Paper Supplement

Contents

1 Model Architecture Supplement 1

2 Dataset Supplement 2
- 2.1 Feature Selection . . . 2
- 2.2 Scaling . . . 3
  - 2.2.1 Precipitation Data . . . 3
- 2.3 Composition . . . 3
- 2.4 OK Datasets . . . 4

3 Spatial Predictions 7

4 Cross Validation 9

5 SHAP 9

6 Ensemble Advantage 10

7 Domain Preference 10

8 Code and Data availability 12

1 Model Architecture Supplement

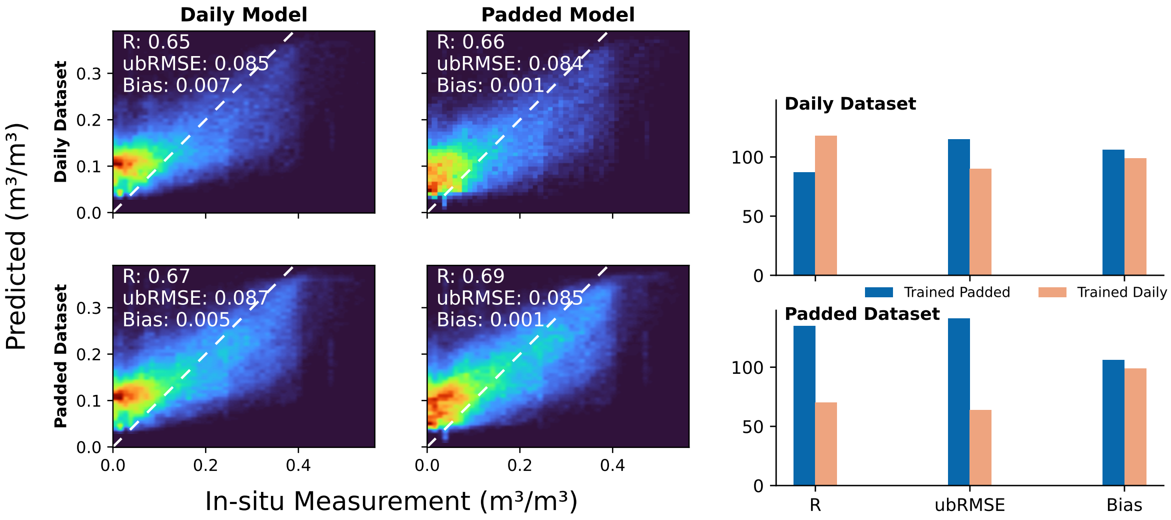
Probabilistic Layer

The hidden probabilistic layer in the Prob model serves to learn and map a posterior multivariate normal distribution onto a prior multivariate normal distribution. The prior multivariate distribution has fixed standard deviations of *0.5* with learnable mean values. The posterior multivariate normal distribution has both learnable mean and standard deviation.

During the training process, the layer tries to learn the prior distribution for all values fed into it. It then tries to learn the distribution for the posterior given the inputs. These two values are then compared via the Kullback-Leibler (KL) divergence to gauge similarity. This is done to ensure the posterior distribution is not overfitting a specific sample and is penalized for deviating too far from the prior distribution. The posterior distribution is then condensed to two values and passed forward in the network to the Independent Normal layer.

Temporal Resolution

The padded model doesn't just out perform the daily model on the dataset as a whole. The padded model also outperformed the daily model on a site by site basis for each metric except for R on the daily validation dataset (Fig. 1b). This is not unexpected as the daily model had greater LST variations in it's dataset than the padded model. The daily training model exhibits slight biased against low SWC readings. This is visible in the heatmaps of Figure 1a. At near zero in-situ SWC measurement readings the daily model has a strong cluster of predictions around 0.1 m^3/m^3. The mechanism for this is unknown, however, it seems apparent that training on additional samples helped the model identify lower SWC trends.



(a)

(b)

Figure 1: a) Predictions for a model trained on a time-padded dataset which contains much more samples (658,000) to learn from and a model trained on a temporally accurate dataset (372,000). Both models predict on the validation sets for each dataset. b) Head to head for these models on sites in each dataset. If a model outperforms the other in a metric the bar increases by one.

2 Dataset Supplement

2.1 Feature Selection

The variables selected *SMAP, NDVI, LST, Precipitation, Sand* and *Clay content, pH, Evapotranspiration, and Topography/Elevation.* are linked to SWC through multiple mechanisms.

NDVI, LST, and ET

Vegetation Index (NDVI), and Evapotranspiration (ET) Land surface temperature (LST) has a very strong coupling with SWC. As LST increases, more energy is available for SWC to harness in order to evaporate and leave the soil. This relationship is well established and exploited to benefit in DisPATCH algorithms. NDVI corresponds to plant greenness and plant cover over an area. Because plants require water for healthy efficient production, NDVI has been correlated to SWC on multiple occasions[1][2]. Evapotranspiration (ET) is also included as a variable as it is directly associated with SWC.

Soil Texture

SWC is directly influenced by the physical properties of the soil, such as texture and composition. Porosity and grain size directly influence the cohesive and adhesive properties of water which permit capillary rise. The greater the surface area by volume, the easier it is for water to adhere to mineral surfaces and resist extracting forces such as the downward flow due to gravity or uptake by evaporative processes. Smaller grained soils offer a greater surface area to volume ratio allowing for large capillarity for soil water. Soils with smaller grain size, e.g., clay, are therefore able to hold more water[3]. For this reason, soil textures and composition were included as variables for prediction.

