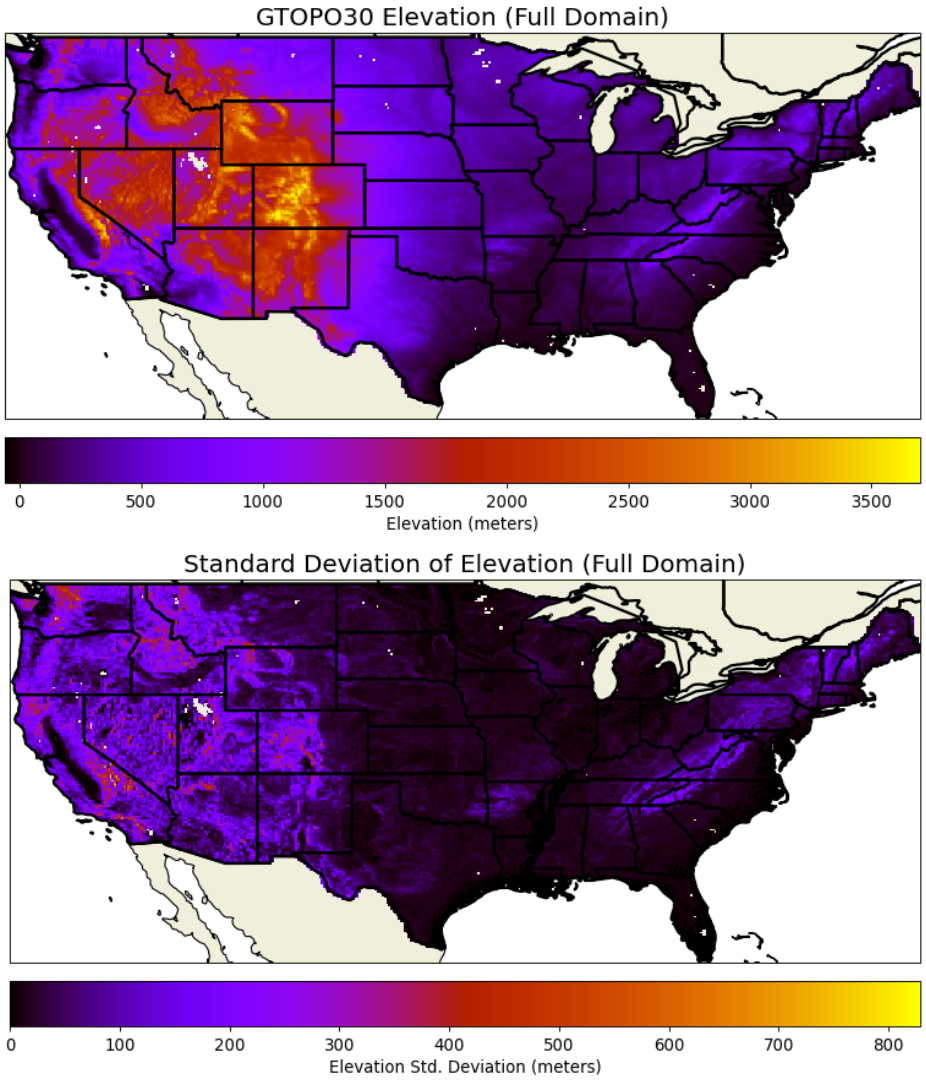

Figure 1.3: Elevation and standard deviation of elevation on the CONUS domain

land surface response is modulated by values for vegetation type, soil texture, elevation, standard deviation of elevation, slope, and slope aspect, which are consistent per-pixel throughout the dataset. In this work, we will only train models informed by the first four. Slope and aspect were left out because within Noah-LSM they are only used within the snowpack parameterization (Barlage et al., 2010), and snow is generally not a target variables for the models presented here. Furthermore, slope correlates strongly with the standard deviation of elevation. In retrospect, however, these inputs may have increased the information available to the models for estimating the transference of water from snow melt to the soil layers, especially in mountainous regions like the Rocky Mountain and Sierra Nevada ranges. An evaluation of the impact of these parameters on model results is left for future work. Elevation is mainly used to perform orographic regressions on pressure, temperature, humidity, and precipitation while resampling forcing data, however it could be useful as a predictor that indicates processes relevant to mountain snowpack dynamics.

