EAEE4000_001_2021_3 Final Project
Machine learning for environmental engineering and science

Reproducing evaporative fraction from a flux tower site at the Great Plains using LSTM

Name: Rong-Yu Gu
UNI: rg3390

Abstract
Surface fluxes is an essential media to link land-atmosphere interactions, and they are affected by both available energy and available water. In order to better predict the surface fluxes, we have to better representation on soil state and land cover type. Evaporative fraction (EF) can be a good indicator as surface wetness, but there are sparse observations due to inhomogeneous distributed flux tower sites. In this project, we predict EF in the Great Plains flux tower site by using LSTM machine learning method and take weather data as input. We excluded EF in which precipitation happens in the past 12 hours. Results show that the LSTM model underestimate EF and the importance to include self-persistency signal of EF in the model. Modification in the input data from the original diurnal cycle to the diurnal anomaly or to the daily timescale might be applied to make the prediction better. Future work can further apply the predicted EF to predict surface fluxes.

Introduction and motivation
Soil moisture-precipitation (SM-P) feedback indicates the memory of water signal over land have a lagged impact on precipitation in particular timescale or regions, and the initiation of such signal is driven by precipitation(Hsu et al., 2017; Koster et al., 2003, 2004; Santanello et al., 2018). Such memory of water signal, differing from regions to regions or even among different timescale, depending on several environmental characteristics(Koster et al., 2004; Tuttle and Salvucci, 2016). For example, in some arid to semi-arid regions, due to soil moisture heterogeneity, the next day precipitation tends to occur over the direr patches (compared to the surrounding soil patches). This negative spatial SM-P coupling can persist after the next day precipitation (Hsu et al., 2017). These dry soil patches become wetter than the surrounding patches, thus creating soil moisture heterogeneity again and sustain the negative coupling relationship. Besides, not only the soil moisture heterogeneity, but also the land cover type can alter the pathway of the feedback. For instance, water can be uptake by plants instead of merely evaporate from soil. Although typical SM-P coupling hotspots on semi-arid regions, by considering the linkage between transpiration and precipitation, the biosphere-atmosphere hotspot can lie in vegetated region, such as savannas (Green et al., 2017; Koster et al., 2004). Recently, scientists seek for models based on the knowledge of land-atmosphere coupling to have better prediction of climate. However, with complicated and turbulence-involved processes, some modelling frameworks still rely on parameterization rather than a direct process-based structure (Koster et al., 2004; Santanello Jr et al., 2018). Thus, taking the advantage of data-driven modelling framework, machine learning could be a possible way to disentangling the correspondence among environmental variables in land-atmosphere coupling.

Surface heat fluxes are crucial components in land-atmosphere coupling to exchange water and heat between land and atmosphere (Santanello Jr et al., 2018). As a partition of surface heat flux, latent heat flux is a form of energy flux of transpiration and it is determined by turbulence, radiation (available energy) and soil-atmospheric dryness gradient (available water) (Gentine et al., 2007, 2011). Typical SM-P coupling hotspots display high sensitivity of latent heat flux to both available energy and available water (Koster et al., 2004). Recent observation strategy on latent heat flux can be either through flux tower observation or satellite retrieval (Liu et al., 2021; Pastorello et al., 2020). The former could be limited to represent the footprint regions, while the latter could be biased in cloudy scenarios. To estimate flux data in regional or global scale, predictions with weather and land characteristic data as input is applicable. Despite a diverse availability to weather data, recent land characteristic data in regional or global scale are yet to thoroughly represent the reality because of high heterogeneity in land cover and soil state.

Evaporative fraction (EF) is a good indicator to soil state because it possesses a feature of memory, which determine the feedback from land to the atmosphere (Gentine et al., 2007, 2011; Zhu et al., 2020). Evaporative fraction is a diagnostic variable for surface energy balance. It is the ratio of latent heat to the available energy (net radiation). While the available energy is a periodic diurnal signal controlled by solar radiation, EF reflects the low frequency control of soil and vegetation on the partitioning of latent heat to the available energy. Therefore, EF has a characteristic of self-persistency and could be highly correlated to soil dryness(Gentine et al., 2007, 2011). Although EF will not dramatically change within a short time period, it is not constant and will be affected by environmental conditions(Gentine et al., 2007), e.g. clouds, precipitation, etc. Since EF has high self-persistency and the processes involved could be with complexity and nonlinearity, a machine learning method, Long-Short Term Memory (LSTM), is applicable to predict EF.