Topography

Elevation and Topography have a strong influence on waterflow and subsequently SWC[4][5]. On a local scale, water naturally moves down the gravitational gradient, draining from higher elevations and accumulating in lower areas[4]. For this reason topographical changes correlate to SWC. At regional scales, topography also informs relative height vs the sea level which is especially relevant for areas that are below sea level or at greater elevations. Greater elevations experience a drop in atmospheric pressure as well as vapor pressure deficit (VPD)[6]. VPD correlates to the rate of evapotranspiration[7] and thus impacts SWC. For these reasons topography is included as a covariate in the dataset.

2.2 Scaling

Variable	Spatial Resolution	Temporal Resolution	Scaling Factors
SMAP	9km	∼3 Days	*N/A*
Precipitation	∼5.5km	Daily	1/1000
LST	1km	8-Days/Daily	(1/5000)-2.7315
Sand/Clay/pH Content	1km	*N/A*	1/100
ET	500m	8-Days	1/1000
NDVI	500m/250m	16-Days	1/10000
Topography	90m	*N/A*	1/10000

Table 1: Variable resolutions and value scaling. All variables are resampled to 90m spatial resolution using nearest neighbor

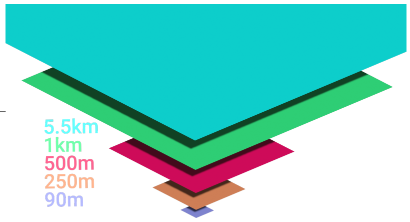


Figure 2: Stacking of readings for one prediction

For the distance based models (WDL, Dense, Prob) the input data/values were normalized to between 0-1 as seen in Table 1. The exceptions being *Topography* which has negative elevation values and **LST** and **Precipitation** values which do not reach all the way to one. The random forest model uses the raw values from the data as the Random Forest algorithm is invariant to distance.

2.2.1 Precipitation Data

Since ensemble predictions will be performed pixel-wise over single timesteps, past precipitation data would be absent from model input. In order to incorporate temporal information into the precipitation input, a rolling window was passed over the time axis summing the last weeks worth of values at each pixel with a decay factor of $e^{(-i/10)}$ where i is the number of days in the past. This served to capture a memory of rain in the days prior while suppressing the impact of rain many days in the past. Any rain from 7 days or further in the past is not included.

2.3 Composition

Dataset	sites	Textures	Climate Class	Land Covers
Washita	20	Lo	Cfa	Grasslands
Fort Cobb	16	Lo	Cfa	Croplands, Grasslands

Table 2: Additional Validation datasets

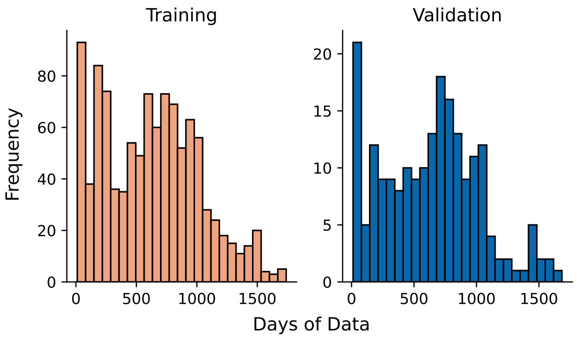


Figure 3: Number of stations with binned number of days of data available

2.4 OK Datasets

Nomenclature *_#
Where * is the prediction method
Where # is the metric

*	Translation
d	Dense
p	Prob
wdl	WDL ensemble
r	RF ensemble
smap	SMAP

#	Translation
r	Pearsons R
ub	ubRMSE
b	Bias

Table 3: short hand writings for metric tables e.g *_# could be *p_b* which means the bias scored by the Prob ensemble

