Applied ML - Final Project Proposal

Michael Gradwohl, Maximilian Losbichler, Wolfgang Preimesberger

Gap-filling (Satellite) Soil Moisture Time Series

Overview

Soil moisture (SM) is the (volumetric) water content stored in the top layer of the soil. Satellite observations in the microwave domain of the electromagnetic spectrum are sensitive to the presence of water and can therefore be used to estimate SM. However, the number of available satellites is limited, and they cannot provide seamless measurements around the globe at any time, as they operate on different orbits, with overpass times of often several days. Soil moisture -- as an environmental variable -- is connected to various other variables, such as precipitation, which drives the available water content, evaporation, which is the process of soil moisture decrease due to factors such as wind and (solar) radiation, temperature, which can be affected by the evaporative cooling effect that soil moisture has on the environment, or infiltration into deeper layers of the soil over time.
Machine learning models are therefore nowadays often used to model changes in this complex environmental system, and based on various available soil moisture covariates, such as the ones mentioned above, some studies have managed to predict soil moisture in case where no satellite measurements are available.

In this project, we will perform a first assessment using ML models, to understand how well we can predict regional soil moisture changes for a small number of study sites / measurement time series using a range of covariate variables. We will use gap-free reanalysis data, to train a model to predict soil moisture. We will perform a train/test split to assess how well our model can restore data in systematic (not random) satellite-like gaps in the originally gap-free reanalysis data.

Finally, if within the scope of this exercise, we would like to apply this model in a real world application, and fill actual data gaps in a satellite time series. For the selected study sites, independent in situ (“ground truth”) measurements are available, which could be used for independent performance assessment.

Proposed datasets

1. **Gap-free reanalysis variables from ERA5-Land**

The main dataset, from which we use time series extracts to train a model to predict soil moisture, is the ERA5-land global reanalysis dataset *(Muñoz-Sabater et al., 2021)*. This dataset provides observation-enhanced simulations (using data assimilation) of soil moisture and other climate variables from a well-established model by the European Center for Medium Range Weather Forecast (ECMWF).

While the data characteristics are not exactly the same as for satellite measurements, the dataset is gap-free and therefore a good candidate to train a model to predict/test soil moisture. This dataset also provides various of the before-mentioned soil moisture covariates, such as surface temperature, vegetation parameters, solar radiation, precipitation, evaporation, etc. Figure 1 shows an example time series.

2. **Optional: Satellite Data - ESA CCI Soil Moisture**

This dataset is developed by TU Wien and among the most widely used satellite soil moisture records *(Dorigo et al., 2017).* It merges the data from currently 19 satellites to create a daily soil moisture record with the best possible spatio-temporal coverage (compare Figure 2). Previous work on gap-filling the data with stand-alone (statistical) models has indicated good performance of the predictions when using relatively simple statistical methods based on available (temporal and spatial) neighborhood information *(Preimesberger et al., 2025)*. ML could further improve these predictions and provide an alternative approach in the future.

3. **Optional: In situ data - International Soil Moisture Network (ISMN)**

ISMN *(Dorigo et al., 2021*) provides a collection of in situ measurement time series of soil moisture. This is considered as the best available reference data to validate soil moisture. However, there can still be issues with this data such as inconsistencies in the time series due to movement of the sensor, or missing observations in winter etc. Figure 2 shows an example of an in situ time series.

Task Summary

<u>Task 1 - Explorative data analysis</u>

Understanding the dynamics in (satellite) soil moisture and the systematics of missing observations. Explore which other variables are most likely to be useful as support to predict soil moisture.

- Inspect the dataset and verify data types, identify missing values, and outliers.
- Compute basic statistics (mean, median, standard deviation, etc.).
- Visualize the distribution of soil moisture from different sources using histograms.
- Analyze correlations and lags between soil moisture and other related variables (precipitation, temperature, evaporation, etc.) using correlation coefficients and appropriate visualizations

<u>Task 2 - Strategy for Model Training and Evaluation</u>

- Understanding the systematics (or randomness) of missing data in satellite measurements and adapt them to select training/test samples from gap-free reanalysis time series
- Design an appropriate framework to predict missing soil moisture values in a satellite-like time series.