In this project, we intend to predict EF with net radiation (available energy) and precipitation (available water) as input by using LSTM method. LSTM can pass the predicted signal from previous timestep, which supposed to fit the self-persistency characteristic of EF. Since both net radiation and precipitation are higher frequency signal than EF, and EF could not well represent soil state in rainy days, we designed three experiment to evaluate the performance of predicted EF:

(1) Exp1: Take net radiation, precipitation and EF in the past 7 days as input, ignore the loss in EF when precipitation happens in the past 12 hours, and predict EF in recent timestep
(2) Exp2: Take net radiation, precipitation and EF in the past 7 days as input, and predict EF in recent timestep
(3) Exp3: Take net radiation and precipitation in the past 7 days as input, ignore the loss in EF when precipitation happens in the past 12 hours, and predict EF in recent timestep

By comparing Exp1 and Exp2, we can see the impact of precipitation on the predicted EF. By comparing Exp1 and Exp3, we can see the impact of self-persistency in EF on the predicted EF.
In the first attempt of running machine learning, we selected a FLUXNET2015 site at the Great Plains where is reported as the hotspot of SM-P coupling.

Data and Method

Site Description and data availability

The US-ARM flux tower site (Biraud et al., 2017) was selected as the study site in this project. Located in the Great Plains, it is covered by grassland. Weather and fluxes data, including radiation, precipitation, latent heat flux (LE), sensible heat flux (H), and volumetric soil water content (SWC), are available from 2003 to 2012. The former 4 variables were used to run the model and the latter was used to do the analysis. These half-hourly data were downloaded from FLUXNET2015, a synthesis dataset of global flux tower sites. EF was calculated through the following equation:

$$EF = \frac{LE}{LE + H}$$

Assuming ground heat flux is relatively small, LE+H can be similar to the available energy for surface layer. Because EF sometimes have spike during dusk and dawn, we simply let EF before 8:00 and after 16:00 as 0. Besides, net radiation was calculated as:

$$Rn = SWin - SWout + LWin - LWout$$

Where Rn is net radiation, SWin is incoming solar radiation, SWout is outgoing solar radiation, LWin is incoming long wave radiation, LWout is outgoing long wave radiation.

LSTM machine learning method

We applied LSTM to predict EF because LSTM can take the self-persistency signal of EF into account. Here we set up 512 neurons, 1e-4 learning rate with 51 epochs in the LSTM model. Dropout and early stopping strategy were applied to avoid overfitting. Both input and output data were in half-hourly time resolution. In Exp1 and Exp3, to exclude the learning of the model from rainy scenario when precipitation happens in the past 12 hours, we do not allow the model to take those prediction into account in the loss function. The loss function is determined by mean absolute error. Observational data in 2003 were used to be training data, and 30% of the training data were used to validate. Observational data from 2004 to 2012 were used to be testing data.

Fig. 1 indicates the training and validation loss in LSTM model. All the loss were in a decaying trend, which show an early stopping is successfully applied to avoid overfitting.