Station	d_r	p_r	wdl_r	r_r	smap_r	d_ub	p_ub	wdl_ub	r_ub	smap_ub	d_b	p_b	wdl_b	r_b	smap_b
A121	0.822	0.788	0.795	0.811	**0.837**	0.051	0.054	0.052	0.056	**0.044**	**0.005**	0.190	0.020	0.056	0.045
A131	**0.844**	0.781	0.798	0.806	0.843	0.038	0.055	0.041	0.044	**0.037**	**0.006**	0.208	0.032	0.047	0.045
A253	**0.766**	0.683	0.663	0.668	0.747	**0.041**	0.059	0.048	0.048	0.045	**0.021**	0.216	0.047	0.034	0.024
A249	0.642	0.596	0.616	0.587	**0.653**	**0.061**	0.072	0.063	0.065	0.061	0.020	0.174	**0.001**	0.077	0.069
A154	0.745	0.698	0.710	0.748	**0.775**	0.052	0.060	0.054	0.055	**0.048**	0.054	0.140	**0.027**	0.099	0.093
A244	**0.797**	0.681	0.734	0.743	0.786	**0.046**	0.061	0.050	0.051	0.046	0.026	0.216	0.050	0.010	**0.006**
A262	0.772	0.645	0.689	0.720	**0.777**	**0.042**	0.064	0.047	0.046	0.043	0.015	0.179	**0.005**	0.055	0.050
A235	**0.661**	0.570	0.554	0.561	0.585	**0.062**	0.075	0.068	0.068	0.069	0.028	0.166	**0.004**	0.066	0.064
A152	**0.720**	0.657	0.668	0.660	0.693	**0.041**	0.061	0.044	0.044	0.049	0.079	0.272	0.098	**0.020**	0.029
A234	0.802	0.749	0.795	0.808	**0.829**	**0.032**	0.054	0.033	0.033	0.036	0.114	0.309	0.136	**0.063**	0.068
A132	0.747	0.654	0.686	0.737	**0.781**	0.029	0.062	0.033	**0.025**	0.044	0.170	0.357	0.193	**0.131**	**0.131**
A282	**0.787**	0.642	0.675	0.704	0.781	0.028	0.068	0.036	**0.028**	0.042	0.143	0.336	0.166	**0.095**	0.102
A250	**0.618**	0.513	0.512	0.540	0.611	**0.050**	0.075	0.057	0.054	0.057	0.046	0.243	0.069	**0.002**	0.004
A148	**0.762**	0.676	0.691	0.705	0.755	**0.046**	0.061	0.051	0.051	0.047	**0.002**	0.192	0.023	0.049	0.044
A124	0.825	0.700	0.744	0.798	**0.859**	0.040	0.058	0.045	0.045	**0.035**	0.119	0.311	0.139	**0.079**	**0.079**
A159	0.813	0.688	0.757	0.770	**0.840**	**0.028**	0.057	0.032	0.029	0.036	0.098	0.283	0.119	**0.067**	0.068
A146	0.817	0.802	0.809	0.820	**0.845**	**0.030**	0.050	0.031	0.030	0.035	0.026	0.227	0.051	0.026	**0.021**
A133	**0.717**	0.635	0.646	0.672	0.709	0.032	0.064	0.036	**0.030**	0.045	0.155	0.351	0.180	**0.106**	0.109
A256	**0.687**	0.546	0.610	0.615	0.661	**0.035**	0.062	0.040	0.037	0.047	0.119	0.310	0.146	**0.081**	0.089
A136	**0.706**	0.513	0.465	0.525	0.539	**0.038**	0.060	0.051	0.046	0.055	0.052	0.239	0.073	0.004	**0.001**
Avg.	**0.752**	0.661	0.681	0.700	0.745	**0.041**	0.062	0.046	0.044	0.046	0.065	0.246	0.079	0.058	**0.057**

Table 4: Washita station metrics. Sorted by stations with most information to least

	d_r	p_r	wdl_r	r_r	smap_r	d_ub	p_ub	wdl_ub	r_ub	smap_ub	d_b	p_b	wdl_b	r_b	smap_b
Station															
F110	**0.762**	0.644	0.623	0.658	0.751	0.032	0.050	0.036	**0.028**	0.046	0.115	0.221	0.166	0.120	**0.104**
F108	**0.767**	0.734	0.689	0.695	0.750	0.064	0.064	0.070	0.075	**0.063**	-0.088	**0.011**	-0.032	-0.083	-0.105
F103	**0.805**	0.681	0.623	0.667	0.778	**0.033**	0.054	0.043	0.038	0.043	**-0.004**	0.094	0.031	-0.018	-0.038
F112	0.716	0.743	**0.752**	0.717	0.717	0.056	**0.054**	**0.054**	0.060	0.056	-0.067	0.034	**-0.016**	-0.069	-0.087
F109	**0.806**	0.745	0.714	0.728	0.782	0.031	0.043	0.033	**0.030**	0.041	0.064	0.153	0.114	0.057	**0.042**
F207	0.761	0.764	0.765	0.782	**0.803**	0.037	0.041	0.035	**0.034**	0.039	0.084	0.177	0.124	0.064	**0.050**
F101	0.743	0.662	0.683	0.722	**0.764**	0.036	0.055	0.036	**0.028**	0.045	0.087	0.187	0.131	0.087	**0.066**
F106	**0.811**	0.744	0.714	0.733	0.792	**0.036**	0.047	0.043	0.042	0.041	**-0.013**	0.089	0.019	-0.027	-0.046
F105	0.844	0.840	0.794	0.831	**0.847**	0.049	0.046	0.056	0.058	**0.045**	-0.046	0.047	**0.003**	-0.053	-0.072
F215	0.769	0.733	0.746	0.760	**0.784**	**0.044**	0.052	0.046	0.048	0.045	**-0.002**	0.098	0.047	-0.012	-0.025
F104	**0.739**	0.661	0.646	0.657	0.730	0.037	0.056	0.040	**0.035**	0.046	0.085	0.185	0.133	0.076	**0.055**
F111	**0.835**	0.809	0.785	0.797	0.827	0.029	0.035	0.029	**0.028**	0.036	0.070	0.158	0.113	0.059	**0.047**
F102	**0.769**	0.746	0.708	0.727	0.767	**0.039**	0.044	0.041	0.040	0.043	**-0.004**	0.084	0.032	-0.030	-0.045
F113	0.634	0.661	0.651	0.663	**0.696**	0.048	0.050	0.044	**0.043**	0.050	-0.132	**-0.039**	-0.091	-0.140	-0.155
F114	0.764	0.717	0.707	0.733	**0.766**	**0.044**	0.051	0.048	0.049	0.045	**-0.006**	0.093	0.036	**-0.006**	-0.021
F214	**0.861**	0.764	0.655	0.688	0.796	**0.048**	0.057	0.067	0.068	0.052	0.072	0.152	0.113	0.060	**0.041**
F213	0.322	0.353	0.174	0.416	**0.425**	0.045	0.043	0.032	**0.029**	0.052	0.075	0.179	0.136	0.087	**0.061**
Avg.	0.748	0.706	0.672	0.705	**0.752**	**0.042**	0.049	0.044	0.043	0.046	0.017	0.113	0.062	0.010	**-0.008**

