UNDERSTANDING HYDROLOGIC FUNCTIONS OF WATER STORAGE AND SOIL MOISTURE AND IMPROVING PREDICTIVE CAPABILITY USING BIG DATA MACHINE LEARNING METHODS

A Dissertation in
Civil Engineering
by
Kuai Fang

© 2018 Kuai Fang

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

December 2018
The dissertation of Kuai Fang was reviewed and approved* by the following:

Chaopeng Shen  
Assistant Professor of Civil Engineering  
Dissertation Advisor, Chair of Committee

Daniel Kifer  
Associate Professor of Computer Science and Engineering

Klaus Keller  
Professor of Geosciences

Xiaofeng Liu  
Assistant Professor of Civil Engineering

Patrick Fox  
Head of the Department of Civil and Environmental Engineering

*Signatures are on file in the Graduate School.
Abstract

Water storages including groundwater and soil moisture are important components in the water cycle, not only for their environmental and societal implications but also as state variables modulating hydrologic fluxes. The latest advance in remote sensing, e.g., Gravity Recovery and Climate Experiment (GRACE) and Soil Moisture Active Passive (SMAP), provided novel observations of subsurface water content. Big data machine learning (BDML) is a powerful approach to extract patterns and identify unrecognized linkages. However, earlier hydrologic studies, including those based on machine learning, tend to focus on the special hydrologic features in a confined geographic region, rather than on the commonality among regions. There is substantial potential for knowledge attainment by examining the connections between hydrologic fluxes and storage states offered by the abovementioned satellite-based datasets, and interpreting the regional gradients in these relationships. In addition, the recent breakthrough in BDML, known as Deep Learning (DL), strongly improved the capability to model data, with higher data efficiencies compared to earlier machine learning methods. While DL has achieved unprecedented success in various disciplines, its potential in water science has not been recognized.

In this dissertation, first, I discuss how hydrologic signatures are extracted from GRACE can help improving water partitioning estimation based on the Budyko hypothesis. Then the relationship between GRACE storage and streamflow is examined via the lens of the newly proposed storage-streamflow correlation spectrum (SSCS), which quantifies how terrestrial water storage monitored by GRACE is correlated with the different annual streamflow percentiles. I ask two questions: (i) 'what are the SSCS patterns that exist over CONUS'; and (ii) 'what factors are controlling such patterns'? The BDML analysis of SSCS patterns presents important theories about controlling factors of the SSCS patterns, including storage on the Appalachian Plateau are limited by thin soils, compacted soils in northern Ohio lead to shallow a water table that limits storage, and streamflow on northern Great Plains and Southeast Atlantic regions are dominated by groundwater.
These theories, based on Classification and regression tree (CART) analysis, can then be corroborated or rejected. Nevertheless, CARTs has low data efficiency. The number of data points at the lower branches are exponentially less comparing to the upper ones. Thus the lower levels are highly unstable to interpret.

In the ensuing section, a novel data-driven technique, time series DL, is proposed to capture how surface soil moisture dynamics respond to atmospheric inputs, human interventions, and subsurface feedbacks. I show that the DL method, Long Short-Term Memory (LSTM), is an effective approach in prolonging satellite-sensed soil moisture spatiotemporally. Compared to the traditional method, this DL algorithm is more robust and more accurate. Furthermore, it created a continental-scale model that unifies the data from the continental United States. This prolonged dataset could be used to related past extreme events to soil moisture dynamics. Furthermore, with the help of LSTM, SMAP product is extended years beyond its lifespan with a similar performance as the training period. A fused product combining this new long-term LSTM projection with model simulation is found to outperform either one of its components or combinations of different models. These findings reveal the potential of DL and the value of SMAP data in the long-term projection of soil moisture dynamics. They also suggest there will be advantages to integrating machine learning, data, and process-based models.
Table of Contents

List of Figures ix
List of Tables xxi
Acknowledgments xxii

Chapter 1
Introduction 1
1.1 Background ........................................ 1
1.2 Dissertation Organization .......................... 6
Bibliography .......................................... 7

Chapter 2
Improving Budyko curve-based estimates of long-term water partitioning using hydrologic signatures from GRACE 13
2.1 Abstract ........................................ 13
2.2 Introduction ...................................... 14
2.3 Methods ......................................... 17
  2.3.1 Data sources and processing procedures ...... 17
  2.3.2 Hydrologic signatures extracted from GRACE .. 19
  2.3.3 Climatic indices ................................ 23
  2.3.4 The general departure model and its rationale ... 24
  2.3.5 Partitioning of variance ........................ 25
  2.3.6 Multi-scale, multi-dataset validation of alternative models 26
2.4 Results and Discussion ........................... 26
  2.4.1 The GRACE-assisted departure model .......... 26
  2.4.2 Factors contributing to Budyko departures and $A/P-E/P$ correlation 29
  2.4.3 Global validation and the relevance of co-evolution to model choosing 31
Chapter 3

Full-flow-regime storage streamflow correlation patterns provide insights into hydrologic functioning over the continental US

3.1 Abstract ................................................. 49
3.2 Introduction ........................................... 50
3.3 Methods ................................................ 53
  3.3.1 Streamflow and GRACE TWSA data sources ............ 53
  3.3.2 The storage-streamflow-correlation spectrum (SSCS) ..... 54
  3.3.3 GRACE signatures .................................... 56
  3.3.4 Exploratory clustering of SSCS ............................ 57
  3.3.5 First-order classification tree analysis of the controls ... 57
3.4 Results and Discussion ................................. 60
  3.4.1 The Streamflow-Storage Correlation Spectrum ............ 60
  3.4.2 Typical SSCS behaviors over CONUS .................... 63
  3.4.3 Continental-scale physical controls of SSCS ............. 67
  3.4.4 SSCS-derived insights and hypotheses about hydrologic behaviors ............................................. 69
    3.4.4.1 General interpretations .............................. 69
    3.4.4.2 Hypothesis 1: Soil-thickness limits storage on the Appalachian Plateau ............................... 72
    3.4.4.3 Hypothesis 2: Groundwater dominates streamflow and large-storage system in the Southeast CONUS 75
    3.4.4.4 Hypothesis 3: Shallow-water-table limits storage in Indiana, Ohio, and Florida ...................... 77
    3.4.4.5 Hypothesis 4: Groundwater dominates streamflow on the northern Great Plains ..................... 79
    3.4.4.6 Hypothesis 5: Gradient in high flow bands along central lowlands is driven by change in energy ...... 80
    3.4.4.7 Unanswered questions ............................... 81
    3.4.4.8 Assessing probability of success for GRACE-based estimates for streamflow .......................... 81
3.5 Conclusions ............................................. 82
3.6 Limitations and Future Work ............................ 83
Chapter 4
Prolongation of SMAP to Spatio-temporally Seamless Long-term Coverage of Continental US Using a Deep Learning Neural Network

4.1 Abstract .............................................. 91
4.2 Introduction ......................................... 92
4.3 Methods and Datasets ............................... 94
  4.3.1 Data sources and inputs ....................... 94
  4.3.2 LSTM setup .................................. 95
  4.3.3 Tests, Conventional Algorithms, and Evaluation Metrics . . 96
4.4 Results .............................................. 98
  4.4.1 Overall test performance .................... 98
  4.4.2 Comparison of generalization capability with other methods 100
4.5 Discussion .......................................... 104
4.6 Conclusion .......................................... 106
4.7 Limitations and Future Work ..................... 106
4.8 Author Contributions ............................. 107
Bibliography ............................................ 107

Chapter 5
The value of SMAP for long-term soil moisture estimation with the help of deep learning

5.1 Abstract .............................................. 113
5.2 Introduction ......................................... 114
5.3 Methods and Data Sets ............................. 117
  5.3.1 Datasets: Targets and Predictors ............ 118
  5.3.2 The time-series deep learning model .......... 119
  5.3.3 Validation: SMAP core validation sites, sparse network, and evaluation metrics .................. 119
5.4 Results and Discussion .......................... 121
  5.4.1 Evaluation at the core sites ................. 121
    5.4.1.1 Statistical comparison between training and hindcast periods ............... 121
    5.4.1.2 Statistical comparison of performances between LSTM, Noah, and Comb .......... 124
    5.4.1.3 Time series comparisons .................. 126
  5.4.2 Evaluation over the sparse network .......... 129
List of Figures

2.1 (Caption on next page.) ................................. 20
2.1 Hydrologic signatures and climatic index computed for the world using data from 2002 Oct to 2010 September: (a) average annual GRACE TWSA amplitude as a fraction of precipitation ($A/P$); (b) the inter-annual GRACE signal variability ratio $\gamma$ (dimensionless ratio between between-year and within-year TWSA variability); and (c) the precipitation-temperature seasonality index following Woods (2009), which is -1 for completely out of phase, 1 for completely in phase, and 0 for uncorrelated. ................................. 21

2.2 Using GRACE TWSA amplitude as a fraction of precipitation ($A/P$) to predict the departure ($\delta$) from the Budyko curve for the HUC4 dataset. $\delta = \frac{p-Q}{p} - f\left(\frac{E_p}{p}\right)$ where $Q$ is observed discharge, $E_p$ is potential evapotranspiration and $f\left(\frac{E_p}{p}\right)$ is the Budyko formula. (a) Without correction, the HUC4 basins scatter around the Budyko curve, some with significant departures; (b) After correcting using $A/P$, basins are now much more closely clustered around the Budyko curve; (c) The negative correlation between $A/P$ vs $\delta$ allows the improvement over Budyko. ................................. 28
2.3 Variance decomposition to surrogate indicators and physical factors using ANCOVA for HUC4 and GRDC datasets. The fraction of the horizontal bars occupied by a factor indicates the variance explained by this factor. Type I Sum of Squares (SS) is “sequential” so the order of factors influence the results, unlike type II, which excludes main effects of other factors. The order in O1 is $A/P$, $\xi$, $\gamma$, $S/P$, $NDVI$, and $acf_{48}$, while O2 is $\xi$, $\gamma$, $S/P$, $NDVI$, $acf_{48}$ and $A/P$. The joint symbol, $\cap$, stands for the part of variance with attributed to more than one factors. As $\xi$ and $S/P$ occupy small fractions in O1 but more noticeable in O2, we conclude that $A/P$ encompasses $\xi$ and $S/P$, while the latter two constitute a large fraction, although not all of the predictive power of $A/P$.

2.5 The scale-dependent performance of 5 models when HUC4-fitted models are migrated to the USGS GAGES-II basins. (a) RMSE of evaporation ratio (blue line is covered by red); (b) $R^2$ between observed vs predicted $E$. The correction formula with $A/P$ are transferrable to >1000 km$^2$ basins while the one without $A/P$ performed only slightly better than Budyko itself. The gradual degradation in performance toward smaller scales with the GRACE-assisted models is similar to that of the Budyko itself, indicating the degradation is due to either data limitation or inherent loss of accuracy of the Budyko equation for smaller basins, rather than due to the correction formula.

2.4 Errors of the models #0 through #13 (Table 2.1) when they are fitted to the HUC4 data (blue line) and their coefficients are migrated to the GRDC dataset ($GRDC^M$ green solid line), compared to when the models that are directly fitted to the GRDC data ($GRDC^F$, green dashed line). Colors of the markers indicate the number of predictors. Note HUC4 and GRDC RMSEs are on two different y-axes. We note that as indicators are added in models 1-6, they slightly reduce errors in HUC4 but increase errors when parameters are migrated to GRDC. In addition, while the model with $A/P$ is transferrable from HUC4 to GRDC, models without $A/P$ performs poorly when migrated. (Evaluation period: 2002/10-2010/09, because GRDC is only available before 2010/09)

2.6 (a) World annual average evapotranspiration, $E^C$, for 2002-2006 based on Budyko and linear model with $A/P$, inter-annual variability index ($\gamma$) and $S/P$, after GRACE error limiting; (b) Difference between $E^C$ and $E^{PJ}$ (white blanks are due to missing data in $E^{PJ}$); and (3) Difference between $E^C$ and $E^{GLDAS}$.
2.7 (a) Comparing GRACE-corrected product ($E^C$) with $E^{PJ}$ and Budyko itself in GRDC basins for 2002/10-2006/09 ($E^{PJ}$ is available only before 2006/09). Budyko has negative biases while $E^{PJ}$ has positive biases in high $E$ regions; (b-c) Compare $E^C$ with $E^{PJ}$ and GLDAS evapotranspiration. The root-mean-squared-difference is 175 mm/yr between $E^C$ and $E^{PJ}$, 131 mm/yr between $E^C$ and $E^{GLDAS}$, and 189 mm/yr between $E^{PJ}$ and $E^{GLDAS}$. We note that red-colored points (large GRACE errors) tend to be the most distant to the 1-to-1 line, indicating in these regions $E^C$ is unreliable due to GRACE error contamination.

3.1 Spatial distributions of some of the predictors. BDave: top soil bulk density; S/P: fraction of precipitation as snow; $\gamma$: TWSA interannual variability index; RH: relative humidity; RockDep: depth to bedrock from soils dataset; Slope: mean slope in the model from GAGES-II; acf: TWSA 1-month-lag auto-correlation function; $\xi$: P-Ep seasonality index which indicates how much in phase they are. We note that many factors, such as rock depth and slope, co-vary.

3.2 Maps of storage-stream correlation spectrum (SSCS) for USGS gauging stations. (Continue on next page)

3.2 (Previous page.) (a) the best correlation across all bands with the histogram and exceedance frequency (1 minus cumulative distribution) in the insert; (b) the percentiles corresponding to the best bands; (c) the nearest class centers of USGS gages based on spectral distances. Each circle represents the location of the gage (basin outlet). As we restricted the size of the catchment areas, these outlets are not far from basin centroids.

3.3 Regions labeled for the convenience of discussion. The different colors of annotations carry no special meaning - they are used only to increase readability of the labels. A: some basins on the low-$\gamma$ belt; B: part of the Southeast region with high storage-streamflow correlations across all bands; C: Ohio/Indiana with Class #6 (non-responsive) basins; D: Superior upland; E: Agricultural basins in Iowa; F: A region intersecting Southern Great Plains, Interior highlands and Western Gulf coastal plains; G: Northern Great plains; H: Northern Rocky Mountains; I: Arizona.
3.4 Boxplots of the SSCS for each class. The boxes contain 25%-75% percentiles, and the crosses are those considered outliers. The 15 bands are 0.5%, 1%, 2%, 5%, 10%, 20%, 50%, 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.5%. The numbers in the parenthesis stand for the number of stations in the class.

3.5 Parameter distribution for each class. Silt Cont: percentage of silt in top soil; P: precipitation; BDAve: average top soil bulk density; WD: number of days with measurable precipitation. Other variables are explained in Figure 3.1 caption. Except for class #3, all other classes are well spread out parameter ranges. This pattern means that a factor may have different impacts on SSCS in different regions.

3.6 (a) Cross validation of potential models. Every line represents a model, and the plot contains top 100 best performing models from cross-validation. The thick black line is the selected model for the tree in Figure 3.6b and Appendix B Figure B.15, and pink and blue lines are two nondominated models without soils data. Note the performance deteriorate when soil data is not available. On the x axis, 20% - 70% is the amount of data randomly held from training; HUC2: One two-digit Hydrologic Cataloging Unit (HUC2) was held out as a testing set; 2× HUC2: Two HUC2s were held out for testing. We looped through all possible combinations of HUC2s, and the error is the average of all cases. We note that withholding data by HUC2 is much more stringent test than randomly withholding data. We select the models that are among the top 10% of all CV tests and are not dominated by any other models. (b) Top three level of the composite-output tree that simultaneously predicts distances to 6 centers (the black line in Figure 3.6a). In the legend, we include the number of basins on each node (#Training Set/#Test Set), the variances (Var) and the composition of classes on the node. We note that S/P, relative humidity (RH), and interannual storage variability (γ) are ranked as most significant parameters for this tree.
3.7 (a and b) Together explain how low-γ belt basins are separated from the Southeast basins, and (c and d) together explain how Indiana/Northern Ohio basins are separated from neighboring basins. (a) Node split map showing how different classes, as separated by depth to bedrock, are distributed in the Eastern region of CONUS. (b) a simple ad hoc tree constructed to find the most sensible split that separate the Class #3 basins on the low-γ belt from the Classes #1 and #2 basins in the Southeast. RockDep (Depth to bedrock, inches) emerges as the most sensible criterion. Note that the GAGES-II data set is derived fromSTATSGO and therefore has a soil bulk density, are distributed in a region surrounding Indiana and Ohio. (d) An ad hoc tree indicates that bulk density is the most apparent factor to separate Class #6 (nonresponsive) basins from adjacent regions. When we construct this tree, the reference flag in GAGES-II data was provided as a candidate attribute, but was not chosen by the tree. The reason is, while Class #6 points here are predominantly agricultural and nonreference, many Class #3 basins (green ones) are also nonreference ones.

3.8 Time series of TWSA for a GRACE cell on the low-γ belt. Dashed lines indicate the mean of the peaks and troughs. $\sigma_p$ and $\sigma_t$, measured in cm, are the standard deviations of the annual peaks and troughs, respectively. $\sigma_p$ is less than $\sigma_t$, meaning the differences in peaks are smaller than the troughs. It suggests that the limited storage capacity is reached (recall the peaks also contain signals from snow) each year regardless of where the minimum is.

3.9 Comparison between streamflow hydrographs from a headwater basin from the high bulk density soil region in Ohio (USGS 04196000 Sandusky River near Bucyrus, Ohio, SSCS Class #6) and a basin from the Southeast (USGS 02390000, Amicalola Creek near Dawsonville, Georgia, SSCS Class #1). Catchment areas of both gages are around 230 km$^2$. The seasonal cycle of the base flow component of the Georgia basin is much more prominent than that of the Ohio basin, which has almost negligible base flow, but large and frequent peaks. We also note that there is a marked interannual difference in the base flow of the Georgia basin. The high storage in 2009 produced more large peaks than 2008.
4.1 Performance of LSTM (evaluated against SMAP) in the test set of the temporal generalization test. (a) $\text{bias}(LSTM) = LSTM - SMAP$. Each pixel in this figure is a patch of 4x4 SMAP pixels. Bias in most parts of CONUS is between -0.02 and 0.015; (b) Change of absolute bias due to LSTM correction. LSTM reduces the absolute bias significantly over CONUS; (c) LSTM anomaly correlation ($R(LSTM)$); (d) change of $R$ due to LSTM correction; (e) $\text{RMSE}(LSTM)$; (f) Since the RMSE improvement over Noah looks similar to panel b, here we show the difference in RMSE between LSTM and NN predictions. Noah’s performance is provided in Figure S3 in SI.

4.2 Comparisons between SMAP observations and soil moisture predicted by LSTM, Noah, and AR at 5 locations. We chose sites around 10-th, 25-th, 50-th, 75-th and 90-th percentiles as ranked by $R(LSTM)$.

4.3 Boxplots comparing LSTM, Noah, AR, LR, $NN_p$ and $LR_p$ in the temporal generalization test and the regular spatial generalization test. Each box and whisker element summarizes SMAP pixels over CONUS with percentiles annotated in the panel. Y-axis limits truncate Noah boxes to focus on the central part of other boxes. Left column: temporal generalization test; Right column: regular spatial generalization test. The three rows are for RMSE, Bias and $R$, respectively. A paired t-test suggests RMSE(LSTM) and RMSE(NN) are significantly different.

4.4 (a) Test biases for cases C1-C4 (distinguished by different colors), which are four combinations of 4 HUC2s. Each group consists of 4 boxes, which are, respectively, LSTM, NN, Noah in the training set, and self-assessed test bias. The self-assessed bias is Noah’s prediction over the test set minus LSTM’s predictions; (b) test $R$. We note significantly better seasonality captured by LSTM in cases C1-C3, but not necessarily in C4; (c) Distributions of Noah’s biases in the training sets for C1-C4.

5.1 Information flow diagram for the models compared in this paper. Both Noah and LSTM use forcing data from North American Land Data Assimilation System (NLDAS). LSTM uses SMAP as the main training target along with land surface characteristic masks from the SMAP mission among the inputs. The static input attributes to LSTM are described in Section 5.3.1.
5.2 Bias, ubRMSE and Pearson correlation of surface and root-zone soil moisture as evaluated by SMAP core-site in-situ data for the after-launch (AL) training period and the pre-launch (PL) validation periods. (next page)

5.2 (Previous page) Comb means the simple average of LSTM and Noah. In general, ubRMSE between in-situ data and SMAP (black) is slightly smaller than that between LSTM and in-situ during the training period (yellow), which is in turn smaller than that between LSTM and in-situ during the hindcast period (red). At sites where $R^{AL}_{(SMAP,in-situ)}$ is higher than $R^{AL}_{(Noah,in-situ)}$ (yellow higher than cyan), LSTM or Comb can have a higher R than Noah during hindcast, with Fort Cobb being an exception. At other sites, Noah is stronger than SMAP and LSTM, but Comb might still be better.

5.3 Same as Figure 5.2, but here we compare the LSTM+Noah (called Comb in Figure 5.2) with Noah+VIC, which seems weaker than Noah+LSTM. We also show the test metrics between LSTM and SMAP when it is trained in one year and tested in another (AL 1 Yr Test).

5.4 The time series comparison of LSTM hindcast, SMAP, and in-situ soil moisture anomalies at six different sites. The means were subtracted from all the series. LSTM hindcast inherits the patterns of discrepancies between SMAP and in-situ data, e.g., faster dry-downs for SMAP. At Carman, SMAP shows high-frequency oscillations, which disagree significantly with the in-situ data. LSTM shows the same oscillations. At St. Josephs, we note some periods of missing SMAP and in-situ data due to frozen ground or snow.

5.5 Time series comparison of LSTM hindcast, SMAP, and in-situ root-zone soil moisture on five core validation sites.

5.6 RMSE, Bias, unbiased-RMSE, and Pearson correlation of surface and root-zone soil moisture between CRN sites, LSTM, and SMAP.

5.7 Map of Unbiased-RMSE and Pearson correlation of surface and root-zone soil moisture between in-situ, LSTM, and SMAP for periods before and after SMAP launch, on CRN in-situ network. LSTM generally has the same spatial pattern as CRN, but there are a few instances of a notable decline in R, as annotated on the figure.
5.8 Comparison of long-term root-zone soil moisture trends between core validation sites, CRN sites, and LSTM hindcast. The correlation between the slopes of core site and LSTM is 0.87. The correlation between CRN and LSTM is 0.73 for all sites. Among all the sites with trends greater than 0.5% per year, the correlation is 0.88. Panels on the right show the time series of several selected sites. Slopes are annotated on the legend with a unit of percent change per year.

A.1 (a) $A/P$ for the HUC4; (b) Locations of USGS gages used in the study; (c) Maps of GRDC basins used in the study, with colors indicating the aridity index.

A.2 Comparisons of annual average fluxes comparison between HUC4 and MOPEX basins for the period 2002/01 to 2013/12.

A.3 GRACE leakage (upper) and measurement (lower) errors. Data from S.C. Swenson. 2012. GRACE monthly land water mass grids NETCDF RELEASE 5.0. Ver. 5.0. PO.DAAC, CA, USA. Dataset accessed [2016-03] at http://dx.doi.org/10.5067/TELND-NC005.

B.1 Comparing an Long Short-Term Memory (LSTM) unit with simple recurrent neural network (RNN). The transformations from inputs to $i$, $f$, $o$ are sigmoidal functions. From inputs to $g$ and from $s$ to $h$ the transformation is $tanh$. $\otimes$ means multiplication by weights. Main point: the conventional design of RNN only iteratively update the hidden state. The design of gates in LSTM allows it to learn when to forget past states, and when to output, thus addressing the issue of slow training of front node with RNN. Figure is modified from [Greff et al., 2015].

B.2 The training and testing errors as functions of the number of predictors in forward (a) and backward (b) feature selections. As we can see, when first few predictors are added to the forward selection, there is a dramatic reduction in error. While the training error continues to decrease after the 8th predictor, the testing error ceases to change and starts to increase after 25th predictor, suggesting overfitting. In backward selection, when least important attributes are removed (starting from the right-hand side of panel b), the error decreases due to the reduction of overfitted parameters. Then the error flattens until the last few important predictors are removed.

B.3 The reference basins for the same region shown in Figure B.13c. A lot of Indiana/Ohio basins are removed.
B.4 Comparisons of storage-streamflow correlations between reference and non-reference basins. We do not observe differences between these two sets.

B.5 Re-created Figure B.13c-d after adding the reference flag from GAGES-II. They are identical to Figure B.13c-d because soil bulk density remains the most influential factor although we added the 'reference' flag. The flag is not important because many class #3 points are also non-reference basins.

B.6 Re-created Figure 3.4 & Figure B.1 using only reference basins. We note Centers #5 and #6 have moved because many basins in the upper right corner have been removed. As a result, panel (g) & (h) look different from Figure 3.4. However, as we have explained previously, non-reference status is not the cause of why some basins were Class #6 in Figure 3.4.

B.7 (Left duplicate) There is no noticeable difference between Figures 3.4 and this one uses 0 for the bands with zero-flows, instead of removing the stations from the analysis. If one compares very carefully with Figure 3.4, she/he may notice a few additional points in the lower right part of the panel (a) of this figure. (Right duplicate) Figures 3.4 and Figure B.1 re-created without removing basins with high dam densities or large dam volumes. Again, the difference is difficult to observe.

B.8 (a-d) Storage-streamflow correlations vs. dam densities for dammed basins and those undammed basins with similar attributes. We fitted a rank-based Sens slope to the data points and found that there is a significant relationship only between low flow bands and dam density. However, the slope is small; (i-iv) the same information presented in boxplot format; (e-h) same as (a-d) but created using dam storage data between 0-200 mm. (v-viii) same as (i-iv) but for dam storage dam. We removed basins with dam density > 2 (100 km$^2$) or dam storage > 500 ML/km$^2$.

B.9 (Left 4 panels) Distribution of dammed vs. non-dammed basins on the SSCS principal component plane. (Right panel) Comparing class assignment pre- and post-correction. We note that most basins with dam storage between 200 and 500 mm retain class assignment.

B.10 Maps of CONUS: (a) Physiographical regions; (b) States.
B.11 (a) Comparison between cumulative distribution functions (CDF) storage-streamflow correlations (high-flow-band averages and low-flow-band averages) using Spearman and Pearson definitions. We note that the Pearson correlation is higher, indicating that there is no strong nonlinear trend between storage-streamflow relationships; (b) comparison between using calendar-year, water-year and our extrema-based windows for extracting streamflow percentiles. The extrema-based window is better than the other two options in all bands, especially in low flow bands.

B.12 Identified TWSA peaks and troughs for an HUC4. There are 10 peaks and, in this case, 11 trough data points.

B.13 Map of spectral distances to class centers.

B.14 SSCS aggregated by physiographic divisions in the US.

B.15 A composite-output tree that predicts distances to 6 centers at the same time. This tree is one of the best performing trees that is selected from cross validation test. End nodes have been annotated with the regions where their points are primarily located in. Other realizations of CART may appear different at lower levels, but many splits are similar, only placed in different branches of the trees.

B.16 (a) Baseflow Index (BFI) map for CONUS estimated using flow separation technique. The data is from Wolock (2003). We note that Southeast CONUS and Northern Great Plains have high BFI, while BFI is low in Indiana/Ohio; (b) Simulated water table depth for CONUS using land surface model-estimated recharge, topography, and a groundwater flow model. Simulated data is from Fan et al. (2013). We can see that the Fan et al.s model captured the shallow water table in Indiana/Ohio and Florida, while the water table is deeper in western Illinois and Iowa. Reprint permission has been obtained.
C.1 Proof-of-concept long-term hindcast tests with Noah-simulated soil moisture as the target. (a) boxplot comparing errors (evaluated against Noah) for the 10-year hindcast. Noah solution is contaminated by a Gaussian noise with a standard deviation ($\sigma$) of 0.04. The RMSEs are calculated for each 2-year period and are grouped over CONUS to form the boxplot. Note that error does not increase as hindcast length increases, i.e., during 2005-2006, the errors are not greater than those in 2013-2014; (b) same as (a) but for a 7% relative noise; (c) time series for Noah, LSTM and AR$_p$ at a pixel. We only show 5 years of hindcast for clarity of the plot. The zoomed-in panel on (c) (corresponding to the brown box in the main plot) highlights how AR$_p$ over-estimates the two soil moisture peaks. Meanwhile, AR$_p$ seems to have dampened small-scale fluctuations.

C.2 SMAP data quality with geographic regions annotated on the map for reference. The values shown is the time-averaged SMAP "recommended quality" flag. We notice that Appalachian Highlands and Southeast Coastal Plains are both mostly flagged as having bad quality, but LSTM's root-mean-squared-error from SMAP L3, RMSE(LSTM) are in the range of 0.02-0.035 in the South Appalachian and Coastal Plains according to Figure 1. Because random error cannot be captured in the test, it suggests SMAP quality may be not as bad as thought. However, this finding and the potential influence from models in the retrieval algorithm need to be thoroughly evaluated.

C.3 Performance of Noah evaluated against SMAP in the testing set of the temporal generalization test. (a) Bias: the time-averaged value of Noah-predicted soil moisture, interpolated to 5 cm, and SMAP L3 product; (b) anomaly correlation coefficient between Noah-predicted soil moisture and SMAP L3 product.

C.4 Same as Figure 3 but the Noah-simulated values are linearly interpolated to 0-5 cm. We tested several interpolation methods: (i) directly using 0-10 cm data; (ii) linear (2-point) and cubic vertical interpolation (3-point) using top layers; and (iii) integral interpolation: we determined a 2nd-order (or 3rd) polynomial whose integral in these layers agree with Noah-simulated values. This Figure shows method (i), whose results are very similar to those reported in Figure 3, while other interpolation methods also generate similar results.
C.5 Comparisons between LSTM models with Noah (L+Noah) and MOS (L+MOS) solutions as inputs. Both Noah and MOS solutions are obtained from North American Land Data Assimilation System (NLDAS). The distinction between 'train' and 'test' for Noah and MOS only means the different time periods for which the metrics are calculated. Noah and MOS have comparable performance in simulating moisture climatology. It appears MOS generally has smaller root-mean-squared error (RMSE) and smaller bias. However, using which model in the inputs does not seem to have a noticeable impact on the test performance of the LSTM models.
List of Tables

2.1 Tested Linear Models With Linear Weights Provided. HUC4 means the model was fitted using HUC4 data. GRDC means the model was fitted using GRDC data. Model #9 was chosen to produce a global-scale comparison with $E^{PJ}$ and $E^{GLDAS}$. 31

3.1 Hydrologic Signals From SSCS and Potential Hypotheses That Explain Them Based on Hydrologic Reasoning 71

B.1 41 factors considered and their rankings in forward and backward feature selections. Factors with a ”” in the name come from GAGES-II dataset. (*) indicates selected for CV tests. 151

C.1 Predictors used in the training of LSTM, lasso-regularized linear regression, and one-layer feedforward neural network 175
Acknowledgments

It took me almost six years to finish this Ph.D. program. There have been several challenging moments. However, the support and encouragement from countless individuals helped me to survive.