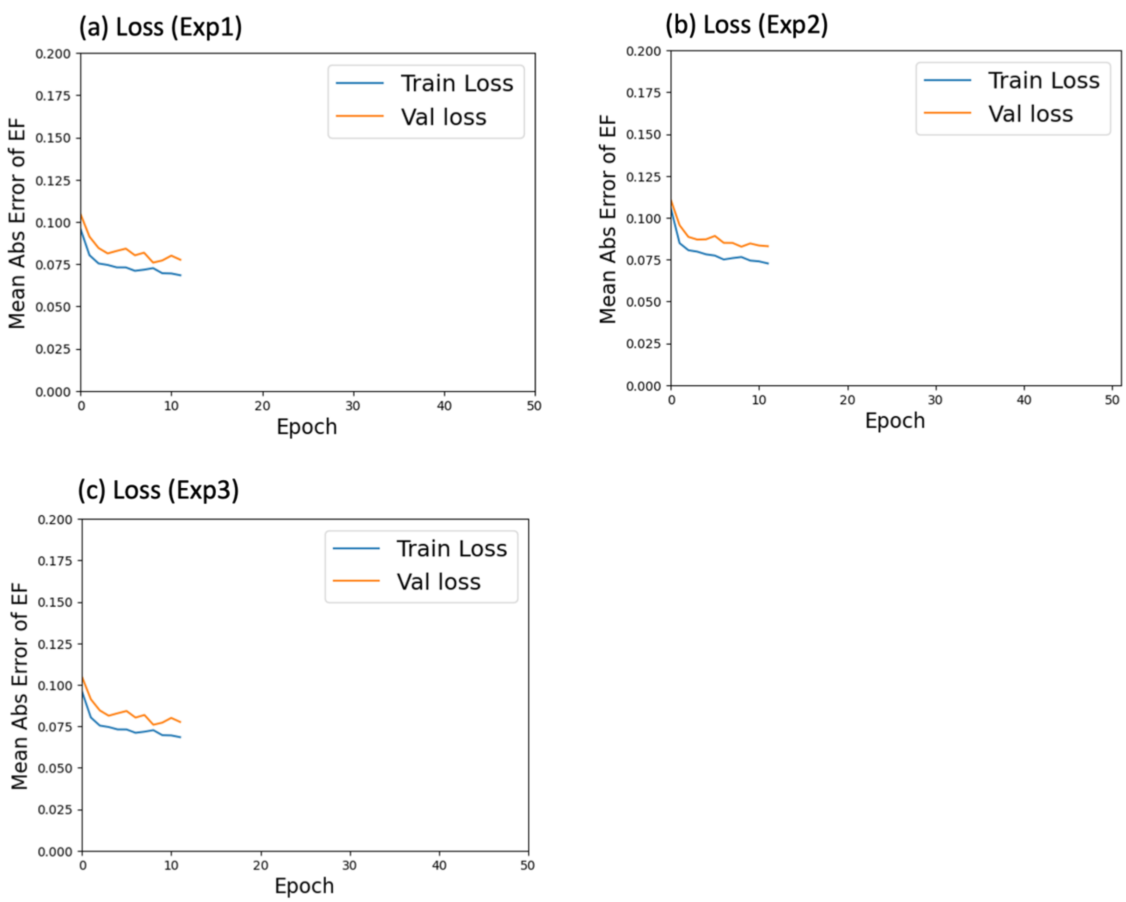

Fig. 1. Training and validation loss of EF among the three experiments.

Results

LSTM Predictions

Fig. 2 display the evaluation of predicted EF in LSTM model. Overall, the predicted EF did not well follow 1:1 line to the true values of EF. A large number of predictions lies on boundaries. For example, the true values of EF were diverse, while the predicted values of EF were 0, and vice versa. Although the correlation coefficients were about 0.8 in both Exp1 and Exp2, the model were unable to capture the EF upper bound as 1. The upper bound of predicted EF were merely about 0.6. Exp3 performed the worst among 3 models with the lowest correlation coefficient and the largest root mean square error (RMSE).