Table 5: Fort Cobb station metrics. Sorted by stations with most information to least

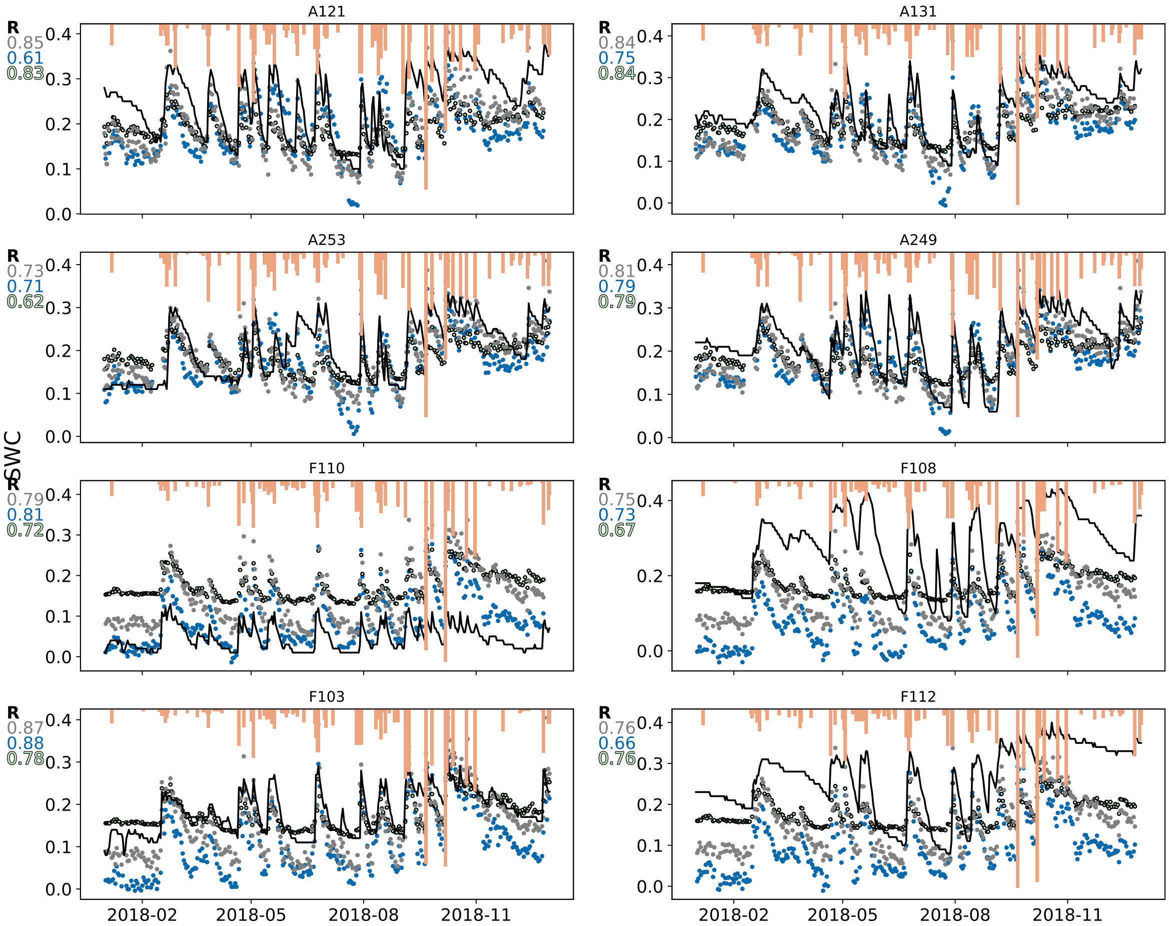


Figure 4: Temporal Predictions for top 4 sites for the Washita and Fort Cobb networks. Grey is SMAP, blue is Dense, and green is RF.

3 Spatial Predictions

Spatial predictions are made using the highest spatial and temporal resolution data. Spatial predictions use the 250m NDVI resolution (MOD13Q1) and the daily resolution LST product (MOD21A2). The time-padded LST product (MOD11A1) can be used to reduce the impact of cloud cover, however, this product can have anomalies if that 8-day period it averages over was very cloudy. This produces aesthetically unappealing artifacts in spatial predictions.

Spatial data is resampled to the highest resolution (90m) using nearest neighbor interpolation. This means that the coordinates of prediction pixels are defined by the highest resolution data pixel and that coarser data contributes the same values to patches of pixels in spatial predictions.

Additional spatial predictions are seen in Figure 5. Here we see that all methods agree on overarching structure of SWC. Tender Foot Creek exhibits an interesting pattern. Here all models identify two moist regions surrounded by dry, wheras in SMAP there exists a smooth gradient between these two regions. This may be a product of the backus-gilbert interpolation.