As described in the background, the vegetation classes encode the properties of the canopy relevant to precipitation interception and land surface shading, the efficiency of plant transpiration at removing water from the soil, and the number of layers from which water is drawn (that is, the rooting depth). Since the vegetation parameter is discretely categorical within the Noah-LSM algorithm, we employ a special method of introducing them into the model called class embedding, which is elaborated upon in the next subsection. The soil texture class corresponds to a variety of hydraulic properties identified by (Cosby et al., 1984),

which include field capacity, hydraulic conductivity, porosity, wilting point, matric potential, and Skempton's pore water pressure ("B" parameter). These describe physical characteristics of the soil-water system including the rate of downward percolation of water, the efficiency and limits of plant water uptake, the speed of infiltration, and the total amount of water that soil can contain per unit volume. The basic observable feature of soil that determines all of the hydraulic properties is the size distribution of its constituent particles, which is often articulated as the mass fraction of sand, silt, and clay components within the soil. In the interest of providing the models with real-valued inputs having relatively low dimensionality, these three texture components will serve as the representation of soil texture for the ANNs trained here.

The interplay between plant water uptake and soil water dynamics as governed by the static inputs represents a considerable source of complexity within Noah-LSM. Furthermore, as Figure 1.1 demonstrates, the distribution of combinations of soil and vegetation categories is extremely non-uniform, which makes it more difficult for ANNs to learn solutions that are general. Figure 1.2 shows the geographic locations of vegetation and soil classes. The most common class combination is silty-loam soil types juxtaposed with cropland, with 3,945 members found dominantly in the Midwest and lower Mississippi river basin, with some contribution from the Columbia Plateau in Washington and Eastern Nebraska. Next most common are the 3,490 pixels in loamy grasslands, which are distributed widely throughout the West US including the high plains, Western Texas and New Mexico, Utah, and Idaho. The remaining combinations all have fewer than 2,500

members within the domain. Sand and clay dominated soils form the upper and lower extremes of soil particle size, respectively, and thus have rather different soil water characteristics. The sandiest soils are found in Southern Coastal Plains, Michigan, Texas, the Nebraska Sandhills, and the desert Southwest. Clay soils are relatively rare compared to silty and sandy soils, and considerably more spatially heterogeneous. They are mainly found in tight groupings around Central Texas, the Mississippi Alluvial Plain, Eerie Lake Plains, and the Missouri River Basin in South Dakota. Despite their infrequency, clay soils span the full range of surface classes.

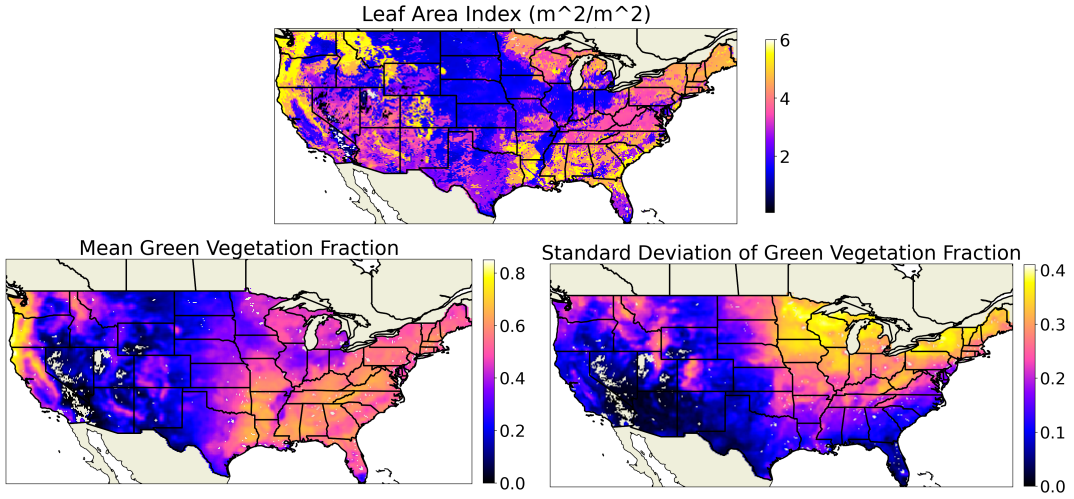

Figure 1.4: Gridded mean and standard deviation of vegetation input parameters (2012-2023)

