AN INTERPRETABLE MACHINE LEARNING MODEL FOR ADVANCING TERRESTRIAL ECOSYSTEM PREDICTIONS

Dan Lu, Daniel Ricciuto, Siyan Liu
Oak Ridge National Laboratory
1 Bethel Valley Rd, Oak Ridge, TN, USA
{lud1,ricciutodm,lius1}@ornl.gov

ABSTRACT

We apply an interpretable Long Short-Term Memory (iLSTM) network for land-atmosphere carbon flux predictions based on time series observations of seven environmental variables. iLSTM enables interpretability of variable importance and variable-wise temporal importance to the prediction of targets by exploring internal network structures. The application results indicate that iLSTM not only improves prediction performance by capturing different dynamics of individual variables, but also reasonably interprets the different contribution of each variable to the target and its different temporal relevance to the target. This variable and temporal importance interpretation of iLSTM advances terrestrial ecosystem model development as well as our predictive understanding of the system.

1 INTRODUCTION AND MOTIVATION

Machine learning (ML) models, Long Short-Term Memory (LSTM) networks [1] in particular, have been demonstrated to improve predictions of carbon fluxes between atmosphere and land [2; 3; 4]. LSTM networks, trained over multivariable time series consisting of exogenous and target variables, capture nonlinear correlation of historical values of environmental drivers and carbon fluxes to predict future carbon fluxes. LSTM models learn system patterns and dynamical behaviors from time series observations; they are not necessarily constrained by the principle of mass, energy conservation or governing equations that describe the carbon cycle related processes a priori. Thus, despite successful applications in terrestrial ecosystem modeling, LSTM models have been criticized for their lack of interpretability [5; 6]. An interpretable ML model should not only provide accurate predictions but also capture the dynamical interdependency between the variables. It should be able to explain, for the carbon flux variable, what are the most important environmental drivers, and at which time scales the environmental drivers have strong impact on the carbon flux estimation? This interpretability on variable importance and variable-wise temporal importance is crucial for improving carbon flux predictions and predictive understanding of the ecosystem.

To achieve interpretability, some studies performed sensitivity or permutation analysis on trained LSTM networks to explain the relative importance of inputs to the outputs [7; 8; 9; 10]. However, this post-hoc interpretability method cannot explain what the model learned in the training process. Realizing the challenges in interpretation of the black-box networks, some studies revert to understandable but non-dynamic ML models such as random forest [11; 12]. However, random forest models have shown inferior performance to LSTM networks, and they cannot learn the temporal dependency between environmental variables and carbon fluxes [13; 3]. So essentially this strategy of applying non-dynamic ML models sacrifices prediction accuracy for predictive interpretability.

In this work, we focus on an interpretable LSTM (iLSTM) method for time series prediction. It explores internal structures of the LSTM network and overcomes the opacity of its hidden states for inherent interpretability. iLSTM jointly learns network parameters, variable and temporal importance with respect to the target prediction. It not only produces an accurate prediction but more importantly provides interpretable insights into the data, which advances our understanding about the influence of environmental variables and their temporal influence on the target predictions. We apply iLSTM to the Morgan Monroe State Forest (MMSF) in central Indiana to learn the importance of seven environmental variables to net ecosystem CO_2 exchange (NEE, a carbon flux variable). We evaluate iLSTM

prediction performance by comparing it with the standard LSTM, and we justify the interpretability of iLSTM through a series of designed numerical experiments.

The main contributions of this work are:

1. We apply an interpretable LSTM for NEE prediction to understand the importance of environmental drivers and their temporal importance to NEE by exploring the network internal structures. This presents a major advancement from previous studies that either lacked importance interpretation or used ad-hoc sensitivity or permutation analysis for a limited explanation.
2. We identify the environmental drivers exhibiting high importance to NEE and their temporal relevance to NEE. This investigation is important for understanding different driver's effects on carbon fluxes and the different memory effects of individual variables.
3. We justify the interpretability of iLSTM and the insights gained from the variable importance help variable selection in modeling and advance terrestrial ecosystem model development.

