# Testing the Complementary Hypothesis: Can Atmospheric Demand Be Used to Estimate Soil Moisture?

*Colin Brust*
*11 December, 2019*
*CSCI 547*

**INTRODUCTION**

Evapotranspiration (ET) is the second largest terrestrial water flux, returning more than 60% of precipitation to the atmosphere, annually (Oki and Kanae, 2006). To estimate ET at the global scale, remote sensing-based (RS) ET models are often used. RS ET models use a combination of satellite surface reflectance data and gridded meteorological data to model ET at the global scale with fine spatial and temporal resolutions (e.g., Fisher et al. (2008); Mu et al. (2011)). Although RS ET models are generally accurate to within 1-2 mm per day at the global scale, all models make simplifying assumptions that can affect accuracy at smaller spatial scales (Purdy et al., 2018; Zhang et al., 2019). One such assumption is that atmospheric conditions such as the vapor pressure deficit (VPD), temperature, and relative humidity (RH) are representative of soil moisture (SM) conditions (Fisher et al., 2008; Mu et al., 2011). This assumption is justified by Bouchet's complementary hypothesis, which posits that at sufficiently large spatial and temporal scales, the evaporative demand of the atmosphere is at equilibrium with the soil water available for ET (Bouchet, 1963; Fisher et al., 2008). This is problematic however, because many RS ET models produce estimates at small spatial and temporal scales, meaning this assumption may not hold true.

One such RS ET model is MOD16, which provides global estimates of ET at an 8-day time step and 500 meter spatial resolution. MOD16 assumes that the VPD, RH and temperature can be used to approximate the SM constraints on soil evaporation and plant transpiration, the two largest components of ET (Mu et al., 2011). Specifically, MOD16 assumes that VPD and RH can be used to approximate the SM constraint on evaporation and that temperature and the VPD can be used to approximate the SM constraint on transpiration. This brings to question whether these meteorological variables can actually be used to approximate SM conditions in the context of MOD16.

Several studies have found that it is indeed possible to accurately approximate SM using various machine learning (ML) approaches (Cai et al., 2019; Morellos et al., 2016; Ge et al., 2019; Padarian et al., 2019). For example, Cai et al. (2019) used temperature, precipitation, RH and a number of other meteorological variables to predict SM at three sites across China, each with distinct soil and landcover characteristics. They used a simple neural network consisting of two hidden layers composed of 100 and 50 nodes to estimate SM with a daily error smaller than 0.5%. While Cai et al. (2019) were able to accurately predict SM using some of the same variables that MOD16 assumes can represent SM conditions, there are two limitations. 1) An input feature into their model was initial SM, meaning some SM dataset was still necessary for the model to work properly. 2) The architecture of the model was set up such that predictions were made at the point scale, not the spatially continuous grid that would be necessary as input into an RS ET model such as MOD16. Ge et al. (2019) developed two separate ML models that overcome these limitations. They used hyperspectral data from a UAV-mounted sensor as inputs into both a neural network and a random forest model to model a continuous grid of SM. They found that while both models had test set root mean square errors (RMSE) of less than 2%, the random forest model consistently was the best performer and was their recommended method for estimating SM. While Ge et al. (2019) didn't use any meteorological variables as input into their model, their study shows that spatially continuous SM can be accurately modeled using a random forest ML model. In a similar study, Padarian et al. (2019) used a convolutional neural network on spectral signals from soil samples to estimate key soil traits such as organic matter content and cation exchange

capacity. They found that compared to the linear modeling methods typically used to estimate these variables, the ML approach significantly reduced errors.

Although none of these studies can directly be applied to the question at hand, they do show that ML is a tool well suited for estimating soil properties such as SM. As such, this study seeks to estimate SM using the same meteorological variables that MOD16 uses as a proxy for SM conditions. If a combination of these variables can be fed into a ML model accurately predicts SM, it would suggest that the approximations of MOD16 are valid and VPD, RH and temperature can be used as a substitute for SM data. However, if the variables are unable to predict SM, it would suggest that these variables alone are not enough to predict SM, and that actual SM observations should instead be included as an input into MOD16.