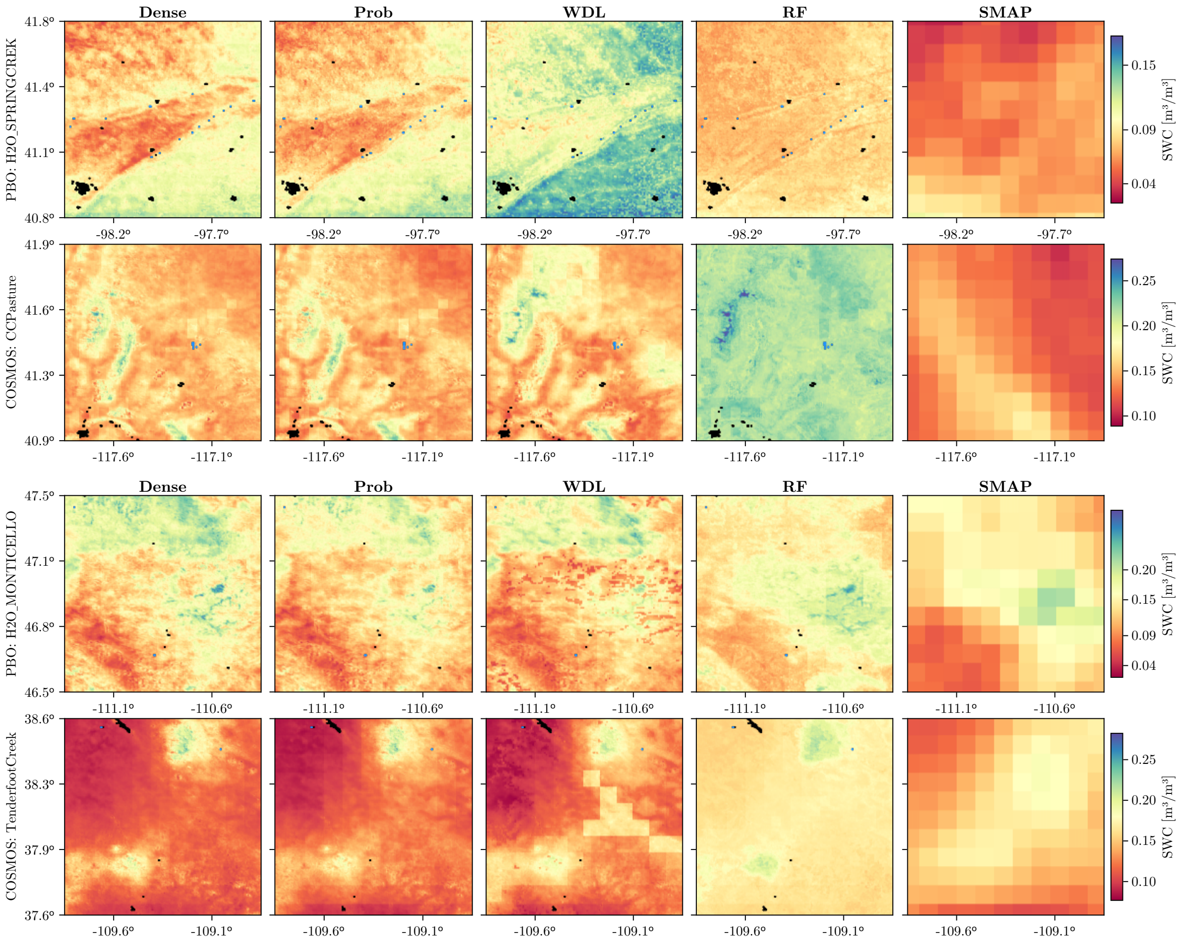
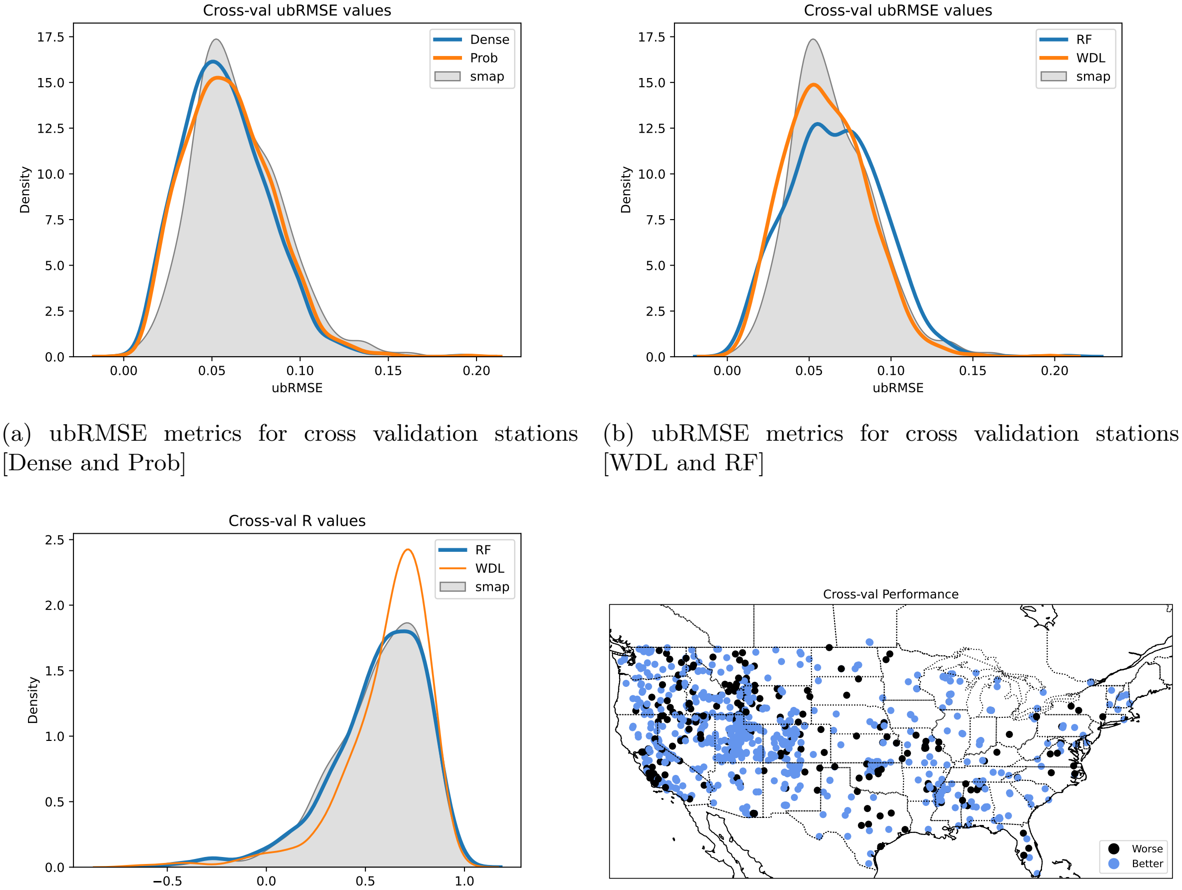