<u>Task 3 - Baseline models</u>

- 1-2 baseline models to predict SM in the introduced gaps without any ancillary data, e.g., linear interpolation, forward filling, or the DCT-PLS smoother *(Garcia 2010*), that was previously used to fill CCI SM gaps *(Preimesberger et al., 2025)*.

Task 4 - Machine Learning models

- If possible, one feature configuration should only use the available soil moisture data (with gaps) and predict SM over time without any ancillary support.
- One feature configuration should use appropriate ancillary variables from ERA5-Land to improve the predictions using a NN model (LSTM).

Task 5 - Hyperparameter tuning

- Improve the performance of the neural network model through hyperparameter tuning. Compare the tuned model against the default neural network and discuss the observed improvements.

Task 6 - Evaluation and Discussion

- Estimate how close to the original (test) data the predictions come (from the baseline models and the NN with and without hyperparameter tuning).
- From a practical perspective, explain which model you would recommend for real-world applications and why. Clearly describe the strengths and weaknesses of each approach

---

Optional Task 1 - Applying the derived model to satellite data (with inherent data gaps)

- Having a model trained and tested with (original gap-free) reanalysis data, it should be used to fill data gaps in the actual satellite observation time series
- Analyze differences in the performance of these predictions with respect to predictions from the previous task. Consider differences betwee satellite and reanalysis time series (e.g., noise) from Task 1.

Optional Task 2 - Comparing the predictions against in situ measurements

- Independent ("ground truth") reference measurements are available and could be used for independent assessment of the predictions. While these data are not expected to be exactly the same as the satellite measurements (they represent a point scale, while satellites measure a large area), they are still considered the best available source of reference data.

All input data are available at https://cloud.geo.tuwien.ac.at/s/sHNDSpFaD4bBZWE (not all provided locations are used in the project)