## MATERIALS AND METHODS

This analysis was conducted in the coterminous United States (CONUS) from January 1st to December 31st 2016. The Gridmet dataset was the source of the VPD, RH and temperature data used in the analysis. Gridmet is a gridded meteorological dataset that provides daily estimates of VPD, RH and temperature for the CONUS domain at a 4km spatial resolution (Abatzoglou, 2013). The Soil Moisture Active Passive Level 4 Soil Moisture (SMAP L4_SM) product was used as the 'ground truth' SM input for the analysis. SMAP L4_SM uses observations from the SMAP satellite paired with land surface model estimates of hydrologic conditions to produce root zone and surface SM (measured as % volumetric water content) estimates at a 3-hour timestep and 9km spatial resolution for the CONUS domain (Reichle and Lannoy, 2014). Although the SMAP L4_SM product isn't as accurate as SM readings from ground stations, the dataset underwent rigorous testing and validation, yielding accuracies at or below 4% volumetric water content(Reichle et al., 2016). The loss in accuracy is a worthwhile trade off for the spatial and temporal fidelity of the product, which allows for model training and testing across all 9km pixels in the CONUS domain, rather than just a small subset of ground stations sparsely scattered across the domain. Because the SMAP data are available at a 3-hour timestep, mean daily values were calculated to upscale to the Gridmet daily timestep. To ensure a uniform spatial resolution between Gridmet inputs and the SMAP data, Gridmet data were bilinearly interpolated to the 9km SMAP grid.

A random forest regression was conducted to determine whether the VPD, RH and temperature can be used to accurately estimate SM. Two separate random forest models were created. The first uses VPD and RH to estimate surface SM. This is because these two variables are used to constrain soil evaporation in MOD16, which only occurs in the surface layer (top 5 cm) of the soil profile (Purdy et al., 2018). Similarly, the second model uses temperature and the VPD to estimate rootzone SM. This is because these two variables are used to constrain plant transpiration in MOD16, which is driven by the amount of water in the soil rootzone (top meter of soil) (Novák et al., 2005; Purdy et al., 2018). For both surface and rootzone predictions, a 'full' model was also created. The full model uses all meteorological variables as well as latitude, longitude and the day of year as predictor variables. These full models were created to determine whether the additional information improved estimates of SM relative to the baseline models.

The analysis was conducted in three main steps:

1. **Data extraction**. For every day in 2016, 500 randomly spaced points were created across the CONUS domain and used to extract underlying Gridmet and SMAP values. This process yielded 182,500 unique data points that were be used for training the model. This process was carried out in Google Earth Engine (GEE) platform, as it had all the datasets necessary for analysis and can very quickly perform these large spatial queries (Gorelick et al., 2017). Once the data were extracted, they were exported as a CSV to use as input into the next step.

2. **Hyperparameter Tuning**. Although Random Forest is a powerful ML algorithm, it is often necessary to make fine scale adjustments to the input parameters to optimize the model's performance (Probst et al.,

| Region | Soil Location | Model | Error Metric | Value |
|---|---|---|---|---|
| Arizona | Rootzone | Partial | RMSE | 0.0564 |
| Arizona | Rootzone | Full | RMSE | 0.0504 |
| Arizona | Surface | Partial | RMSE | 0.0407 |
| Arizona | Surface | Full | RMSE | 0.0355 |
| Arizona | Rootzone | Partial | R2 | 0.1920 |
| Arizona | Rootzone | Full | R2 | 0.2664 |
| Arizona | Surface | Partial | R2 | 0.3213 |
| Arizona | Surface | Full | R2 | 0.3468 |
| Midwest | Rootzone | Partial | RMSE | 0.0834 |
| Midwest | Rootzone | Full | RMSE | 0.0513 |
| Midwest | Surface | Partial | RMSE | 0.0709 |
| Midwest | Surface | Full | RMSE | 0.0469 |
| Midwest | Rootzone | Partial | R2 | 0.0110 |
| Midwest | Rootzone | Full | R2 | 0.1873 |
| Midwest | Surface | Partial | R2 | 0.0408 |
| Midwest | Surface | Full | R2 | 0.1406 |

**Table 1.** Accuracies of final random forest models calculated using the mean annual rootzone and surface SM. 'Region' refers to the sub-region the error metric was calculated for; 'Soil Location' is the location in the soil profile the model was attempting to predict; 'Model' refers to whether or not all possible predictors were used in the model; 'Error Metric' refers to the statistic used to assess the accuracy, RMSE is in units of % SM

2019). This hyperparameter tuning process was carried out using the Scikit-Learn Python package, as it has built in functionality to test a variety of model hyperparameters and find the combination that yields the highest accuracy (Pedregosa et al., 2011). The hyperparameters were tuned using a random sample of 10,000 points from the original 182,500 to minimize the local computing power needed at this step.