Figure 5: Spatial Preds

4 Cross Validation



(a) ubRMSE metrics for cross validation stations [Dense and Prob]

(b) ubRMSE metrics for cross validation stations [WDL and RF]

(c) R scores for excluded models

(d) Locations where Dense outperforms SMAPs ubRMSE value on the cross validation dataset

Figure 6: Cross Validation metrics and performance

5 SHAP

Validating the spatial predictions is difficult due to the lack of ground truth data at the same high resolution. Timeseries data provides the only pure validation source. One analysis method involves observing the learned sensitivities of the models on their spatial predictions to see if they have learned patterns that align with empirical observations and expectations. In other words, "do the models find the same correlations between variables that we expected before training them?" To accomplish this, the spatial predictions made earlier were analyzed to observe prediction sensitivity to changes in input variables.

In Fig. 7 we can see these varying sensitivities to specific variables as well as the trends/correlations that are being drawn. The most apparent source of concern is the inversely correlated NDVI sensitivities for all of the models. This trend is the opposite from what we would expect from previous literature. Generally, an increase in NDVI correlates with increases in SWC. However, as seen in Fig. **??**, that positive relationship is not readily apparent in the training data and in fact, the negative correlation is present at low SWC. Besides this trend in NDVI, all of the models appear to have learned appropriate relationships between the variables and SWC. The RF ensemble has a more noisy signal on many of the lower sensitivity variables. The Dense model has a noisy signal on pH showing no clearly identified trend. Overall, the Prob ensemble demonstrates the clearest adherence to learned trends.

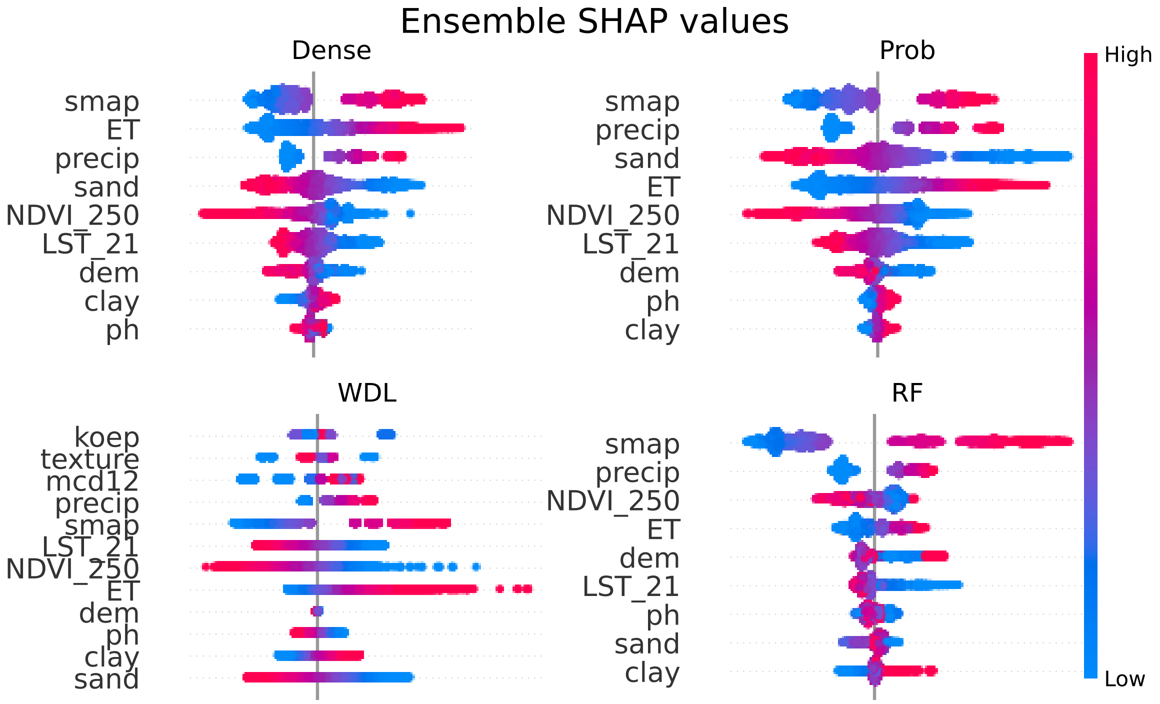


Figure 7: SHAP values for each input variable for each Ensemble. WDL takes three categorical inputs and these three exhibit the strongest sensitivity due to their categorical nature.