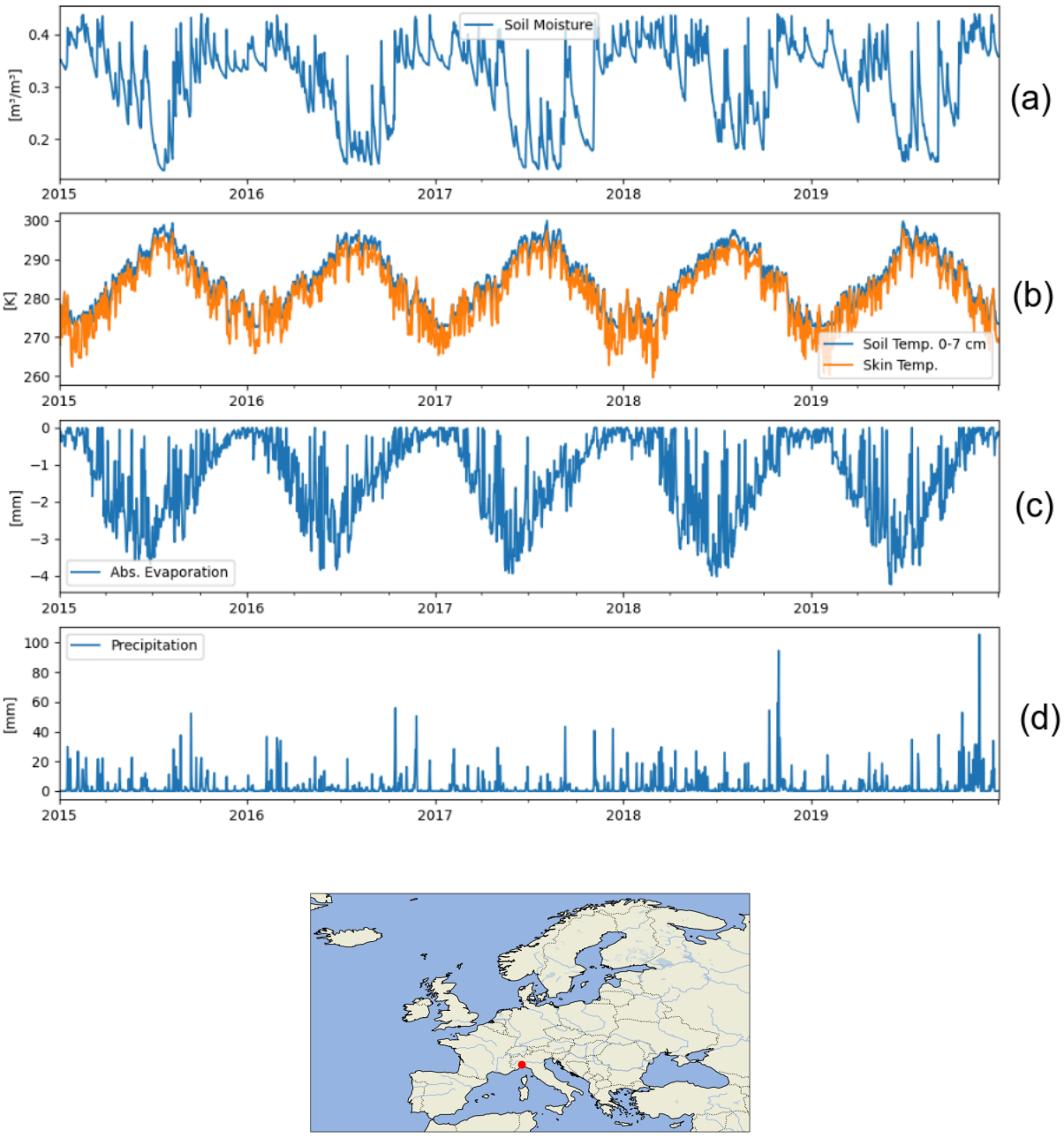

Figure 1 - Reanalysis variables: (a) soil moisture, (b) temperature, (c) evaporation, (d) precipitation, for one test location in Italy (e).