3. **CONUS-scale Model Training and Testing**. Once the optimal hyperparameters for the model were found, they were used to train a random forest model on the GEE platform. The model was trained on a random sample of 30,000 points from the original 182,500 point dataset. After training, the models were used to predict surface and rootzone SM for the entire CONUS domain for every day of 2016.

To assess the accuracy of the final models, mean annual root mean squared error (RMSE) and r-squared correspondence were calculated at two sub-regions. Both regions were approximately 500km x 500km and spanned two geographically and climatically diverse areas. The first covered most of the state of Arizona while the second covered a large portion of the midwest centered on Kentucky. Accuracy was calculated at each area to determine whether the models performed better in arid or moist regions of the CONUS.

**RESULTS**

Across the two sub-regions assessed in this analysis, all models can predict both rootzone and surface SM with RMSEs ranging from 3 - 8% (Table 1). R-squared correspondence for all models across both sub-regions ranges from 0.01 to 0.35. Disregarding the error metric, region and soil location, the 'full' models that use all meteorological variables as well as geographic and temporal information as predictors always outperform the models that exclusively use the meteorological predictors. When looking at just rootzone and surface SM prediction accuracy, the surface SM estimates consistently outperform the rootzone SM estimates. When looking at error metrics between the two sub-regions, rootzone and surface SM errors are consistently lower in the Arizona region than the Midwest region.

At the CONUS scale, both the partial and full models miss out on the fine-scale patterns and variations in SM seen in the SMAP L4_SM product (Figure 1) In Figure 1, the SMAP product displays rootzone and surface SM values that are heterogeneous at fine spatial scales. However, the partial and full models predict SM values that