Although they vary smoothly on an hourly basis, the LAI and GVF parameters are similar to static parameters in that they cycle consistently per-pixel on an annual basis (rather than dynamically changing based on variable atmospheric conditions), and modulate the soil water dynamics via through their effect on the

vegetation parameterization. As Figure 1.4 indicates, the densest annual-averaged canopy cover corresponds to evergreen needleleaf surface types, and there is almost no canopy over croplands and grasslands of the Midwest, California Valley, and the Great Plains. The greenest satellite-derived vegetation covers the West Coast and Sierra ranges, followed by the South and Northeast. The standard deviation of GVF indicates the regions of most significant seasonal variability, which corresponds to deciduous-dominant locales.

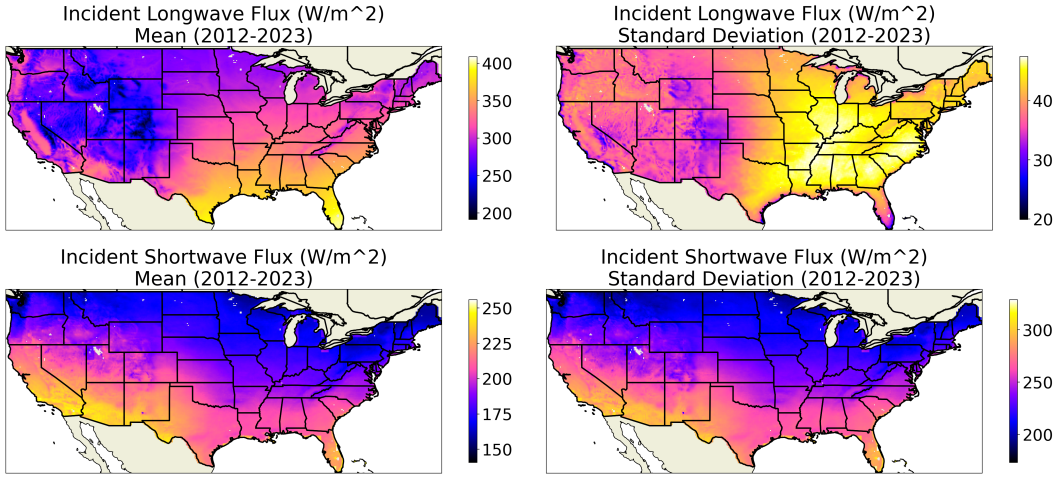

Figure 1.5: Gridded mean and standard deviation of radiative input forcings (2012-2023)