2 Interpretable Long Short-Term Memory networks

LSTM network is specifically designed to learn the dynamic temporal dependence structure within the data. Its hidden states dynamically add and store memory from the multiple input sequences to infer the output. Standard LSTM models blindly blend the information of all input variables into the hidden states used for prediction. It is therefore difficult to distinguish the relative contribution of individual inputs to the output, and the mixed multi-variable data in the hidden states neglect the different dynamics of the individual input sequence [14]. In this work, we introduce an interpretable LSTM (iLSTM) [15] for accurate prediction and importance interpretation.

iLSTM enables interpretability by exploring the internal structure of LSTM networks. First, it enables hidden states to encode individual variables, such that the contribution from individual inputs to the prediction can be distinguished. For example, the standard LSTM uses a hidden state vector to summarize the input information (where the vector length is the neuron size of the hidden layer). In contrast, iLSTM uses a hidden state *matrix* which includes D rows of the standard hidden vector; each row encapsulates information exclusively from one of the D input variables. Second, iLSTM uses a mixture attention mechanism to summarize the variable-wise hidden states and jointly learns the network parameters for prediction and the importance weights for interpretation. Specifically, temporal attention is first applied to the sequence of hidden states corresponding to each input variable, in order to obtain the summarized historical information of each input time series. Then variable attention is derived to merge the variable-wise states. Next, we assemble the temporal attention weights and variable attention weights into a probabilistic mixture model for learning [16] to calculate the temporal and variable importance weights. The structure of iLSTM is analogue to a set of parallel standard LSTMs, each of which processes one variable series and then merges via the mixture attention mechanism. By learning both the temporal and variable attention weights in the training, iLSTM largely leverages each input sequence's information for prediction and captures their individual dynamics. Therefore, iLSTM can not only achieve accurate prediction by optimizing the learning, but also enables interpretability by calculating the variable and temporal importance. The variable importance weights β (sum to one) reflects the relative importance of the corresponding inputs with respect to the target output. For each input variable, the temporal importance weights α (sum to one) explains the relative importance of the time instant of that variable to the prediction.

3 Terrestrial ecosystem prediction

The dataset is from the Morgan Monroe State Forest (MMSF) in south-central Indiana, which represents one of the most comprehensive records of leaf- and canopy-scale processes in a mature forest collected in situ [17]. The MMSF is a managed deciduous broadleaf forest where the lands are dominated by woody vegetation with a percent cover >60% and height exceeding 2 meters. It consists of broadleaf tree communities with an annual cycle of leaf-on and leaf-off periods. The average age of the trees are 80-90 years. Since 1998, a 46-m AmeriFlux tower has been operating continuously at MMSF [18] for collecting eddy covariance fluxes and meteorological data. The time series dataset includes NEE ($gCm^{-2}d^{-1}$) and seven environmental drivers, i.e., nighttime temperature (Tn ($deg\ C$)), daytime temperature (Td ($deg\ C$)), shortwave radiation (Ra (Wm^{-2})),

vapor pressure deficit (VPD (hPa)), precipitation (P (mm)), soil water content (SWC (%)), and atmospheric CO_2 concentration (CO_2 (ppm)) at a daily timescale from 1999 to 2014.

Our purpose here is to use iLSTM to learn and interpret the relationship between the seven environmental drivers and NEE for the NEE prediction. The multiple environmental variables carry different patterns (see Figure in Appendix A). Properly modeling individual variables and their interactions is crucial for accurate prediction and understanding of NEE dynamics. The two temperature variables, Tn and Td, have more similar patterns with NEE, and theoretically they would have more impact on NEE compared to other environmental features. The climate in MMSF is humid subtropical, mild with no dry season and has hot summers. Year 2012 is anomalous; it is a dry year with the annual total precipitation of $782mm$ which is $631mm$ less than the wet year of 2008. A warm and dry spring caused an abnormally early start to the growing season in 2012. The date at which weekly averaged NEE first crossed zero (meaning a switch from a net source to a net sink of CO_2 by the ecosystem) in 2012 occurred about 3 weeks earlier than average. The pattern of enhanced CO_2 uptake in the spring reversed in the summer to reduced CO_2 uptake compared to other years. During the peak of the growing season, the absolute value of measured NEE (sum of hourly fluxes for July up to and including mid-August) in 2012 was reduced by $102gC$, or 55%, relative to baseline (1999–2010) mean NEE. In the iLSTM simulation, we use the first 11 years (i.e., 1999-2009) of data for training and the remaining 5 years of data (i.e., 2010-2014) for out-of-sample testing. The details of training are summarized in Appendix B. Given the lack of extremely dry years in the training data, the different NEE dynamics in 2012 from the training data may cause challenges in prediction.