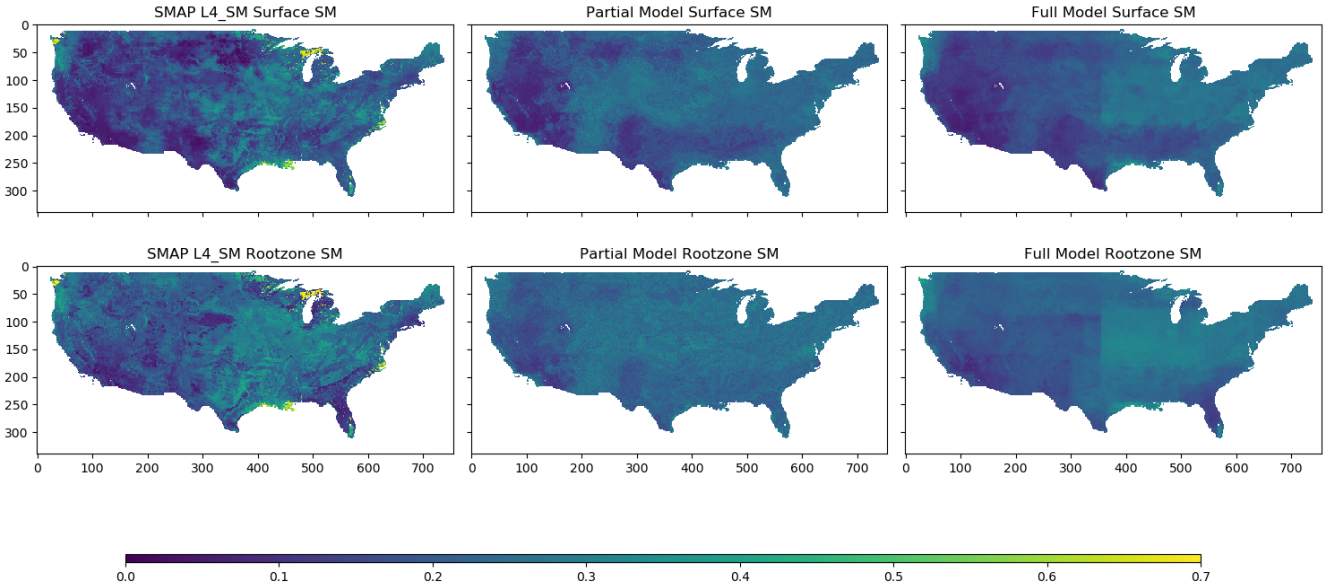


**Figure 1.** Spatial patterns of SM on July 1st, 2016 for the CONUS domain. The first row shows surface SM observations/estimates and the second row shows rootzone SM observations/estimates. The first columns shows SMAP observations, the second column shows the partial model estimates and the third column shows the full model estimates.

are very homogeneous across much of the CONUS domain.

When looking at a time series of modeled SM at a single point, both models reflect the general seasonal trends of the SMAP L4_SM data (Figure 2). Both surface and rootzone SM show a general decreasing trend up to approximately the 200 day mark, when SM begins to steadily increase. This is consistent with the SMAP SM patterns, which follows the same general temporal trends. While the modeled SM matches the large-scale, seasonal trends of the SMAP SM, there is variability in the model estimates on a day to day basis, which is notably different as compared to the SMAP L4_SM timeseries. Across the timeseries, model rootzone RMSE and r-squared was 0.0196 and 0.1702, respectively, and model surface RMSE and r-squared was 0.0353 and 0.3758, respectively. These results are similar to the results seen in Table 1, suggesting that the model performance is similar at daily and annual timescales.

## DISCUSSION

Across both study sites and both surface and rootzone SM models, the full models had consistently higher r-squared values and lower RMSE values than the models that exclusively used the VPD, RH and temperature. Intuitively, this makes sense, as the climate of the CONUS changes significantly across both time and space. The importance of the spatial input data is clear in the full model results of Figure 1, where there is a vertical dividing line across the middle of the CONUS. To the west of the line, there is generally lower SM and to the east of the line, there is generally higher SM. This spatial pattern matches the well known west to east aridity gradient seen in the CONUS.

Another way to assess whether or not the VPD, RH and temperature are sufficient for predicting SM is to leverage