1.2 Model Architectures

1.3 Training Paradigm

1.4 Evaluation System

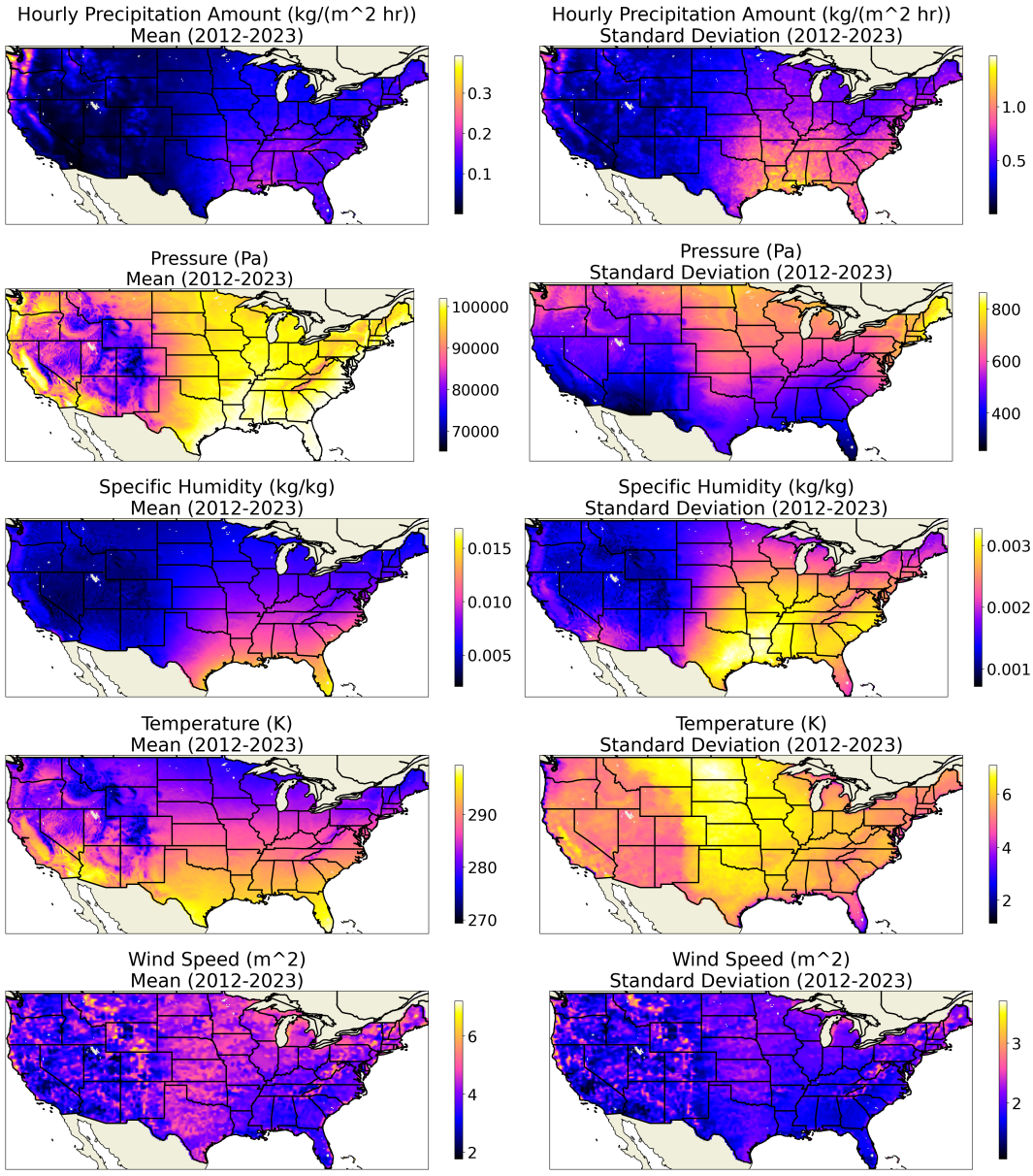

Figure 1.6: Gridded mean and standard deviation of input forcings (2012-2023)