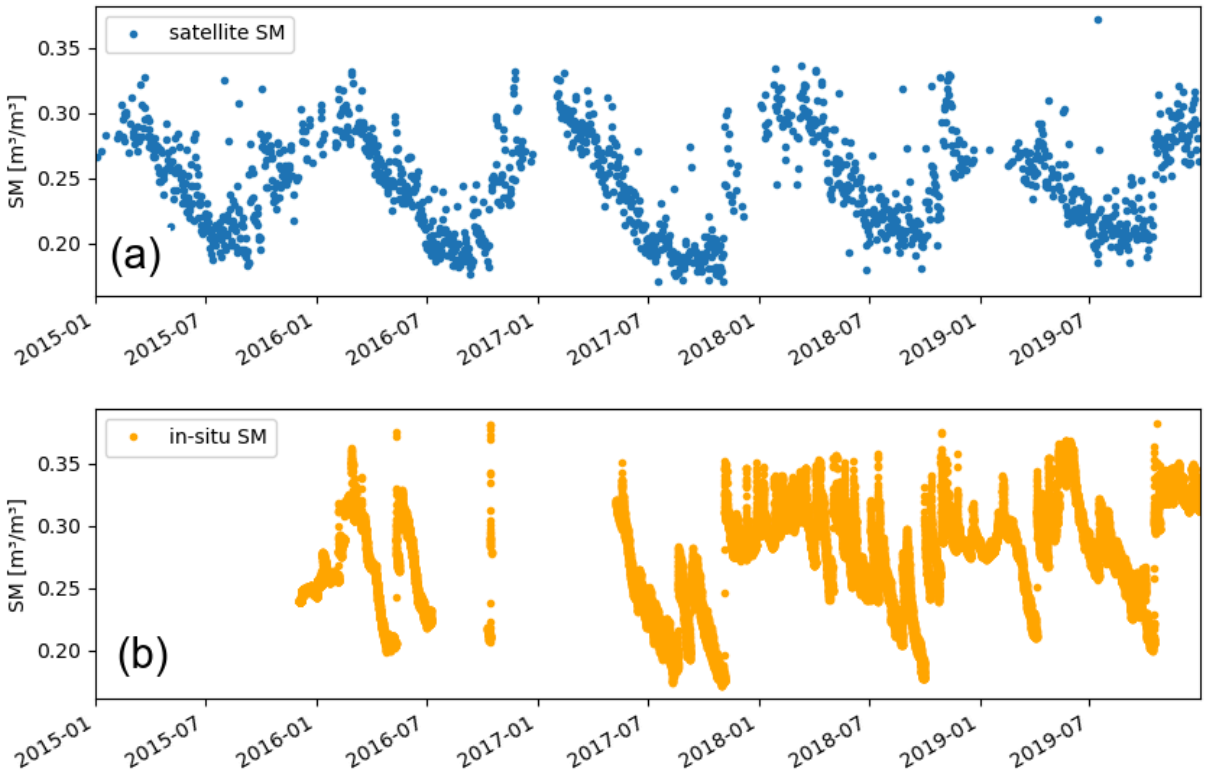

Figure 2 - Observed soil moisture in CCI satellite measurements with real data gaps (a) and collocated in situ measurements (b) for the same location as indicated in Figure 1e.

References

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S. I., Smolander, T., & Lecomte, P. (2017). *ESA CCI soil moisture for improved Earth system understanding: State-of-the-art and future directions*. *Remote Sensing of Environment, 203*, 185–215. https://doi.org/10.1016/j.rse.2017.07.001

Dorigo, W., Himmelbauer, I., Aberer, D., Schremmer, L., Petrakovic, I., Zappa, L., Preimesberger, W., Xaver, A., Annor, F., Ardö, J., Baldocchi, D., Bitelli, M., Blöschl, G., Bogena, H., Brocca, L., Calvet, J.-C., Camarero, J. J., Capello, G., Choi, M., Cosh, M. C., van de Giesen, N., Hajdu, I., Ikonen, J., Jensen, K. H., Kanniah, K. D., de Kat, I., Kirchengast, G., Kumar Rai, P., Kyrouac, J., Larson, K., Liu, S., Loew, A., Moghaddam, M., Martínez Fernández, J., Mattar Bader, C., Morbidelli, R., Musial, J. P., Osenga, E., Palecki, M. A., Pellarin, T., Petropoulos, G. P., Pfeil, I., Powers, J., Robock, A., Rüdiger, C., Rummel, U., Strobel, M., Su, Z., Sullivan, R., Tagesson, T., Varlagin, A., Vreugdenhil, M., Walker, J., Wen, J., Wenger, F., Wigneron, J. P., Woods, M., Yang, K., Zeng, Y., Zhang, X., Zreda, M., Dietrich, S., Gruber, A., van Oevelen, P., Wagner, W., Scipal, K., Drusch, M., and Sabia, R.: The International Soil Moisture Network: serving Earth system science for over a decade, Hydrol. Earth Syst. Sci., 25, 5749–5804, https://doi.org/10.5194/hess-25-5749-2021, 2021.

Garcia, D., 2010. Robust smoothing of gridded data in one and higher dimensions with missing values. Computational Statistics & Data Analysis, 54(4), pp.1167-1178. Available at: https://doi.org/10.1016/j.csda.2009.09.020

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Preimesberger, W., Stradiotti, P., and Dorigo, W.: ESA CCI Soil Moisture GAPFILLED: an independent global gap-free satellite climate data record with uncertainty estimates, Earth Syst. Sci. Data, 17, 4305–4329, https://doi.org/10.5194/essd-17-4305-2025, 2025.