6 Ensemble Advantage

The RF model ensemble members are trained only seeing a specific subset of data. As a result all members are missing some contexts. Although some ensemble members can outperform the ensemble on bias and ubRMSE, these predictions are not consistent through time and there exists a significant ensemble advantage for the R metric across all members (Figure 8)

7 Domain Preference

To further explore areas of strengths and weakness', metrics are calculated across each of the three categorical static variables: **texture**, **climate class**, and **land cover**. These static variables are further broken down into the subclasses previously shown in Table 1 of the main text. A significant drop in metric performance in one of these subclasses may indicate an inability for a model to fully generalize SWC from the input variables. To search for these preferences/weaknesses we compute the average metric score for a model on each station in the 40 subclasses from Table 1. We then divide this by the average performance for all models on that subclass. This final value gives us the relative performance of a model compared to all others. If any models performance is at least 10% better or worse than the mean score for all models on that subclass, then that model is deemed to have a bias for that subclass. These instances are seen in Table 6. The Bias metric was excluded as the RF model consistently exhibited poor bias. The only instance where a model demonstrates a negative or positive performance on both ubRMSE and R was on Sand. Here, the Dense R value is 40% the mean R value and the ubRMSE is 124% the mean ubRMSE value. This category constitutes only one stations worth of data and so no conclusions can be made about the models performance on sand overall.

Although there doesn't appear to be any strong or negative biases for any single static variables, what if there exists a combination of inputs that exhibit difficulties?

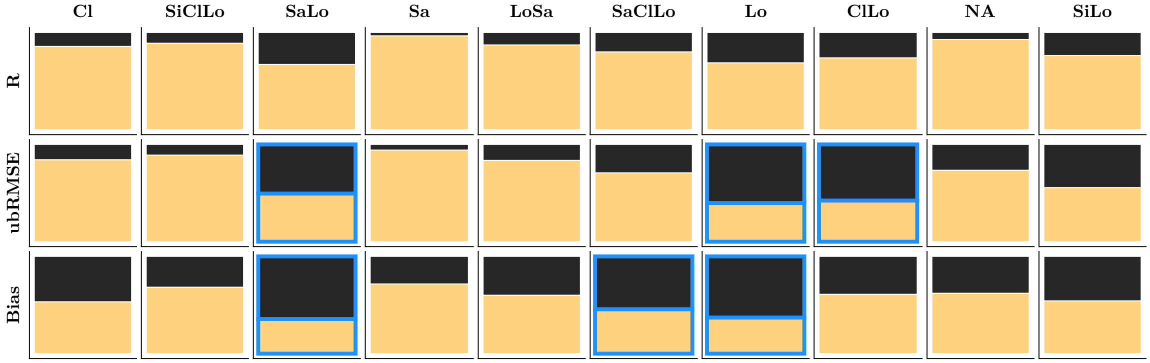


Figure 8: Ensemble head-to-head on normalized domains with RF. Blue highlighted boxes represent a metric where an ensemble member outperformed the ensemble on more domains stations weighted by domains