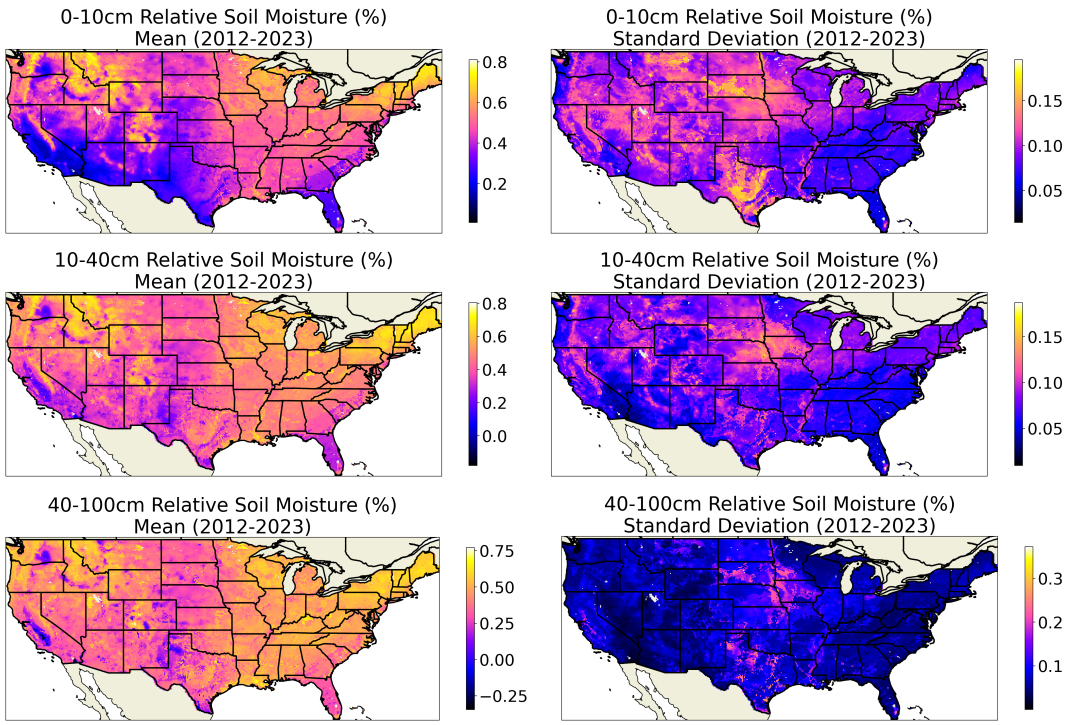

Figure 1.7: Gridded mean and standard deviation of RSM (2012-2023)

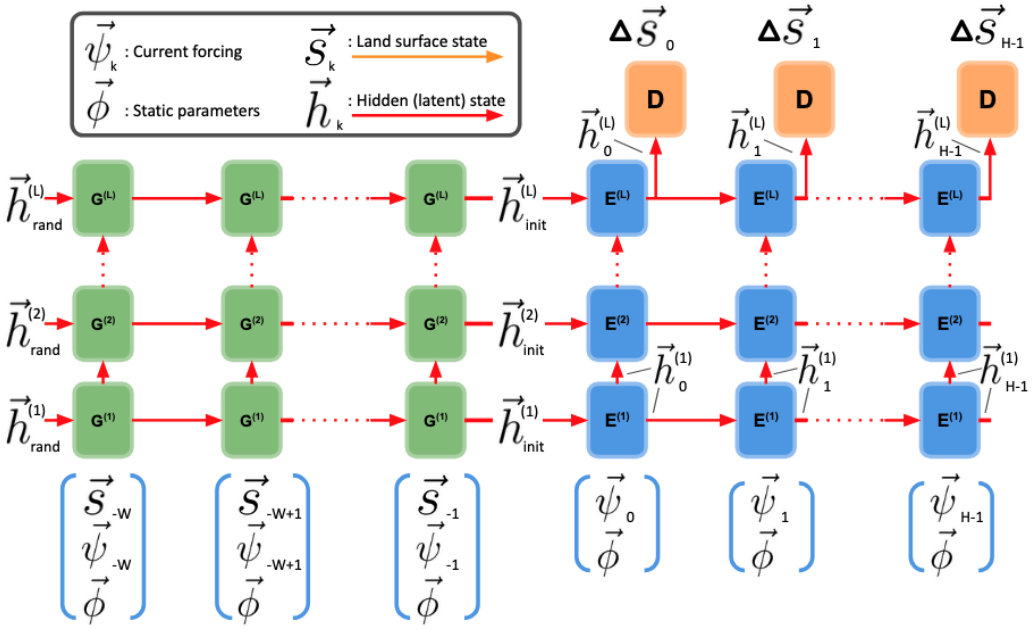

Figure 1.8: Sequence-to-sequence RNN architecture with spin-up window cells **G** for initializing first-step weights, and fully-connected decoder **D**.