4 Results and Discussion

We first evaluate prediction performance and then assess the interpretability of iLSTM by analyzing the temporal and variable importance. Next, we quantitatively evaluate the efficacy of variable importance through lens of numerical experiments design. The prediction performance of iLSTM is evaluated in comparison with the standard LSTM. We consider two evaluation metrics, coefficient of determination (R^2) and root mean squared error (RMSE) which are defined in Appendix B.

Figure 1 shows the prediction of the standard LSTM and iLSTM in the testing period. iLSTM produces a high R^2 value of 0.82 which suggests that iLSTM learns the seasonal pattern of NEE very well, including a model of leaf phenology that is a strong control on carbon cycling in deciduous forest systems. Additionally, iLSTM performs better than the standard LSTM both over the five testing years and for the individual years. A closer examination of the dry year 2012 shows that iLSTM can accurately simulate the abnormally early start of the growing season while LSTM underestimates the NEE values in April and does not accurately capture the timing and magnitude of reduced uptake in late summer that is caused by the water shortage. The superior performance of iLSTM in simulating NEE, especially in an abnormal event, is attributed to its capability of learning the relative importance of each inputs and their dynamics of the importance over time with respect to the prediction.

iLSTM not only produces good prediction performance but also empowers the interpretability on variable importance and temporal importance. Figure 2(a) depicts that variables Tn, Td, and Ra are recognized as the most important drivers for NEE estimation, followed by a less relevant P and SWC, and a much less significant VPD and CO_2. Figure 2(b) illustrates the importance of considering memory in the system and additionally indicates that these important variables have different patterns in the temporal importance of the last 60 days which are used to predict the NEE at the current day. For example, the two temperature variables, Td and Tn, have relatively long-term correlation to NEE and the short-term data of Ra contributes relatively more to the NEE prediction. These variable and temporal importance analysis is in line with our domain knowledge. The study site MMSF is a deciduous forest where the carbon flux is normally more sensitive to temperature than to water, and the temperature can have a long-term memory effect on NEE [3]. The radiation variable Ra affects latent and sensible heat fluxes and also has an impact on carbon fluxes by controlling photosynthetic uptake at shorter time scales. This interpretability of iLSTM in variable and temporal importance can guide terrestrial ecosystem model development. Process-based terrestrial ecosystem models implement environmental variable related processes and specifies the physics that consider the memory effect of the environmental drivers to the carbon fluxes. From the prediction analysis of iLSTM, we can gain insights about which environmental features are more important and which variable has a stronger memory effect on NEE. The iLSTM network may eventually also be applied to model output to better understand whether model equations and assumptions lead to consideration