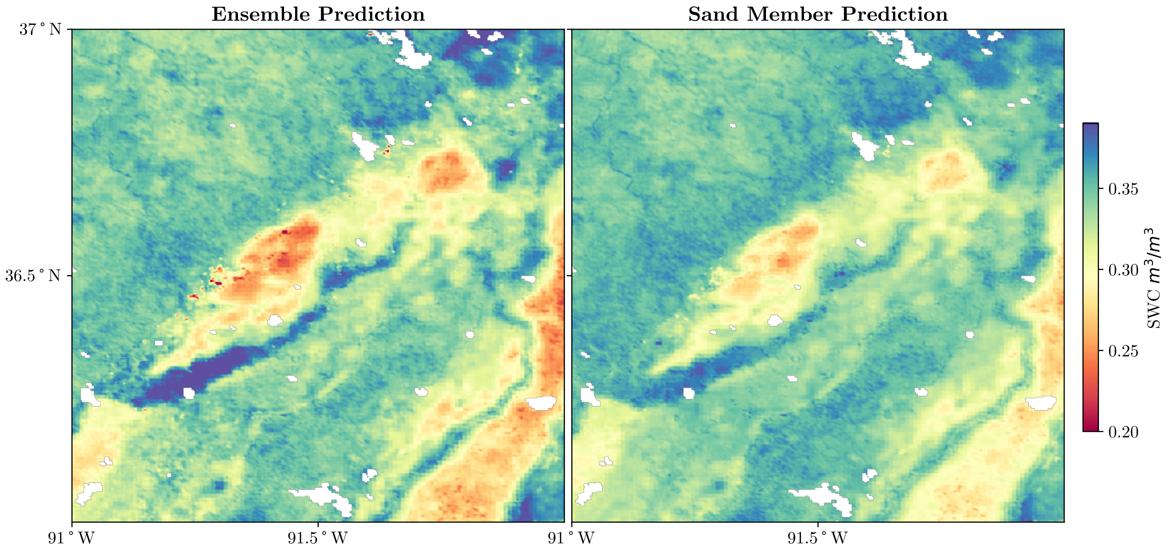


Figure 9: Prob Ensemble spatial predictions vs top performing member

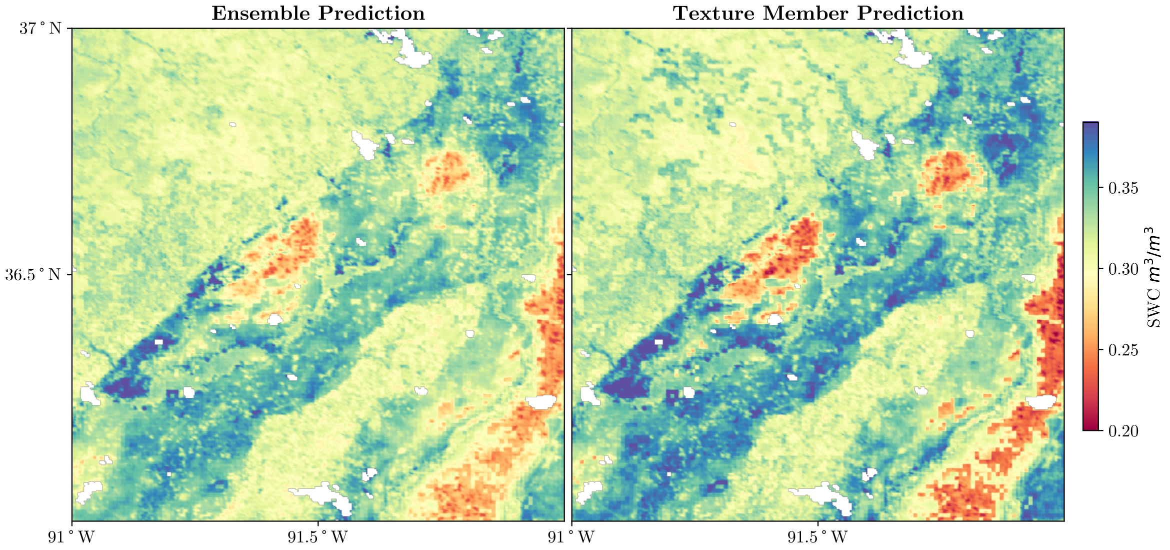


Figure 10: WDL Ensemble spatial predictions vs top performing member

Characteristic	Dense	Prob	WDL	RF	No. of Stations
			R		
SiClLo	1.07	1.05	**0.83**	1.05	3
Mxd Frsts	1.08	0.98	**0.89**	1.04	3
Bsh	1.04	1.05	**0.88**	1.02	2
Sa	**0.44**	1.21	1.17	1.18	1
			ubRMSE		
Csa	0.92	0.99	**1.10**	0.98	24
Opn Shrblnds	0.94	1.01	**1.14**	0.91	6
SaClLo	1.03	1.04	1.04	**0.89**	3
Bsh	0.95	**1.14**	0.94	0.91	2
ET	1.00	**1.14**	0.94	0.92	2
BWh	0.99	1.13	0.99	**0.90**	1
Sa	**1.24**	**0.71**	1.05	1.00	1
Cl	**0.85**	1.03	1.09	1.03	1

Table 6: Static classes where one model displays a bias (an average metric score on that class which deviates 10% or more from the mean of all models) for that specific class. For R, values greater than 1.0 outperform the mean, for ubRMSE values below 1.0 outperform the mean. No. of stations represents number of locations possessing that characteristic

8 Code and Data availability

All relevant code and data can be accessed at https://github.com/TheJeran/ensemble-downscaling The RF model files are not available as they are too large for github. However, a new set can easily be trained by running through the available code.

References

[1] Marta Chiesi, Piero Battista, Luca Fibbi, Lorenzo Gardin, Maurizio Pieri, Bernardo Rapi, Maurizio Romani, Francesco Sabatini, and Fabio Maselli. Spatio-temporal fusion of NDVI data for simulating soil water content in heterogeneous Mediterranean areas. *European Journal of Remote Sensing*, 52(1):88–95, January 2019.

[2] Hongxue Zhang, Jianxia Chang, Lianpeng Zhang, Yimin Wang, Yunyun Li, and Xiaoyu Wang. NDVI dynamic changes and their relationship with meteorological factors and soil moisture. *Environmental Earth Sciences*, 77(16):582, August 2018.

[3] C. Brouwer, A. Goffeau, and M. Heibloem. CHAPTER 2 - SOIL AND WATER. In *Irrigation Water Management: Training Manual No. 1 - Introduction to Irrigation*. 1985.

[4] Jesper E Moeslund, Lars Arge, Peder K Bøcher, Tommy Dalgaard, Mette V Odgaard, Bettina Nygaard, and Jens-Christian Svenning. Topographically controlled soil moisture is the primary driver of local vegetation patterns across a lowland region. *Ecosphere*, 4(7):art91, July 2013.

[5] H.J. Tromp-van Meerveld and J.J. McDonnell. On the interrelations between topography, soil depth, soil moisture, transpiration rates and species distribution at the hillslope scale. *Advances in Water Resources*, 29(2):293–310, February 2006.

[6] Sybil G. Gotsch, Kenneth Davidson, Jessica G. Murray, Vanessa J. Duarte, and Danel Draguljić. Vapor pressure deficit predicts epiphyte abundance across an elevational gradient in a tropical montane region. *American Journal of Botany*, 104(12):1790–1801, December 2017.

[7] Wenping Yuan, Yi Zheng, Shilong Piao, Philippe Ciais, Danica Lombardozzi, Yingping Wang, Youngryel Ryu, Guixing Chen, Wenjie Dong, Zhongming Hu, Atul K. Jain, Chongya Jiang, Etsushi Kato, Shihua Li, Sebastian Lienert, Shuguang Liu, Julia E.M.S. Nabel, Zhangcai Qin, Timothy Quine, Stephen Sitch, William K. Smith, Fan Wang, Chaoyang Wu, Zhiqiang Xiao, and Song Yang. Increased atmospheric vapor pressure deficit reduces global vegetation growth. *Science Advances*, 5(8):eaax1396, August 2019.