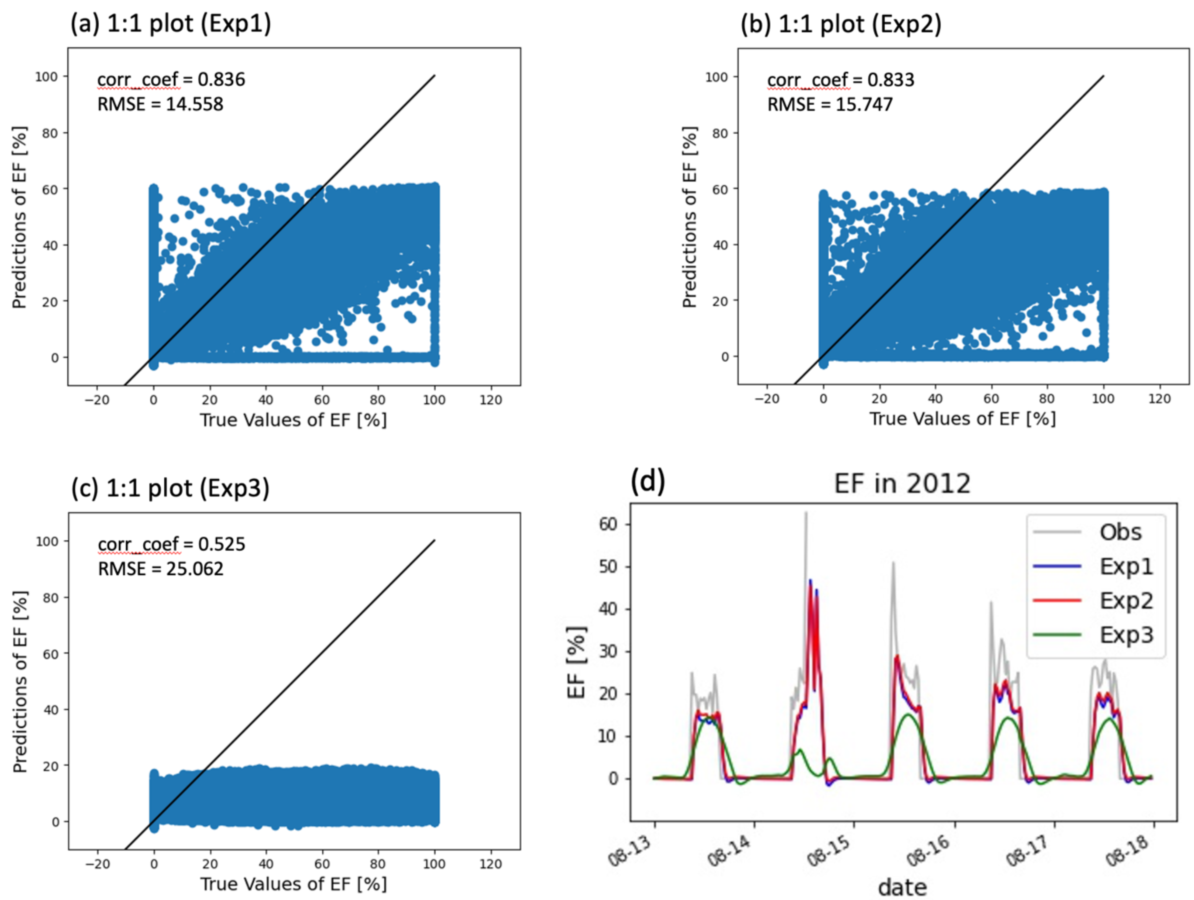

Fig. 2. The comparison between the true values of EF (Obs) and the predictions of EF among the three experiments.

The impact of excluding EF in which precipitation happens in the past 12 hours in the loss function

Comparing Exp1 with Exp2, we can see the impact of rainy hours on the predicted EF. Without considering precipitation in the past 12 hours, Exp1 can have better prediction in EF than EXP2 (Fig. 3a). Noted that here the SWC anomaly is the anomaly of SWC to its 10-year mean. Correlation coefficient increased and RMSE decreased in Exp1, compared to Exp2 (Fig. 2a; Fig. 2b).

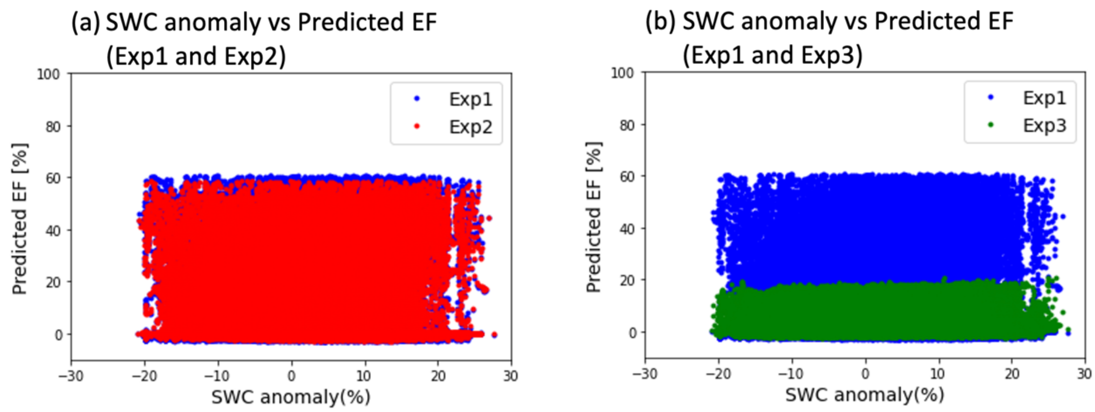

Fig. 3. The relation of SWC anomaly to predicted EF. (a) The comparison of the relation between Exp1 and Exp2. (b) The comparison of the relation between Exp1 and Exp3. Noted that here the SWC anomaly is the anomaly of SWC to its 10-year mean.

The impact of excluding EF self-persistency in LSTM model predictions

Comparing Exp1 with Exp3, we can see the impact of self-persistency on the predicted EF. Without considering EF in the past 7 days as input, Exp3 can have worse prediction in EF than EXP1 (Fig. 3b). The upper bound of predicted EF in Exp3 was far less than that in Exp1. Not only did Exp3 show a much lower correlation coefficient, but a much larger RMSE than Exp1. (Fig. 2a; Fig. 2c).

Discussion and conclusion

LSTM is applicable to predict EF because this strategy is able to pass the self-persistent signal. Exp1 performed better among our three experiments, which means excluding the effect of precipitation on EF and allowing self-persistency of EF can better capture the characteristics of true EF. However, despite about 0.8 of correlation coefficient, the predicted EF does not well fit the 1:1 line to the true value of EF. To improve the predictions in the LSTM model, the following ways can be further applied:

1. Since there is still a dominant frequency of diurnal cycle in our input data, if we would like to set EF in nighttime as 0, we should not allow LSTM to learn those 0 values. Also, removing the diurnal signal of the weather data in the corresponding month may prevent the dominant diurnal signal. That is, we may change the input data from the original observation to the anomaly signal in the corresponding month.
2. If we would like to use EF as an indicator to soil state, perhaps changing the input and output time resolution as daily timescale is enough for EF to represent soil wetness. No relationship had found in fig.3 may also due to the seasonal signals in both SWC and EF. We did not remove seasonal cycle of SWC when calculating the anomaly. Instead, we only remove the annual mean.