First of all, I would like to express my gratitude to my academic advisor, Dr. Chaopeng Shen, for the continuous support and guidance of my Ph.D. study and related research. You can read my thoughts, and I am appreciated your understanding and maybe tolerance of my work plan. I also learn and am affected a lot from your knowledge, passion, and vision. It has always been a pleasure working with you.

I would like to thanks to my committee members: Dr. Daniel Kifer, Dr. Klaus Keller, and Dr. Xiaofeng Liu, for helpful advice on the scope and implementation of my research. I am grateful for your insightful comments and suggests, as well as hard questions and criticisms. I learned from every meeting we had. I am thankful that you pushed me to think critically and toward more perspectives.

This dissertation would be impossible without the help and support form lots of individuals. Thanks to everyone I worked with in Shen’s group, Xinye, Shilong, Kurt, Nuan, Tasnuva, and Wenping, as well as other lab-mates at Sackett 4th floor, for the countless help I asked. Many thanks to my teammates at the CSA, for all those matches we wined or lost, for all the tackles, fights, and beer. I want to give special thanks to my friends. Thanks to coming every Saturday night the big night. There are also those who would rather not to be mentioned here, and I am grateful that you are there.

I would like to acknowledge the support from Penn State through Cyber-Science seed grant; U.S. National Science Foundation through award EAR-1832294; U.S. Bureau of Land Management through Interagency Agreement L11PG00354; Lawrence Berkeley National Lab and the U.S. Department of Energy, Office of Science, under Award Number DE-AC02-05CH11231.

Last but not least, thank my parents and families for supporting me for my entire life so far. Their support is endless, and they have given me all they have.
Chapter 1  |  Introduction

1.1 Background

Water is stored on land in several forms. Globally, groundwater, soil moisture, reservoirs, and streams store approximately 36.1%, 0.2%, 0.4%, and 0.005%, of the freshwater, respectively (Dingman, 2015). Not only do these stores provide water for human and ecosystem consumption, but they also serve as the state variables in the hydrologic system and are associated to land surface fluxes such as evapotranspiration, vapor/heat/momentum transfers, streamflow, and soil respiration (Anderson, 1982). On the one hand, the state variables change as a result of fluxes that occur; on the other hand, the state variables nonlinearly modulate the fluxes: the magnitudes of these fluxes depend on the amount of storages in various compartments (Wittenberg, 1999). For example, groundwater storage determines baseflow (its contributions to rivers) while soil moisture modulates the rate of evaporation. Long-term changes in these stores can induce changes in related processes such as land-atmosphere interactions, ecosystem health and agricultural productivity. Indeed, the impacts of global change and climate variability have marked imprints on these stores, e.g., groundwater depletion (Wada et al., 2010; Famiglietti, 2014) and soil moisture droughts (Wang et al., 2011; Robock, 2014), which, in turn, have implications for physical processes and sustainability of water resource. Thus, comprehensively understanding the connections and feedbacks between these stores and other processes is of fundamental importance to improving our understanding and predictive capability in in hydrology.

Significant research in the past decades from observational, analytical and mod-
eling perspectives, e.g., surrounding the critical zone initiatives, have revealed hydrologic dynamics for basins around the world. For example, hydrologists now understand groundwater storage modulate hydrologic connectivity which has threshold-like controls on rapid runoff production (McGlynn et al., 2004; Jencso et al., 2009). However, the effects of such control differ widely in different geographic regions, as controlled by many interweaving dynamics. Hydrologists have found it difficult to generalize their theories across a diversity of landscapes. While each study was an attempt to elucidate the unique features of their study basin, the knowledge of the domain collectively remained disorganized and scattered. Despite local success in various sub-branches, there clearly lacked a global unity to the findings and a sense of order (Wagener et al., 2007).

The lack of uniform understanding is reflected in the existence of myriad hydrologic models, which were often formulated based on observations from an extensively studied region, and then extrapolated to other regions with the requirement of heterogeneous input data. The TOPMODEL (Beven, 1997), for example, was developed in England where groundwater plays a predominant role. It emphasizes groundwater convergence and assumes the hydraulic gradient is equal to the surface slope. This assumption, along with others used by the model, are invalid for arid, infiltration-excess dominated basins. In contrast, the popular curve number method (Boughton, 1989) developed from small agricultural watersheds in the U.S., completely ignores the impacts of groundwater on runoff production. Due to structural inadequacies, conceptual models need to be calibrated with observations in each region to obtain effective parameter values, often at the cost of distorting model dynamics and realism. Even more recent physically-based models, e.g., Maxwell and Miller (2005); Shen (2013); Christian Refsgaard et al. (2010), could not be exempt from the needs for calibration.

Process-based models (PBMs) are also a primary tool for testing hypotheses about system functioning. For example, using an integrated surface-subsurface hydrologic model, Shen et al. (2016) examined how the representation of channels and channel bathymetry could impact simulated catchment dynamics via impacting groundwater levels. In this approach, hydrologists first form a hypothesis, then implement the representations for the hypotheses, and test them via the model. Although PBMs are highly useful in this regard, there are nonetheless philosophical issues with respect to non-uniqueness and subjectivity. To give a concrete exam-
ple, consider a classical problem of rainfall-runoff modeling. Suppose a hydrologist found that hydrologic responses in several nearby basins are different. Some basins produce flashier peaks while others have smaller peaks in summer, large seasonal fluctuation and large peak streamflows only in winter. Taking a modeling approach, the hydrologist might invoke a hydrologic model. However, the model results may not adequately describe the observed heterogeneity in the rainfall-runoff response. The hydrologist might hypothesize that the different behaviors are due to heterogeneity in soil texture, which is not well represented in the model. The hydrologist may add in processes that represent soil spatial heterogeneity, such as modified soil pedo-transfer functions that can differentiate between the soil types in different regions. Perhaps with some parameter adjustment, this model can provide streamflow predictions that are qualitatively similar to the observations. This procedure then increases the hydrologists confidence that the heterogeneity in soil hydraulic parameters is responsible for their different hydrologic responses. However, this improvement is not conclusive due to process equifinality: there can be alternative processes that can also result in similar outcomes, e.g., the influence of soil thickness, Karsted geology, terrain or drainage density. The identification of potential improvement might be dependent on the hydrologists intuition or preconceptions, which are nonetheless important but potentially biased. Furthermore, incorporating all the physics into the model may prove technically challenging or too time-consuming.

Drowned in the complexity and heterogeneity of the hydrologic systems and a variety of models, hydrologists have started to take top-down approaches based on synthesis of large-scale data. They have examined how to classify basins based on their hydrologic signatures (Sivapalan et al., 2003) which were metrics extracted from streamflow data (Sawicz et al., 2014; Berghuijs et al., 2014). Classifications in this case are not only plain catalogues compiled to reduce complexity, but theories about the basis of natural order (Gould, 1990), ways to organize knowledge. Nevertheless, there are critical limitations. First, it is substantially challenging to measure water storage in a basin, which prevents us from linking fluxes to basin-scale water storage. Second, if taken a modeling approach as suggested earlier, hydrologists would be subject to all explicit or implicit assumptions associated with the given model.

Recently there have been several critical advances that start to support a more
holistic (or "panoramic") view of hydrologic dynamics. With progress in remote sensing, large-scale measurements of hydrologic states such as water storage and soil moisture are now available. Measurements of water storage is provided by the Gravity Recovery and Climate Experiment (GRACE) twin satellites (Wahr et al., 2006). GRACE measures changes in Earths gravity field, through which it records the changes in total water storage on land, termed the terrestrial water storage anomalies (TWSA) (Famiglietti and Rodell, 2013). TWSA peaks can help estimate maximum storage capacity (Reager and Famiglietti, 2009), watershed storage trends [Thomas et al., 2016], help estimate flood potential (Reager et al., 2014), and help constrain or validate hydrologic models (Lo et al., 2010; Li et al., 2012; Niu et al., 2014). Related to water storage, there are also improved measurements of soil moisture over large scales via satellites such as SMOS (Kerr et al., 2010) and SMAP (Entekhabi et al., 2010). The accuracy of these satellites have progressed significantly from earlier satellites. It is further anticipated that various forms of new remote sensing data, e.g., from CubeSat (Heidt et al., 2000) to Unmanned Aerial Vehicles, will become more accessible in the future, providing critical support for data-driven analysis of hydrologic dynamics.

On a different front, significant advances in big data machine learning (BDML) techniques have been made in the past decade. Data-driven method extract relations from analysis of concurrent inputs and outputs (Solomatine and Ostfeld, 2008). Compared to the classical hypothesis-driven avenue, the data-driven approach allows us to more efficiently and more objectively explore a larger set of hypotheses. Although it cannot be said that the BDML algorithms present no human bias (because inputs are human-defined and some hyper-parameters are empirically adjusted), the larger set of hypotheses presented will at least greatly reduce that risk. For the rainfall-runoff example discussed above, hydrologists could start with physiographic data for many basins in this region, including terrain, soil type, soil thickness, etc. Then one can use data mining to more objectively highlight key factors that are different between the two basins. Data-driven analysis might present multiple plausible explanations, which the researcher can delve into and examine in greater details.

There is a long history of using BDML in hydrology (Babovic, 2005). Earlier-generation BDML methods, including Artificial Neural Network (ANN) (Aires et al., 2014; Hsu et al., 2002), Support Vector Machine (Xing et al., 2016), Classi-
fication and Regression Tree (CART) (Sawicz et al., 2014) have been applied extensively in hydrology with moderate success. For stream-flow forecasting, Yaseen et al. (2015) reviewed AI-based models in past 15 years and numerous studies are listed. In addition to predicting streamflow or other system responses, BDML has also been used to study many interesting flood damage (Merz et al., 2013), soil map disaggregation (Nauman and Thompson, 2014), drought monitoring (Tadesse et al., 2004), hydrology-fish community linkages (Yang et al., 2008), topographic controls on vegetation (White et al., 2005), etc. However, many of these studies were done only for a study region or site where data is available for training (Yaseen et al., 2015). When the method is applied to regions outside of training set, they tend to degrade heavily in performance, which is commonly known as 'overfitting' in BDML. Earlier generation of BDML methods are lack of skill to learn cross-region or large scale patterns of hydrologic system.

Recently, there have been rapid advances in a breed of BDML called deep learning (DL). Breakthroughs in DL arrived at 2006 with three milestone papers (Hinton et al., 2006; LeCun et al., 2006; Bengio et al., 2007). The main element that separates a deep network from a non-deep network is an unsupervised learning layer in front of supervise learning network. This unsupervised unit will automatically generate hidden features from data and enhance dataset volume. For example, Hinton et al. (2006) pre-trained a Restricted Boltzmann Machines (RBM) before transfer them into a Deep Belief Networks (DBN), by which the error from DBN was reduced faster. Recurrent neural network (RNN) is a widely used DL network underlies sequence learning applications such as handwriting recognition (Graves et al., 2013) and speech recognition (Beaufays, 2015). Vanilla RNN was discovered to have difficulties learning long-term dependencies refered as vanishing gradient problem. Hochreiter et al. (1997) proposed a Long-short-term memory (LSTM) which added memory cells and gate unit to RNN. Memory cell will keep long time memory while the gate controls the flow that current state enters the memory cell. LSTM is proved to be effective in dealing with vanishing gradient issue and its efficiency is further improved by recently developed variations including Gated Recurrent Units (GRU) (Cho et al., 2014) and fast LSTM (Beaufays et al., 2014). Besides advances in DL, scientists also tried to improve the structure of neural network to deal with their deficiencies, which is well-known as 'overfitting'. Dropout (Srivastava et al., 2014) which randomly disable units in a neural network while
training can reduce overfitting by preventing complex co-adaptations on training data. Glorot et al. (2011) suggested a domain adaptation strategy which trained independent network on different domains and avoided the ‘overfitting’ problem. In addition to above-mentioned improvements in software, Hardware advances such as Graphical Processing Units (GPUs) also play an major role in BDML advances. However, in contrast to the long history of using BDML methods in hydrologic data mining, and also in contrast to the rapid and significant achievement of DL in other disciplines, there are only a limit set of application of DL in water science.

1.2 Dissertation Organization

The basic premise of this dissertation is that the combination of big data machine learning and large-scale remotely-sensed hydrologic data, from diverse physiographic regions will provide a holistic, unbiased perspective of the hydrologic functioning of water storage including soil moisture. The advances in deep learning should also lend to stronger predictive tools for these states that can help us better project future changes. The research covered in this dissertation addressed the following questions.

- What is the relationship between basin terrestrial water storage and basin evapotranspiration?
- What variety of relationships exist between basin terrestrial storage and streamflow, and, from a data mining perspective, what are the factors that control those relationships?
- Can data-driven deep learning improve our predictive capability of soil moisture dynamics, as judged by SMAP data, and help extend SMAP data to spatio-temporally seamless coverage?
- Can SMAP data, with the prolongation provided by deep learning, present added value beyond land surface models? Does combining DL with LSMs help improve the long-term soil moisture predictions, as verified by in-situ data?
In this thesis, BDML methods are explored to assist perusing above questions. Four studies are presented focus on above four questions, as chapter 2, 3, 4, and 5 correspondingly.

In Chapter 2 the departure between the Budyko hypothesis and real-world basins behaviors are interpreted and estimated by signatures extracted from GRACE data and other physical factors. The main hypothesis is that the Budyko formula describes the evapotranspiration of a standard reference basin, and the actual evapotranspiration changes smoothly surrounding it as a function of the hydrologic signatures.

In Chapter 3, the storage-streamflow correlation spectrum (SSCS) is proposed to examine macroscopic gradients in the relationships between basin terrestrial storage and streamflow. The main hypothesis is that SSCS patterns vary substantially over space as the underlying relationships between storage and flow are modulated by climatic, land surface, and geologic conditions. By analyzing controlling factors of SSCS patterns, insights about basin hydrologic behaviors can be inspired, corroborated or rejected.

In Chapter 4, a time-dependent deep learning method, Long-short Term Memory (LSTM), is employed to simulate and prolongate SMAP data to spatial-temporal seamless coverage of CONUS, using climatic forcings and model simulations as input. The primary hypothesis is that with two years of SMAP data, LSTM can learn patterns in soil moisture dynamics and LSM errors, and by utilizing them, extend SMAP data over long time spans.

In Chapter 5 a long-term soil moisture product is produced by extending SMAP product using LSTM network. The LSTM-extended SMAP product is of similar performance to SMAP data, and a simple average to a land surface model (LSM) simulation frequently outperforms both SMAP itself and the LSM alone. The central hypothesis is that LSTM-prolonged SMAP data presents added value beyond LSMs and combining it with LSMs will help improve the long-term soil moisture predictions, as verified by in-situ data.
Bibliography


Chapter 2 | Improving Budyko curve-based estimates of long-term water partitioning using hydrologic signatures from GRACE


2.1 Abstract

The Budyko hypothesis provides a first-order estimate of water partitioning into runoff (Q) and evapotranspiration (E). Observations, however, often show significant departures from the Budyko curve; moreover, past improvements to Budyko curve tend to lose predictive power when migrated between regions or to small scales. Here, to estimate departures from the Budyko curve, we use hydrologic signatures extracted from Gravity Recovery And Climate Experiment (GRACE) terrestrial water storage anomalies. The signatures include GRACE amplitude as a fraction of precipitation (A/P), inter-annual variability, and 1-month-lag autocorrelation. We created a group of linear models embodying two alternate hypotheses that departures can be predicted by (a) Taylor series expansion based
on deviation of physical characteristics (seasonality, snow fraction and vegetation index) from reference conditions; and (b) surrogate indicators co-varying with E, e.g., A/P. These models are fitted using a mesoscale USA dataset (HUC4) and then evaluated using world datasets and USA basins <1105 km². The model with A/P could reduce error by 50% compared to Budyko itself. We found that seasonality and fraction of precipitation as snow account for a major portion of the predictive power of A/P, while the remainder is attributed to unexplained basin characteristics. When migrated to a global dataset, type (b) models performed better than type (a). This contrast in transferability is argued to be due to dataset limitations and catchment co-evolution. The GRACE-base correction performs well for USA basins >1000 km² and, according to comparison with other global datasets, is suitable for data fusion purposes, with GRACE error as estimates of uncertainty.

2.2 Introduction

The Budyko hypothesis (Budyko, 1961; Arora, 2002; Gerrits et al., 2009; Wang and Tang, 2014) describes the long term partitioning of precipitation (P) between evapotranspiration (E) and runoff (Q), as a function of the ratio between potential evapotranspiration (Ep) and P, also called the aridity index (Ep/P), i.e.,

\[
\frac{E}{P} = f\left(\frac{E_p}{P}\right),
\]

where \( E/P \) is termed the evaporation ratio and \( f \) stands for the Budyko curve, for which many formulations exist, e.g., the Turk-Pike equation (Pike, 1964; Yang et al., 2008), modified by Chen et al. (2013) with the addition of an abscissa-intercept term:

\[
\frac{E_p}{P} = \left[1 + \left(\frac{E_p}{P} - \varphi\right)^{-2}\right]^{-1/2},
\]

where, \( \varphi \) is the intercept added by Chen et al. (2013). Recently, the Budyko curve has found wide applications as a reference condition (Istanbulluoglu et al., 2012; Berghuijs et al., 2014a; Carmona et al., 2014) or providing a framework to understanding the hydrologic controls (Gentine et al., 2012) under climate change (Berghuijs et al., 2014a; Zhang et al., 2016; Yang et al., 2014). On a theoretical level, the Budyko curve initiated an interesting hypothesis that various catchment
characteristics co-evolved with climate to manifest such a simple water partitioning pattern (Troch et al., 2013; Li and Sivapalan, 2014). On a practical level, the Budyko curve provides an independent first-order estimate of $E$ for predictions in ungauged basins (PUB) (Hrachowitz et al., 2013), without the use of any hydrologic models, which is useful for the evaluation of land surface hydrologic models, e.g., Xia et al. (2012); Shen et al. (2013, 2014, 2016); Clark et al. (2015); Fatichi et al. (2016).

However, recent studies have focused on often-observed noticeable deviations from the traditional theoretical Budyko curve, i.e., there can often be a significant departure:

$$\frac{E_p}{P} = f\left(\frac{E_p}{P}\right) + \delta,$$

where $\delta = \frac{P - Q}{P} - f\left(\frac{E_p}{P}\right)$ is the departure of actual evaporation ratio ($E/P$) from the Budyko-predicted value (for long-term water balance, $E$ can be well approximated by $P - Q$). First, attention was paid to climatic pattern, in particular, the phase difference between $E$ and precipitation in an arid region (aridity index > 1) (Chen et al., 2013; Berghuijs et al., 2014b). If precipitation and $E_p$ peaks are “in-sync” throughout a year, nearly all precipitation will become $E$ and $\frac{E}{P}$ will be close to 1; in contrast, if $P$ concentrates in the winter when there is little $E_p$, actual $E$ will be significantly less than that of the uniform case. Besides climatic factors, researchers also examined influence of physical characteristics, especially vegetation control. Li et al. (2013) used the normalized difference vegetation index ($NDVI$) to parameterize Fu’s version of the Budyko equation (Fu, 1981; Zhang et al., 2001), and found that vegetation control is more apparent for larger basins (>$300,000$ km$^2$) but diminishes for basins smaller than $50,000$ km$^2$. Xu et al. (2013) parameterized Fu’s equation at different scales with $NDVI$, topography, latitude, longitude and elevation. Their calibrated coefficients are also different for basins of different sizes. The performance degradation and the required change of coefficients for small scales found in the above studies need to be better understood. For prediction in ungauged basins, special attention is needed for the generality of the method to avoid overfitting. A generalized formula that is portable across scales and regions can also help advance process-level understanding of water partitioning.
Although remote sensing methods have made large strides recently, there is not yet direct measurement of . Satellite-based products for , e.g., MOD16A2 (Mu et al., 2011) rely on assumptions and empirical formulations. The GRACE mission (Tapley et al., 2004; Wahr, 2004) records terrestrial water storage anomalies (TWSA, storage deviation from the long-term mean) for the world and has been shown to be useful for a variety of applications include monitoring groundwater resources (Famiglietti et al., 2011; Scanlon et al., 2012; Döll et al., 2014; Huang et al., 2015), model calibration and testing (Lo et al., 2010; Niu et al., 2014) and flood forecasting (Reager et al., 2014) and drought monitoring (Long et al., 2014a). Since storage is a competing process of runoff and , we expect TWSA to contain signals relevant to runoff and . For example, the amplitude of the TWSA (, average peak height from the mean) as fraction of is an indicator of the relative strengths of storage and release of the system (climate and catchment). Relevant to , the use of GRACE is ordinarily in a data assimilation/model calibration setting, e.g., Long et al. (2014b). No effort, to the authors’ best knowledge, examines how GRACE TWSA is related to in a Budyko framework.

Hydrologic "signatures" (Vogel and Sankarasubramanian, 2003; Gupta et al., 2008; Yilmaz et al., 2008) are statistics variables extracted from hydrologic time series to highlight certain distinguishing behaviors of the hydrologic systems. Streamflow derived hydrologic signatures have been used in model calibration (Yilmaz et al., 2008), parameter regionalization (Yadav et al., 2007) and catchments classification (Sawicz et al., 2014). In this paper, we attempt to answer the following questions: Does there exist a relationship between basin and hydrologic signatures, such as GRACE as a fraction of , which are useful for improving estimates of water partitioning, and, if so, what factors contribute to the relationship? Is such a relationship general enough to be portable across different regions? At what scale is the equation valid? In the following, we first describe data sources, processing procedures, signatures and indicators computed, and the Analysis of Covariance (ANCOVA) used to partition the predictive power of to different factors. Then we introduce the linear models for the departure term and their different underlying hypotheses. After that, we show the performance of the models across scales and regions and its control factors through variance partitioning. Finally, a new long-term global average dataset with error estimates is introduced. This dataset is a new "hydrologic-model-free", independent validation.
datasets useful in data fusion.

2.3 Methods

2.3.1 Data sources and processing procedures

We employed three sets of basins (Figures A.1 in the Appendix A) in order to comprehensively examine model portability and scale-dependence issues: (1) 179 Hydrologic Cataloging Unit 4-digit (HUC4) basins, which seamlessly cover the conterminous USA; (2) 605 basins from the global runoff data center (GRDC) that are within $10^4$ to $10^5$ km$^2$ and (3) 4627 US. Geological Survey (USGS) gauged basins from the GAGES-II dataset [USGS, 2011] which has been used to analyze climate change imprint on alluvial rivers [Slater and Singer, 2013] and hydraulic geometry [Shen et al., 2016]. We screened the data for temporal data coverage (sites with less than 90% for the period 2002-Oct-01 to 2012-Dec-31 are removed) and spatial coverage (sites with area bigger than the HUC4 they are located in are removed, because they are downstream gages in major rivers).

The three sets of basins have different forcing data sources. For the HUC4 basins, hourly precipitation, temperature, radiation, wind speeds and humidity from the North American Land Data Assimilation System (NLDAS) (Xia et al., 2012) were aggregated to the basins for the calculation of $E_p$ using the Shuttleworth equation (Section 2.3.3). Monthly runoff was obtained from USGS website 1.

This version of HUC4 runoff was computed by aggregating flow from data-sufficient gages located within a HUC4. For the USGS GAGES-II basins, NLDAS climate forcing was distributed into the basin boundaries, similar to the HUC4 dataset. Daily discharge from USGS websites were downloaded from USGS website 2 and aggregated to analysis time periods/scales. We compared HUC4 data with intersecting MOPEX datasets (Duan et al., 2006) and the corresponding $P$ and $E_p$ are similar (Appendix A Figure A.2).

For the GRDC basins, long-term annual average discharge was obtained from the GRDC dataset (3). Precipitation was derived from two datasets. For tropical and mid-latitude regions (between latitudes $-50^\circ$ and $50^\circ$), we employed

the precipitation product from the Tropical Rainfall Measuring Mission (TRMM) 
(Huffman et al., 1997) (3B42V7 derived). For high latitude basins, we used pre-
cipitation forcing data from the Global Land Data Assimilation System (GLDAS) 
version 2 (Rodell et al., 2003). The version 2 GLDAS product is only available 
until the end of 2010. Other climatic inputs are extracted from GLDAS for all 
basins.

Following Li et al. (2013), we obtained the 10 km resolution NDVI from Global 
Inventory Modeling and Mapping Studies (GIMMS) (Buermann et al., 2002). We 
used the average NDVI for the period 2002-2013 in our analysis. For E com-
parison, we used the PT-JPL product described in Fisher et al. (2008), hereafter 
termed $E^{PJ}$. This approach utilizes ecophysiological constraint functions to down-
scale $E_p$ to actual $E$ using remotely sensed observations of land and atmosphere 
properties. The algorithm and product have been widely used and independently 
validated extensively throughout the scientific literature, showing top performance 
across multiple inter-comparisons (e.g., Vinukollu et al. (2011); Chen et al. (2014); 
McCabe et al. (2016); Miralles et al. (2016)). In addition, we also compared our 
$E$ estimate with simulated $E$ from GLDAS version 2 with the NOAH land surface 
model (Ek, 2003).

Currently, three GRACE solutions [JPL, 2014] are provided at monthly time 
intervals and 1 degree spatial resolution for the world. GRACE TWSA mass 
grids level 3 version 5.0 data, processed using University of Texas Center for Space 
Research (CSR) algorithm was downloaded from GRACE Tellus website (Landerer 
and Swenson, 2012a). The GRACE product uses a destriping filter and a 300 km 
wide Gaussian filter as well to minimize North-South stripes in the monthly maps. 
The scaling factor based on land surface models, proposed by Swenson and Wahr 
(2006); Landerer and Swenson (2012b), was applied to the original gridded data to 
restore signal losses due to surface mass variations at small spatial scales tend to 
be attenuated by the low-pass filtering of GRACE spherical harmonics. GRACE 
TWSA data and hydrologic signatures are averaged by area to the HUC4 and 
GRDC basins. For the USGS basins, since some of them are too small, they are 
assigned the GRACE data from the HUC4 in which they are located.
2.3.2 Hydrologic signatures extracted from GRACE

We extracted three hydrologic signatures from GRACE monthly time series (2002/10 and 2014/09), based on reasoning about the hydrologic systems and statistical significance: (i) the average annual maximum TWSA amplitude ($A$) as a fraction of precipitation ($A/P$) (Figure 2.1a). This signature is chosen because the fraction of precipitation stored reflects the competition between storage, runoff and ET. Water stored through infiltration or snowpack accumulation is more likely to be released as runoff. Therefore, we anticipate that higher $A/P$ is correlated with lower $E/P$. More explanations are provided in Section 2.3.4; (ii) the ratio of TWSA variance explained by inter-annual variability and intra-annual variability ($\gamma$) (Figure 2.1b). $\gamma$ is chosen because basins with higher inter-annual variability in storage (normalized to intra-annual variability) tend to have smaller long-term average $E/P$ compared more evenly-distributed ones: in years with extraordinary precipitation, there is a higher chance of water partitioned as storage and runoff as opposed to $E$. Inter-annual variability was also found to be important for long-term water balance (Sivapalan et al., 2011; Li, 2014); and (iii) $acf_D$, the one-month-lag, piecewise-detrended autocorrelation function of GRACE TWSA based on $D$-month-long segments. $acf_D$ describes the smoothness of the monthly TWSA signal and reflects the seasonal distribution pattern of storage. As with $\gamma$, higher concentration of storage in a few months will likely result in smaller $E/P$ compared to average conditions. Apart from climate pattern, higher memory in storage may indicate the ability of the system to hold water from runoff, hypothetically leading to higher $E/P$ compared to low-memory systems.
Figure 2.1. (Caption on next page.)
Figure 2.1. Hydrologic signatures and climatic index computed for the world using data from 2002 Oct to 2010 September: (a) average annual GRACE TWSA amplitude as a fraction of precipitation \(A/P\); (b) the inter-annual GRACE signal variability ratio \(\gamma\) (dimensionless ratio between between-year and within-year TWSA variability); and (c) the precipitation-temperature seasonality index following Woods (2009), which is -1 for completely out of phase, 1 for completely in phase, and 0 for uncorrelated.

\(A\) is estimated by first applying a Fourier-transform to the GRACE time series, and then take the maximum amplitude for frequencies between 0.8 to 1.2 cycles/year. Previous research showed that most mega-basins in the world have the highest peaks at annual periods (Reager and Famiglietti, 2013). Although in that work the Yule-Walker autoregressive method (Thomson and Emery, 2014) was used, we are only interested in the annual amplitude and therefore a Fourier transform is sufficient. The band window of 0.8 to 1.2 is for numerical stability of the method and slight changes of the window did not change our results. We can see hotspots of \(A/P\) in large river areas, e.g., Mississippi, lower Nile and Amazon (Figure 2.1a). As described previously, in these regions \(A/P\) is not reflective of land surface runoff and storage, but rather, seasonal river stage fluctuations, so they are removed from model fitting. When working with HUC4 dataset, since the Mississippi River induce large seasonal mass changes and leakage errors which are unrelated to nearby land surface runoff/water storage dynamics, the basins that contain the Mississippi River are removed from analysis. In the USA (a bigger map of \(A/P\) for HUC4 basins is presented in Figure A.1 in the Appendix A), the high \(A/P\) regions are in the west, where precipitation is winter-dominant (Berghuijs et al., 2014b) and has a big phase difference with temperature.