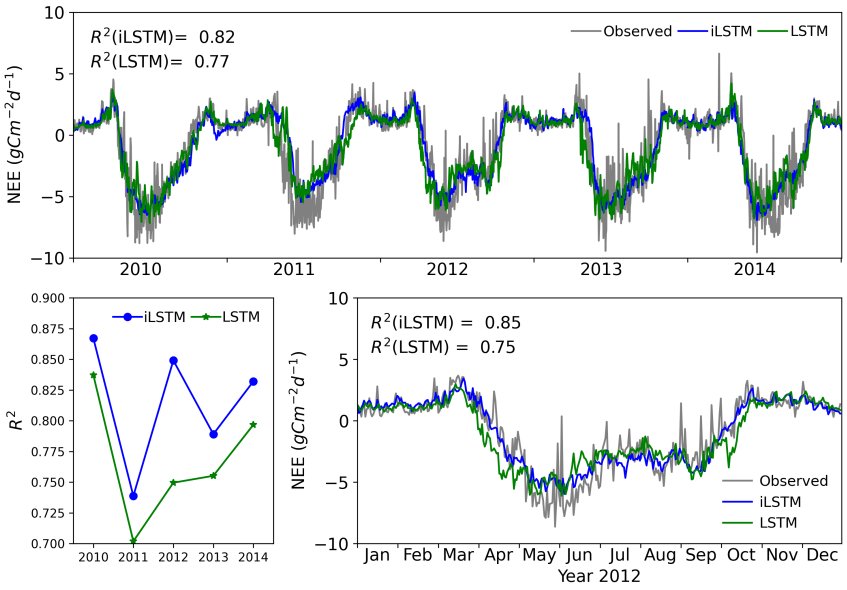

Figure 1: Comparing prediction performance of NEE between iLSTM and the standard LSTM in testing period.

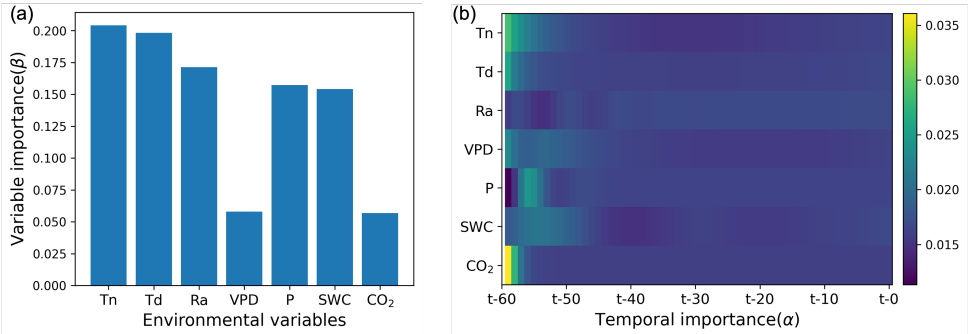

Figure 2: (a) Variable importance β measures the relative importance of the seven environmental variables to NEE and (b) temporal importance α measures relative contribution of the time steps of a certain variable to the NEE prediction. The larger the value, the higher the importance.

of memory effects that are consistent with the observations. Such information is crucial to guide process development and physics implementation in terrestrial ecosystem models and to advance our predictive understanding of the system.

To quantitatively evaluate iLSTM's efficacy of variable importance, we design four new models for NEE prediction based on iLSTM suggested variable selections. Table 1 describes these four new models along with the original model that considers all the input variables. We make the following comparisons between the five model setups.

- Comparing Model I and II evaluates iLSTM's efficacy of variable importance. Similar results would suggest that iLSTM can accurately identify the important variables for NEE prediction.
- Comparing Model II and III investigates the influence of two water-related variables, P and SWC, on NEE prediction. If the two models produce different results in dry year of 2012, it will suggest that P and SWC are important to estimate NEE dynamics in drought seasons.
- Instantaneous values of SWC already include memory effects of precipitation and water use by the ecosystem; therefore comparing Model IV to V tests whether iLSTM can effectively learn a soil hydrology model without SWC by considering memory effects of precipitation and other drivers that are relevant for carbon fluxes. If the two models produce similar results, it will suggest that we may be able to use observations of P instead of SWC to predict NEE. Observations and

reanalyses of P are readily available while reliable SWC data are sparse and usually involve large measurement errors due to spatial heterogeneity.

Table 1: Comparison of iLSTM prediction performance on NEE between five models which use different input variables. R^2 and RMSE are calculated for the entire testing period (2010-2014) and the dry year 2012.

Model	Input variables	R^2 (2010-2014 / 2012)	RMSE (2010-2014 / 2012)	Insights
I	Tn, Td, Ra, P, SWC, VPD, CO_2	0.82 / 0.85	1.35 / 1.14	Comparing Model I & II suggests that iLSTM can accurately identify the important inputs.
II	Tn, Td, Ra, P, SWC	0.81 / 0.84	1.36 / 1.19	Comparing Model II & III suggests that water-related variables P and SWC are critical for NEE estimation in drought (year 2012).
III	Tn, Td, Ra	0.81 / 0.77	1.36 / 1.41	
IV	Tn, Td, Ra, P	0.82 / 0.83	1.34 / 1.20	Comparing Model IV & V suggests that iLSTM can learn soil hydrology model; using either P or SWC produces similar NEE prediction.
V	Tn, Td, Ra, SWC	0.81 / 0.83	1.36 / 1.21	