To sum up, we may improve the underestimation of predicted EF by removing diurnal signals or to use a courser timescale, such as daily mean, as input. If the improvement works well and the predicted EF can better represent soil wetness, future work can apply the same strategy to different land cover type. Taking the better predicted EF and suitable weather data, we can further predict surface fluxes. In order to make the prediction more applicable to the place at which flux and EF data are not easily observed, low frequency net radiation (removing the diurnal signal in the corresponding month) plus precipitation data might hopefully fill the gap. This can provide a data-driven regional or global product to fit land heterogeneity, rather than using homogeneous parameterization over a coarse grid scale.

Code availability

Codes are available in Github: https://github.com/rongyugu/EAEE4000_001_2021_3_ML.git

References

Gentine, P., Entekhabi, D., Chehbouni, A., Boulet, G., & Duchemin, B. (2007). Analysis of evaporative fraction diurnal behaviour. *Agricultural and Forest Meteorology*, *143*(1–2), 13–29. https://doi.org/10.1016/j.agrformet.2006.11.002

Gentine, P., Entekhabi, D., & Polcher, J. (2011). The diurnal behavior of evaporative fraction in the soil-vegetation-atmospheric boundary layer continuum. *Journal of Hydrometeorology*, *12*(6), 1530–1546. https://doi.org/10.1175/2011JHM1261.1

Green, J. K., Konings, A. G., Alemohammad, S. H., Berry, J., Entekhabi, D., Kolassa, J., et al. (2017). Regionally strong feedbacks between the atmosphere and terrestrial biosphere. *Nature Geoscience*, *10*(6), 410–414. https://doi.org/10.1038/ngeo2957

Hsu, H., Lo, M. H., Guillod, B. P., Miralles, D. G., & Kumar, S. (2017). Relation between precipitation location and antecedent/ subsequent soil moisture spatial patterns. *Journal of Geophysical Research*, *122*(12), 6319–6328. https://doi.org/10.1002/2016JD026042

Koster, R. D., Suarez, M. J., Higgins, R. W., & van den Dool, H. M. (2003). Observational evidence that soil moisture variations affect precipitation. *Geophysical Research Letters*, *30*(5). https://doi.org/10.1029/2002gl016571

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., et al. (2004). Regions of strong coupling between soil moisture and precipitation. *Science*, *305*(5687), 1138–1140.

Liu, N., Oishi, A. C., Miniat, C. F., & Bolstad, P. (2021). An evaluation of ECOSTRESS products of a temperate montane humid forest in a complex terrain environment. *Remote Sensing of Environment*, *265*. https://doi.org/10.1016/j.rse.2021.112662

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y. W., et al. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, *7*(1), 225. https://doi.org/10.1038/s41597-020-0534-3

Santanello, J. A., Dirmeyer, P. A., Ferguson, C. R., Findell, K. L., Tawfik, A. B., Berg, A., et al. (2018, June 1). Land-atmosphere interactions the LoCo perspective. *Bulletin of the American Meteorological Society*. American Meteorological Society. https://doi.org/10.1175/BAMS-D-17-0001.1

Santanello Jr, J. A., Dirmeyer, P. A., Ferguson, C. R., Findell, K. L., Tawfik, A. B., Berg, A., et al. (2018). Land–atmosphere interactions: The LoCo perspective. *Bulletin of the American Meteorological Society*, *99*(6), 1253–1272.

Sebastien Biraud, Marc Fischer, Stephen Chan, Margaret Torn (2021), AmeriFlux BASE US-ARM ARM Southern Great Plains site- Lamont, *Ver. 11-5, AmeriFlux AMP*, (Dataset). https://doi.org/10.17190/AMF/1246027

Tuttle, S., & Salvucci, G. (2016). Empirical evidence of contrasting soil moisture–precipitation feedbacks across the United States. *Science*, *352*(6287), 825–828.

Zhu, W., Jia, S., Lall, U., Cheng, Y., & Gentine, P. (2020). An observation-driven optimization method for continuous estimation of evaporative fraction over large heterogeneous areas. *Remote Sensing of Environment*, *247*. https://doi.org/10.1016/j.rse.2020.111887