In addition, we propose an inter-annual variability index, \(\gamma\) (Figure 2.1b), which quantifies the ratio of between-year variability and within-year variability in TWSA. If \(\sigma_{W,i}(S)\) is the standard deviation of TWSA in i-th year (based on monthly data), the average within-year standard deviation is \(\overline{\sigma_W} = \frac{1}{n} \sum_{i=1}^{n_y} \sigma_{W,i}(TWSA)\), where \(n_y\) is the number of years. If the mean of each year’s TWSA is \(TWSA_i\), then \(\sigma_b\) is the standard deviation of \(\overline{TWSA}_i\). We define \(\gamma\) as

\[
\gamma = \frac{\sigma_b}{\overline{\sigma_W}}. \tag{2.4}
\]

We calculated the indices using GRACE data from 2002 October to 2014
September. Some data gaps in GRACE has been filled using spline interpolation. We evaluated using data from 2002-2012 which has less gaps. This was not found to have significant influence on our results.

\[ \text{acf}_D \] is the cross-correlation of the GRACE data with itself with 1-month lag, after applying piece-wise detrending in every \( D \) months. Higher \( \text{acf}_D \) indicates higher similarity between data points and their next-month neighbors and thus higher smoothness. Lower \( \text{acf}_D \) curves have more abrupt changes. We first divided the time series into multiple segments, each consisting of \( D \) months of data (in our analysis we used \( D = 48 \)). Without piecewise detrending, non-stationary trends may interfere with the extraction of autocorrelation function. It well-known that different time periods can exhibit different trends in the GRACE data (e.g., see Famiglietti et al. (2011); Voss et al. (2013)). Here to simplify analysis, we only examined 1-lag \( \text{acf} \) with 48 month as the segment length for detrending. We tested 2-month-lag and 2-month-lag (data not shown here), which did not provide much additional predictive power.

GRACE data are influenced by different sources of errors (Wahr et al., 2006). Signal degradation due to measurement noises are called measurement errors (Swenson and Wahr, 2006) and the contamination of signal by nearby region (due to spectral truncation and filtering) is termed leakage errors (Figure A.3 in Appendix A). We employed measurement and leakage estimated using the approach in Landerer and Swenson (2012b) and provided at http://grace.jpl.nasa.gov to calculate the combined error as the quadratic mean of the two errors. The errors are used to determine regions where GRACE-based signatures have low reliability, which are over-written by interpolation. After our initial testing, we found that where the combined error is large (>74 mm, which is two times the global mean combined error), the hydrologic signatures from GRACE is no longer usable. For world datasets (GRDC and world-gridded products), regions with errors larger than 74 mm obtain their \( A/P \) via interpolation from other regions. In fact, few GRDC basins fall into this category. \( A/P \) for cells with major rivers such as Amazon, lower Nile and Mississippi are automatically interpolated from neighboring regions.
2.3.3 Climatic indices

$E_p$, which was calculated using the Shuttleworth equation (Maidment and Others, 1993; Zhou et al., 2006; Li et al., 2013):

$$
\lambda E_p = \frac{\Delta \cdot R_n}{\Delta + \gamma_p} \times \left( \frac{6.43\gamma_p}{\Delta + \gamma_p} \times (1 + 0.5361\mu) \times \frac{e_s - e_a}{C_p} \right), \quad (2.5)
$$

where $\Delta$ is rate of change of saturation specific humidity (kPa°C$^{-1}$), $\gamma_p$ is Psychrometric constant (kPa°C$^{-1}$), $R_n$ is net radiation (MJ m$^{-2}$ d$^{-1}$), $\mu$ is wind speed (m s$^{-1}$), $C_p$ is specific heat of evaporation (MJ kg$^{-1}$°C$^{-1}$), $e_s$ is saturation vapor pressure (kPa), $e_a$ is near-surface air vapor pressure (kPa) and the unit of $E_p$ from this equation is mm d$^{-1}$. The daily $E_p$ was aggregated to mm yr$^{-1}$ for the calculation of the aridity index.

We calculated the fraction of $P$ as snow ($S=P$), which is known to be important for runoff (Berghuijs et al., 2014a). We also calculated a seasonality index (Figure 2.1c) that quantifies the phase difference between precipitation and temperature following previous work (Woods, 2009; Berghuijs et al., 2014b), who derived it based on sine curve assumptions (Miralles et al., 2016; Potter et al., 2005):

$$
P(t) = \overline{P} \times [1 + \Delta_p \sin(2\pi(t - s_p)/\tau)],
$$

$$
T(t) = \overline{T} + \Delta_t \sin(2\pi(t - s_t)/\tau),
$$

$$
\xi = \Delta_p \times \text{sign}(\Delta_t) \times \cos(2\pi(s_p - s_t)/t), \quad (2.6)
$$

where $t$ is the time (months), and are a phase shift for precipitation and temperature, respectively (months), $\tau$ is the duration of the seasonal cycle (12 months), and $\Delta_p$ and $\Delta_t$ are dimensionless seasonal amplitudes for precipitation and temperature, respectively.

In theory, the seasonality index varies within $[-1, 1]$ and -1 is for completely in-phase $E_p$ and $P$, 0 for uniform precipitation with seasonal temperature and 1 for completely opposite-phase $E_p$ and $P$. However, in practice because some basins have very dry seasons, the fitted sine curves can have >1 amplitudes. Note that while $\xi$ is negatively correlated with the $A/P$ map (e.g., the western USA has high $A/P$ and most negative $\xi$, the correlation is not perfect (e.g., the northern central lowland has relatively large $A/P$ and also relatively large $\xi$).
2.3.4 The general departure model and its rationale

Our linear formula for the departure term is

\[ \delta^*(x) = \delta(x) + \varepsilon = a^T x, \]  

(2.7)

where \( \delta^* \) is an approximation to \( \delta \), \( x \) is a vector of independent physical or surrogate factors (except 1 is the first element for the intercept term), or a mixture of both, \( a \) are the corresponding linear coefficients and \( \varepsilon \) is the error term.

Although Equation 2.7 seems a simple linear regression model, it in fact embodies different hypotheses with different predictors. Here we make the distinction between physical factors (\( \xi, acf_{48}, S/P \) and \( NDVI \)), which are independent variables that vary in space, and surrogate factors (\( A/P, \gamma \) and \( acf_{48} \)), which are dependent variables potentially influenced by the former. When Equation 2.7 involves only physical factors, it is an approximation to the relationships between potentially causal factors and outcome (departure from Budyko). Further, it can be interpreted as a first-order Taylor Series expansion of the perturbation from the reference state (Budyko curve) due to changes in physical factors (Appendix A Section A.1). However, when a surrogate factor is involved in Equation 2.7, there is no causal or controlling relationship: \( E/P \) does not change because of changes in \( A/P \). Thus a Taylor Series expansion interpretation is inappropriate. Rather, the surrogate factors and \( E/P \) co-vary due to changes in some common factors, and Equation 2.7 captures their co-variation. When we test the different models we are also testing different hypotheses, i.e., whether the departure from Budyko is better modeled by the difference in a list of physical factors or co-varying surrogate indices. More mathematical discussion of the different hypotheses is provided in Appendix A Section A.1.

There are multiple reasons behind choosing \( A/P \). The most important is the following observation: if the climate is such that \( P \) and \( E_p \) reach peaks and lows at approximately the same time, e.g., in the US central high plains, they are "in-phase". For "in-phase" and water-limited basins, \( P \) immediately evaporates, leaving very little water for storage and runoff and thus high \( E/P \), but at the same time the \( A/P \) ratio is also small because little water can be stored. On the other hand, if \( P \) and \( E_p \) are "out-of-phase", as in the case of US southwest, \( P \) during winter times "evades" the peak of \( E_p \) and has the chance to be stored. The
storage of water in winter times leads to higher $A/P$, and at the same time a low $E/P$. Therefore, there should be negative co-variation between the $\delta$ and $A/P$.

There are also many other potential ‘negative-$A/P$-$E/P$-correlation-inducing’ (NAECI) physical factors. Whether a factor is NAECI depends on how, when it is varied, it shifts the competition between $E$, $Q$ and $S$. NAECI factors in general should favor infiltration against evaporation, for example, high vertical soil hydraulic conductivity in deep-water-table regions: more water infiltrate below plant-accessible zone, increasing groundwater storage while reducing $E$. Conversely, if a region has low vertical conductivity, it inhibits groundwater storage while enhancing $E$, still contributing to a negative $A/P$-$E/P$ correlation. For another example, an area with consistently light rainfall patterns can promote infiltration and inhibit $E$. Some other processes may in fact cause positive $A/P$ and $E/P$ correlations. For example, high terrain slope potentially boosts runoff while decreasing both storage and evaporation, thus causing a positive correlation between the two. A factor’s influence on $A/P - E/P$ correlation can also be complicated and climate-dependent. For example, unusually high vegetation cover is a NAECI factor for an arid catchment as it boosts $E$ and reduces infiltration (and thus storage); however, for low-aridity, energy-limited basins, vegetation interception primarily modulates runoff and storage, with little impact on $E$.

2.3.5 Partitioning of variance

We used analysis of covariance (ANCOVA) to attribute the variability of to various physical factors or surrogate indices. Compared to the Analysis of Variance (ANOVA), which is designed for categorical data, ANCOVA builds linear models for continuous variables to attribute variance to predictors (Keppel et al., 1992). When the data are unbalanced (uneven sampling of data in different parts of the viable ranges of factors), as is the case with our data, there are three different ways of attributing the variance (sum of squares, SS): type I (sequential), II (main effects excluded) and III (main and interaction effect excluded) (Fox, 2008). With type I, SS are attributed to the factors in the order of what is supplied, and only residuals are attributed to the next factors. Thus, the earlier factors will claim part of the SS of subsequent correlated factors. For example, if two factors are perfectly correlated, the first factor in the sequence will be attributed all the SS.
they can explain and the second will be attributed none. With type II SS, the main
effects of other factors are excluded so the only the part of SS that can be solely
attributed to one factor is reported. As a result, type II SS is not influenced by
the order of factors. Type III is similar to type II, but further excluded interaction
terms. Here we examined type I and II to see how different variables overlap. In
our ANCOVA we examined a total of six factors: A/P, ξ, γ, S/P, NDVI and
acf48, and this order is called order #1 (O1). In the second order (O2), A/P is
placed as the last argument: ξ, γ, acf48, S/P, NDVI, and A/P.

2.3.6 Multi-scale, multi-dataset validation of alternative models

Since an important goal of the proposed method is to estimate $E$ for ungauged
basins, understanding how the formula performs when coefficients are migrated
between regions and across scales is of great importance. We tested the perfor-
mance of a total of 13 linear models (Table 1), each with a different combination
of the above-listed predictors, when their coefficients are estimated from the HUC4
dataset and then migrated to the GRDC and USGS basins. The root mean squared
error (RMSE) of $E/P$ was used as a measure of error of the models:

$$RMSE = \left\{ \frac{\sum_{j=1}^{n_b}[f(E_p/P) + \delta^* - (P - Q)/P]^2}{n_b} \right\}^{1/2},$$ (2.8)

where $n_b$ is the number of basins and $\delta^*$ is the model-predicted departure term.

When validating the models using the USGS dataset, we rank the basins in de-
sceding order by the areas. Based on this ranking the basins were evenly divided
into 28 area classes. A separate RMSE was calculated for each class. Each plotted
point is the average of two neighboring classes.

2.4 Results and Discussion

2.4.1 The GRACE-assisted departure model

From the HUC4 dataset, we note many basins, especially those with larger aridity
index, deviate significantly from the Budyko curve (Figure 2.2a). For some arid
basins, the ratio of Budyko-predicted and actual water partitioning, $\frac{P-Q}{P} : f\left(\frac{E_a}{P}\right)$,
can be as large as 50%-100%. As expected, these points are always accompanied by large $A/P$, showing strong influence from seasonality. For wetter basins, on the other hand, there may be negative or positive departures, which may be small but could potentially lead to large errors in the estimation of $E$ due to the large $P$. We can clearly see there is a negative co-variation between $\delta$ and $A/P$ (Figure 2.2c), which is exploited by the predictive formula. After we subtract the corrector term, based solely on $A/P$, the points now clusters much more closely to the Budyko curve (Figure 2.2b).
Figure 2.2. Using GRACE TWSA amplitude as a fraction of precipitation ($A/P$) to predict the departure ($\delta$) from the Budyko curve for the HUC4 dataset. $\delta = \frac{P-Q}{P} - f\left(\frac{E_p}{P}\right)$ where $Q$ is observed discharge, $E_p$ is potential evapotranspiration and $f\left(\frac{E_p}{P}\right)$ is the Budyko formula. (a) Without correction, the HUC4 basins scatter around the Budyko curve, some with significant departures; (b) After correcting using $A/P$, basins are now much more closely clustered around the Budyko curve; (c) The negative correlation between $A/P$ vs $\delta$ allows the improvement over Budyko.
2.4.2 Factors contributing to Budyko departures and $A/P-E/P$ correlation

The ANCOVA results show that, for the HUC4 dataset, Precipitation seasonality ($\xi$) and $S/P$ are important but not the sole factors contributing to $A/P$ and the departure. The sequential (type-I) sum of squares (SS) attributed to $A/P$, when factors are laid out in O1, is more than 61% of the total SS, leaving only about 3% to $\xi$ (Figure 2.2). When they are laid in O2, in which $A/P$ is placed last, both and $S/P$ explain much more variance than in O1, but there is still around 14% of SS attributed to $A/P$ that is unexplained by any other factor. Therefore, a large fraction ($\frac{61\%-14\%}{61\%} = 77\%$) of the explanatory power of $A/P$ is attributed to correlation with $\xi$ and $S/P$. This pattern suggests that $P - E_p$ phase shifts and snowpack accumulation are two major reasons causing $A/P$ in the USA. In previous sections we reason that when $P$ and $E_p$ are "out-of-phase", rain water has more opportunity to infiltrate, rather than becoming $E$, so that $E/P$ is small compared to Budyko prediction while $A/P$ is larger than average, giving rise to the negative correlation. $S/P$ has a similar effect: when larger fraction of precipitation falls as snow, snowpack accumulation produces larger $A/P$, while the snowmelt water becomes runoff or infiltration more easily compared to average conditions. In addition, there is little independent explanatory power in $\xi$ and $S/P$ that cannot be replaced by $A/P$, as reflected in type-I SS in O1. This means $A/P$ is an excellent surrogate index to represent their aggregate effects on long-term water partitioning. On another note, the inter-annual storage variability index has little correlation with $A/P$ or $\delta$ as evidenced by the small change in its attributed SS from O1 to O2. When there is large inter-annual variability in precipitation, rainfall in wet years can create much higher runoff than average years and thus causes overall negative $\delta$. 
Figure 2.3. Variance decomposition to surrogate indicators and physical factors using ANCOVA for HUC4 and GRDC datasets. The fraction of the horizontal bars occupied by a factor indicates the variance explained by this factor. Type I Sum of Squares (SS) is “sequential” so the order of factors influence the results, unlike type II, which excludes main effects of other factors. The order in O1 is $A/P$, $\xi$, $\gamma$, $S/P$, $NDVI$, and $acf_{48}$, while O2 is $\xi$, $\gamma$, $S/P$, $NDVI$, $acf_{48}$ and $A/P$. The joint symbol, $\cap$, stands for the part of variance with attributed to more than one factors. As $\xi$ and $S/P$ occupy small fractions in O1 but more noticeable in O2, we conclude that $A/P$ encompasses $\xi$ and $S/P$, while the latter two constitute a large fraction, although not all of the predictive power of $A/P$.

There are myriad processes that interact in complex ways to influence $A/P$-$E/P$ correlation and the departure from Budyko, and they are not easily described by a small number of indices. However, at the scales of basins in the US and the world, such effects seem to be muted and we observe primarily a negative $A/P$-$E/P$ correlation, suggesting their influence on $E$ is limited. Overall, $A/P$ appears to be an effective way of capturing the lumped effects of these myriad processes, while the residual effects are in the error terms.
2.4.3 Global validation and the relevance of co-evolution to model choosing

The HUC4-fitted models with more predictors and slightly better accuracy dataset decrease in performance when migrated to GRDC, although most of them are still better than the original Budyko (Figure 2.4). In the 14 models inspected (the 0-th is Budyko itself), models #1-6 have 1-6 predictors, respectively, all of which include $A/P$. Model #1 with $A/P$ as the only predictor contributes the most significant error reduction, while additional predictors have limited impact. When migrated to GRDC ($GRDC^M$), model #1 still reduces the error significantly, and this reduction is similar to that achieved by directly fitting the model to the GRDC dataset ($GRDC^F$). With each addition predictor added (models #2-6), RMSE decreases slightly for HUC4 and somewhat more notably for $GRDC^F$. However, RMSE increases with the same models for $GRDC^M$. When $A/P$ is absent (Models #10-13), RMSE increases significantly for both HUC4 and GRDC. RMSE($GRDC^M$) even exceeds the Budyko model itself. From Table 2.1), the coefficient for in GRDC-fitted model #10 is 0, suggesting seasonality plays different roles in different regions so that no consistent pattern exists on a global scale. In addition, for many models in Table 2.1), the coefficients for some other factors, e.g. $\gamma$, $\xi$ and $S/P$, could switch signs between HUC4 and GRDC. This means their influences on $E$ are region-specific. Model #13 is most interesting: with all 5 predictors other than $A/P$, this model is able to bring RMSE(HUC4) down to a level similar to model #1, but when migrated to GRDC, the error is even larger than original Budyko. We conclude that the coefficients estimated for these physical parameters are not portable across regions, consistent with previous research (Xu et al., 2013). In contrast, $A/P$ is an effective and crucial predictor whose coefficient is transferrable between regions.

Table 2.1: Tested Linear Models With Linear Weights Provided. HUC4 means the model was fitted using HUC4 data. GRDC means the model was fitted using GRDC data. Model #9 was chosen to produce a global-scale comparison with $E^{PJ}$ and $E^{GLDAS}$.

<table>
<thead>
<tr>
<th>Model #</th>
<th>Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Budyko itself</td>
</tr>
<tr>
<td>---</td>
<td>----------------</td>
</tr>
<tr>
<td>1</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>1</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>2</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>2</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>3</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>3</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>4</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>4</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>5</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>5</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>6</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>6</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>7</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>7</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>8</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>8</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>9</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>9</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>10</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>10</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>11</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>11</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>12</td>
<td>(HUC4)</td>
</tr>
<tr>
<td>12</td>
<td>(GRDC)</td>
</tr>
<tr>
<td>13</td>
<td>(HUC4)</td>
</tr>
</tbody>
</table>
When migrating HUC4-fitted models to the USGS basins, the performance gradually degrades for smaller basins (Figure 5). At the largest scale ($20 \times 10^3 km^2$), $R^2$ is 0.85 for model #1 (with $A/P$ only) while it is just 0.69 for Budyko itself. At around $4 \times 10^2 km^2$, $R^2$ decreases rapidly and for the original Budyko it drops below 0.3. In terms of RMSE for $E/P$, the gap between model #1 and Budyko curve remain similar across scales, suggesting that the deterioration is not primarily due to overfitting or using GRACE signal from the enclosing HUC4. Such decrease may be due to the scale limitation of the Budyko hypothesis itself, but may also be partly attributed to the decrease in quality of climate input data. Our input data were obtained from NLDAS with a 1/8 degree spatial resolution (approximately 100 km$^2$ at 40°N and 144 km$^2$ at 30°N), which corresponds to the point when RMSE drops below 0.3.

![Migrating parameters from HUC4 to USGS basins and the influence of scales](image)

**Figure 2.5.** The scale-dependent performance of 5 models when HUC4-fitted models are migrated to the USGS GAGES-II basins. (a) RMSE of evaporation ratio (blue line is covered by red); (b) $R^2$ between observed vs predicted $E$. The correction formula with $A/P$ are transferrable to $>1000$ km$^2$ basins while the one without $A/P$ performed only slightly better than Budyko itself. The gradual degradation in performance toward smaller scales with the GRACE-assisted models is similar to that of the Budyko itself, indicating the degradation is due to either data limitation or inherent loss of accuracy of the Budyko equation for smaller basins, rather than due to the correction formula.
Figure 2.4. Errors of the models #0 through #13 (Table 2.1) when they are fitted to the HUC4 data (blue line) and their coefficients are migrated to the GRDC dataset (GRDC\textsuperscript{M}, green solid line), compared to when the models that are directly fitted to the GRDC data (GRDC\textsuperscript{F}, green dashed line). Colors of the markers indicate the number of predictors. Note HUC4 and GRDC RMSEs are on two different y-axes. We note that as indicators are added in models 1-6, they slightly reduce errors in HUC4 but increase errors when parameters are migrated to GRDC. In addition, while the model with $A/P$ is transferrable from HUC4 to GRDC, models without $A/P$ performs poorly when migrated. (Evaluation period: 2002/10-2010/09, because GRDC is only available before 2010/09)

Adding predictors to $A/P$ does not noticeably improve performance for either GRDC or USGS. The overfitting problem is mild with models containing $A/P$, but significant for those without it. We notice that the model with $A/P + \gamma + S/P + acf_{18}$ remains the best model across all USGS basin scales, while when migrated to GRDC the overfitting is noticeable. We offer three explanations for the overfitting when migrating to GRDC: (1) catchments in the world experience more diverse climatic, geologic and human modification conditions than USA, so the coefficients learned from HUC4 for the attributes other than $A/P$ are not general enough. This is most apparent from the seasonality index, which has a coefficient of 0 in
the GRDC dataset; (2) different datasets (NLDAS vs GLDAS + TRMM) have different biases in various regions; (3) catchment co-evolution invalidates a linear correction model based on the physical factors.

The co-evolution theory may help explain contrasts between the overfitting of physical factors and the portability of $A/P$-based models. The abiotic and biotic systems (e.g., vegetation, soil, topography and landforms) co-evolve and adapt to each other and climate conditions such as $P - E_p$ phase differences (Gentine et al., 2012). Troch et al. (2013) argued that catchment characteristics co-evolve with climate to produce the manifested $E$ pattern such as the Budyko curve. Assuming this theory has validity, when climate changes, the system responds with a variety of intertwined and initial-value and path-dependent changes in different physical factors. We have limited capability in capturing changes in all these factors. It is challenging to observe or too numerous to analyze robustly, when each factor only accounts for a small part of the variability. For example, climate change may cause a change in $\gamma$, and then co-evolution induces changes in other factors such as soil and vegetation (leaf and root). While changes in leaf status can be measured by $NDVI$ and captured in our linear model, soil and root are much more difficult to measure. Linear correction models may fail to generalize due to missing factors. However, $A/P$ as a surrogate captures some of the effects of all these responses because $A/P$ co-varies with the departure. Therefore, the existence of partial co-evolution might be the reason why the surrogate model is favored over linear models with physical factors. As a side note, the co-evolution theory above leads to the surmise that a departure from Budyko indicates incomplete co-evolution, that is, if given sufficiently long time, these basins will eventually return to the Budyko curve.

2.4.4 Comparison with other evapotranspiration products

We use model #9 calibrated based on GRDC data, i.e., with $A/P$, inter-annual variability ($\gamma$) and fraction of $P$ as snow ($S/P$) as predictors, to produce a world long-term annual $E$, termed $E^C$ (Figure 2.6a). Overall, $E^C$ had slighter higher correlation coefficient than $E^{PJ}$, while both perform much better than the original Budyko estimates (Figure 2.7a). Toward the higher $E$ ranges (in the Amazon basin), $E^{PJ}$ tends to over-estimate $E$ while $E^C$ tends to under-estimate $E$. The
gridded product was compared with $E^{PJ}$ and $E^{GLDAS}$ (Figures 2.6b-c and Figures 2.7b-c). We note that in many places in the world the differences are small, and the difference between $E^C$ and the two data products are smaller than that between themselves (Figures 2.7b-c). Regions with noticeable differences include the Amazon forest, Southern Brazil, central Africa, Southeast Asia, Northern Australia and Japan. In some of these regions (central Amazon, central and central-south Africa), three datasets all differ, but $E^C$ is between $E^{GLDAS}$ and $E^{PJ}$, and it agrees with one more than the other. In some regions (Southwest Amazon, Southeast Asia, Japan, Northern Australia), $E^C$ is likely to be in error due to GRACE error contamination. We discuss these differences in the following.

In most parts of the Amazon, $E^C$ is substantially (0 - 400 mm/yr) smaller than $E^{PJ}$ but somewhat larger (0 - 250 mm/yr) than $E^{GLDAS}$, except in the southwest. Judging from the limited GRDC comparisons (Figure 2.7a toward to the high $E$ range), true $E$ in this region is likely in the middle between $E^C$ and $E^{PJ}$, and $E^{GLDAS}$ is likely the most biased potentially because of under-estimating evaporation of canopy interception. Both $E^C$ and $E^{GLDAS}$ can suffer from errors in precipitation which can be significant in this region (Zhou et al., 2012). In addition, as shown previously, using NDVI as physical predictors did not help with this issue. It is possible that the satellite-based sensing of NDVI and LAI are saturated in the Amazon, invalidating a linear correction formula. In the southwest Amazon, $E^C$ is slightly less than both, and is likely to be in error here. As we can see from the GRACE error map (Appendix A Figure A.3), the southwest Amazon have large leakage errors. In fact, it is clear from Figures 2.7b-c that regions with large combined GRACE errors are likely to be biased.

In central and central-south Africa (immediately beneath the Sahara), $E^C$ is slightly less than $E^{PJ}$ yet it is noticeably larger than $E^{GLDAS}$. The agreement between $E^C$ and $E^{PJ}$, the small GRACE error in this region and the large difference from $E^{GLDAS}$ suggest that $E^{GLDAS}$ is in error here. In contrast, in Southeast Asia and Northern Australia, $E^C$ is bigger than both $E^{PJ}$ and $E^{GLDAS}$. Even after the interpolation attempt, error is still large in this region. Given the large leakage error in this region, $E^C$ most likely has a large positive bias here. $E^{PJ}$ data is missing for Japan, but $E^C$ is also unreliable here due to large measurement error.
Figure 2.6. (a) World annual average evapotranspiration, $E^C$, for 2002-2006 based on Budyko and linear model with $A/P$, inter-annual variability index ($\gamma$) and $S/P$, after GRACE error limiting; (b) Difference between $E^C$ and $E^{PJ}$ (white blanks are due to missing data in $E^{PJ}$); and (3) Difference between $E^C$ and $E^{GLDAS}$.

2.5 Limitations and future work

As discussed above, in regions where GRACE errors are large (e.g., Southeast Asia and Japan), the correction formula might be contaminated by errors. In the future, it should be possible to merge the GRACE-assisted product with others.
Observed $P-Q$ (mm/year) vs Predicted $E$ (mm/year) for GRDC basins

Figure 2.7. (a) Comparing GRACE-corrected product ($E^C$) with $E^{PJ}$ and Budyko itself in GRDC basins for 2002/10-2006/09 ($E^{PJ}$ is available only before 2006/09). Budyko has negative biases while $E^{PJ}$ has positive biases in high $E$ regions; (b-c) Compare $E^C$ with $E^{PJ}$ and GLDAS evapotranspiration. The root-mean-squared-difference is 175 mm/yr between $E^C$ and $E^{PJ}$, 131 mm/yr between $E^C$ and $E^{GLDAS}$, and 189 mm/yr between $E^{PJ}$ and $E^{GLDAS}$. We note that red-colored points (large GRACE errors) tend to be the most distant to the 1-to-1 line, indicating in these regions $E^C$ is unreliable due to GRACE error contamination.

Using data fusion techniques that weigh GRACE-corrected data using combined GRACE errors. With advances in algorithms, we might be able to obtain improved GRACE estimates with lower errors, or new algorithms to reduce the area of regions with large errors. Because all inputs are available on a monthly timescale, it should be possible to produce annual $E$ dataset, instead of a long-term average. However, extracting meaningful amplitude for each year requires additional work, which we leave for the next stage.
2.6 Conclusions

Given the large number of $E$ products available for the world and leakage/measure-
ment errors facing GRACE, our estimate is more useful in a data-fusion setting than being used as a standalone product. However, the novelty of this work resides with the advances in understanding the relationship between storage amplitude and $E$, using Budyko as a reference condition, the mechanisms influencing storage am-
plitude, the hydrologic signatures from GRACE and the transferability of the mod-
els embodying different hypotheses. To our knowledge the present work is the first one examining the physical significance of TWSA amplitude as a fraction of precipita-
tion ($A/P$) and its controls by seasonality and snow, and the first one to link it to the departure from the Budyko curve. The surrogate indicator $A/P$ is a pow-
erful and portable predictor for the departure from the Budyko curve. Migrating models from HUC4 to GRDC and USGS datasets, the models with $A/P$ are more transferrable than those without it. We argue the overfitting with physical-factor-
based linear models may be due to partial co-evolution of basin characteristics with climate, which can be difficult to fully capture. Although GRACE has coarse spatial resolution, the methodology works well for basins above 1000 km$^2$ in the USA, which is much smaller than the GRACE footprint. The gradual degradation in performance toward smaller scales is not due to the GRACE-based correction formula, but Budyko itself and data limitations. Compared to two different global $E$ products, in many regions in the world our improved estimate ($E^C$) is either similar to both, between the two, or agree with one more than the other. The errors with $E^C$ is related to GRACE measurement and leakage errors. Southeast Asia, Southwest Amazon, Northern Australia and Japan are regions where $E^C$ most likely has large biases.

2.7 Author Contributions

K. F. collected data, implemented the algorithms, analyzed the results and draft the manuscript. C. S. provided general oversights and instruction, assisted with the method implementing, and critically revised the article. J. F. provided the comparing evapotranspiration product. J. N. helped with the GRACE data processing. All authors discussed the results and commented on the manuscript. K.
F. is the first author and has contributed the majority of work.


Fox, J. (2008), Applied regression analysis and generalized linear models.


Chapter 3  
Full-flow-regime storage streamflow correlation patterns provide insights into hydrologic functioning over the continental US

This chapter is published as: Fang, K., & Shen, C. (2017). Full-flow-regime storage-streamflow correlation patterns provide insights into hydrologic functioning over the continental US. Water Resources Research, 53(9), 80648083. https://doi.org/10.1002/2016WR020283

3.1 Abstract

Inter-annual changes in low, median and high regimes of streamflow have important implications for flood control, irrigation, and ecologic and human health. The Gravity Recovery and Climate Experiment (GRACE) satellites record global terrestrial water storage anomalies (TWSA), providing an opportunity to observe, interpret, and potentially utilize the complex relationships between storage and full-flow-regime streamflow. Here we show that utilizable storage-streamflow correlations exist throughout vastly different climates in the continental US (CONUS) across low to high flow regimes. A panoramic framework, the storage-streamflow correlation spectrum (SSCS), is proposed to examine macroscopic gradients in these relationships. SSCS helps form, corroborate or reject hypotheses about basin
hydrologic behaviors. SSCS patterns vary greatly over CONUS with climate, land surface and geologic conditions. Data mining analysis suggests that for catchments with hydrologic settings that favor storage over runoff, e.g., a large fraction of precipitation as snow, thick and highly permeable soil, SSCS values tend to be high. Based on our results, we form the hypotheses that groundwater flow dominates streamflows in Southeastern CONUS and Great Plains, while thin soils in a belt along the Appalachian Mountains impose a limit on water storage. SSCS also suggests shallow water table caused by high-bulk density soil and flat terrain induces rapid runoff in several regions. Our results highlight the importance of subsurface properties and groundwater flow in capturing flood and drought. We propose that SSCS can be used as a fundamental hydrologic signature to constrain models and to provide insights that lead us to better understand hydrologic functioning.