Comparison results are summarized in Table 1 and detailed performance is shown in Figure 4 of Appendix C. First, Model I and II have minor difference in the prediction performance. This finding justifies iLSTM's effectiveness in variable selection. It suggests that among the seven inputs, Tn, Td, Ra, P, and SWC are sufficient to make a good prediction of NEE, and inputs of VPD and CO_2 bring too little new information to significantly boost the prediction in the MMSF ecosystem. This effective variable selection of iLSTM not only reduces the requirement of training data but also enhances cause-effect process understanding. Second, Model II and III produces similar R^2 and RMSE over the entire testing period. But for the dry year of 2012, Model III, which does not consider water-related inputs such as P and SWC, shows worse prediction than Model II with a lower R^2 and a higher RMSE. This suggests that the water-related variables are critical for NEE estimation in drought. iLSTM realizes the importance of P and SWC by giving them a relatively high weight (Figure 2(a)), but meanwhile their weights are relatively small compared to those of the temperature inputs Tn and Td which are generally more important to NEE in all the seasons and the importance of P and SWC is particularly reflected in drought event. Third, Model IV and V presents similar prediction performance in both entire testing period and in year 2012, which indicates that including either P or SWC results in comparable NEE prediction. This finding is significant. It is known that SWC relates to vegetation functioning and therefore NEE more directly than P. Plant growth and development directly depend on SWC and plant water uptake. Sometimes, rainfall records can be misleading as intense precipitation events do not always result in a proportional increase of SWC either because of low infiltration rates or a relatively small soil water holding capacity. Additionally, the occurrence of precipitation is not always associated with the onset of the vegetation growth, exhibiting lags of several weeks and even months. The similar prediction of NEE from SWC or P suggests that our trained iLSTM includes an accurate model of how soil hydrology impacts NEE learned implicitly from the precipitation data and its different temporal relevance to the carbon flux. We will investigate this point further in future studies.

5 Conclusion and Future Work

Incorporating ML into earth sciences is not only for improvement of prediction but more importantly for enhancing predictive understanding. In this work, we apply an interpretable LSTM network for NEE prediction. The iLSTM model not only results in an accurate prediction of NEE by capturing different dynamics of individual environmental drivers, but also interprets the relative importance of these drivers to NEE as well as their timescales of influence. This insight into ecological dynamics guides the process-based terrestrial ecosystem model development. Leveraging the iLSTM interpretability alongside model hypothesis testing, we can determine the process mechanisms about the effective drivers and their legacies of past events on carbon fluxes.

This study is the first and important step towards understanding the sensitivity of global terrestrial ecosystems to environmental variability. In the future, we will apply iLSTM to a variety of forest sites with different plant functional types to explore the distinct variable importance and memory effects from both climate and vegetation in quantifying spatio-temporal variations in forest NEE.

ACKNOWLEDGMENTS

This research is supported as part of the Terrestrial Ecosystem Science-Science Focus Area (TES-SFA) project and part of the Artificial Intelligence Initiative of the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US DOE under contract DE-AC05-00OR22725. It is also sponsored by the Data-Driven Decision Control for Complex Systems (DnC2S) project funded by the US DOE, Office of Advanced Scientific Computing Research.

REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.

[2] M. Reichstein, S. Besnard, N. Carvalhais, F. Gans, M. Jung, B. Kraft, and M. Mahecha, "Modelling landsurface time-series with recurrent neural nets," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7640–7643, 2018.

[3] S. Besnard and et al., "Memory effects of climate and vegetation affecting net ecosystem co2 fluxes in global forests," *PloS one*, vol. 14, 2019.

[4] B. Kraft, M. Jung, M. Körner, C. Requena Mesa, J. Cortés, and M. Reichstein, "Identifying dynamic memory effects on vegetation state using recurrent neural networks," *Frontiers in Big Data*, vol. 2, p. 31, 2019.

[5] A. Pérez-Suay, J. E. Adsuara, M. Piles, L. Martínez-Ferrer, E. Díaz, A. Moreno-Martínez, and G. Camps-Valls, "Interpretability of recurrent neural networks in remote sensing," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3991–3994, 2020.