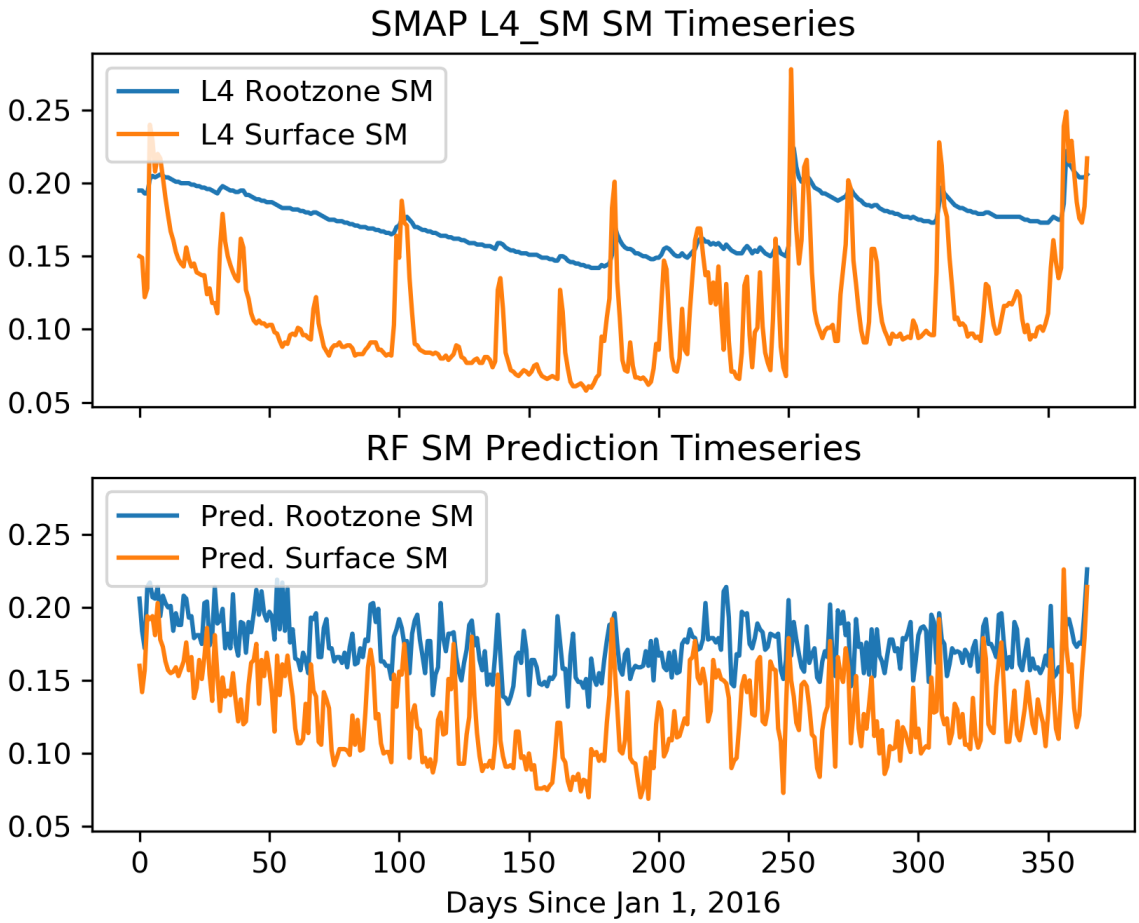


**Figure 2.** Timeseries of SMAP L4_SM and modeled SM from one pixel in southeastern Arizona. The top plot shows the rootzone and surface from the SMAP L4_SM product and the bottom plot shows the predicted rootzone and surface SM from the 'full' model.

the variable importance output of the random forest model. The random forest algorithm computes the relative importance of all variables used as predictors, showing which variables are most important for estimating SM. For this analysis, latitude, longitude and day of the year consistently came out as the most important predictors of SM in the full model. This variable importance held true for both the surface and rootzone SM estimates. This further suggests that the VPD, RH and temperature alone are not sufficient for predicting SM.

It is also important to note the difference in model performance in the Arizona and Midwest study regions. The model predictions were more accurate in the Arizona region than the Midwest region for almost all soil location, model and error metric combinations (Table 1). This suggests that the assumption of equilibrium between the atmospheric demand and SM made in MOD16 is more applicable in some regions than others. This is consistent with the literature, which suggests that the relationship between VPD and SM is spatially variable and should not be considered constant across the CONUS domain (Novick et al., 2016).

There was also a discrepancy between the model accuracies of rootzone and surface SM estimates. In general, model estimates of surface SM were more accurate than estimates of rootzone SM. This is possibly because rootzone SM represents the top meter of the soil whereas the surface SM represents the top 5cm of the soil. Because

the top 5cm of the soil are directly exposed to the atmosphere, it stands to reason that it will be more influenced by atmospheric conditions such as RH and the VPD. Additionally, the SMAP L4_SM surface SM observations have more variability than the rootzone observations, which better matches the high day to day variability seen in the rootzone and surface SM models (Figure 2).

Another notable issue with the models is the inability to predict high values of SM. In Figure 1, there are regions of extremely high SM in the Olympic Peninsula and Michigan's upper peninsula. In both the partial and full models, these high values are missed. This trend can also be seen in Figure 2, where there is a large spike in SMAP 4_SM surface SM around the 250 day mark. Although there is a slight spike in the modeled surface SM at the same time, the magnitude is nowhere near as large as the SMAP magnitude. A possible solution to this issue might be to manually add training points in these high SM regions to ensure that the model has sufficient training data to learn what combination of variables creates zones of high SM.