3.2 Introduction

The relationships between streamflow and water storage of a catchment may yield important insights about its hydrologic functioning (Spence, 2010; Berghuijs et al., 2016). Streamflow is the most accessible component of the water cycle, and it is also under the significant influence of climate change Stocker et al. (2013). The full range of streamflow variability can be described by flow regimes, ranging from high flows, the periods with the largest flow rates in a year (possibly following snowmelt or heavy storms), to low flows, the periods of low but steady flow after floodwaters recede. Flow magnitudes at different flow regimes, measured by percentiles, can be sampled using corresponding exceedance probabilities from the flow duration curve (a curve plotting flow magnitude versus daily flow exceedance probability) (Mays, 2010). Tracking changes in the full range of streamflow regimes are important for practical socio-economic and ecological purposes. The changes in high flows are of great importance for flood control and mitigation, engineering design of water infrastructure (Mays, 2010) and geomorphological studies (Castro and Jackson, 2001). The low flows are of special importance to competing needs including fish habitat and ecosystem health (Poff and Allan, 1995), human health (Nilsson and Renöfält, 2008) and power plant cooling needs (Giuliani et al., 2014). The median and mean flows, on the other hand, can be important for estimating available water for irrigation. Nevertheless, a large number of basins in the world remain
ungauged or inaccessible (Sivapalan et al., 2003). There is still no practical way of measuring streamflow changes on large scales until potentially 2020 (Pavelsky et al., 2014). Therefore, our ability to monitor and understand the changes is far from sufficient.

The full range of streamflow is linked to water storage in complex manners. High soil wetness or large snowpack are both potential precursors for saturation excess and flooding (Reager and Famiglietti, 2009). Medium range flows co-vary with storage and precipitation inputs. Baseflow, on the other hand, is contributed by groundwater and is often a function of subsurface storage, e.g., see Wittenberg (1999); Brutsaert (2008). However, these relationships are modulated by the landscape and geologic settings. For example, peak streamflow for a groundwater-dominated or saturation-excess-dominated region is preceded by a large increase in storage, whereas in an infiltration-excess-dominated basin, storage and streamflow can be disconnected. Therefore, observing streamflow-storage relationships, especially their gradients on a large spatial scale, may reveal novel patterns and how the hydrologic system functions. Further, if these relationships are well understood, they may potentially provide mutual predictive capabilities between storage and streamflow.

While the connection between storage and streamflow has long been recognized, the measurement of water storage at large spatial scales was not available until the advent of the Gravity Recovery and Climate Experiment (GRACE) twin satellites Wahr et al. (2006). GRACE measures changes in Earth’s gravity field, through which it records the changes in total water storage on land, termed the terrestrial water storage anomalies (TWSA) Famiglietti and Rodell (2013). TWSA peaks can help estimate maximum storage capacity Reager and Famiglietti (2009), watershed storage trends Thomas et al. (2016), help estimate flood potential Reager et al. (2014), and help constrain or validate hydrologic models Li et al. (2012); Lo et al. (2010); Niu et al. (2014). However, few studies explored of the direct relationships between TWSA and baseflow over a variety of basin conditions. There are also substantial knowledge gaps regarding how storage and streamflow are interrelated in different geographic areas and across a wide range flow regimes.

To concisely describe catchment hydrologic functioning, the literature has introduced two concepts: "Hydrologic signatures" (Sivapalan et al., 2003; Vogel and Sankarasubramanin, 2003; Yadav et al., 2007; Yilmaz et al., 2008; Hrachowitz
et al., 2013; Troch et al., 2013; Gupta et al., 2014) and "catchment classification" Wagener et al. (2007). "Hydrologic signatures" refers to statistics or summaries that reflect certain aspects of hydrologic systems. For example, the slope of the flow duration curve (Sawicz et al., 2014) is extracted from daily streamflow and indicate flow variability. "Catchment Classification" groups catchments into a number of classes, perhaps using hydrologic signatures, so that we can recognize and understand their behaviors. Sawicz et al. (2014) used six hydrologic signatures, some extracted from daily streamflow records, to group basins into interpretable classes. Berghuijs et al. (2014) grouped 321 basins into several clusters with similar seasonal water balance behaviors and how seasonal water balance influences streamflow variability signatures. Such data mining exercises are well suited to help answer the question of why catchments are hydrologically similar or dissimilar (McDonnell and Woods, 2004), and allow a mapping between climate, physical characteristics and hydrologic behavior (Sawicz et al., 2014). However, the series of work above built mainly on streamflow data. The inclusion of GRACE storage data can provide a new observational dimension describing how basins store water, leading to fresh new insights. On the other hand, the limited spatiotemporal resolutions of satellite-based observations pose a different set of challenges than earlier catchment classification work using primarily land-based data.

This work achieves the following goals: 1. Present an overview of the typical storage-streamflow correlation patterns and their spatial distributions across the contiguous United States (CONUS); 2. Propose a novel, panoramic framework of storage-streamflow relationships and a first-order analysis of its controlling factors; 3. Demonstrate the use of this framework, along with data mining techniques, to inspire, corroborate and reject hypotheses about hydrologic behaviors in several regions of CONUS. Here, a 'hypothesis' refers to an explanation for storage-streamflow correlation behaviors. They can be human-conceived, but may also be derived based on classification tree analyses (section 3.3.5). Some hypotheses are given in Section 3.4.4.
3.3 Methods

3.3.1 Streamflow and GRACE TWSA data sources

A list of 9063 United States Geological Survey (USGS) streamflow gages was obtained from the GAGES-II dataset (Falcone, 2011), which was used in hydrology as well as geomorphological studies (Slater and Singer, 2013; Shen et al., 2016a). GAGES-II has not only station numbers but also a wide range of climate, physiographic and geologic attributes. Daily discharges from USGS were downloaded from USGS website\(^1\) and aggregated to analysis time periods and scales as discussed later. The stations with more than 10\% of the daily streamflow records missing in the 2002-2012 periods are removed from further analysis. We removed stations with catchment areas greater than the HUC4 basins in which they are located, to restrict our analysis within small to mesoscale catchments. We removed \(\sim 165\) stations that have more than 0.5\% daily flows that are zeros, as zero flows are insensitive to storage changes, leading to unreliable interpretation of storage-streamflow correlations. Also removed are around 303 stations whose catchments have more than 2 major dams per 100 km\(^2\), or whose dam volume is greater than 500 Mega L/km\(^2\) (or mm). We conducted sensitivity study to examine the impacts of these choices. After the screening, there are 4476 stations (Figure 3.2a). Furthermore, for comparison, we selected 974 stations from this list which are reference stations, meaning with little human disturbance. We compared the results between using reference stations only and using the 4476 to examine the effects of human disturbance on the analyses.

We obtained the monthly, 1 degree, TWSA mass grids level 3 version 5.0 data (JPL, 2014), processed using University of Texas Center for Space Research (CSR) algorithm (Swenson, 2012). The GRACE product employs a destriping filter and a 300 km wide Gaussian filter to minimize North-South stripes. However, the low-pass filtering of GRACE solutions causes attenuation of small-scale surface mass variations. We used a scaling factor based on land surface models (Landerer and Swenson, 2012) to restore signal losses. GRACE TWSA data and hydrologic signatures are averaged by area to the USGS 4-digit Hydrologic Cataloging Unit (HUC4) basins (202 basins, average 42,000 km\(^2\)). A map of HUC4 is shown in

\(^1\)http://waterdata.usgs.gov/nwis/
Figure S3 in the Appendix B. GRACE data are impacted by measurement noises (Landerer and Swenson, 2012; Swenson and Wahr, 2006) and the contamination of signal by nearby regions (due to spectral truncation and filtering), termed leakage errors (Figure S4 in Appendix B). Over CONUS, both errors are relatively small and uniform compared to other regions in the world, except for near the mouth of the Mississippi River. We excluded 4 HUC4s at the mouth of the Mississippi River from analysis due to relatively large GRACE leakage error from the river. Although GRACE has a large spatial footprint (200,000 km²), basin size was not found to be a limiting factor for hydrologic analysis in past studies, e.g., in Fang et al. (2016). Importantly, our work focuses on the collective patterns emerging from spatial clusters of basins, rather than individual basins. We analyzed the impacts of basin size on our results in this study. For each USGS gage, storage-streamflow correlation was calculated between the streamflow from this gage and the GRACE data in the HUC4 basin in which this gage is situated.

3.3.2 The storage-streamflow-correlation spectrum (SSCS)

Concisely, SSCS is the collection of correlations between TWSA annual extrema and different streamflow percentiles in a window around them. We define $\rho^p_x$ (or $\rho^t_x$) as the Pearson’s correlation coefficient between TWSA annual peaks (or troughs) and x-percentile flow in an annual observation window (the selection of window is discussed below) surrounding the extrema in the corresponding years. The superscripts $p$ and $t$ indicate correlations with TWSA peaks and troughs, respectively. We have also conducted the same analysis using Spearman’s rank correlation coefficient, which shows slightly lower values than Pearson’s (Figure S5 in the Appendix B). Moreover, since here we are also interested in evaluating the strength of a simple linear model to estimate inter-annual streamflow changes using GRACE, we focus on Pearson’s correlation coefficient. Referring to x-percentile as a band, the storage-streamflow correlation spectrum (SSCS) is then the collection of $\rho^p_x$ across different flow percentiles. The correlations are calculated on an annual scale. Throughout the study period (2002-Oct-01 to 2012-Sep-31) we have 10 data points of TWSA peaks and 10 points of matching $Q_x$, where $x$ is the percentile. There are at least 10 annual TWSA troughs, but if the last trough occurs before 2012, we add one additional trough to the series, to maximally use the GRACE data before 2013.
(illustrated in Appendix B Figure B.12). The number of flow data points match the number of troughs. From 2013, GRACE data has more gaps. Based on the extrema selected, we selected a window of analysis for streamflow. Within each window the 15 percentiles extracted are: 0.5%, 1%, 2%, 5%, 10%, 20%, 50%, 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.5%. For narrative convenience, we also denote high-flow (≥ 80%), mid-flow (≥ 40% but ≤ 60%), and low-flow (≤ 20%) by ρH, ρM and ρL, respectively. The SSCS gives an instant snapshot of the correlations across all bands as a whole, as compared to previous studies that focused only on flood regimes. As such, it provides a panoramic rather than focused view of the storage-streamflow relationships. On the other hand, while we have identified the potential of storage-streamflow correlations, we consider the predictions of inter-annual changes in streamflow using GRACE to be outside of the scope of this study.

One can use water years as a time window to select the peaks, troughs, and streamflow percentiles, but we found this window to be sub-optimal: If a peak occurred in October 2014 in an HUC4 and we used October as the start of the accounting cycle, then September of the same year might be chosen as the peak for the water year 2013. Such a choice would not be reasonable as these two months belong to the same annual cycle. Instead, we found an extrema-based procedures improved the correlations (Appendix B Figure B.11b provides a comparison) noticeably. First, we need to determine the position of the TWSA annual extrema, as the extrema can occur in all of 12 months across CONUS. To achieve this, for each HUC4 we selected a month that minimized the standard deviation of the Julian days of identified peaks in the 10 years of analysis. If multiple months produced similar results, we selected the month that maximized the absolute value of the identified extrema. Different starting months were selected for peaks and troughs. With this approach, we were able to choose the most hydrologically meaningful extrema.

Then, we searched for a period enclosing $m_1$ months before and until $m_2$ months after the month of TWSA extrema that produce high correlations. We found that for peaks, $m_1^p = 3$ and $m_2^p = 3$ (superscript $p$ stands for peaks). For troughs, a whole year was better, $m_1^t = 8$ and $m_2^t = 8$ (superscript $t$ stands for troughs). Streamflow percentiles were selected from this analysis period. This ideal window also means the streamflow percentiles may precede the TWSA extrema.
3.3.3 GRACE signatures

Three hydrologic signatures from GRACE TWSA monthly time series were proposed in Fang et al. (2016) and included here among the predictors for SSCS: (i) the average annual TWSA amplitude ($A$) as a fraction of precipitation ($A/P$), which reflects the competition between storage, runoff and evapotranspiration ($ET$). Higher $A/P$ means more water has the opportunity to infiltrate and boost storage as compared to immediately becoming runoff or evapotranspiration; (ii) an inter-annual storage variability index, $\gamma$, which represents the ratio of inter-annual to intra-annual TWSA variabilities. $\gamma$ is defined as the following. Let $\sigma_{w,i}(TWSA)$ represent the standard deviation of monthly TWSA data in i-th year, and let $\sigma_w = \frac{1}{n_y} \sum_{i=1}^{n_y} \sigma_{w,i}(TWSA)$ be the average within-year standard deviation, where $n_y$ is the number of years. Also, if the mean of each year’s TWSA is $TWSA_i$, we use $\sigma_b$ to represent the standard deviation $TWSA_i$. We then define $\gamma$ as

$$\gamma = \frac{\sigma_b}{\sigma_w}$$

We calculated this index using GRACE data from 2002 October to 2014 September. Some data gaps in GRACE has been filled using spline interpolation. The larger this index is, the higher proportion of TWSA variance is explained by inter-annual variability. $\gamma$ concerns both climate and the system’s ability to hold water. For example, a system with thin soil will dampen the influence of inter-annual precipitation variability and reduce $\gamma$ because it limits storage; (iii) the 1-month-lag, piecewise-detrended autocorrelation function, $acf$, which describes the seasonal distribution pattern, or smoothness, of TWSA. The $acf$ is calculated using the $autocorr$ function in Matlab. The coastal zones, especially Florida catchments, have the lowest $acf$. We also calculated a precipitation-potential evapotranspiration ($Ep$) seasonality index, $\xi$ (Milly, 1994; Fang et al., 2016), which equals 1 when $P$ and $Ep$ are completely in phase (cycle peaks arrive at the same time), 0 for uniform precipitation with seasonal temperature, and -1 when the two are completely opposite in phase. Conceptually, when $\xi = 1$, rainfall reaches peaks when energy is also at its peak. As a result, there is maximum water availability when evaporative demand is also high. On the other hand, when $\xi = -1$, such as in the Sierra California, precipitation falls when the energy input is the smallest. There is relatively little water to satisfy the higher evaporative demand compare
to larger $\xi$ cases. Precipitation and climate forcing data used to calculate these indices were obtained from the North American Land Data Assimilation System (NLDAS) (Xia et al., 2012).

### 3.3.4 Exploratory clustering of SSCS

Since the volume of data is large (30 bands total at each station), its direct interpretation is difficult. Inspired by the catchment classification work, we employed k-means clustering technique [Seber, 1984] to group all the stations into a small, interpretable number (6) of characteristic classes (#1 to #6). A clustering algorithm like k-means depends on a user-defined notion of distance. The SSCS spectral distance between catchments $a$ and $b$ is defined as the Euclidean distance $(d_{a,b})$ in the space of SSCS bands:

$$d_{a,b} = \sqrt{\sum_{x \in \rho(a) \cup \rho(b)} (\rho(a) - \rho(b))^2}$$

where $\rho \in (\rho^S \cup \rho^L)$ is all the SSCS bands. Using a specified notion of distance, k-means clustering analysis partitions the data points into $k$ clusters with the objective of minimizing the within-cluster sum of distance between points and the cluster center. While the correct choice of $k$ is sometimes ambiguous [Han et al., 2012], one should in general prefer a small $k$ that achieves the minimization and the desired cluster resolution. We used principal component analysis (PCA) to visualize the distribution of catchments in the storage-streamflow correlation spectral space and to assign class numbers. Details are provided in the Appendix Text S1.

### 3.3.5 First-order classification tree analysis of the controls

Classification and Regression Tree algorithms (CART) is employed in this paper for two purposes: (1) based on all CONUS basins, we used CART to rank important factors influencing SSCS. CART can analyze factor importances even if their influences are non-linear and non-monotonic; (2) based on basins in a region, we construct ad-hoc, two-level trees to identify the main differences between several sub-regions. In machine learning terminology, each classification tree is a hypothesis because it attempts to provide an explanation, based on splits of attribute
values, of why the outcome variable takes a certain value (Mitchell, 2006). We employed a two-steps regression algorithm combining classification tree and linear regression. Attributes chosen from the feature selection procedure (described below) were input variables, while spectral distances to class centers are the outputs. The first step is to train a classification tree which hierarchically splits the data points using binary decision rules. Each node at a certain level in the tree may be further split based on one of the attributes. The algorithm efficiently searches through many possible splits to minimize the total information entropy (Shannon, 1948) of the tree, which in effect is similar to the variance of calculated spectral distances. At the bottom node, a multiple linear regression is used to predict the spectral distance.

Eight important attributes are shown in Figure 3.1. We note a continental-scale East-West gradients in $\xi$ (Figure. 3.1h) and relative humidity (RH), the North-South gradient in S/P, and high S/P values in the mountainous regions in the West. Along the Appalachian Plateau, the depth to bedrock is shallow (Figure. 3.1e), and the inter-annual storage variability index is low (Figure. 3.1c), while the soils to the North and South of this region are thicker than 60 inches ($\sim 150$ cm, the maximum reported depth in soils surveys). We also note high average soil bulk density (BDAve) in Ohio/Indiana.
Figure 3.1. Spatial distributions of some of the predictors. BDave: top soil bulk density; S/P: fraction of precipitation as snow; \( \gamma \): TWSA inter-annual variability index; RH: relative humidity; RockDep: depth to bedrock from soils dataset; Slope: mean slope in the model from GAGES-II; acf: TWSA 1-month-lag auto-correlation function; \( \xi \): P-Ep seasonality index which indicates how much in phase they are. We note that many factors, such as rock depth and slope, co-vary.

In order to rank the important physical controls of SSCS, we used CART to predict the spectral distances of basins to class centers with a number of attributes. Data was divided into two bins: the training set, the portion of data used to train the model; and the testing set, the portion of data withheld from training.
but used to evaluate the trained algorithm. The testing error is the difference between regression-tree-predicted spectral distance and the 'observed'. There are also two modes of prediction. In the composite mode, the tree will have six outputs that match the spectral distances to six class centers. In the single-class mode, a tree is only responsible for predicting the distance to one of the class centers. The composite tree is useful for identifying sensitive parameters while the single-distance tree is useful for understanding how a class splits from others.

We conducted forward and backward feature selections to screen for important predictors from a large list of potential attributes (details are in Appendix B Text B.2, Figure B.2, and the list of predictors in Table B.1). The selected features are still too many, and some of them are ranked as important because they are correlated with each other. We then did extensive cross-validation tests to identify the most robust models in terms of testing error, from a large number of potential trees with different combinations of the selected predictors. Briefly, we varied the amount of available data, both by random sampling and by holding out data from some regions for testing. We selected the trees with the smallest error heuristically for our interpretation. Details are in Text B.3.

3.4 Results and Discussion

In the following, we first present SSCS distribution over CONUS, the SSCS classes, and their typical patterns. We then examine the important physical controls of SSCS. The most interesting results, which are using interpreting SSCS to understand basin hydrologic behaviors, are presented last in Section 3.4.4.

3.4.1 The Streamflow-Storage Correlation Spectrum

From Figure 3.2, we note that the best correlations amongst all SSCS bands, $\rho_{mx} = \max_j \rho_{t,x} (\max_j \rho_t \mid \rho_i^2)$, are high, suggesting that TWSA is strongly correlated with streamflow in some bands in many parts of CONUS. 80% of the stations have $\rho_{mx}$ of greater than 0.72 and 42% greater than 0.85. High $\rho_{mx}$ (>0.75) points are found throughout the CONUS, and surprisingly, under varied temperature and precipitation regimes. Especially, high values are found in the southeast CONUS around the region labeled as B in Figure 3.3 (thereafter abbreviated as F3-B.
Similar notation is used for other labeled regions, northern half of central lowland (F3-I) and most western half of CONUS except for along the northwestern coast (maps of CONUS states and physiographic divisions are provided in Appendix B Figure B.10). However, there are several regions with noticeably lower correlations ($\rho_{mx} < 0.75$): there is a belt extending from Northeast to South-central CONUS, hereafter we term the low-$\gamma$ belt (F3-A, compare with Figure 3.1c), which has low $\rho_{mx}$. We will later describe that there are several different reasons that cause the low-$\gamma$ and low $\rho_{mx}$ in different parts of this belt; in addition, Eastern Colorado, Northwestern coast, Southern Florida, and the Superior upland in the north (F3-D) also have points of low $\rho_{mx}$.
Figure 3.2. Maps of storage-stream correlation spectrum (SSCS) for USGS gaging stations. (Continue on next page)
Figure 3.2. (Previous page.) (a) the best correlation across all bands with the histogram and exceedance frequency (1 minus cumulative distribution) in the insert; (b) the percentiles corresponding to the best bands; (c) the nearest class centers of USGS gages based on spectral distances. Each circle represents the location of the gage (basin outlet). As we restricted the size of the catchment areas, these outlets are not far from basin centroids.

Figure 3.3. Regions labeled for the convenience of discussion. The different colors of annotations carry no special meaning - they are used only to increase readability of the labels. A: some basins on the low-\(^\gamma\) belt; B: part of the Southeast region with high storage-streamflow correlations across all bands; C: Ohio/Indiana with Class #6 (non-responsive) basins; D: Superior upland; E: Agricultural basins in Iowa; F: A region intersecting Southern Great Plains, Interior highlands and Western Gulf coastal plains; G: Northern Great plains; H: Northern Rocky Mountains; I: Arizona.

3.4.2 Typical SSCS behaviors over CONUS

SSCS is best treated as continuous variables as opposed to distinct classes based on the lack of clear separation of clusters (Appendix B Text B.1 and Figure B.1 in Appendix B). As a result, our classification is only used for reducing the number of data points to manageable groups for easy recognition and convenience of the narrative. In addition, our ad-hoc trees which are constructed to find physical controls use the continuous distance to class centers as their training targets. Labeling a catchment as a certain class only means that it is spectrally nearest to the center
of that class.

Keeping this point in mind, we can proceed to examine typical SSCS patterns one can expect over CONUS. Our k-means clustering (see Section 3.3) divides the catchments into 6 classes (Figure 3.4), and Figure 3.2c shows their locations. \( k \) is set to 6, not only because the average within-class sum of squared distances changes slowly after 6 clusters, but also because it is a suitable resolution that aids our appreciation of SSCS gradients over CONUS. Class #1 can be called 'full-spectrum responsive' since it has the highest \( \rho_x \) and the smallest variability across all bands (Figure 3.4a, Also Figure B.13a in Appendix B presents the spatial distribution of spectral distance to this class center). As discussed earlier, class center #1, which means the first class, has overall the highest correlations. These catchments are located in the Northwestern quadrant (above 36°N and west of 90°W, except for along northwestern coast), center (Iowa, Illinois, Missouri), Southeast (Alabama, Georgia, and Virginia), Southern Texas, and scattered elsewhere. Class #2 (Figure 3.4b) is similar to #1, but is generally lower in peak-TWSA bands, while trough-TWSA bands weaken only very slightly. Clusters #1 and #2 are generally interspersed in space. Compared to #1, catchments with SSCS close to Center #2 appear more often in the coastal plains. Class #3 catchments are concentrated in the low-belt along norther Appalachian Plateau. Perplexingly, \( \rho_x^p \) can be negative in low-flows bands. This class has \( \rho_x^p \) that rises with \( x \) and \( \rho_x^t \) that declines with \( x \). The trough-low-flow bands are the most strongly correlated in this class. Class #4 are scattered throughout CONUS, and can be found in higher concentrations in the northwestern quadrant and California. They have higher correlation in high-flow bands, with high \( \rho_x^p \) and \( \rho_x^t \) in high-flow bands. In contrast, Class #5 catchments have higher correlations in low flow bands, and are concentrated in northeastern coast, south-central CONUS (Oklahoma, Arkansas, Louisiana and part of northern Texas). Class #4 and #5 are similar but #4 has higher values in high-flow bands. Class #5 are scattered widely over CONUS and have a larger spread in most attributes than #4 (Figure 3.5). Lastly, Class #6 are non-responsive catchments that are concentrated in Indiana and the Superior Upland. \( \rho_x \) is mostly low across all bands. Judging from the distance to centers, class #6 also has some overlap with Class #3 (Appendix B Figure B.13f). Except for Class #3, all classes are found in disparate environmental conditions, showing that similar SSCS patterns could arise from different climatic and physiographic
combinations (Figure 3.5). Except for the inter-annual storage variability $\gamma$ which separates out class #3, no other parameter can clearly separate classes. Readers can also find the SSCS aggregated by geographic divisions in Figure B.14 in Appendix B.

![Boxplots of the SSCS for each class](image)

**Figure 3.4.** Boxplots of the SSCS for each class. The boxes contain 25%-75% percentiles, and the crosses are those considered outliers. The 15 bands are 0.5%, 1%, 2%, 5%, 10%, 20%, 50%, 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.5%. The numbers in the parenthesis stand for the number of stations in the class.
Figure 3.5. Parameter distribution for each class. Silt Cont: percentage of silt in top soil; P: precipitation; BDAve: average top soil bulk density; WD: number of days with measurable precipitation. Other variables are explained in Figure 3.1 caption. Except for class #3, all other classes are well spread out parameter ranges. This pattern means that a factor may have different impacts on SSCS in different regions.

According to our sensitivity tests (Text B.4 in the Appendix B), our treatment of zero-flow gages has no noticeable impact on the clustering. On the other hand, Pearson’s correlation is in general higher than Spearman’s, suggesting the lack of systematic nonlinear trends in SSCS (Appendix B Figure B.11b). With respect to the impacts of major dams, when major dam density or dam volume are high, they tend to lower correlation in the trough-low flow bands (Appendix B Figure B.8), potentially due to human-controlled release during low flows. This result led
us to remove basins with high major dam density (> 2 dams/100 km$^2$) or large dam volumes (>500 mm averaged over basin area). This removal has little impact on the clustering because high-dam-density points are evenly distributed in the SSCS space (Appendix B Figure B.9). Between 200 and 500 mm, dam storage has a mild and non-consequential influence on SSCS clustering (Appendix B Text B.4, Figures B.8 and B.9). Thus we retained the 355 basins in this bin for analysis. Its impact on the clustering is also negligible.

### 3.4.3 Continental-scale physical controls of SSCS

Through cross validation-tests (Figure 3.6a, details in Text B.3 in Appendix B) we identified the most robust CART models. Because CONUS spans a large variety of climates and physical landscapes, a CART analysis of SSCS over CONUS is very complex. We show the top three levels of the selected tree in Figure 3.6b, but even at deeper levels, we can still see some meaningful separations (one of the best performing realizations is provided for completeness of the paper in Figure B.15 in Appendix B. We warn that this full tree is complex). From both our feature selection analysis and tree selected from cross-validation, several variables consistently emerge as top-level decision variables. They include snow as a fraction of precipitation ($S/P$), long-term average relative humidity and inter-annual TWSA variability ($\gamma$) and P- Ep seasonality index ($\xi$) (Appendix B Table B.1). We note that the first 2 levels of splits in Figure 3.6b remains identical even if we add the dammed basins. Some other important factors include top soil texture (Silt or Sandy percentages. Clay may also be used), depth to bedrock, annual precipitation, mean topographic slope, agricultural land use. However, their positions in the tree may be different in alternative CART realizations as many of these factors co-vary. Soil parameters’ importance comes only next to climatic ones: the best performing tree without soil parameters ranked the 88th tree in our cross-validation in terms of average testing accuracy. The best two trees without soil parameters are dominated by other trees with soil parameters (Figure 3.6a).
Figure 3.6. (a) Cross validation of potential models. Every line represents a model, and the plot contains top 100 best performing models from cross-validation. The thick black line is the selected model for the tree in Figure 3.6b and Appendix B Figure B.15, and pink and blue lines are two nondominated models without soils data. Note the performance deteriorate when soil data is not available. On the x axis, 20% - 70% is the amount of data randomly held from training; HUC2: One two-digit Hydrologic Cataloging Unit (HUC2) was held out as a testing set; 2× HUC2: Two HUC2s were held out for testing. We looped through all possible combinations of HUC2s, and the error is the average of all cases. We note that withholding data by HUC2 is much more stringent test than randomly withholding data. We select the models that are among the top 10% of all CV tests and are not dominated by any other models. (b) Top three level of the composite-output tree that simultaneously predicts distances to 6 centers (the black line in Figure 3.6a). In the legend, we include the number of basins on each node (#Training Set/#Test Set), the variances (Var) and the composition of classes on the node. We note that S/P, relative humidity (RH), and interannual storage variability (γ) are ranked as most significant parameters for this tree.

The CART analysis suggests that higher $S/P$, higher $\gamma$, and other mechanisms
that promote infiltration as compared to runoff all tend to generate higher SSCS values and boost the proportion of #1 catchments. \( S/P \) is typically higher in the north but it is not just a function of latitude. Higher \( S/P \) means larger snowpack and snowmelt-induced high flows, and therefore it will likely lead to higher values in peak-high flow SSCS bands. However, \( S/P \) and \( RH \) take such prominent roles also because the relationships between physical controls and SSCS are conditional on them. In other words, these relationships are characteristically different in different ranges of \( S/P \) and \( RH \). Regarding \( \gamma \), as our definition of SSCS is based on inter-annual changes, high inter-annual variability means high signal-to-noise ratio which tends to yield high correlation. Also, low \( \gamma \), as will be described later, may indicate limited storage capacity that disconnect storage from streamflow.

Surprisingly, despite the very coarse spatial footprint of GRACE, catchment area has not turned out to be particularly important to SSCS. There is a mild trend for larger catchments to belong to class #1, and for class #6 catchments (non-responsive) to more likely be small catchments (Figure 3.5a). Presumably, in some small catchments, the water storage patterns are not similar to aggregated ones in the corresponding HUC4s. However, both forward and backward selections remove catchment area from top 16 influential predictors. Moreover, basins with similar SSCS tend to aggregate regionally and manifest spatial patterns. These results suggest that GRACE signal can be interpreted as a macroscopic environmental variable relevant to a region and a collection of basins much smaller than its footprint. We should be cautious using it for individual small catchments. A better practice is to examine regional emergent patterns from a cluster of basins, which is our focus.