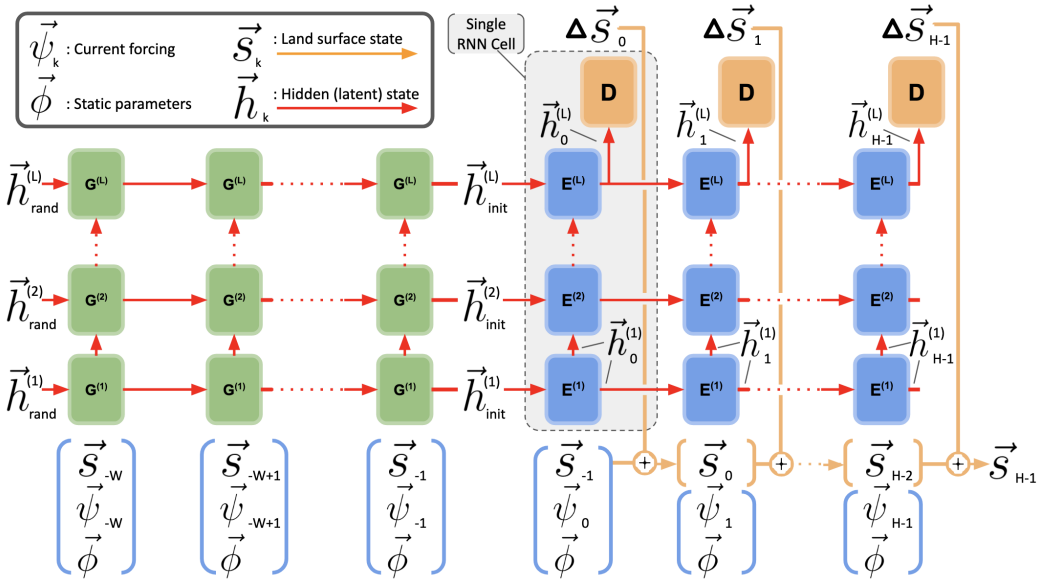

Figure 1.9: Sequence-to-sequence RNN with explicit output state accumulation.

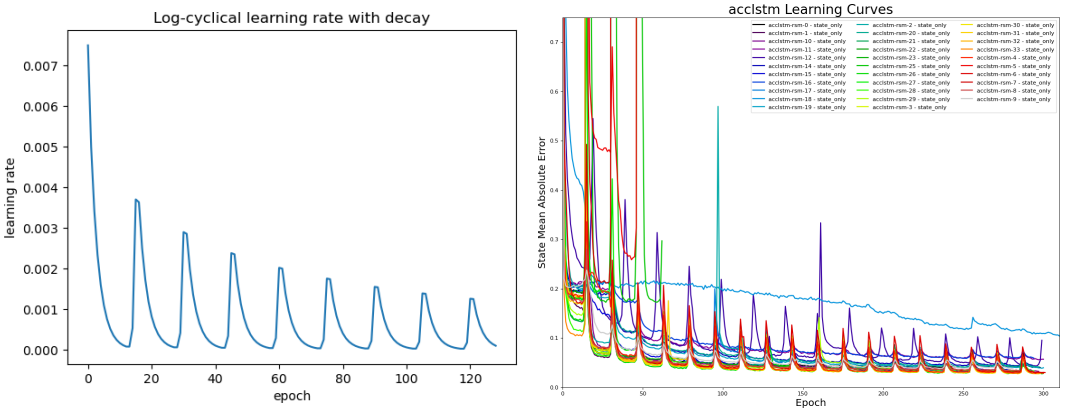

Figure 1.10: Sequence-to-sequence RNN with explicit output state accumulation.

References

Barlage, M., Chen, F., Tewari, M., Ikeda, K., Gochis, D., Dudhia, J., Rasmussen, R., Livneh, B., Ek, M., and Mitchell, K. (2010). Noah land surface model modifications to improve snowpack prediction in the Colorado Rocky Mountains. *Journal of Geophysical Research: Atmospheres*, 115(D22). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2009JD013470.

Cosby, B. J., Hornberger, G. M., Clapp, R. B., and Ginn, T. R. (1984). A Statistical Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils. *Water Resources Research*, 20(6):682–690. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/WR020i006p00682.

Nguyen, T. T., Trahay, F., Domke, J., Drozd, A., Vatai, E., Liao, J., Wahib, M., and Gerofi, B. (2022). Why Globally Re-shuffle? Revisiting Data Shuffling in Large Scale Deep Learning. In *2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 1085–1096. ISSN: 1530-2075.