Although the goal of this analysis wasn't to develop a model that accurately predicts SM, these results do bring to question what would be necessary to accurately predict SM across the CONUS domain. An obvious first step would be to add more predictors. As seen in this analysis, adding three additional predictor variables led to a notable increase in model accuracy. Precipitation, incoming radiation, and landcover are all key inputs into many SM models, and adding them into this framework would likely further improve accuracy (Reichle et al., 2016). Another possible option could to use a different modeling approach. As seen in Figure 2, there is significant day to day variability in the modeled estimates of rootzone and surface SM. In contrast, there are generally smooth day to day transitions in SM seen in the SMAP L4_SM product. Using a recurrent neural network could be a promising way to smooth out this model variability. This is because recurrent neural networks have a 'memory' and model predictions are dependent on the historical outputs of the model (Hochreiter and Schmidhuber, 1997). This means that for a given day, the model would use both the predictor variables as well as the previous model state to make a prediction, potentially leading to a smoother time series of SM.

Although the best performing models were able to predict the SMAP L4_SM SM with an RMSE around 4%, it is important to note that SMAP L4_SM is itself a model. This means that the reported accuracy is likely lower than it would be if the models were compared to ground observations. Further, the spatial and temporal patterns of modeled SM did not emulate the SMAP L4_SM product. In order to produce a model that very accurately predicts SM, more input predictor variables and an updated model architecture are necessary. Because the surface and rootzone SM models weren't able to accurately predict SMAP L4_SM SM, it is recommended that the MOD16 algorithm incorporates a SM input to constrain soil evaporation and plant transpiration. Doing so will better physically represent these key processes and potentially lead to more accurate estimates of ET.

**REFERENCES**

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1):121–131.

Bouchet, R. J. (1963). Évapotranspiration Réelle Et Potentielle Signification Climatique. *International Association of Science and Hydrology*, pages 134–142.

Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., and Xue, X. (2019). Research on soil moisture prediction model based on deep learning. *PLoS ONE*, 14(4):1–19.

Fisher, J. B., Tu, K. P., and Baldocchi, D. D. (2008). Global estimates of the land-atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. *Remote Sensing of Environment*, 112:901–919.

Ge, X., Wang, J., Ding, J., Cao, X., Zhang, Z., Liu, J., and Li, X. (2019). Combining UAV-based hyperspectral imagery and machine learning algorithms for soil moisture content monitoring. *PeerJ*, 7:e6926.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Morellos, A., Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., and Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, 152:104–116.

Mu, Q., Zhao, M., and Running, S. W. (2011). Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment*, 115(8):1781–1800.

Novák, V., Hurtalová, T., and Matejka, F. (2005). Predicting the effects of soil water content and soil water potential on transpiration of maize. *Agricultural Water Management*, 76(3):211–223.

Novick, K. A., Ficklin, D. L., Stoy, P. C., Williams, C. A., Bohrer, G., Oishi, A. C., Papuga, S. A., Blanken, P. D., Noormets, A., Sulman, B. N., Scott, R. L., Wang, L., and Phillips, R. P. (2016). The increasing importance of atmospheric demand for ecosystem water and carbon fluxes. *Nature Climate Change*, 6(11):1023–1027.

Oki, T. and Kanae, S. (2006). Global Hydrological Cycles and World Water. *Science*, 313(August):1068–1073.

Padarian, J., Minasny, B., and McBratney, A. B. (2019). Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional*, 16:e00198.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and Bertand, T. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(1):2825 – 2830.

Probst, P., Wright, M. N., and Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):1–19.

Purdy, A. J., Fisher, J. B., Goulden, M. L., Colliander, A., Halverson, G., Tu, K., and Famiglietti, J. S. (2018). SMAP soil moisture improves global evapotranspiration. *Remote Sensing of Environment*, 219(December 2017):1–14.

Reichle, R., Lannoy, G., and Liu, Q. (2016). Global Modeling and Assimilation Office Soil Moisture Active Passive Mission L4 _ SM Data Product Assessment ( Version 2 Validated Release ). 12(12):1–55.

Reichle, Rolf Wade Crow, R. K. J. K. and Lannoy, G. D. (2014). Soil Moisture Active Passive ( SMAP ) Project Algorithm Theoretical Basis Document ( ATBD ) SMAP Level 4 Surface and Root Zone Soil Moisture ( L4 _-SM ) Data Product. *Review Literature And Arts Of The Americas*.

Zhang, K., Zhu, G., Ma, J., Yang, Y., Shang, S., and Gu, C. (2019). Parameter Analysis and Estimates for the MODIS Evapotranspiration Algorithm and Multiscale Verification. *Water Resources Research*, 55(3):2211–2231.