### 3.4.4 SSCS-derived insights and hypotheses about hydrologic behaviors

#### 3.4.4.1 General interpretations

Based on our basic understanding of the hydrologic cycle, generally, if flows in certain bands are disconnected from storage, we expect a low correlation in those bands. The average of all SSCS (The black line in Appendix B Figure B.1b) is the 'baseline' basin behavior. As inter-annual changes in both storage and streamflow are correlated with the amount of total precipitation in each year, the 'baseline'
basin has a moderate correlation in all bands. By default, we expect high-flow bands to have lower storage-stream correlations than middle and low bands because peak streamflows result from large storms whose magnitudes are poorly correlated with annual total precipitation. The most informative insights have come from significant deviations of peak-high flow ($\rho^p_H$) and trough-low flow ($\rho^t_L$) bands from the national average. Table 3.1 provides some general interpretations of these signals based on our reasoning.
Table 3.1. Hydrologic Signals From SSCS and Potential Hypotheses That Explain Them Based on Hydrologic Reasoning

<table>
<thead>
<tr>
<th>ρ_H^p</th>
<th>... Is Low</th>
<th>ρ_H^p</th>
<th>... Is High</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>What it means: The generation of high flows is disconnected from water storage, i.e., heavy storms cause large runoff without leaving signals in the storage</td>
<td>What it means: High flows are connected to water storage. Even large storms leave a signal in storage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Possibilities: (i) Shallow water table in a wet climate with little seasonal fluctuation; (ii) Very dampened groundwater response; (iii) water storage exhausted during dry years</td>
<td>Possibilities: (i) Snowmelt-dominated flow peaks; (ii) Saturation excess with large storage variation among seasons; (iii) Groundwater-dominated peak streamflow</td>
<td></td>
</tr>
<tr>
<td>ρ_L^L</td>
<td>Base flow does not depend on storage</td>
<td>Many regions by default have a ρ_L^L of &gt;0.5. It also suggests groundwater is not depleted during dry years.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Possibilities: (i) Shallow water table in a wet climate with little seasonal fluctuation; (ii) Very dampened groundwater response; (iii) water storage exhausted during dry years</td>
<td>Possibilities: Storage-dependent base flow</td>
<td></td>
</tr>
<tr>
<td>ρ_L^p</td>
<td>Annual storage peaks are not related to low flows, which is what we expect and appears to mean little</td>
<td>The pattern that annual storage peaks are correlated to low flow appears to mean little, but it does suggest that the region has not reached storage capacity during wet years</td>
<td></td>
</tr>
<tr>
<td>ρ_H^L</td>
<td>Annual storage minimum is not related to peak flows, which is what we expect and appears to mean little</td>
<td>Annual storage minimum is related to peak flows. High flows depend on the entire history of storage accumulation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Possibilities: It suggests a groundwater-dominated system and groundwater availability during dry years</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It must be noted that SSCS and its CART-based analyses are meant to inspire and corroborate hypotheses about how the hydrologic system behaves from a storage-streamflow-correlation perspective. For this reason, these hypotheses are discussed after SSCS is introduced. We note that these hypotheses are already con-
sistent with observations to the extent deemed acceptable by the CART analyses. SSCS alone may contain limited information about a system, but combining it with the predictive CART analysis and other sources of information can eliminate competing hypotheses. Below we focus on a few regional examples of SSCS-inspired hypotheses (Sections 3.4.4.2 through 3.4.4.6), and then we list a few unanswered questions.

3.4.4.2 Hypothesis 1: Soil-thickness limits storage on the Appalachian Plateau

A large part of the low-γ belt (F3-A) coincides with a region with shallow soil/regolith thickness (Figure 3.7c & e) along Appalachian Plateau and the Valley and Ridge physiographies (see physiographic provinces in Appendix B Figure B.10a). Here, we find the thin soils to explain both the class #3 SSCS behavior and the low values of γ. In a two-level adhoc tree constructed to predict distance to class center #1, soil thickness is chosen as the criterion to separate the low-γ belt from its neighbors in the South that are spectrally nearest to Centers #1 and #2 (Figure 3.7a & b). In this region, soils on the plateau and ridgetop are thin (Figure 3.1e) because they are developed in shale and sandstone. Our SSCS-derived hypothesis here is that shallow soils and the underlain low-permeability bedrock limits soil water storage capacity, leading to saturation excess arising from a perched water table, low values and mid-to-low SSCS values in high flow bands. In the wetting phase of the annual water cycle, the pore space fills up. The storage capacity appears to be reached regardless of initial storage at the beginning of the wetting cycle. We can see supporting evidence from the GRACE time series for one of the GRACE mass grid cells (Figure 3.8). TWSA peaks, which also contain signal from the snowpack, are more evenly distributed than the troughs. When peak flows occur, most of the surface soil of the basin is already close to saturation. Thus extra water quickly becomes surface runoff and exit the system without leaving a signal in GRACE monthly TWSA. As a result, the high flows are not highly correlated with either storage peaks or troughs (ρ_H is mediocre), even though snow storage likely helps increase ρ_H. On the other hand, the median ρ_L values are above 0.5, suggesting baseflow in the summer is controlled by storage. In accord with our findings, soil thickness limitation was found to significant influence TWSA in mountainous regions in a recent modeling study (Brunke et al., 2016). Our finding also suggests
that accurately describing soil thickness is important for flood prediction in this region.
Figure 3.7. (a and b) Together explain how low-γ belt basins are separated from the Southeast basins, and (c and d) together explain how Indiana/Northern Ohio basins are separated from neighboring basins. (a) Node split map showing how different classes, as separated by depth to bedrock, are distributed in the Eastern region of CONUS. (b) a simple ad hoc tree constructed to find the most sensible split that separate the Class #3 basins on the low-γ belt from the Classes #1 and #2 basins in the Southeast. RockDep (Depth to bedrock, inches) emerges as the most sensible criterion. Note that the GAGES-II data set is derived from STATSGO and therefore has a soil bulk density, are distributed in a region surrounding Indiana and Ohio. (d) An ad hoc tree indicates that bulk density is the most apparent factor to separate Class #6 (nonresponsive) basins from adjacent regions. When we construct this tree, the reference flag in GAGES-II data was provided as a candidate attribute, but was not chosen by the tree. The reason is, while Class #6 points here are predominantly agricultural and nonreference, many Class #3 basins (green ones) are also nonreference ones.
Figure 3.8. Time series of TWSA for a GRACE cell on the low-γ belt. Dashed lines indicate the mean of the peaks and troughs. $\sigma^p$ and $\sigma^t$, measured in cm, are the standard deviations of the annual peaks and troughs, respectively. $\sigma^p$ is less than $\sigma^t$, meaning the differences in peaks are smaller than the troughs. It suggests that the limited storage capacity is reached (recall the peaks also contain signals from snow) each year regardless of where the minimum is.

3.4.4.3 Hypothesis 2: Groundwater dominates streamflow and large-storage system in the Southeast CONUS

In contrast to the low-γ belt, basins in the Southeast CONUS (region F3-B, Center #1) have much thicker soil and high correlations across all bands. This region is among the highest in annual precipitation, is heavily vegetated, and has little to no snow. The high values in SSCS means that streamflow, including high flows, is tightly linked to non-snow storage. The soil and regolith thickness serves as a good criterion to separate this region from the Class #3 basins on the low-γ belt (Figure 3.7a). The CART analysis suggests that the SSCS values here are higher than Northern neighbors because of sufficient storage capacity to absorb even large storms. These traits lead us to hypothesize that streamflow in this system is dominated by groundwater contributions, and peaks are generated by groundwater-induced saturation excess. In such systems, most un-intercepted rainfall (even moderately heavy storms) first infiltrates and becomes storage, which
then exits as ET or groundwater contributions to rivers or riparian zones. Large streamflow peaks occur only after high water storage and groundwater flow saturate lowland areas, producing large runoff source areas. Therefore, high streamflow leaves clear signals in storage, producing high $\rho$. Our hypothesis is corroborated by three pieces of evidence: First, Sawicz et al. (2014) showed that these basins have characteristically high permeability (Class 3 in Figure 1 in their paper), providing the infiltrating capacity required by our hypothesis. Second, earlier estimates of baseflow index (BFI, the fraction of streamflow contributed by baseflow) are between 50 to 90 percent for this region (Wolock, 2003) (Appendix B Figure B.16a). Third, the stream hydrograph (Figure 3.9) for a basin in Georgia shows that there is a large seasonal baseflow fluctuation. For a dry year (2008), few peaks were generated. For a wetter year (2009), the baseflow is much larger and many more peaks occurred.
Figure 3.9. Comparison between streamflow hydrographs from a headwater basin from the high bulk density soil region in Ohio (USGS 04196000 Sandusky River near Bucyrus, Ohio, SSCS Class #6) and a basin from the Southeast (USGS 02390000, Amicalola Creek near Dawsonville, Georgia, SSCS Class #1). Catchment areas of both gages are around 230 km$^2$. The seasonal cycle of the base flow component of the Georgia basin is much more prominent than that of the Ohio basin, which has almost negligible base flow, but large and frequent peaks. We also note that there is a marked interannual difference in the base flow of the Georgia basin. The high storage in 2009 produced more large peaks than 2008.

3.4.4.4 Hypothesis 3: Shallow-water-table limits storage in Indiana, Ohio, and Florida

Some basins in Northern Indiana and Northern Ohio (F3-C) are classified as #6, in which correlation values are low for all bands. These basins, also on the low-γ belt, are surrounded by Class #2 and #3 basins. However, they have notably lower correlations in low-flow bands. Our hypothesis of the main contributor to low full-spectrum correlations is the shallow water table caused by compacted soil and flat terrain in this region. We came to our hypotheses by examining the differences between this region and other basins the low-γ belt basins, as described below. This region is characterized by glacial tilt surficial geology and large depth to bedrock.
Thus, the pore space is not a limiting factor as on the Appalachian Plateau, and soil thickness is not the explanation for the low full-spectrum correlations. A CART criterion that separates this region from others is the bulk density (Figures 3.7c & d), which is characteristically higher here than other regions (see Figure 3.1a). We have not identified the reason why bulk density is so high here, but it means the soils here appear to be heavily compacted which leads to poor drainage (Letey, 1958; Carter, 1990). It also has flat terrain and a wet climate soils, all of which contribute to poor drainage, and a very shallow water table. Past studies showed that >17% of the area has a high water table within 1 foot from the ground surface [ONR, 2016].

Therefore, the full-spectrum low correlation (and low $\gamma$) found in this region could only be attributed to the shallow water table. The steady state groundwater table simulated in the literature (Fan et al., 2013) (reprinted in Appendix B Figure B.16) agrees with this observation. Based on SSCS and knowledge about the physical landscape, our hypotheses are that (i) water storage capacity is limited by shallow water table resulting from flat terrain, wet climate, and poor soil drainage. The small storage capacity disassociates streamflow generation from storage; (ii) low flow is not related to storage trough also because the terrain is too flat to generate topography-induced baseflow. The soils also drain too slowly for baseflow to be significant. Indeed, the hydrographs show streamflow peaks that very easily generated from frequent storms, with a negligible baseflow component compared to that of a Southeast basin (Figure 3.9). Our SSCS-inspired hypothesis is different from that in Sawicz et al. (2014). Their results classified basins in this region as "Small and energy limited catchments along Appalachian range with 50/50 streamflow/ET release function" (Figure 1, Class 1 in their work), the same as region F3-A. While a 50/50 release function may be similar for both regions, it is true for different reasons and SSCS suggests they are hydrologically distinct. Thus, the new observation dimension offered by GRACE leads to new insights and a macroscopic appreciation of the spatial gradients in storage-streamflow relationships.

It is worth noting that most of the basins in this region are non-reference basins as they are predominantly agricultural. While some earlier analysis, e.g., Sawicz et al. (2014), removed non-reference basins from the analysis, doing so will remove most of the Class #6 basins here, which we avoided (Appendix B Figure B.3). Hence one might be tempted to think landscape modification is a significant
factor causing low storage-streamflow correlations. However, when we inserted the reference flag as an input for CART, it was not picked as the criterion in the first two levels. The reason is that while most Class #6 basins are non-reference, so are many Class #3 basins in nearby regions. Therefore, the reference status is not the cause for the low correlations in this region (Text S4 in the Appendix B). In addition, the agricultural land cover percent was one of the 44 attribute considered but was not ranked among top 15 in Table S1 in Appendix B.

With flat terrain and shallow water table (Appendix B Figure B.16b), Florida basins also show a pattern of disconnection between storage peaks and high flows. The shallow water table also tends to prevent infiltration and cause rapid runoff. However, its are noticeably higher (nearest to Centers #3 and #5), potentially because the seasonal decline of groundwater is stronger in Florida due to stronger heat input. On a side note, both Florida and Indiana appear to defy the descriptions of topography-controlled or recharge-controlled water tables that were introduced in previous work (Gleeson et al., 2011), because their terrain is too flat to be topography-controlled and their water table is too shallow to be recharge-controlled. Thus their theory may need to be extended.

3.4.4.5 Hypothesis 4: Groundwater dominates streamflow on the northern Great Plains

The pattern of high SSCS values across all bands is intriguing for the northern Great Plains (F3-G). These catchments are dry, with hot summer, cold winter, annual precipitation amounts in most sub-regions less than 550 mm/year, and potential ET that much exceeds rainfall. Moreover, the rainfall and energy cycles are in-phase, meaning that precipitation peak occurs when energy inputs are the highest (Berghuijs et al., 2014; Fang et al., 2016). As a result, most of the rainfall is immediately released as ET. Thus, the high correlations invite an interesting hypothesis that a portion of the precipitation manages to escape ET, recharges the groundwater, and exits as the sole source of streamflow. The rationale is that if high flows were contributed by surface flow, in such an arid and hot (during rainfall season) environment it must be Hortonian/infiltration-excess runoff, which is disconnected from storage. Hortonian runoff would result in low $p_H^*$ which is in conflict with the observed SSCS. Indeed, a local study (Driscoll et al., 2002) suggests many streams in this region primarily originates as springflow in higher
latitudes, and some streams gains baseflow discharge along its path (some stream segments may also lose water when flowing across outcrops). The same work also showed that this region is underlain by layers of aquifer formations sandwiched by confining units, and these aquifers are recharged by orographic precipitation in the high latitude outcrops. These findings all lend support to our SSCS-derived hypothesis. In addition, in the Northern half of North Dakota, where the BFI map (Appendix B Figure B.16a) shows lower values, $\rho_H^p$ is lower, which also agrees with our hypothesis. While local hydrologic studies, e.g., Driscoll et al. (2002) and Naus et al. (2001) already discussed groundwater-fed streams, the regional patterns of storage-dependencies of these streams across flow regimes become apparent only with the help of SSCS. SSCS also points the possibility of monitoring inter-annual streamflow changes using GRACE for similar systems.

A similar theory applies to southern Texas, which has a similar seasonality, more precipitation but also stronger potential ET. In comparison to the low-$\gamma$ belt, the Great Plains also has shallow soil thickness (Figure 3.1e), which appears to have no impact on SSCS, because water storage takes place in deeper aquifers here.

### 3.4.4.6 Hypothesis 5: Gradient in high flow bands along central lowlands is driven by change in energy

The spectrally nearest centers change from #1 (F3-E) to #2 and then to #5 (F3-F) going south along the central lowlands, as the peak-high flow correlations decrease along this path. First, although the F3-E region (Iowa, western Illinois, and northern Missouri) also has flat terrain and thick soil similar to F3-C, the simulated water table is 5-10 m deeper than F3-C (Appendix B Figure B.16b). It appears when the water table is this deep, it no longer limits water storage. Reasons for the deeper water table include smaller precipitation and less bulky soil compared to F3-C.

As we proceed South (F3-F), the spatial trends reflect increasing energy inputs and the decreasing fraction of precipitation as snow. As basins get drier, the predominant runoff mechanism shifts from saturation excess to infiltration excess, leading to lower correlations in the high-flow bands.
3.4.4.7 Unanswered questions

There are many patterns that call for explanations. Why does the Superior Upland physiographic province (F3-D) show poor correlations despite its high $S/P$? Similarly, why do basins in Maine show low correlations in high flow bands? Why do basins in the middle Rocky Mountains (F3-H) show low correlations in the low flow bands (they are nearest to Center #4), is it because of too much dampening of groundwater flow signal by the mountain blocks, residual-snow-dominated baseflow, or that the groundwater tank is depleted in some years? Why do some basins in southern Texas show good correlations? Why is there such a variety of SSCS patterns in Arizona (F3-I)? Is it because of the large variety of streamflow generation mechanism, e.g., snowmelt near Flagstaff and water management near Phoenix, or is it because that GRACE footprint is too coarse? We call on the community to answer these questions, using SSCS as a hypothesis generator to advance our understanding of these systems.

3.4.4.8 Assessing probability of success for GRACE-based estimates for streamflow

Earlier studies showed that GRACE-based flood potential index is effective in the Mississippi River basin, especially in Iowa and Northern Kansas (Reager et al., 2014; Molotkova et al., 2016) and it could be useful in other basins in the world (Molotkova et al., 2016). One region in Molotkova et al. (2016) where GRACE data successfully predicted floods matches well with a cluster of class #1 basins in our study (close to F3-E), although their flood records analyzed were limited to 2007. Our analysis suggests that, TWSA should provide great predictive power for floods in regions where values in storage-high-flow SSCS bands are high, such as Southeast coastal plains, Mid-Atlantic (North to F3-B), and Northern Great Plains. However, this index will not work well in the Appalachian mountains (F3-A), Maine, Upper Michigan Peninsula (F3-D), or in Arizona-New Mexico (around the F3-I), where infiltration excess causes flash floods. For baseflow, Lo et al. (2010) studied water table correlation with GRACE in Illinois, which is also in region E. On the other hand, our results suggest GRACE-based estimates may not work well for baseflow in class #3 and #6 basins, such as F3-A. As a result, SSCS informs us, before running the analyses, where GRACE-based methods will likely
succeed for streamflow analysis.

### 3.5 Conclusions

Utilizable yet undemonstrated storage-streamflow patterns show strong yet heterogeneous linkages between inter-annual changes in storage and streamflow across different flow percentiles, summarized as the SSCS. SSCS patterns vary substantially over space as the basic relationships between storage and flow are modulated by climatic, land surface, and geologic conditions. As a result, SSCS patterns can be used as an effective tool to inspire, corroborate or reject hypotheses about basin hydrologic behaviors such as runoff generation mechanisms, infiltration and storage capacities, and the groundwater cycle. Overall, when the hydrologic settings favor storage as compared to runoff, SSCS tends to be higher, and catchments tend toward class #1. High $S/P$, thick and permeable soils may all contribute to higher values in SSCS. The climatic factors are the most important controls while soil parameters and slopes rank next. On the other hand, where storage is limited, e.g., by thin soils or very shallow water table, correlations tend to be low. Our analysis highlights the dominant roles of groundwater in the hydrologic cycle in the Southeast, Great Plains, and lowland plains.

For several regions in the US, we interpreted the SSCS patterns, which corroborate local hydrologic studies. These results agree in general with previous analyses through baseflow separation techniques, but they also differ. However, the continental-scale SSCS presents an opportunity to appreciate the spatial gradient of hydrologic behaviors on a macroscopic scale, which is not available before from earlier local studies. We argue that SSCS patterns can serve as a potentially fundamental hydrologic signature to characterize different systems. Process-based hydrologic models, e.g., summarized in (Shen et al., 2014; Shen, 2013; Shen et al., 2016b; Clark et al., 2015; Fatichi et al., 2016), should aim to reproduce SSCS patterns. Disagreement between models and observations on SSCS will be an important sign of model deficiencies. On a side note, we demonstrate the usefulness of data mining technique in presenting patterns that emerge from hydrologic data and in stimulating hypotheses.

Subsurface configurations and lateral groundwater flow are very important for SSCS, e.g., in Southeast, low-$\gamma$ belt, Iowa, and Great Plains. Therefore, to ad-
equately predict flood or capture response to drought, representing subsurface features and hydrologic responses is of great relevance. For example, without adequate representation of soil thickness, a model may not be able to distinguish between Southeast and low-γ belt. Long-range groundwater flow can be important for drought resilience in arid regions like the Great Plains. Accurate predictions of floods in Ohio and Florida, on the other hand, will apparently depend on capturing water table depth.

3.6 Limitations and Future Work

GRACE has a very coarse spatial footprint and limited history. We used 10-11 years of GRACE data because there are more gaps after 2012 (fortunately, GRACE follow-on and GRACE2 are prepared to be launched). The data limitation means we should be cautious about using correlation values at an individual station, or using the correlation as a predictive tool for streamflow, but put more focus on spatial gradients of the patterns. In addition, GRACE leakage and measurement errors are relatively small and uniform over CONUS. In other regions where GRACE error is more significant, the error can be included in CART analysis.

When dam storage volume is large (> 500 mm averaged over basin area) or dam densities are high (>2 major dams per 100 km²), dams tend to reduce the trough-low flow storage-streamflow correlations. Therefore, we should be careful with interpreting SSCS for basins controlled by large dams.

As the concept of SSCS is proposed here only for the first time, more work will be needed to better understand its best use. For example, the effects of the selection of flow bands, its applicability at smaller scales, alternative definitions of spectral distances that address attribute correlations, e.g., Mahalanobis distance (De Maesschalck et al., 2000), etc., will need to be further studied. In regions where storage-streamflow correlations are strong (> 0.9) in certain bands, there is indeed potential in using these relationships in a prediction/hindcasting mode. Can we use streamflow records to hindcast GRACE extrema in densely gauged regions such as CONUS? On the other hand, can we predict streamflow inter-annual changes in ungauged basins in the world, by combining machine learning techniques with global estimates of soil thickness (Pelletier et al., 2016), rock permeabilities (Gleeson et al., 2011), soil properties (Chaney et al., 2016), and
water table depth (Fan et al., 2013)? Can we extend SSCS to relationships between groundwater-levels and streamflow? Also, can we use SSCS to estimate the fraction of baseflow contributions in the groundwater-dominated regions? All of these tasks will be of high scientific and practical value, but we may have to wait for more data to be available.

CART analysis can be effective in ruling out hypotheses. Nevertheless, similar to the fact that correlations alone cannot prove causality, CART analysis alone does not distinguish between causal and associative relationships. Because of factor co-variation and latent variables, one cannot draw definitive conclusions about the causal controls of SSCS. Ultimately, it provides hypotheses, not conclusions. However, SSCS indeed proves to be an effective tool to help to form hypotheses (third goal of the paper mentioned in the Introduction). Process-based models and numerical experiments can be utilized to further distinguish between causal and co-varying factors.

### 3.7 Author Contributions

K. F. collected and processed the data, carried out the experiments, and draft the manuscript. C. S. conceived and directed this study, and critically revised the manuscript. K. F. is the first author and has contributed the majority of work.
Bibliography


Clark, M. P., Y. Fan, D. M. Lawrence, J. C. Adam, D. Bolster, D. J. Gochis, R. P. Hooper, M. Kumar, L. R. Leung, D. S. Mackay, R. M. Maxwell, C. Shen,


Chapter 4  |
Prolongation of SMAP to Spatio-temporally Seamless Long-term Coverage of Continental US Using a Deep Learning Neural Network


4.1 Abstract

SMAP mission has delivered valuable sensing of surface soil moisture since 2015. However, it has a short time span and irregular revisit schedule. Utilizing a state-of-the-art time-series deep learning neural network, Long Short-Term Memory (LSTM), we created a system that predicts SMAP level-3 soil moisture data with atmospheric forcing, model-simulated moisture, and static physiographic attributes as inputs. The system removes most of the bias with model simulations and improves predicted moisture climatology, achieving small test root-mean-squared error (<0.035) and high correlation coefficient >0.87 for over 75% of Continental
United States, including the forested Southeast. As the first application of LSTM in hydrology, we show the proposed network avoids overfitting and is robust for both temporal and spatial extrapolation tests. LSTM generalizes well across regions with distinct climates and physiography. With high fidelity to SMAP, LSTM shows great potential for hindcasting, data assimilation, and weather forecasting.

4.2 Introduction

Soil moisture is a key variable that controls various hydrologic processes, including infiltration, evapotranspiration and subsurface flow. It is important to drought monitoring (Narasimhan and Srinivasan, 2005), flood prediction (Norbiato et al., 2008), weather forecasting (Koster, 2004), irrigation planning, and many other scientifically- and socially-important applications. Launched in 2015, NASA’s Soil Moisture Active Passive (SMAP) satellite mission (Entekhabi et al., 2010) is designed to measure top 5 cm soil moisture globally with a standard deviation of 0.04 cm$^3$/cm$^3$ volumetric ratio when vegetation water content (VWC) is less than 5 kg/m$^2$ (O’Neill et al., 2012). It achieved this goal in most core evaluation sites (Colliander et al., 2017; Jackson et al., 2016). Notwithstanding its great value, SMAP passive radiometer-based observations only have a short time span (since April 2015) with an irregular revisit time of 2-3 days, which makes it difficult to observe soil moisture responses immediately after storms or snowmelt.

Compared to SMAP’s limited resolution and time span, land surface models (LSMs), e.g., VIC (Nijssen et al., 2001), Noah (Ek, 2003), CLSM (Koster et al., 2000) and MOS (Koster and Suarez, 1994), simulate soil moisture seamlessly over much longer time spans. Despite their frequent use, these models may differ considerably from observations (Leeper et al., 2017; Yuan and Quiring, 2017; Dirmeyer et al., 2016; Xia et al., 2015). Biases (mean difference from the observed) are notable in all models evaluated in Xia et al. (2015). Their error patterns generally vary by region, model, season, and soil depths, yet there are systematic patterns in them. For example, moisture tends to be over-estimated in the arid western CONUS and under-estimated in wetter eastern US (Yuan and Quiring, 2017); the Noah model tends to under-estimate moisture in wet seasons and over-estimate in dry seasons (Xia et al., 2015). These systematic error patterns could be exploited to improve predictions.
To correct systematic model errors, we turn to deep learning (DL), a rebranding of artificial neural network. DL has made revolutionary strides in recent years and helped to solve problems that have resisted artificial intelligence for decades (LeCun et al., 2015). With earlier-generation machine learning methods, human experts extract features from data that are strongly correlated with dependent variables. DL, on the other hand, automatically extracts abstract features via their hidden layers. Two highly successful network structures are convolutional neural networks (CNN) for image-domain tasks (Krizhevsky et al., 2012), and Long Short-Term Memory (LSTM) (Hochreiter et al., 1997; Greff et al., 2017) for time-domain tasks, although the separation is not absolute. No study, to our knowledge, used recurrent time series deep learning in hydrology, especially for large datasets. Earlier work applied non-recurrent deep belief network on trend, period and random elements extracted from a dam inflow time series, without directly learning from raw observations or forcings (Bai et al., 2016). Given DL’s success in other disciplines (Voosen, 2017), it is plausible that DL can capture model error patterns that humans have yet come to explicitly formulate.

The parameter space of deep networks is substantially large in order to provide the flexibility in mapping diverse, complex functions. Thus one might be concerned about overfitting, which means coefficients are fitted to noise rather than meaningful information. However, there are recent breakthroughs in regularization techniques, e.g., Dropout (Srivastava et al., 2014), which penalize overfitting and reduce mutually dependent coefficients. Nevertheless, since LSTM has not been applied to hydrology, it is important to examine its robustness compared to conventional statistical methods.

The central hypothesis of this work is that with two years of SMAP data, LSTM can learn patterns in soil moisture dynamics and LSM errors, and by utilizing them, extend SMAP data over long time spans. Our objectives are: (1) to produce a seamless top-surface soil moisture dataset for continental United States (CONUS) with high fidelity when compared with SMAP data; (2) to provide an initial investigation of LSTM’s capability in correcting process-based model errors; (3) to compare LSTM’s generalization capability to conventional methods in spatial and temporal extrapolation tests. Here, by 'high fidelity', we mean a high consistency with the target data resulting in its faithful reproduction. SMAP’s retrieval algorithm for the passive product derives soil moisture from brightness
temperature readings using radiative transfer and soil dielectric models (O’Neill et al., 2012), thus it also incurs biases (Colliander et al., 2017). Nevertheless, a high-fidelity hindcast product has a wide range of applications, e.g., data mining of past fire hazards, calibration of hydrologic models, or benchmarking satellite product with historical in-situ data.

4.3 Methods and Datasets

As an overview, we trained an LSTM network to predict SMAP L3 product with, as inputs, atmospheric forcing time series, LSM-simulated surface soil moisture and static physiographic attributes. We compared LSTM to regularized multiple linear regression, auto-regressive models, and a simple one-layer feedforward neural network. Their performances were tested by (i) temporal generalization test: training over one year and testing over another; (ii) regular spatial generalization test: training over a uniformly down-selected subset of SMAP pixels and testing over other cells; and (iii) regional holdout test: training on some sub-regions of CONUS and test on the rest. All data sources are aggregated to a daily time scale and interpolated to SMAP L3 grid. Each SMAP pixel is an instance.

4.3.1 Data sources and inputs

For the learning target, we focus on the L3 passive radiometer product (L3_SM_P) which combines swaths available in each day. The spatial resolution of L3_SM_P is 36 km. For inputs, we obtained atmospheric forcing data including precipitation, temperature, radiation, humidity and wind speed from North-American Land Data Assimilation System phase II (NLDAS-2) (Xia et al., 2015). NLDAS-2 also provides, from 1979 to present, simulations of land surface states and fluxes by several LSMs. We chose Noah’s (and also compared with MOS’s) outputs (Ek, 2003) because it ranks in the middle among models (Xia et al., 2015) and is not as extensively calibrated as some other models, e.g., SAC. Our work does not require the best LSM, as we can observe how LSTM and other methods correct LSM errors. Noah has 4 soil layers which are of depths 0-10, 10-40, 40-100 and 100-200 cm, respectively. To match with the 0-5 cm sensed by SMAP, we tested: (i) directly using 0-10 cm data; (ii) polynomial interpolation; and (iii) integral interpolation.
where we find polynomials whose integrals agree with Noah-simulated values.

Static physiographic attributes (Table S1 in SI) include sand, silt and clay percentages, bulk density and soil capacity from ISRIC-WISE (Batjes, 1995). County-level annual-average irrigation data for 2010 (USGS, 2016) was overlaid with landuse data to assign irrigation in each county to agricultural land uses. The values are then aggregated to SMAP grid. Also among inputs are SMAP product flags that indicate mountainous terrain, land cover classes, VWC, urban area, water body fraction and data quality (time-averaged). SMAP product flags indicate lower data quality in dense forested areas. However, instead of removing all regions labeled as low-quality, we hypothesize that including the flags as inputs allows LSTM to implicitly assign less focus to high-uncertainty regions.

4.3.2 LSTM setup

As a type of Recurrent Neural Network (RNN), LSTM makes use of sequential information by updating states based on both inputs of the current time step \( x(t) \) and network states of previous time steps, as illustrated in Figure S1 in Supporting information (SI). Following the notation in Lipton et al. (2015), we can write an LSTM as \( \text{LSTM} : x(t), h(t-1), s(t-1) \rightarrow h(t), s(t) \). The sequential update formula are:

\[
\begin{align*}
\text{(input node)} & \quad g(t) = \tanh(W_{gx}x(t) + W_{gh}h(t-1) + b_g), \\
\text{(input gate)} & \quad i(t) = \sigma(W_{ix}x(t) + W_{ih}h(t-1) + b_i), \\
\text{(forget gate)} & \quad f(t) = \sigma(W_{fx}x(t) + W_{fh}h(t-1) + b_f), \\
\text{(output gate)} & \quad o(t) = \sigma(W_{ox}x(t) + W_{oh}h(t-1) + b_o), \\
\text{(cell state)} & \quad s(t) = g(t) \odot i(t) + s(t-1) \odot f(t), \\
\text{(hidden gate)} & \quad h(t) = \tanh(s(t) \odot o(t)), \\
\text{(output layer)} & \quad y(t) = W_{hy}h(t) + b_y,
\end{align*}
\]

where \( \sigma \) is the sigmoidal function, \( \odot \) is element-wise multiplication, \( x(t) \) is the input vector (forcings and static attributes) for the time step \( t \), \( W \)'s are the network weights, \( b \)'s are bias parameters, \( y \) is the output to be compared to observations, \( h \) is the hidden states, and \( s \) is called the cell states of memory cells, which is unique to LSTM. Readers are referred to the literature for the detailed functionality of
these units. Summarized briefly, \( i, f, o \) control, respectively, when the input is significant enough to use, how long the past states should be remembered for, and how much the value in memory is used to compute the output. During training, \( W' \)’s and \( b' \)’s are adjusted using back-propagation through time (BPTT). In BPTT, the network is first unrolled over a prescribed length before the difference between the output and target propagates into the network. We used the LSTM implemented in Recurrent Neural Network library for Torch7 (Léonard et al., 2015), which is a scientific computing framework for the programming language Lua. We employed Dropout regularization, which randomly sets a fraction (dropout rate, \( dr \)) of its operand to 0. Dropout prevents the co-adaptation of neurons and thus reduces overfitting. We used dropout regularization to non-recurrent links as in Zaremba et al. (2015), a constant dropout mask to recurrent connections as in Gal and Ghahramani (2015). We also implemented dropout for the memory cell as described in Semeniuta et al. (2016).

At each time step, the network outputs one scalar value \( (y^{(t)}) \), which is compared to SMAP L3 passive product. The loss function to be minimized is the mean-squared error calculated for the time series:

\[
L = \frac{1}{\rho} \sum_{t=1}^{\rho} 1_{o}(t)[y^{(t)} - y^{*(t)}]^2, \tag{4.8}
\]

where \( 1_{o}(t) \) is 1 when time step \( t \) has SMAP observation and is 0 otherwise, \( \rho \) is the length of the time series, and \( y^{*(t)} \) is SMAP observation. For computational efficiency and stability reasons, the training is done through "mini-batches": for each batch, a number of instances, or SMAP pixels, are randomly collected from the training set. The loss function is then averaged over all the instances in a batch.

4.3.3 Tests, Conventional Algorithms, and Evaluation Metrics

In our temporal generalization test, the training set is SMAP data from April 2015 to March 2016. For computational efficiency, we picked 1 pixel from every 4 x 4 patch, resulting in a 1/16 coverage of CONUS. The test set is SMAP data for the same pixels, but for the period from April 2016 to March 2017. In the regular
spatial generalization test, the training set is the same as described above, but the
test set is the neighboring cells for the same period. In the regional holdout test, we
trained models over 4 of the 18 2-digit Hydrologic Cataloging Units (HUC2s) and
tested on others. This test challenges the ability of different methods to generalize
across characteristically different climates and physiographic conditions. There
are a large number of such 4-HUC2 combinations. As an initial investigation, we
chose 4 of such combinations (C1-C4). Two combinations have a broad coverage
of the range of Noah’s bias over CONUS, while the other two cover only part
of that range. These tests inform us about the effect of biased training sets on
generalization.

LSTM predictions are compared to three conventional methods: the least ab-
solute shrinkage and selection operator (lasso), auto-regressive model (AR), and
a single-layer feedforward Neural Network (NN), given the same inputs. Lasso,
shorthanded as LR here, is multiple linear regression with a regularization term
that penalizes large regression coefficients (Tibshirani, 1994). NN can construct
nonlinear transformations of inputs, but does not have memory, and therefore cannot retain time dependencies. LR and NN are operated in (1) the CONUS-scale
mode, where a single model is trained for the entire training set; and (2) point-by-
point mode, with subscript "p", where a separate model is trained for each pixel.
AR models are trained only point-by-point (ARp). More details are provided in
SI Text S1. Three statistical metrics, bias (the time-averaged difference), root-
mean-squared error (RMSE) and Pearson’s correlation (R) are calculated between
between predicted and SMAP-observed soil moisture on training and test sets sep-
arately. R measures the agreement between simulated and observed climatology.

While short-term forecast employs observations to continuously update solution, long-term hindcast has no observations to use. As a "proof-of-concept" test
of LSTM’s appropriateness for long-term hindcast, we trained LSTM and ARp using 2 years of Noah-simulated soil moisture as the target, to hindcast to 10 years back.
4.4 Results

4.4.1 Overall test performance

For the temporal generalization test, we note substantial improvement with respect to both bias and $R$ compared to Noah (Figure 4.1). We report results from directly using the 0-10 cm Noah layer, although other choices are similar, as will be discussed later. The Noah solutions have a significant, spatially-varying bias in many parts of CONUS, as shown in Figure S3a in SI, especially in Southeastern Coastal Plains (annotated in Figure S2 in SI). The LSTM correction reduces the bias by an order of magnitude, and mostly removed the spatial pattern of bias (Figure 4.1a). We note there is a CONUS-scale spatial trend of larger reduction of absolute bias in the eastern CONUS, except the Southeast Coast (Figure 4.1b). The gradient appears to be related to the CONUS annual precipitation map, as it corrects Noah’s bias to over-estimate in the arid West while over-estimate in the humid East (Figure S3a, also noted in (Xia et al., 2015)).
Figure 4.1. Performance of LSTM (evaluated against SMAP) in the test set of the temporal generalization test. (a) $bias(LSTM) = LSTM - SMAP$. Each pixel in this figure is patch of 4x4 SMAP pixels. Bias in most parts of CONUS is between -0.02 and 0.015; (b) Change of absolute bias due to LSTM correction. LSTM reduces the absolute bias significantly over CONUS; (c) LSTM anomaly correlation ($R(LSTM)$); (d) change of $R$ due to LSTM correction; (e) $RMSE(LSTM)$; (f) Since the RMSE improvement over Noah looks similar to panel b, here we show the difference in RMSE between LSTM and NN predictions. Noah’s performance is provided in Figure S3 in SI.
LSTM does not only reduce bias. The improvement in predicted climatology is much more noticeable, according to $R$ (Figure 4.1c,d). LSTM $R$ is mostly above 0.8 and 50% pixels are over 0.9, significantly above Noah. In most CONUS the $R$ improvement is greater than 0.1, while it can be 0.4 ∼ 0.6 in the Eastern CONUS, especially the agricultural regions in Central Lowland and on the Appalachian Plateau (Figure 4.1d). We note this map is no longer similar to the bias map of Noah, suggesting mechanisms that correct seasonality are different from those correcting bias. We hypothesize LSTM significantly improves soil moisture dynamics in agricultural regions, e.g., irrigation, and the influence of shallow soils on the Appalachian highlands (Fang et al., 2017). On the other hand, over the majority of CONUS, the RMSE of LSTM is lower than 0.035 (Figure 4.1e). A continental-scale west-to-east increasing trend in RMSE(LSTM) is apparent. The higher errors in the East may result from higher annual precipitation, which results in (i) higher annual-mean soil moisture, and (ii) high VWC, which reduces SMAP data quality. However, the Southeast regions facing the Atlantic has a medium-to-low RMSE(LSTM). Figure 4.1f suggests the improvement of LSTM over the one-layer NN is obvious, especially in the central lowland region and coastal plains.

### 4.4.2 Comparison of generalization capability with other methods

In the temporal generalization test, time series prolonged by LSTM compares favorably against AR$_p$ across a wide range of LSTM performance levels $R(LSTM)$ (Figure 4.2). For the 10-th to even 75-th percentile pixel, LSTM is able to closely follow SMAP, except that peaks are under-estimated in the 50-th percentile pixel in 2016. The frequent rain events in April-May 2016 in Figure 4.2a and their recessions are well captured. For the 75-th percentile pixel, all peaks are captured, but we notice some over-estimation near the troughs in August 2016. In contrast, while AR$_p$ is also behavioral, we notice it often noticeably under-estimates the troughs, over-estimates the seasonal rising limbs and overshoots some peaks. In the 10-th percentile cell, AR$_p$ performs poorly between October 2016 and early 2017.
Figure 4.2. Comparisons between SMAP observations and soil moisture predicted by LSTM, Noah, and ARp at 5 locations. We chose sites around 10-th, 25-th, 50-th, 75-th and 90-th percentiles as ranked by $R(LSTM)$.

Summarized over CONUS, LSTM shows the lowest test RMSE and bias, and the highest $R$ (Figure 4.3), followed by NN, LRp, ARp, LR, NNp and lastly Noah. Neither the vertical interpolation procedure nor the choice of LSM (MOS or Noah) has much impact on LSTM’s prolongation performance (see Figures S4 and S5 in SI). The test RMSEs of LSTM are 0.022, 0.027, 0.036 and 0.057 for the 25-th, 50-th, 75-th and 90-th percentile pixel, respectively (Figure 4.3a). With lasso regularization, LR has similar training and test RMSEs, but its 25-th percentile test RMSE is similar to the 75-th percentile of LSTM’s. Therefore, the more complex relationships permitted by LSTM are beneficial. The LRp improves from LR as it specializes in each pixel, and the lasso regularization appears to have prevented overfitting, but its error is still larger than the CONUS-scale LSTM. NNp and ARp appear more overfitted than LRp. LSTM’s test bias is only moderately smaller than that of NN, ARp and LRp, but $R$ is much higher. 75% and 50% of $R(LSTM)$ are greater than 0.80 and 0.87, respectively.
Figure 4.3. Boxplots comparing LSTM, Noah, NN, AR, LR, $NN_p$ and $LR_p$ in the temporal generalization test and the regular spatial generalization test. Each box and whisker element summarizes SMAP pixels over CONUS with percentiles annotated in the panel. Y-axis limits truncate Noah boxes to focus on the central part of other boxes. Left column: temporal generalization test; Right column: regular spatial generalization test. The three rows are for RMSE, Bias and $R$, respectively. A paired t-test suggests $\text{RMSE}(\text{LSTM})$ and $\text{RMSE}(\text{NN})$ are significantly different.

Note $AR_p$ has sub-par performance in both training and test periods. The test RMSE box for $AR_p$ is wider, suggesting its formulation works well for some pixels but not so well in others. Furthermore, the extended proof-of-concept long-term hindcast experiment shows a similar contrast. LSTM has robust prolongation performance at a 10-year hindcast scale while $AR_p$ generates larger errors. Errors
for both methods are independent of hindcast lengths, i.e., 10-year-prior hindcast error is not much different from 2-year hindcast (Text S2 and Figure S6 in SI). Meanwhile, in the regular spatial generalization test, LSTM again exhibits the smallest RMSE and bias (Figure 4.3c-d). The contrast in bias is smaller than the temporal test, but the $R$ comparisons are similar.

In 4-HUC2 combinations 1 and 2 (C1 and C2), the Noah bias covers a wide range from -0.25 to 0.15 cm$^3$/cm$^3$, which appears to be the whole range of the Noah biases we see over CONUS. In both cases, LSTM is able to greatly reduce the bias and improve soil moisture climatology (much higher $R$) compared to both NN and Noah (Figure 4.4a-b). The boxes corresponding to LSTM bias are very narrow, and its centers are nearly 0. For the case C3, we note that it has few points with bias <-0.2, so for this HUC2 combination, the training set under-samples the Noah errors that lead to strong negative biases. As a result, LSTM’s bias is no longer near 0, although still much better than NN’s and Noah’s. On the other hand, for C4, the training set is strongly biased. It lacks any basin with a Noah bias of <-0.1. We note the narrow box corresponding to Noah’s bias in the training set for C4. Unsurprisingly, LSTM’s performance deteriorates: LSTM is no longer able to correct the bias, and its range of bias is large. NN, similarly, also fails to correct the bias. We obtained LSTM’s self-assessment of Noah bias by subtracting Noah’s solution from LSTM’s prediction. The self-assessed bias (Figure 4.4a) has little overlap with the Noah bias in the training set. This may be a signal we can utilize in the future to identify biased training sample and large predictive uncertainty.
Figure 4.4. (a) Test biases for cases C1-C4 (distinguished by different colors), which are four combinations of 4 HUC2s. Each group consists of 4 boxes, which are, respectively, LSTM, NN, Noah in the training set, and self-assessed test bias. The self-assessed bias is Noah’s prediction over the test set minus LSTM’s predictions; (b) test $R$. We note significantly better seasonality captured by LSTM in cases C1-C3, but not necessarily in C4; (c) Distributions of Noah’s biases in the training sets for C1-C4.

4.5 Discussion

In many parts of CONUS, LSTM’s RMSE is smaller than SMAP’s design measurement accuracy. It appears even 1 year of CONUS-scale data, when grouped together, has enough information to train an LSTM to hindcast SMAP data. A
factor that contributes to such performance is the short memory length of soil moisture, which was found to range between 5 to 40 days (Orth and Seneviratne, 2012) in previous work. As a result, two years of data, when grouped together, contain many complete wet-dry cycles. The hindcast quality should improve as SMAP data increases. On a side note, because true random noise cannot be predicted in the test set, it follows that SMAP L3’s RMSE could be below 0.027 in 50% of CONUS. Also, the official SMAP data quality flag labels the forested Southeast Coastal Plains and South Atlantic regions as "not recommended" quality (C.2). Our LSTM has a RMSE of 0.02-0.035 there, which suggest SMAP may be functional in these regions, but the impacts of the retrieval algorithm should be carefully examined.

It seems non-recurrent NN can already remove a large part of bias by capturing how environmental factors lead to certain type of biases. However, NN cannot maintain time dependencies, which may explain its performance difference from LSTM. Therefore, we argue an advantage of LSTM originates from its recurrent nature. Meanwhile, alternative recurrent models, e.g., AR and moisture loss functions (Koster et al., 2017), are useful due to their interpretability, parsimony and great value in 'nowcasting' or short-term forecasting (see Koster et al. (2017) for a robust application), but they require continuous updates by observations to avoid drift from true dynamics. At one-year scale, injected data already has little effects on hindcast solutions. For longer-term hindcast, pattern-based methods like LSTM appear to be more suitable.

Previous soil moisture comparisons mainly focused on anomalies, but the prevalent bias with Noah-simulated moisture can introduce large errors to downstream applications such as weather modeling (Massey et al., 2016). The continental-scale bias pattern suggests some systematic errors with Noah’s model structure/parameters. Some hypotheses include (i) Noah’s soil pedo-transfer functions are fundamentally inadequate in resolving regionally heterogeneous soil responses to rainfall, which could explain the need for calibration in most large-scale flood forecasting systems; or (ii) groundwater flow, which is important in thick-soiled, high-infiltrating capacity regions like the southeast (Fang et al., 2016), is not properly simulated in LSMS (Clark et al., 2015). However, LSTM appears to be able to integrate information from raw data and compensate for the inadequate representation uniformly over CONUS.
Conventional statistical wisdom suggests that simpler models are more robust and models with high degrees of freedom may be easily overfitted. However, the present work shows the CONUS-scale deep learning networks have smaller test errors than three alternative methods trained point-by-point. In fact, we hypothesize an important strength of LSTM originates from its flexibility to simultaneously learn from a large and heterogeneous collection of data and identify commonalities and differences. Its generalization capability stems from building internal models (in the attribute space) to capture biases and temporal fluctuations. Despite widely different climates, topography, landcover and soils over CONUS, just a small subset of these combinations seem sufficient for training accurate LSTM models, provided the sample is not too biased, e.g., C1-C2 in Figure 4.4.

4.6 Conclusion

We have trained a CONUS-scale LSTM network to predict SMAP data. This network is capable of correcting spatially-heterogeneous model bias as well as climatological errors between Noah-simulated and SMAP-observed top-surface soil moisture, creating a CONUS-scale seamless moisture product that has high fidelity to SMAP. Despite having high degrees of freedom, when properly regularized, LSTM exhibits better generalization capability, both in space and time, than linear regression, auto-regressive models, and a one-layer neural network. Its test error approaches the instrument accuracy limit with SMAP. LSTM will be helpful in long-range soil moisture hindcasting or forecasting, weather modeling, and data assimilation. Its generalization capability arises from building internal models from physical attributes and synthesis of climate forcing. It does not necessarily require similar examples in the training set. Unless the training set is strongly biased, LSTM has a good chance of success.

4.7 Limitations and Future Work

As a first paper using LSTM in hydrology, this work is by no means a thorough or optimized investigation. Our work does not address the question about the accuracy of SMAP data, which is addressed by other studies. The hindcast performance with respect to capturing soil moisture during drought should be further
examined with *in-situ* data. We should further assess LSTM’s performance in comparison with regionally-trained models. The implications of low LSTM RMSEs in forested region warrants further investigations.

### 4.8 Author Contributions

K. F. collected and processed the data, implemented the computer codes, analyzed and visualize the results, and draft the manuscript. C. S. conceived and directed the study, and critically revised the article. D. F. provided general instruction of the algorithm architecture. X. Y. helped with the code implementation and algorithms that were compared against DL. All authors contributed substantially to the study design and final approval of this version to be published. K. F. is the first author and has contributed the majority of work.
Bibliography


Chapter 5  
The value of SMAP for long-term soil moisture estimation with the help of deep learning


5.1 Abstract

The Soil Moisture Active and Passive (SMAP) mission measures important soil moisture data globally. SMAP’s products might not always perform better than land surface models when evaluated against in-situ measurements. However, we hypothesize that SMAP presents added value for long-term soil moisture estimation in a data fusion setting as evaluated by in-situ data. Here, with the help of a time series deep learning (DL) method, we created a seamlessly-extended SMAP dataset to test this hypothesis and, importantly, gauge whether such benefits extend to years beyond SMAP’s limited lifespan. We first show that the DL model, called Long Short-Term Memory (LSTM), can extrapolate SMAP for several years and the results are similar to the training period. We obtained prolongation results with low performance degradation where SMAP itself matches well with in-situ data. Inter-annual trends of root-zone soil moisture are surprisingly well captured
by LSTM. In some cases, LSTM’s performance is limited by SMAP, whose main issue appears to be its shallow sensing depth. Despite this limitation, a simple average between LSTM and a land surface model Noah frequently outperforms Noah alone. Moreover, Noah combined with LSTM is more skillful than when it is combined with another land surface model. Over sparsely instrumented sites, the Noah-LSTM combination shows a stronger edge. Our results verified the value of LSTM-extended SMAP data. Moreover, DL is completely data-driven and does not require structural assumptions. As such, it has its unique potential for long-term projections and may be applied synergistically with other model-data integration techniques.

5.2 Introduction

Soil moisture (θ) critically controls various environmental and ecosystem processes, such as photosynthesis, evapotranspiration, runoff, soil respiration (Orchard and Cook, 1983), flood (Norbianto et al., 2008), land-atmosphere interactions (Koster, 2004), etc. For agricultural planning and other purposes, soil moisture is routinely measured by in-situ monitoring networks. However, these networks have limited coverage in space and are typically sparsely instrumented. Soil moisture has a high spatial variability so that the scale of these measurements poorly matches the needs of large-scale modeling and monitoring.

To offer large-scale monitoring of soil moisture, multiple soil moisture sensing satellite missions have been launched. Notable missions include the Advanced Microwave Scanning Radiometer for EOS (AMSR-E) (Njoku et al., 2003), the Advanced Scatterometer (ASCAT) (Wagner et al., 1999), the Soil Moisture and Ocean Salinity (SMOS) (Kerr et al., 2010), and the Soil Moisture Active Passive (SMAP) (Entekhabi et al., 2010), among others. The data from these missions have been utilized for a variety of applications in hydrology, ecology, meteorology, and agriculture (Dorigo et al., 2017). Among these missions, SMOS was reported to suffer significantly from unexpected radio frequency interference (RFI) (Oliva et al., 2012, 2016). Launched in the year 2015, SMAP mitigates the RFI effects by utilizing the combination of spaceflight instrument hardware and ground-based science processing algorithms (Piepmeier et al., 2014). Initial evaluations suggest that SMAP’s passive product has reached its design accuracy (≈ 0.04) at most validation sites.
In some parts of the world, the quality of SMAP product might not exceed the modeled soil moisture from state-of-the-art LSMs forced with high-quality meteorological data (Pan et al., 2016; Karthikeyan et al., 2017). However, SMAP provides an independent source of observation derived from a different physical mechanism (radiative transfer). Independent information often helps reducing errors and uncertainties in a data fusion setting (Pan and Wood, 2010).

While SMAP has a global coverage, its uneven revisit schedule and short timespan significantly limit its utility. Longer records are needed to study drought trends or relate soil moisture to past environmental events, e.g., wildfires and landslides. Satellite data can also improve the current soil moisture simulations via data assimilation (DA) (Koster et al., 2018; Liu et al., 2011), but DA has little impact on long-term soil moisture estimation. Previously, Liu et al. Liu et al. (2012) merged multiple satellite products by matching earlier satellite data’s cumulative distribution function (CDF) to AMSR-E’s. They also replaced older product’s seasonality based on AMSR-E’s when overlaps exist. Later work further weighed different satellite data based on their random error variance (Dorigo et al., 2017; Gruber et al., 2017). While this series of work provided valuable satellite soil moisture estimates back to 1979, the data quality is still affected by the retrieval skills of earlier satellites. The retrieval skills have been much improved in the more recent, dedicated soil moisture missions (SMOS and SMAP) (Karthikeyan et al., 2017). As some earlier satellites have limited or no overlap with SMOS or SMAP, it is difficult to apply data fusion techniques such as CDF-matching with SMOS or SMAP.

In our earlier work, we developed a deep learning-based dynamical modeling system, which predicts the SMAP surface soil moisture product with high fidelity (Fang et al., 2017). Deep learning (DL) refers to large-sized neural networks with the ability to automatically extract features using its intermediate layers (LeCun et al., 2015). DL is rapidly transforming various industries and scientific disciplines (see a review in (Shen et al., 2018)). The inputs to our model include atmospheric forcing data, static physiographic attributes, and (optionally) simulated fluxes and states from land surface models (LSMs). The model was built on a form of deep recurrent neural network called the Long Short-Term Memory (LSTM) architecture, which is typically employed for natural language processing. We showed that, when
trained with a year of data, this system was capable of predicting SMAP Level 3 passive data product in another year. LSTM can extend SMAP to spatiotemporally seamless coverage of continental US (CONUS) with high fidelity to SMAP. Despite its large parameter space, the system had higher accuracy and better robustness than simpler statistical methods. We called our method a prolongation of SMAP via deep learning. However, it must be noted that DL only captures the systematic soil moisture response pattern as observed by SMAP. It cannot extend the stochastic signal in satellite observations that could be used to correct forcing errors.

Our main hypothesis is that LSTM-prolonged SMAP data presents added value beyond LSMs and combining it with LSMs will help improve the long-term soil moisture predictions, as verified by in-situ data. Additionally, we seek to answer several important technical questions. First, we study whether the DL model can perform well in a multi-year scale projection setting. We check if it can avoid performance degradation over time like data assimilation (DA) would. When the training data only contain two or three years of records, it is unclear whether multi-year trends can be learned and predicted in case they exist. Second, all the previous evaluation metrics of LSTM were calculated against SMAP data. The DL system has yet to show that it is reasonably close to the in-situ data. If SMAP data have significant noise, the extent to which such errors would misinform the training process and amplify during long-term simulation has not been explored. Finally, SMAP also offers a root-zone soil moisture (RZSM, 0-100 cm depth) product via assimilating surface brightness temperature into model simulations (De Lannoy and Reichle, 2016). Past evaluations show that the RZSM product has achieved the design accuracy of 0.04 cm³/cm³ (Reichle et al., 2017). We ask whether LSTM can also reproduce RZSM, which has a much longer system memory.

This paper uses LSTM to hindcast SMAP surface and root-zone the soil moisture products. In contrast to our earlier work (Fang et al., 2017), here we (1) prolonged the SMAP data to several years beyond the available data to test the model’s ability to maintain good performance in long-term estimation, (2) additionally evaluated the root-zone soil moisture hindcast product, (3) combined LSTM with Noah to test our main hypothesis above, (4) evaluated different products against in-situ data instead of against SMAP, and (5) removed Noah simulation from the list of inputs. The in-situ soil moisture data include both SMAP core
validation sites (Jackson et al., 2012), where dense measurements provide a means of upscaling, and a network of sparsely instrumented sites. We examined whether some important characteristics of data, e.g., inter-annual trends, are reproduced in the hindcasts. Here, we chose to test the hindcast product because it can be evaluated using in-situ data for its long-term behaviors. Given the future climate projections, a similar system can be similarly employed to make future soil moisture projections.

5.3 Methods and Data Sets

The surface and root-zone LSTM models were both trained from April 2015 (SMAP launch) to April 2018. Inputs to the deep learning models include atmospheric forcings and constant geophysical attributes (Figure 5.1). We then ran the trained models in forward mode using atmospheric forcings from 2010 to 2018. The hindcast products are validated over SMAP core validation sites (from 2012) and sparse networks (using available data since 2010) over CONUS.

Figure 5.1. Information flow diagram for the models compared in this paper. Both Noah and LSTM use forcing data from North American Land Data Assimilation System (NLDAS). LSTM uses SMAP as the main training target along with land surface characteristic masks from the SMAP mission among the inputs. The static input attributes to LSTM are described in Section 5.3.1.
5.3.1 Datasets: Targets and Predictors

We used the SMAP Level 3 radiometer product (L3SMP, version 4) as the target of surface soil moisture. This product is based on L-band passive observations of surface brightness temperature and was designed to represent top 5 cm soil moisture on a 36 km Equal-Area Scalable Earth Grid (EASE-Grid). As the root zone soil moisture cannot be directly sensed by satellite, SMAP level 4 soil moisture product (L4SM, version 3) integrates SMAP L-band brightness temperature observations into the Catchment land model in Goddard Earth Observing System (GEOS-5) (Koster et al., 2000; Ducharne et al., 2000) through data assimilation. The model outputs column soil moisture of roughly 0-100 cm depth on a 9 km EASE-Grid, which is used as the target of our root-zone LSTM model.

LSTM receives three kinds of inputs, which have been selected from the list of inputs used in Fang et al. (Fang et al., 2017) via sensitivity analysis. First, atmospheric forcings, including precipitation, temperature, long-wave and short-wave radiation, specific humidity, and wind speed, were obtained from North American Land Data Assimilation System phase II (NLDAS-2) (Xia et al., 2012). Second, for static physiographic attributes, we obtained soil properties like sand, silt and clay percentages, bulk density, and soil water capacity from World Soil Information (ISRIC-WISE) database (Batjes, 1995). The third type of inputs are SMAP masks that indicate mountainous terrain, ice, surface roughness, urban areas, water bodies, land cover classes, and vegetation density. Incorporating these masks as inputs is hypothesized to enable LSTM to implicitly assign less importance to high-uncertainty regions. LSTM is thought to automatically analyze, based on the data, which mask is useful and discards the unimportant masks. In our earlier work (Fang et al., 2017), we included time-averaged SMAP data quality mask as an input. However, since our preliminary sensitivity test suggests this field was not influential, we did not include it in the model reported here. Its information has already been encompassed in the other inputs.

NLDAS-2 provides several land surface models which simulate land surface states and fluxes. We selected Noah for comparison and model fusion purposes (Ek, 2003). Since here one of our objectives is to measure the value of the independent information from LSTM, we did not include the Noah-simulated variables as LSTM inputs.
5.3.2 The time-series deep learning model

We trained the same time-series deep learning model proposed in Chapter 4 from 2015/04 to 2017/04 to learn SMAP dynamics. For the surface moisture product, we trained a network with the SMAP data over CONUS at the 36-km resolution on SMAP EASE-Grid, forming a training set of 6321 pixels. The time series for each SMAP pixel was split into multiple training instances for the network. For the 9-km resolution root-zone product, we took one pixel out of every 4 \times 4 pixels to obtain a training set of 7365 pixels. Such spatial down-sampling was performed for the sake of computational efficiency. The training was conducted on NVidia 1080 Ti Graphical Processing Units.

5.3.3 Validation: SMAP core validation sites, sparse network, and evaluation metrics

Due to high spatial heterogeneity inherent with soil moisture (Famiglietti et al., 2008), the point-scale measurements cannot be directly compared to the SMAP data. SMAP partnered with in-situ networks that have sufficiently dense soil moisture sensors to be reliably aggregated to a larger spatial scale. These sites are called core validation sites (Jackson et al., 2012). Nine core sites are distributed over CONUS and each of them contains multiple pixels matching the SMAP footprint. Consistent with past calibration and validation procedures, we created Thiessen polygons for measurement sensors on validation pixels, and in-situ observations are the average of all the sensors using weights determined by the fraction of area covered by their Thiessen polygons. As sensors may stop working temporarily, especially under cold weather, we only retain the in-situ records when at least half of the pixel is covered by the Thiessen polygons of the operational sensors. We also removed the records when the measured in-situ soil temperatures were lower than 4°C, as sensors’ measurements are not reliable under cold weather.

We also employed the U.S. Climate Reference Network (CRN) (Bell et al., 2013), which are sparser, to validate the models. CRN sites were first deployed in 2000 and completed full deployment of 114 sites in 2011. We removed ten coastal sites because they are not covered by SMAP footprint or the SMAP signals are missing for more than 75% days on the corresponding pixel. Out of 104 CRN sites employed in validation, 82 installed soil moisture probes at standard WMO levels
(5, 10, 20, 50, and 100 cm), whereas 22 of them were instrumented only at the top two layers. Both types of sites were used for the evaluation of surface soil moisture, while only those with deeper sensors were used for evaluating RZSM. At each CRN site and every point of depth, three probes were planted 1.5 meters away from the base at 0°, 120°, and 240° compass direction, and the soil moisture measurement employed was the average of those three sensors.

We evaluated three products: LSTM, Noah, and a simple average of LSTM and Noah, denoted by “Comb.” Ideally, to fuse the two datasets, we can obtain both of their variances. We can weigh their values using the inverse of these variances during averaging, i.e., using a Bayesian model averaging scheme (Hoeting et al., 1999) or instrumental variable regression (Crow et al., 2015). However, at this stage, the uncertainty estimation methods for the DL model are not developed enough for such an application. The simple average here only represents an initial evaluation of the value of LSTM-extended SMAP data. Better schemes can certainly be evaluated in the future.

For each site, we calculated several metrics for two pairs of data: LSTM-predicted vs. in-situ data and SMAP vs. in-situ data. The metrics included the root-mean-square error (RMSE), bias (the difference between the means), the Pearson’s correlation coefficient, R, and the unbiased RMSE (ubRMSE), which is RMSE calculated for two time series after means have been subtracted. The latter two metrics were unaffected by model bias. We calculated the metrics for the training period (2015/04/01 - 2017/04/01) and test period (site’s operation start time - 2015/04/01) separately. When added, the \( AL \) superscript denotes the after-launch training period, while the \( PL \) superscript denotes the pre-launch test period. Also, \( \Delta \text{ubRMSE} \) denotes the difference in ubRMSE between training and hindcast periods, i.e., \( \text{ubRMSE}^{PL}(\text{LSTM, in-situ})-\text{ubRMSE}^{AL}(\text{LSTM, in-situ}) \). It must be noted here again that the RZSM product results from assimilating SMAP brightness temperature into the Catchment land model with GEOS-5 FP forcing data (Reichle et al., 2014). Therefore, it is not as independent from model dynamics as the L3 product. Simultaneously, the innovations still take into account the actual SMAP observations.

We examined how inter-annual trends observed in in-situ networks were captured by LSTM. We focused our trends analysis on the root-zone product, as it has longer memory than the surface soil moisture. The trend comparison was calcu-
lated for the maximum period of available data since 2005, which varies from site to site. For RZSM, we performed the non-parametric Mann-Kendall test \cite{Helsel2002} and computed the Sen’s slopes using the longest available data. The starting date of the data between 2000 and 2011. The Mann-Kendall test examines the null hypothesis that there is no trend in the data.

For RZSM, the SMAP root-zone product represents the accumulated water content in 0-100 cm, while in-situ probes are located only at several depths. Following the guidelines mentioned in \cite{Reichle2017}, we extrapolated the in-situ measurements using weights that were proportional to sensor depths within the root-zone layer. Soil moisture was discretized using the moisture probes as layer centers, and the probe-measured moisture was used as the mean value for the corresponding layer. For example, if a station has sensors at depth $d_1, d_2, \ldots, d_n$ cm, the weight of those sensors will be proportional to $$(d_1 + d_2)/2, (d_2 + d_3)/2 - (d_1 + d_2)/2, \ldots, 100 - (d_{n-1} + d_n)/2.$$ There are only five core sites that have soil moisture measurements at multiple depths. For the CRN network, we selected sites that had soil moisture sensors at all design depths (5, 10, 20, 50, and 100 cm). Seventy-two sites remained after removing the ones with more than 50% of missing measurements or less than one year of records before the SMAP launch.

\section*{5.4 Results and Discussion}

\subsection*{5.4.1 Evaluation at the core sites}

\subsubsection*{5.4.1.1 Statistical comparison between training and hindcast periods}

We first focus on the comparison of LSTM error statistics between the training period (yellow bars in Figure 5.2) and the hindcast period (red bars) for the surface product. For most of the sites, we found only a small increase in ubRMSE from the training to the hindcast period (Figure 5.2b). Five sites showed ubRMSE$^{AL}$(SMAP, in-situ) values that are $\leq 0.04$, and for all of these sites, ubRMSE$^{PL}$(LSTM, in-situ) are also $\leq 0.04$. For more than half of the sites, $\Delta$ubRMSE is less than 0.006. At Little River and TxSON, the LSTM-to-in-situ error actually reduces from training to test period, which is attributed to randomness and some missing data. Four sites have $\Delta$ubRMSE of more than 0.006: ubRMSE increases from
0.028 to 0.035 at Walnut Gulch, from 0.059 to 0.069 at Carman, from 0.038 to 0.048 at St. Josephs, and from 0.054 to 0.061 at South Fork.

**Figure 5.2.** Bias, ubRMSE and Pearson correlation of surface and root-zone soil moisture as evaluated by SMAP core-site in-situ data for the after-launch (AL) training period and the pre-launch (PL) validation periods. (next page)
Figure 5.2. (Previous page) Comb means the simple average of LSTM and Noah. In general, ubRMSE between in-situ data and SMAP (black) is slightly smaller than that between LSTM and in-situ during the training period (yellow), which is in turn smaller than that between LSTM and in-situ during the hindcast period (red). At sites where $R^{AL}(\text{SMAP, in-situ})$ is higher than $R^{AL}(\text{Noah, in-situ})$ (yellow higher than cyan), LSTM or Comb can have a higher $R$ than Noah during hindcast, with Fort Cobb being an exception. At other sites, Noah is stronger than SMAP and LSTM, but Comb might still be better.

As it is normal for the test error to be larger than the training error, a small and positive $\Delta_{\text{ubRMSE}}$ is to be expected. We note that when SMAP itself has a relatively large disagreement with the in-situ data (>0.04), as in Carman, St. Josephs and South Fork, $\Delta_{\text{ubRMSE}}$ also tends to be larger. Presumably, when SMAP has a large difference from the in-situ data, there might be some systematic deviations at those sites, e.g., the cold season behaviors do not match the in-situ data as well, which is discussed next. When such discrepancies exist, there are less detectable patterns to learn and more noise to interfere with the training. As a result, the effectiveness of learning weakens. The SMAP-in-situ discrepancies mostly affect the temporal fluctuations, as witnessed by the degrading ubRMSE and anomaly correlation during the training, and it has little impact on the hindcast bias. In contrast, the bias between SMAP and LSTM appears to have little impact on $\Delta_{\text{ubRMSE}}$. If we examine Little River, which has a large positive bias, the ubRMSE and Pearson anomaly correlation in the hindcast period are as good as the training period. Despite the mismatch between SMAP and in-situ and a mild amplification of error, none of the increase in ubRMSE is greater than 0.011.

LSTM hindcast of root-zone soil moisture also shows similar error statistics to SMAP level 4 root-zone product. ubRMSEs for root-zone soil moisture are generally smaller than the surface product due to smaller soil moisture temporal variability in the root zone. Hindcast ubRMSEs at all the five sites are smaller than 0.04. For three sites (Little Washita, Little River, and South Fork), the $\text{ubRMSE}^{PL}(\text{LSTM, in-situ})$ is almost identical to $\text{ubRMSE}^{AL}(\text{SMAP, in-situ})$, but the correlation degraded at Little Washita and TxSon (Figure 5.2).
5.4.1.2 Statistical comparison of performances between LSTM, Noah, and Comb

We now switch our attention to the comparison between LSTM, Noah, and their averages. It should be recalled that (i) LSTM learns from SMAP. Thus, in theory, SMAP is the “performance ceiling” of LSTM when evaluated against in-situ data. (ii) SMAP senses a moisture-dependent depth that is shallower than 5 cm. For the surface soil moisture, We found that for four of the nine core sites, SMAP has a higher R than Noah during the training period (Carman, Fort Cobb, Little River, and St. Josephs) (compare black to cyan bars). Outside of these sites, LSTM has little chance of outperforming Noah. However, SMAP might still bring in new information to help improve the Comb. For the root-zone soil moisture, SMAP has advantages for only two of the sites (Little River and South Fork).

With the above information in mind, we found that while LSTM has limited chance to surpass Noah for some sites, Comb shows noticeable advantages over Noah. During the training period, Comb has a higher R than Noah at six out of nine core sites for the surface soil moisture. For the remaining three sites, Comb and Noah are also quite close. In contrast, LSTM only outperforms Noah at two sites (Little River and St. Josephs), where SMAP shows noticeable advantage (Figure 5.2). For the hindcast period, Comb is higher at five of nine sites. For the other four sites, Comb is similar to Noah. Comb’s advantage is more clear with the root-zone soil moisture. During the training period, Comb outperforms Noah at four of the five sites during the training period, even though LSTM only surpasses Noah at South Fork, which is the only site where SMAP’s root-zone product has a higher R. However, during the hindcast period, Comb is more skillful at all of the five sites, while LSTM is higher than Noah at three of the five sites.

It perhaps is not surprising that the ensemble model average works better than each of the individual models. However, we tested the average of Noah and the Variable Infiltration Capacity (VIC) model, which showed lower R than that of LSTM+Noah for six of the nine sites (Figure 5.3). Therefore, simply combining the models together does not guarantee better skills. Since LSTM not only helps reduce bias but also improves the temporal dynamics as reflected in R, the results suggest SMAP and LSTM have brought in unique and valuable information. In time series analysis (next Section), we examine how LSTM, despite being generally lower than R, offers this unique contribution.
Figure 5.3. Same as Figure 5.2, but here we compare the LSTM+Noah (called Comb in Figure 5.2) with Noah+VIC, which seems weaker than Noah+LSTM. We also show the test metrics between LSTM and SMAP when it is trained in one year and tested in another (AL 1 Yr Test).

The mixed comparison is in contrast to our previous analysis in Fang et al. (2017), where LSTM was always found to outperform Noah when SMAP was used as the benchmark. Indeed, when we trained LSTM in one year and test it in another, the 1-year test R(SMAP, LSTM), which describes the match between LSTM and SMAP in prolongation period, is mostly 0.05-0.1 higher than R (SMAP,
in-situ), making it the highest among all the bars for most of the sites (Figure 5.3). This comparison suggests while LSTM is sufficiently close to SMAP, there are some significant differences between SMAP and in-situ data, which limited the LSTM’s performance. These results show that the land surface model is a powerful simulator for the temporal dynamics of soil moisture, which confirms the earlier findings (Xia et al., 2014). Currently, it is difficult for a remote-sensing product to completely surpass Noah where high-quality meteorological forcings are available.

5.4.1.3 Time series comparisons

To gain further insights beyond the statistical comparisons, Figure 5.4 shows the time series comparisons of surface soil moisture at five core sites. In general, LSTM mimics the well-performing characteristics of SMAP as well as some systematic discrepancies concerning the in-situ data. For example, at Carman, LSTM inherits the high-frequency, high-amplitude fluctuations from SMAP. These fluctuations appear unphysical because much smaller fluctuations can be observed with Noah and in-situ data. Another example is, at Walnut Gulch, both the peaks and troughs are well captured by LSTM, but both the drydowns of SMAP/LSTM tend to be faster than the in-situ data (Figure 5.4). Similar faster drydowns can also be clearly observed with LSTM at Little Washita, Fort Cobb, and Little River. As noted previously, SMAP senses’ moisture between the surface and a moisture-dependent depth is often 5 cm or less, whereas the probes are often centered at 5 cm and measure soil water between 3.5 and 6.5 cm (Rondinelli et al., 2015). Moisture near the surface dries faster. Thus the lower sensing depths of SMAP could explain lower R with SMAP and LSTM.
Figure 5.4. The time series comparison of LSTM hindcast, SMAP, and in-situ soil moisture anomalies at six different sites. The means were subtracted from all the series. LSTM hindcast inherits the patterns of discrepancies between SMAP and in-situ data, e.g., faster drydowns for SMAP. At Carman, SMAP shows high-frequency oscillations, which disagree significantly with the in-situ data. LSTM shows the same oscillations. At St. Josephs, we note some periods of missing SMAP and in-situ data due to frozen ground or snow.

In contrast, the drydown in the Noah-simulated moisture is more in line with the in-situ data. This is most obvious at Walnut Gulch and Little Washita, where Noah has higher R than LSTM. It seems Noah’s simulation better matches the depths of the in-situ probes. However, at Walnut Gulch, we noted that Noah tends to predict too much drydown toward the end of the dry spell, while LSTM and
in-situ both have a flatter bottom. It is possible that soil water holding parameters in Noah were not correctly specified here. SMAP data were better able to capture such trends. This pattern is also why, at this site, averaging Noah with LSTM helps Comb outperform Noah.

When there are fewer reliable data points for winter in the North, LSTM is not well trained like the other periods. At St. Josephs, we witnessed the periods of missing SMAP data during the winter freeze. During this period, LSTM shows high oscillations. During hindcast, in-situ data could be missing in the winter as well. When there is snow on the ground, SMAP’s retrieval quality flag would indicate that a retrieval was attempted, but the quality cannot be guaranteed. The data quality might not be reliable during when snow is present. Also, the mission does not report soil moisture when it suspects that the ground is frozen.

Similar to the surface soil moisture hindcast, root-zone moisture hindcast also inherits the systematic discrepancies between SMAP and in-situ. At Little Washita, the soil moisture lows of in-situ data are lower than that of the SMAP’s. We see a similar pattern for the hindcast period (Figure 5.5). At Little River, LSTM and SMAP tend to overestimate the peaks. At South Fork, both the LSTM and SMAP tend to underestimate soil moisture peaks. In contrast, Noah could dry down too much as well as overestimate the peaks, in fact, more prominently than the surface soil moisture.
Figure 5.5. Time series comparison of LSTM hindcast, SMAP, and in-situ root-zone soil moisture on five core validation sites.

5.4.2 Evaluation over the sparse network

We evaluated LSTM and SMAP against CRN in-situ network over CONUS. As there is a scale disparity between a SMAP pixel and point-scale CRN measurements, we expect their differences to be significantly larger than those between SMAP and core sites. Indeed, the median ubRMSE between SMAP and surface in-situ (CRN) data as well as between LSTM hindcast and in-situ are now around 0.05, which is notably higher than that at the core sites (Figure 5.6 upper panels). The correlation with the surface soil moisture is slightly lower than that at the core sites. At some sites, where LSTM has high correlations, during the training period, R decreased during the test period, as evidenced by the lower 75-percentile box edge for the hindcast period. Despite this decrease, the median LSTM R values for all the CRN sites is above 0.72 for the surface hindcast, which is comparable
to the median of SMAP. For the root-zone data, LSTM only declines slightly from training to hindcast period, and they are both limited by SMAP’s relatively low R.

![Surface Soil Moisture Comparison of CRN Network](image)

![Rootzone Soil Moisture Comparison of CRN Network](image)

**Figure 5.6.** RMSE, Bias, unbiased-RMSE, and Pearson correlation of surface and root-zone soil moisture between CRN sites, LSTM, and SMAP.

Noah, in general, compares more favorably with CRN than LSTM and SMAP, again attesting to its good performance. However, Comb’s R is higher than that of Noah’s alone for the surface sites during both the training or the hindcast period. CRN probes are installed in triplicate redundancy at 5, 10, 20, 50, and 100 cm depths (sometimes only 5 and 10 cm are available). Thus, CRN sites have the same
sensing depth discrepancy from SMAP and in turn LSTM. However, integrating information from SMAP data may nonetheless help correct the model errors. For the root-zone soil moisture, it seems Noah has more better-performing sites, as it has a higher 75-percentile R.

The CRN sites can demonstrate spatial patterns of LSTM and SMAP performances and help us understand where the correlations have dropped from the training to the test periods (Figure 5.7a). LSTM surface hindcast resembles the spatial pattern as SMAP. However, we found several pockets of sites where LSTM is noticeably lower than that of SMAP’s. Test R for some northeastern sites (eastern ellipsoid in Figure 5.7b) have decreased significantly, i.e., below 0.5. These sites are located in cold and heavily-vegetated regions, where SMAP does not work properly due to large vegetation water content and snow. Thus, LSTM could not find useful relationships in this region. Another pocket of notable decrease as compared to SMAP lies at the boundaries between Utah, Colorado, and Wyoming (the western ellipsoid). Of these three, the two western sites are located on the Rocky mountains, while the easternmost site is on a plain east to the mountains. We think the large discrepancy here might be due to the poor quality of forcings in this region. For CONUS, NLDAS-2 precipitation integrates Climate Prediction Center gauge data and hourly radar data (NASA, 2018). This region has the lowest weather station density and yet a large spatial variability in rainfall due to orographic effects. All of these factors also challenge the radar precipitation sensing. Therefore, hindcast quality here is likely adversely impacted by lower forcing quality.
Figure 5.7. Map of Unbiased-RMSE and Pearson correlation of surface and root-zone soil moisture between in-situ, LSTM, and SMAP for periods before and after SMAP launch, on CRN in-situ network. LSTM generally has the same spatial pattern as CRN, but there are a few instances of a notable decline in R, as annotated on the figure.

5.4.3 Evaluation of inter-annual trends

Because SMAP has only three years of data, we employed both the CRN and the core sites to investigate the extent to which LSTM can reproduce the inter-annual trends with root-zone soil moisture. While we expect large differences between hindcast and CRN data due to scale disparity, their inter-annual trends should be more comparable. Many of these sites showed only small trends (with an absolute
value of less than 0.5% per year), and LSTM mostly predicts similarly small trend magnitudes (Figure 5.8). When the trend magnitudes are this small in value, the statistical significance of the slopes are low. For the sites that have larger inter-annual trends, LSTM also predicts larger trends. R between LSTM hindcast and in-situ (both CRN and core sites) trends is 0.88 for the sites that have large trends (> 0.5% annually), indicating that SMAP prolongation by LSTM is a promising approach that can capture the inter-annual variability in root-zone soil moisture. It should be recalled that the model was only trained using three years of data. The increase in hindcast length does not appear to strongly influence the results.

Figure 5.8. Comparison of long-term root-zone soil moisture trends between core validation sites, CRN sites, and LSTM hindcast. The correlation between the slopes of core site and LSTM is 0.87. The correlation between CRN and LSTM is 0.73 for all sites. Among all the sites with trends greater than 0.5% per year, the correlation is 0.88. Panels on the right show the time series of several selected sites. Slopes are annotated on the legend with a unit of percent change per year.

5.4.4 Further discussion and future research

Although we noted some issues, such as winter soil moisture, the LSTM results are presented “as is” (without any further corrections) for the sake of thoroughly exploring the strengths and limitations of the LSTM-based long-term predictions. When SMAP itself poorly matches the in-situ dynamics, the LSTM hindcast also
degrades. The inaccuracies in SMAP might amplify, which were found to be mild. As per other studies, a main difference between SMAP and in-situ data appears to be related to the different sensing depths. SMAP appears to be capturing shallower soil moisture than the in-situ probes. In contrast, the process-based model Noah seems to simulate similar depths concerning the in-situ data, which has led to Noah’s better performance at some sites. It also raises the question as to how SMAP data should be calibrated against the core site data. A corrective step accounting for the variable sensing depths seems necessary for the incremental improvement of SMAP’s retrieval algorithm.

In contrast to other model-data integration methods, such as DA and calibration, the power of LSTM is to learn the relationship between inputs and soil moisture without explicit assumptions concerning the underlying process. Thus, LSTM can largely avoid systematic (parametric or model structural) errors due to such assumptions. Our results show the learned knowledge is permanent and does not decay with time, when judged by the learning target. As previously mentioned, LSTM cannot correct stochastic forcing errors as DA could. For example, if a rainfall event was missing in the forcing, LSTM would be unable to predict the corresponding soil moisture rise. However, while DA can correct for stochastic forcing errors, its effects wane with time, leading to little impact on long-term projections. The model states would drift away from the corrected states some time after data ingestion (De Lannoy et al., 2007). On the other hand, calibration only addresses parametric errors. Too much calibration might in fact distort the parameters’ values and model dynamics to compensate for the structural inadequacies. Therefore, all these techniques have their respective advantages and disadvantages, and might be applied synergistically to address different sources of errors.

As discussed in the introduction, extending SMAP records back in time (hindcast) with high fidelity to SMAP observations might provide considerable value to myriad applications. For example, a high-fidelity data-driven historical reconstruction of SMAP can help benchmarking or calibrating the SMAP retrieval algorithms. For another example, future research may examine combining data merging techniques, e.g., (Liu et al., 2012), with DL to exploit stochastic signal measured by older satellites. In Liu et al.’s work (Liu et al., 2012), we cannot merge SMAP and older satellites because they had no overlaps. However, the LSTM hindcast may help create that overlap to enable CDF matching and separation of seasonality.
and random errors, among other outcomes.

Here, we only showed a simple average between LSTM and Noah to demonstrate the value of integrating information from SMAP via a DL model. This simple approach is obviously not the optimal way to exploit the value of SMAP data. Currently, uncertainty estimation methods for DL models are still in the developing stage. As these methods become more matured, we can more adaptively combine the models, given our knowledge about their location-dependent and time-dependent accuracy. Further, the LSTM-extended data can be injected into land surface models through either DA or model calibration to improve the model’s internal dynamics other than soil moisture. On the other hand, we can construct more informative training data for winter periods, perhaps by combining observations with simulations by process-based models, to fill the gaps due to frozen soil. We envision the process-based model will continue to be a valuable constraint for temporal extrapolation as well.

The good performance of LSTM suggests that it learns the soil moisture evolution dynamics, as modulated by the land surface characteristics rather than a simple relationship between climate forcings and soil moisture. For forecast purposes, we have the additional information from present and near-past observations, which are expected to help improve forecast skills. The forecast model will require a slight reformulation of our network training procedure, where the current and past soil moisture observations are included as inputs. It must be noted that forecasting skills will depend heavily on the skill of the weather forecasts. We saw that the hindcast quality appears to deteriorate for sites that are surrounded by lower weather station density.

### 5.5 Conclusions

Our results indicate that the long-term LSTM predictions have mostly retained the quality of the SMAP data. LSTM has a solid performance in regions where SMAP has good skills. For the root-zone product, the hindcast is able to capture and reproduce multi-year trends correctly, even if there are only three years of training data. LSTM’s performance is limited by SMAP as a performance ceiling and is outperformed by Noah at some sites. A major factor is that SMAP has a different sensing depth in comparison to the probes measuring at 5 cm depth.
More importantlly, despite the limitations, averaging LSTM and LSM predic-
tions often produces better predictions. Even at sites where this is not true, Comb’s performance is typically very close to Noah’s. As a comparison, we have also shown that randomly combining models, e.g., Noah+VIC, do not always produce better results. Thus, SMAP extended by LSTM has added value and helped correct er-
ors in Noah. There could be many novel uses of this approach, as described in the Further Discussion section.

In this application, the LSTM model aimed at reproducing a model-free, seam-
less replica for the information contained in SMAP. LSTM predictions are entirely data-driven. As such, we expect that as SMAP data accrue and better retrieval algorithms are implemented, the DL models will also improve. Therefore, there is substantial value in the LSTM model as an ensemble model member for the recon-
struction of the past or the prediction of future trends. Meanwhile, we confirm the land surface model Noah as a valuable contributor of information in ensemble predictions.

As LSTM was trained with SMAP as the target, it inherited systematic discrep-
ancies between SMAP and in-situ data. When the discrepancy is large, the error seems to be mildly amplified, leading to an increase in ubRMSE, which is around 0.01. This amplification is attributed to less detectable patterns to learn and more noise in the SMAP data, which can contaminate the learning process. Hindcasts have larger errors during winter for cold regions. SMAP data can be missing and less reliable under snow and frozen soil conditions, which reduces the available training data.

5.6 Author Contributions

K. F. implemented the computer codes, analyzed and visualize the results, and draft the manuscript. M. P. provided the in-situ dataset and discussion. C. S. con-
ceived the work, provided general oversights, and critically revised the manuscript. All authors contributed substantially to the study design, paper writing, and fi-
nal approval of this version to be published. K. F. is the first author and has contributed the majority of work.
Bibliography


NASA (2018), NLDAS-2 Forcing Dataset Information.


Appendix A
Supporting Information for Chapter 2

A.1 Different interpretations of the linear regression equation

When \( x \) is composed of only independent physical factors, Equation 2.7 can be interpreted as having made the following hypothesis: the Budyko formula describes the \( E \) of a standard reference basin, \( \bar{x} \), (\( \bar{x} \) stands for the reference values of a comprehensive set of physical factors, e.g., vegetation cover, terrain slope, phase shift between \( P \) and \( Ep \), etc), surrounding which the actual \( E \) changes smoothly as a function of the physical factors. The reference basin represents average conditions, which may vary as a function of the aridity index, of world catchments. The smoothness assumption allows us to approximate the deviation from the reference state using a linear formula of the change in the factors, i.e., employing Taylor Series expansion:

\[
\delta(x) = \frac{E}{P} - f\left(\frac{E_p}{P}\right) = \mathbf{a}^T(x - \bar{x}) + \varepsilon. \tag{A.1}
\]

When \( \bar{x} \) are constants, they need not be estimated independently, but can be lumped into the one constant in the linear regression, i.e., in terms of data fitting, Equation A.1 is equivalent to Equation 2.7.

When \( x \) also contains surrogate indices, e.g., \( A/P \) and \( \gamma \), it has a different physical meaning. If, in addition to influencing \( E/P \), \( x \) also influence the surrogate
indices, say \( A/P \), we can write an equation similar to Equation A.1:

\[
\frac{\Delta A}{P} = \frac{A}{P} - \frac{A_0}{P} = b^T(x - \bar{x}) + \varepsilon, \tag{A.2}
\]

where \( A_0 \) is a reference amplitude, at which the amplitude-based correction is 0. We can split \( \bar{x} \) into two components, \( \bar{x} = \bar{x}_0 + \bar{x}_c \), so the above equations can be re-written as:

\[
\delta = a_0^T(x_0 - \bar{x}_0) + a_c^T(x_c - \bar{x}_c) + \varepsilon,
\]

\[
\frac{A}{P} - \frac{A_0}{P} = b_0^T(x_0 - \bar{x}_0) + b_c^T(x_c - \bar{x}_c) + \varepsilon, \tag{A.3}
\]

where \( x_c \) are major climate or basin characteristics that can be conveniently computed with available data for ungauged basins, e.g., phase shift between \( P \) and \( E_p \), or fraction of precipitation as snow, fraction of precipitation falling as snow \( (S/P) \) or vegetation indices \( (NDVI) \). We have tested using the \( NDVI \) as an aridity-dependent variable, i.e., fitting the mean \( NDVI \) to aridity value, but this did not improve our results. \( x_0 \) is central to our method: these are a set of factors that influence \( \delta \) both \( \Delta A/P \) and and we can find an approximate, effective ratio \( \beta \) between \( b_0 \) and \( a_0 \), i.e., \( \beta = a_0/b_0 \). As a result, \( \Delta A/P \) can be used as a surrogate for \( x_0 \), and we can re-write Equation A.3 as:

\[
\delta^*(x) = \beta \frac{\Delta A}{P} + (a_c^T - \beta b_c^T)(x_c - \bar{x}_c). \tag{A.4}
\]

This formula gives some flexibility in assigning a factor into either \( x_0 \) or \( x_c \). Although a fixed ratio \( \beta = a_0/b_0 \) seems a strong assumption, in practice, many factors exert weak controls, or they have strong controls but with limited variability, and we can lump them into an effective parameter to be represented by \( \Delta A/P \). At the extreme, we can merge all of \( x_E \) and \( x_c \) into \( x_0 \). And Equation A.4 becomes \( \delta(E_p/\bar{x}, X) = \beta \frac{\Delta A}{P} + \varepsilon \). When we test Equation A.4, the prediction of \( \delta \) turns into a linear regression problem between \( \Delta A/P \) and \( \delta \) so again it is equivalent to Equation 2.7 for parameter estimation purpose.
A.2 Supporting figures for variables discussed

The boundaries of HUC4 and GRDC datasets, longterm average aridity index and $A/P$ for HUC4 are presented in Figure A.1. To bridge the communities that use MOPEX and NLDAS datasets, we show their comparisons in Figure A.2. Finally, Figure A.3 presents GRACE leakage and measurement errors. From this figure, we notice that Northwestern coast of North America, Andes, South Asia, Japan, Indonesia, Madagascar and Northern Australia are all regions with relatively large GRACE errors.

Figure A.1. (a) $A/P$ for the HUC4; (b) Locations of USGS gages used in the study; (c) Maps of GRDC basins used in the study, with colors indicating the aridity index.
Figure A.2. Comparisons of annual average fluxes comparison between HUC4 and MOPEX basins for the period 2002/01 to 2013/12.
Figure A.3. GRACE leakage (upper) and measurement (lower) errors. Data from S.C. Swenson. 2012. GRACE monthly land water mass grids NETCDF RELEASE 5.0. Ver. 5.0. PO.DAAC, CA, USA. Dataset accessed [2016-03] at http://dx.doi.org/10.5067/TELND-NC005.
Appendix B
Supporting Information for Chapter 3

B.1 Visualization and numbering of clusters using Principal Component Analysis

Principal component analysis (PCA) [Seber, 1984] was employed to visualize the pattern of SSCS spectral clustering. PCA finds orthogonal basis called principal components (PCs) for the observation data in a way that the first PC explains the largest possible variance, and each succeeding orthogonal PC, in turn, explains the highest remaining variance. Using PCA, we reduced SSCS data from 30 total bands (to the two most important PCs (the principal plane). The data points (catchments) are projected to the plane of the first two PCs to facilitate visualization.

We also used PCA analysis to assign numbering to the cluster centers described above. Our numbering scheme first groups centers into 2 bins based on PC2 values. It sequentially assigns #1, #2, etc., to centers in the first bin (with lower PC2 value) in ascending order of PC1. Then we continue to label those in the second bin also in ascending order of PC1. This numbering scheme has a minor but practical advantage: during our testing, when different datasets and algorithmic parameters were involved, the class centers might shift in positions, and their orders might change after k-means algorithm finished. Our numbering scheme allows the centers to be consistently identified and interpreted during these trials. For example, #1 always corresponds to the center with the lowest PC1 and PC2 values, and they
have overall the highest correlation values amongst all basins. On the other hand, the last class has the lowest correlations amongst all. Note that here as our main goal is to interpret the SSCS behaviors and their controls, we did not use principal components in the k-means clustering, but exclusively as a tool for visualization and center-number-assignment.

The projection of SSCS onto the principal plane gives us a rough idea of how basins are clustered in the SSCS space. The plane is densely spanned by catchments, showing no noticeable class separation nor clear clustering. The first and second principal components explained 46.68% and 15.8% of total variance, respectively. The explained variance suggests there is a significant correlation between the SSCS bands, which is expected as flows across nearby flow bands are naturally correlated. PC1 has negative loadings from all flow bands, and therefore, a basin with low PC1 value implies it has high storage-streamflow correlations. On the principal plane, the cloud is dense in the left-lower part, which has high storage-streamflow correlations, and thinner in the right-upper part, which has low correlation values in all bands.

**Figure B.1.** Comparing an Long Short-Term Memory (LSTM) unit with simple recurrent neural network (RNN). The transformations from inputs to $i$, $f$, $o$ are sigmoidal functions. From inputs to $g$ and from $s$ to $h$ the transformation is $tanh$. ⊗ means multiplication by weights. Main point: the conventional design of RNN only iteratively update the hidden state. The design of gates in LSTM allows it to learn when to forget past states, and when to output, thus addressing the issue of slow training of front node with RNN. Figure is modified from [Greff et al., 2015].
B.2 Feature (attribute) selection of SSCS analysis

44 attributes from GAGES-II, Gravity Recovery and Climate Experiment (GRACE) Terrestrial Water Storage Anomalies (TWSA) signatures, and climate forcing data extracted from North American Land Data Assimilation System (NLDAS) (Xia et al., 2012) have been tested as predictors for spectral distances to class centers and the top 40 of them are listed in Table B.1 in Appendix B. They include climate factors such as climate, basin morphological, and physiographic attributes. Spatial distributions of selected attributes are presented in Figure 3.1.

We first conducted both forward and backward feature selection. Forward feature selection starts with one best predictor and consecutively adds the next most important predictors. The testing error of the Classification and Regression Tree (CART) model will first drop as more predictors become available, but at some point, the decline becomes slow, flat and the error may even rise, indicative of overfitting (Figure B.2a). Backward feature selection, on the other hand, starts with the full list of predictors and removes the next least important predictor. The error will initially decline as overfitting are reduced, then stay constant, but will rise significantly after important predictors are removed (Figure B.2b). We identified 15 attributes by picking 11 from backward feature selection, 7 top ranking attributes from forward selection and removing the redundant ones.
Figure B.2. The training and testing errors as functions of the number of predictors in forward (a) and backward (b) feature selections. As we can see, when first few predictors are added to the forward selection, there is a dramatic reduction in error. While the training error continues to decrease after the 8th predictor, the testing error ceases to change and starts to increase after 25th predictor, suggesting overfitting. In backward selection, when least important attributes are removed (starting from the right-hand side of panel b), the error decreases due to the reduction of overfitted parameters. Then the error flattens until the last few important predictors are removed.

Table B.1: 41 factors considered and their rankings in forward and backward feature selections. Factors with a ’*’ in the name come from GAGES-II dataset. (*)& indicates selected for CV tests.

<table>
<thead>
<tr>
<th>Factor Name (*)</th>
<th>Forward</th>
<th>Backward</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/P (*)</td>
<td>2</td>
<td>1</td>
<td>Fraction of precipitation as snow</td>
</tr>
<tr>
<td>ξ(*)</td>
<td>3</td>
<td>2</td>
<td>P-Ep seasonality index</td>
</tr>
<tr>
<td>acf (*)</td>
<td>4</td>
<td>3</td>
<td>1-month lag TWSA auto-correlation</td>
</tr>
<tr>
<td>σ(*)</td>
<td>1</td>
<td>4</td>
<td>TWSA Inter-annual variability index</td>
</tr>
<tr>
<td>Soils.SANDAVE (*)</td>
<td>5</td>
<td>5</td>
<td>Percent of sand of soil from STATSGO</td>
</tr>
<tr>
<td>Soils.RockDep (*)</td>
<td>9</td>
<td>6</td>
<td>Depth to bedrock from STATSGO</td>
</tr>
<tr>
<td>Topo. ELEV_STD_30M (*)</td>
<td>41</td>
<td>7</td>
<td>Standard deviation of elevation</td>
</tr>
<tr>
<td>Climate.RH (*)</td>
<td>39</td>
<td>8</td>
<td>Basin mean relative humidity</td>
</tr>
<tr>
<td>P (*)</td>
<td>8</td>
<td>9</td>
<td>Average annual precipitation (cm)</td>
</tr>
<tr>
<td>LC06_Basin. PLANTNLCD06 (*)</td>
<td>41</td>
<td>10</td>
<td>Watershed percent agriculture</td>
</tr>
<tr>
<td><strong>Amp/P or A/P (*)</strong></td>
<td>40</td>
<td>11</td>
<td>TWSA Amplitude as a fraction of P</td>
</tr>
<tr>
<td>----------------------</td>
<td>----</td>
<td>----</td>
<td>---------------------------------</td>
</tr>
<tr>
<td><strong>Soils.HGB</strong></td>
<td>12</td>
<td>12</td>
<td>Fraction of Soil hydraulic group B</td>
</tr>
<tr>
<td><strong>Soils.BDave(*)</strong></td>
<td>7</td>
<td>13</td>
<td>Average soil bulk density</td>
</tr>
<tr>
<td><strong>Slope (*)</strong></td>
<td>6</td>
<td>14</td>
<td>Average 8-neighbor slope calculated from 30m DEM.</td>
</tr>
<tr>
<td><strong>NDVI</strong></td>
<td>36</td>
<td>15</td>
<td>Time-Averaged Normalized Difference Vegetation index</td>
</tr>
<tr>
<td><strong>WATERNLCD06</strong></td>
<td>11</td>
<td>16</td>
<td>Watershed percent open water</td>
</tr>
<tr>
<td><strong>DEVNLCD06 (*)</strong></td>
<td>12</td>
<td>16</td>
<td>Watershed percent developed</td>
</tr>
<tr>
<td><strong>Soils.AWCAVE</strong></td>
<td>12</td>
<td>16</td>
<td>Soil average water capacity from STATSGO</td>
</tr>
<tr>
<td><strong>Soils.OMAVE</strong></td>
<td>12</td>
<td>16</td>
<td>Soil average organic matter percentage</td>
</tr>
<tr>
<td><strong>Soils.Silt (*)</strong></td>
<td>12</td>
<td>16</td>
<td>Percent of silt of soil from STATSGO</td>
</tr>
<tr>
<td><strong>Soils.HGA</strong></td>
<td>12</td>
<td>16</td>
<td>Fraction of Soil hydraulic group A</td>
</tr>
<tr>
<td><strong>Soils.HGAD</strong></td>
<td>12</td>
<td>16</td>
<td>Fraction of Soil hydraulic group AD</td>
</tr>
<tr>
<td><strong>Soils.HGD</strong></td>
<td>12</td>
<td>16</td>
<td>Fraction of Soil hydraulic group D</td>
</tr>
<tr>
<td><strong>Soils.HGBC</strong></td>
<td>12</td>
<td>16</td>
<td>Fraction of Soil hydraulic group BC</td>
</tr>
<tr>
<td><strong>RRMEAN_30M</strong></td>
<td>12</td>
<td>16</td>
<td>Dimensionless elevation - relief ratio</td>
</tr>
<tr>
<td><strong>ASPECT_NORTHNESS</strong></td>
<td>12</td>
<td>16</td>
<td>Basin aspect &quot;northness&quot;</td>
</tr>
<tr>
<td><strong>Soils.HGC</strong></td>
<td>36</td>
<td>16</td>
<td>Fraction of Soil hydraulic group C</td>
</tr>
<tr>
<td><strong>LCD01_Bas. Forest</strong></td>
<td>41</td>
<td>16</td>
<td>Percent of forest cover</td>
</tr>
<tr>
<td><strong>T_AVG_BASIN</strong></td>
<td>10</td>
<td>29</td>
<td>Annual average temperature</td>
</tr>
<tr>
<td><strong>DRAIN_SQKM</strong></td>
<td>12</td>
<td>29</td>
<td>Drainage area (km²)</td>
</tr>
<tr>
<td><strong>BAS_COMPACTNESS</strong></td>
<td>12</td>
<td>29</td>
<td>Basin compactness ratio (m)</td>
</tr>
<tr>
<td><strong>PET</strong></td>
<td>12</td>
<td>29</td>
<td>Mean-annual potential evapotranspiration (mm/year)</td>
</tr>
<tr>
<td><strong>REEDBUSH_DOM</strong></td>
<td>12</td>
<td>29</td>
<td>Dominant geology types in the basin</td>
</tr>
<tr>
<td><strong>REEDBUSH_DOM_PCT</strong></td>
<td>12</td>
<td>29</td>
<td>Percentage of the watershed covered by the dominant geology type</td>
</tr>
<tr>
<td>Geology. REEBUSH_SITE</td>
<td>12</td>
<td>29</td>
<td>Geology type at the gage location</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----</td>
<td>----</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Geology. BEDROCK_PERM</td>
<td>12</td>
<td>29</td>
<td>Bedrock permeability (in/hour)</td>
</tr>
<tr>
<td>Hydro. STREAM_KM_SQ_KM</td>
<td>12</td>
<td>29</td>
<td>Stream density (km/km²)</td>
</tr>
<tr>
<td>Dams. MAJ_DDENS_2009</td>
<td>12</td>
<td>29</td>
<td>Density of major dams in the basin</td>
</tr>
<tr>
<td>Dams. STOR_NID_2009</td>
<td>12</td>
<td>29</td>
<td>Dam storage in watershed (ml/km²)</td>
</tr>
<tr>
<td>Soils.Clay</td>
<td>12</td>
<td>29</td>
<td>Soil clay percentage</td>
</tr>
</tbody>
</table>

B.3 Cross-validation and CART model selection

15 attributes are too many for interpretation, and we find that models with more predictors can have small training error but large testing error, suggesting overfitting. Therefore, we constrained ourselves to 8 predictors. There can be $C_{15}^8 = 6435$ of such models from the 15 attributes selected. We then conducted two types of cross-validation (CV) tests with varying amount of training data to identify the most robust ones. First, the training set is randomly selected but the percentage of data used for training is varied between 80% to 20%. The second one is a more stringent regional holdout test, in which one (or two) 2-digit Hydrologic Cataloging Unit(s) (HUC2, there are 18 HUC2s over CONUS) were withheld from the training set. The trained models are evaluated using data from withheld HUC2s. The test loops through all possible selections of one (or two) HUC2s, and the average testing error was calculated. We then picked models that perform well when confronted with instances never seen in the training set. The criteria are trees that are within the top 10% in all CV tests and are non-dominated by any other trees.
B.4 Impacts of dam density, human disturbance, zero-flow gages, type of correlations and number of data points on analyses and conclusions

Here we examine the impacts of various choices on our analyses. Mainly, we examined their impacts in the light of (i) how they change the values of storage-streamflow correlations; (ii) how they influence clustering; and (iii) how they influence our interpretation of physical controls. Overall, after removing a small set of basins with major dams, these choices do not have significant impacts on our analyses and conclusions.

B.4.1 Human disturbance of landscapes.

Firstly, in our study design, human disturbance should not be excluded because we would like to use SSCS to understand hydrology for both undisturbed and disturbed landscapes. Land use/land cover does not appear to be a major factor in modulating SSCS, as least not as important as soil properties. As a result, we included non-reference basins in our analyses in order to obtain more data points, and also to show the relative importance of human disturbances. We show 4 pieces of evidence: (i) There is no statistically significant difference in storage-streamflow correlations between the undammed reference and non-reference gages (Figure B.4). (ii) We included forest cover, agricultural land use, and the ratio of developed areas in the list of attributes for CART, but they are not highly ranked in feature selection (Table B.1, the percent of agricultural land use was tested and was not even ranked among top 41). We additionally included the 'reference' flag from the GAGES-II dataset as an attribute, but it also does not appear to be important. (iii) The most illuminating demonstration comes from re-creating Figure B.13c-d while including the reference flag as an attribute. We note that most of the class #6 points in Indiana/Ohio are non-reference, agricultural basins. Therefore, if we had removed non-reference points, they would have all been removed (Figure S B.3). As a result, one might suspect that the reference flag might be the most important criterion in separating class #6 from class #3 basins. However, surrounding these basins, there are a large number of non-reference basins that are also removed. In
fact, as we re-run the CART analysis for Figure B.13b with the reference flag as an attribute, soil bulk density again emerges as the top-level split instead of the reference flag (Figure B.5). This result strongly supports the argument that soil properties are much more important than human disturbance for SSCS.

Figure B.3. The reference basins for the same region shown in Figure B.13c. A lot of Indiana/Ohio basins are removed
Figure B.4. Comparisons of storage-streamflow correlations between reference and non-reference basins. We do not observe differences between these two sets.
Figure B.5. Re-created Figure B.13c-d after adding the reference flag from GAGES-II. They are identical to Figure B.13c-d because soil bulk density remains the most influential factor although we added the "reference" flag. The flag is not important because many class #3 points are also non-reference basins.

(iv) In Figure B.6, we also re-created Figure 3.4 using only reference basins. We note that as the number of basins is reduced, classes #1-#4 remain unchanged. Classes #5 and #6 have moved slightly, which results from the removal of many Class #6 basins, such as those in Indiana/Ohio shown above. We do not wish to remove these basins as they contain valid signals. In summary, human disturbance of landscapes does not significantly change the patterns of SSCS clustering or the analyses of controlling factors for Classes #1-#4. While the removal of non-reference basins shifted the center of Class #5 and #6, human disturbance is not the reason why Class #6 points are different from other classes. In order to retain more data and also to study the relative importance of human disturbance on SSCS, we chose to include non-reference basins in the main study.
Figure B.6. Re-created Figure 3.4 & Figure B.1 using only reference basins. We note Centers #5 and #6 have moved because many basins in the upper right corner have been removed. As a result, panel (g) & (h) look different from Figure 3.4. However, as we have explained previously, non-reference status is not the cause of why some basins were Class #6 in Figure 3.4.

B.4.2 Zero-flow gages.

There are only about 200 stations that have 0 flows for more than 0.5% of the daily flows. In our sensitivity test, we added these stations into the analyses and assigned a value of 0 for bands that have 0 flows. The resulting figure is nearly identical to our Figure 3.4 and Figure 3.7. This choice does not impact any other Figure or conclusion in the paper.
Figure B.7. (Left duplicate) There is no noticeable difference between Figures 3.4 and this one uses 0 for the bands with zero-flows, instead of removing the stations from the analysis. If one compares very carefully with Figure 3.4, she/he may notice a few additional points in the lower right part of the panel (a) of this figure. (Right duplicate) Figures 3.4 and Figure B.1 recreated without removing basins with high dam densities or large dam volumes. Again, the difference is difficult to observe.

B.4.3 Dams and reservoirs.

We first examined how dam density and volume influences storage-streamflow correlations. With "major dam density" data from the GAGES-II dataset, we extracted basins with major dam densities $> 0$ (dams/100 km$^2$). Then, in the attribute space formed by the top 8 most influential predictors (listed in Table B.1), we searched for 6 basins nearest to each dammed basin. This search allows us to obtain undammed basins with similar attributes to the dammed ones. We fitted a rank-based Sens slope to the dam density data, as well as the dam storage data.

We found that dam density has little influence on correlations in the high-flow bands, but a small, negative influence on the trough-low-flow bands (Figure B.8iv & viii). Such influence arises from the human-controlled release during the low flows.

Based on these results, we first removed basins with dam density $> 2$ major dams/100 km$^2$ or dam storage $> 500$ MLitres/km$^2$ (or 500 mm averaged over the basin area). Below 200 mm, dam storage has no statistically significant impacts on $\rho'_L$ (Figure B.8h), meaning these smaller dams are not significant enough to
alter SSCS patterns. However, between 200 mm to 500 mm, dam storage exerts some minor influence (Figure B.8viii), but it is not certain whether such influence is important enough to justify removing them (355 basins) from the analysis of controls. In the interest of retaining more data for analysis, we further evaluate this point using a linear correction model fitted to dam storage and $\rho_L^I$. We applied this correction to the low-flow bands of these basins and then we used the trained k-means classifier to classify the 'corrected' data points. Comparing their pre- and post-correction class assignment, we found only 15 out of 355 have changed their classes due to the correction (Figure B.9 the right panel). As a result, we decided to retain these basins for analysis. In total, we removed 303 basins due to dams.
Figure B.8. (a-d) Storage-streamflow correlations vs. dam densities for dammed basins and those undammed basins with similar attributes. We fitted a rank-based Sens slope to the data points and found that there is a significant relationship only between low flow bands and dam density. However, the slope is small; (i-iv) the same information presented in boxplot format; (e-h) same as (a-d) but created using dam storage data between 0-200 mm. (v-viii) same as (i-iv) but for dam storage dam. We removed basins with dam density > 2 (100 km$^2$) or dam storage > 500 ML/km$^2$. 
Figure B.9. (Left 4 panels) Distribution of dammed vs. non-dammed basins on the SSCS principal component plane. (Right panel) Comparing class assignment pre- and post-correction. We note that most basins with dam storage between 200 and 500 mm retain class assignment.

B.5 Other supporting figures for Chapter 3

1. Figure B.10. Maps of CONUS: (a) Physiographical regions; (b) States

2. Figure B.11. Comparisons between CDFs of storage-streamflow correlations using Spearman and Pearson definitions, and between water year, calendar year and extrema-based windows.

3. Figure B.12. Identified peaks and troughs in a GRACE TWSA time series for an HUC4

4. Figure B.13. Map of spectral distances to class centers.

5. Figure B.14. SSCS aggregated by physiographic divisions in the US.

6. Figure B.15. A composite-output tree that predicts distances to 6 centers simultaneously.

7. Figure B.16. Figures recreated from literature: (a) baseflow index; (b) simulated water table depth
Figure B.10. Maps of CONUS: (a) Physiographical regions; (b) States
Figure B.11. (a) Comparison between cumulative distribution functions (CDF) storage-streamflow correlations (high-flow-band averages and low-flow-band averages) using Spearman and Pearson definitions. We note that the Pearson correlation is higher, indicating that there is no strong nonlinear trend between storage-streamflow relationships; (b) comparison between using calendar-year, water-year and our extrema-based windows for extracting streamflow percentiles. The extrema-based window is better than the other two options in all bands, especially in low flow bands.
Figure B.12. Identified TWSA peaks and troughs for an HUC4. There are 10 peaks and, in this case, 11 trough data points.

Figure B.13. Map of spectral distances to class centers.
Figure B.14. SSCS aggregated by physiographic divisions in the US.
**Figure B.15.** A composite-output tree that predicts distances to 6 centers at the same time. This tree is one of the best performing trees that is selected from cross validation test. End nodes have been annotated with the regions where their points are primarily located in. Other realizations of CART may appear different at lower levels, but many splits are similar, only placed in different branches of the trees.
Figure B.16. (a) Baseflow Index (BFI) map for CONUS estimated using flow separation technique. The data is from Wolock (2003). We note that Southeast CONUS and Northern Great Plains have high BFI, while BFI is low in Indiana/Ohio; (b) Simulated water table depth for CONUS using land surface model-estimated recharge, topography, and a groundwater flow model. Simulated data is from Fan et al. (2013). We can see that the Fan et al.s model captured the shallow water table in Indiana/Ohio and Florida, while the water table is deeper in western Illinois and Iowa. Reprint permission has been obtained.
Bibliography


Appendix C
Supporting Information for Chapter 4

C.1 Technical Details about Conventional Methods

We compared the Long-Short Term Memory (LSTM) network to the least absolute shrinkage and selection operator (lasso), auto-regressive moving average model (AR), and a single-layer feedforward Neural Network (NN), given the same inputs. Because lasso is essentially a regularized linear regression, it is shorthanded as LR in our paper. The equation for estimating the parameters for LR is:

\[
\beta_{LR} = \arg\min_{\beta_0^{LR}, \beta_{LR}} \left( \frac{1}{2N} \sum_{i=1}^{N} \left( \theta_i^o - \beta_0^{LR} - x_i^T \beta_{LR} \right)^2 + \lambda \sum_{j=1}^{n} |\beta_{LR}| \right), \tag{C.1}
\]

where \( \theta^o \) is the SMAP soil moisture product, \( \beta_{LR} \) are coefficients for the LR model, \( \lambda \) is a regularization parameter that determines how much penalty is applied on large coefficients, and \( x \) contains exogenous inputs including temperature, precipitation, wind, downward shortwave and long wave radiations, specific humidity, and Noah-simulated potential evapotranspiration, evaporation, and runoff. In alternative models that we examined, we also tested removing the list of Noah outputs. The regularization parameter (\( \lambda \)) is determined experimentally to minimize the test error, and a value of 0.002 is found to be appropriate for LR, and point-by-point LR (LR\(_p\)).

We have added point-by-point auto-regressive model with exogenous inputs into the comparisons, meaning a separate model is trained for each SMAP pixel. We
did not consider moving average models because our focus is on the potential of the method for long-term forecast, while moving-average models require observations to calculate residuals. The equation for the AR is:

\[ \theta_t = c + \epsilon_t + \sum_{i=1}^{p} \alpha_i \theta_{t-i} + \sum_{k=1}^{r} \gamma_k x_{k,t}, \]  

where \( c \) is a constant, \( t \) is the time step, \( \theta \)'s are soil moisture observations, \( p \) is the order of the auto-regression, \( \alpha \) and \( \gamma \) are coefficients that will be estimated for each SMAP pixel, and \( x_t \) are \( r \) forcing inputs as indicated above. For our long-term hindcast test. We could include static attributes in this equation but since they are static in time they will be absorbed by the constant \( c \), and because we are training point-by-point there is no reason to consider them. During parameter estimation (training) stage, observations are used to update the past states (\( \theta_{t-i} \)). In the long-term hindcast (testing) stage, because there is no observation, \( \theta \) are the AR-predicted values. The model has to recursively apply the forecast equation to proceed in time. We varied \( p \) from 0 to 5 and identified the value that gave the smallest testing error for each site.

The one-layer feedforward neural network (NN) is simply a linear combination of inputs \( x \) and a transformation:

\[ \theta^{NN}(t) = f(W_{NN}x + b), \]  

where \( W_{NN} \) is the weights of the neural network, \( b \) is a constant coefficient and \( f \) is a nonlinear transformation, in this case tan-sigmoid (\texttt{tansig}). We regularized NN using early stopping and L2-norm regularization. A regularization parameter of 0.002 was found to be give the smallest test root-mean-squared error (RMSE). NN and its point-by-point version, \( NN_p \), have a linear hidden layer of size 100 and 30, respectively, as larger hidden size results in more over-fitting.
C.2 Proof-of-concept test for the potential of LSTM for long-term hindcast

Since SMAP has a limited time span, we conducted a proof-of-concept experiment that examines the potential of LSTM for multi-year-scale soil moisture hindcasting and compare it to point-by-point auto-regressive models (AR$_p$). These synthetic experiments are not thorough in performance optimization, as true hindcasting will involve auxiliary satellite-based observations and in-situ data. Both LSTM and AR$_p$ are trained in 2015-2016 with Noah-simulated soil moisture as the target and climate forcing as the inputs. Based on the temporal generalization test described in the main text, we removed all Noah-simulated fields from the inputs, and, since AR$_p$ does not require any static attributes like topography and soil texture, we also removed such attributes from LSTM’s inputs. We added two types of synthetic noise to Noah solutions: a Gaussian white noise (with standard deviation $= 0.04$) and a relative error. Neither types of noise is auto-correlated. The formulae for the relative error is:

$$\theta_s = \theta_{Noah} \ast (1 + \epsilon), \quad (C.4)$$

where $\theta_s$ is the synthetic observation to be treated as the learning target, $\theta_{Noah}$ is the top 10 cm soil moisture simulated by Noah, and $\epsilon \sim N(0, 0.07)$ is a Gaussian relative error term.

The results show that with two years of training data, LSTM can well learn the soil moisture dynamics of Noah (Figure C.1a-b). The median error for the white-noise case only slightly increases from 0.04 in the training period, which is almost equal to the added noise) to 0.043 in 2005-2006. Importantly, the hindcast noise does not increase as a function of hindcast length, i.e., distance from the first synthetic observation. The AR$_p$ also works decently, with a median error around 0.049 in 2005-2006. Its error is also not influenced by hindcast length, perhaps because soil moisture dynamics simulated by Noah has only limited memory length. However, LSTM is still noticeably stronger as 85-th percentile of LSTM’s error in 2005-2006 is less than 25-th percentile error of AR$_p$. The LSTM boxes are much narrower than those of AR$_p$. Also, note that we only created one LSTM model for the continental U.S. (CONUS). In addition, LSTM can make use of static attributes to differentiate between locations with different soil textures and land
covers, but ARp cannot. Therefore, the performance of LSTM may further improve as these attributes are included. LSTM may compensate for the lack of attributes by summarizing information from climate forcings, as climate features co-vary with physical attributes. Figure C.1c compares the hindcast time series at a pixel. We note that ARp tends to over-predict major peaks but under-predict the rise limbs. LSTM well captures the troughs but ARp may over-predict the troughs.

As Noah has simpler dynamics and less unknown variables than real systems, it is easier to learn so it is not surprising the errors are close to the added Gaussian noise. The larger error of ARp during the training period suggest its formulation is not flexible enough to completely reproduce the dynamics of Noah. These results shown here mainly illustrates that LSTM has a great potential for long-term hindcasting. It appears from our results that since soil moisture has short memory, hindcasting to one year is not very different from hindcasting to 10 years. However, the training data should adequately sample plausible soil moisture dynamics.
Figure C.1. Proof-of-concept long-term hindcast tests with Noah-simulated soil moisture as the target. (a) boxplot comparing errors (evaluated against Noah) for the 10-year hindcast. Noah solution is contaminated by a Gaussian noise with a standard deviation ($\sigma$) of 0.04. The RMSEs are calculated for each 2-year period and are grouped over CONUS to form the boxplot. Note that error does not increase as hindcast length increases, i.e., during 2005-2006, the errors are not greater than those in 2013-2014; (b) same as (a) but for a 7% relative noise; (c) time series for Noah, LSTM and ARp at a pixel. We only show 5 years of hindcast for clarity of the plot. The zoomed-in panel on (c) (corresponding to the brown box in the main plot) highlights how ARp over-estimates the two soil moisture peaks. Meanwhile, ARp seems to have dampened small-scale fluctuations.
## C.3 Table of all predictors

Table C.1: Predictors used in the training of LSTM, lasso-regularized linear regression, and one-layer feedforward neural network.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBDO</td>
<td>Albedo</td>
<td>RCSOL</td>
<td>Soil moisture parameter in canopy conductance</td>
</tr>
<tr>
<td>ARAIN</td>
<td>Liquid precipitation</td>
<td>RCT</td>
<td>Temperature parameter in canopy conductance</td>
</tr>
<tr>
<td>ASNOW</td>
<td>Frozen precipitation</td>
<td>RSMACR</td>
<td>Relative soil moisture availability control factor</td>
</tr>
<tr>
<td>AVSFT</td>
<td>Average surface skin temperature</td>
<td>RSMIN</td>
<td>Minimal stomatal resistance</td>
</tr>
<tr>
<td>BGRUN</td>
<td>Subsurface runoff (baseflow)</td>
<td>SBSNO</td>
<td>Sublimation (evaporation from snow)</td>
</tr>
<tr>
<td>CCOND</td>
<td>Canopy conductance</td>
<td>SHTFL</td>
<td>Sensible heat flux</td>
</tr>
<tr>
<td>CNWAT</td>
<td>Plant canopy surface water</td>
<td>SNOD</td>
<td>Snow depth</td>
</tr>
<tr>
<td>GFLUX</td>
<td>Ground heat flux</td>
<td>SNOHF</td>
<td>Snow phase-change heat flux</td>
</tr>
<tr>
<td>LAI</td>
<td>Leaf area index</td>
<td>SNOM</td>
<td>Snow melt</td>
</tr>
<tr>
<td>LHITFL</td>
<td>Latent heat flux</td>
<td>SNOWC</td>
<td>Snow cover</td>
</tr>
<tr>
<td>LSOIL</td>
<td>Liquid soil moisture content (non-frozen)</td>
<td>SOILM</td>
<td>Soil moisture content</td>
</tr>
<tr>
<td>MSTAV</td>
<td>Moisture availability</td>
<td>SSRUN</td>
<td>Surface runoff (non-infiltrating)</td>
</tr>
<tr>
<td>NLWRS</td>
<td>Net longwave radiation flux (surface)</td>
<td>TRANS</td>
<td>Transpiration</td>
</tr>
<tr>
<td>NSWRS</td>
<td>Net shortwave radiation flux (surface)</td>
<td>TSOIL</td>
<td>Soil temperature</td>
</tr>
<tr>
<td>PEVPR</td>
<td>Potential latent heat flux</td>
<td>VEG</td>
<td>Vegetation</td>
</tr>
<tr>
<td>RCQ</td>
<td>Humidity parameter in canopy conductance</td>
<td>WEASD</td>
<td>Water equivalent of accumulated snow depth</td>
</tr>
<tr>
<td>RCS</td>
<td>Solar parameter in canopy conductance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
C.4 Other supporting figures for Chapter 4

1. Figure C.2 Map of SMAP quality flag over CONUS

2. Figure C.3 Map of Noah Performance Component Analysis

3. Figure C.4 Performance of training using extrapolated 0-5cm Noah soil moisture

4. Figure C.5 Comparisons between LSTM performance using different Land
Surface Models as inputs.

**Figure C.2.** SMAP data quality with geographic regions annotated on the map for reference. The values shown is the time-averaged SMAP "recommended quality" flag. We notice that Appalachian Highlands and Southeast Coastal Plains are both mostly flagged as having bad quality, but LSTM’s root-mean-squared-error from SMAP L3, RMSE(LSTM) are in the range of 0.02-0.035 in the South Appalachian and Coastal Plains according to Figure 1. Because random error cannot be captured in the test, it suggests SMAP quality may be not as bad as thought. However, this finding and the potential influence from models in the retrieval algorithm need to be thoroughly evaluated.
Figure C.3. Performance of Noah evaluated against SMAP in the testing set of the temporal generalization test. (a) Bias: the time-averaged value of Noah-predicted soil moisture, interpolated to 5 cm, and SMAP L3 product; (b) anomaly correlation coefficient between Noah-predicted soil moisture and SMAP L3 product.
Figure C.4. Same as Figure 3 but the Noah-simulated values are linearly interpolated to 0-5 cm. We tested several interpolation methods: (i) directly using 0-10 cm data; (ii) linear (2-point) and cubic vertical interpolation (3-point) using top layers; and (iii) integral interpolation: we determined a 2\textsuperscript{nd}-order (or 3\textsuperscript{rd}) polynomial whose integral in these layers agree with Noah-simulated values. This Figure shows method (i), whose results are very similar to those reported in Figure 3, while other interpolation methods also generate similar results.
Figure C.5. Comparisons between LSTM models with Noah (L+Noah) and MOS (L+MOS) solutions as inputs. Both Noah and MOS solutions are obtained from North American Land Data Assimilation System (NLDAS). The distinction between 'train' and 'test' for Noah and MOS only means the different time periods for which the metrics are calculated. Noah and MOS have comparable performance in simulating moisture climatology. It appears MOS generally has smaller root-mean-squared error (RMSE) and smaller bias. However, using which model in the inputs does not seem to have a noticeable impact on the test performance of the LSTM models.
Vita

Kuai Fang

Education

2013–Present: Ph.D. student in Civil Engineering, Pennsylvania State University, PA, USA

2008–2012: B.S. in Geography, Beijing Normal University, China

Publications