[6] G. Camps-Valls, M. Reichstein, X. Zhu, and D. Tuia, "Advancing deep learning for earth sciences: From hybrid modeling to interpretability," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3979–3982, 2020.

[7] W. L. Zhao, P. Gentine, M. Reichstein, Y. Zhang, S. Zhou, Y. Wen, C. Lin, X. Li, and G. Y. Qiu, "Physics-constrained machine learning of evapotranspiration," *Geophysical Research Letters*, vol. 46, no. 24, pp. 14496–14507, 2019.

[8] M. Shahhosseini, G. Hu, I. Huber, and S. V. Archontoulis, "Coupling machine learning and crop modeling improves crop yield prediction in the us corn belt," *Scientific Reports*, vol. 11, Jan 2021.

[9] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.

[10] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, p. 3145–3153, JMLR.org, 2017.

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[12] Y. Everingham, J. Sexton, D. Skocaj, and I.-B. G., "Accurate prediction of sugarcane yield using a random forest algorithm," *Agron. Sustain. Dev.*, vol. 36, no. 27, 2016.

[13] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, "Rainfall–runoff modelling using long short-term memory (lstm) networks," *Hydrology and Earth System Sciences*, vol. 22, no. 11, pp. 6005–6022, 2018.

[14] L. Zhang, C. Aggarwal, and G.-J. Qi, "Stock price prediction via discovering multi-frequency trading patterns," KDD '17, (New York, NY, USA), p. 2141–2149, Association for Computing Machinery, 2017.

[15] T. Guo and T. Lin, "Exploring the interpretability of LSTM neural networks over multi-variable data," 2019.

[16] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International Conference on Learning Representations*, 2018.

[17] D. Roman, K. Novick, and E. e. a. Brzostek, "The role of isohydric and anisohydric species in determining ecosystem-scale response to severe drought," *Oecologia*, vol. 179, pp. 641–654, 2015.

[18] D. Baldocchi and et al, "Fluxnet: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities," *Bulletin of the American Meteorological Society*, vol. 82, no. 11, pp. 2415 – 2434, 2001.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[20] A. Seddon, M. Macias-Fauria, P. Long, and et al., "Sensitivity of global terrestrial ecosystems to climate variability," *Nature*, vol. 531, pp. 229–232, 2016.

APPENDIX

A OBSERVATION DATA

The following Figure 3 shows the daily data of the seven environmental variables and NEE in 1999 to 2014 which are used in the LSTM simulation.

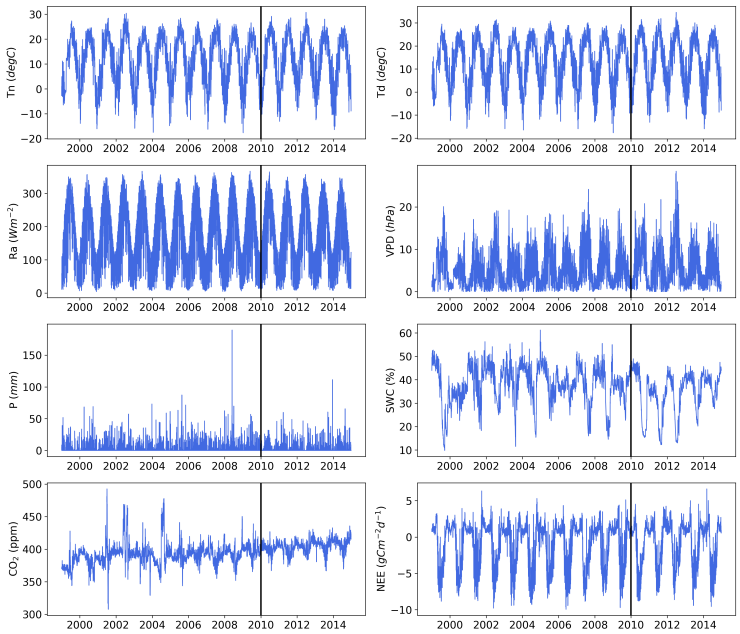

Figure 3: Daily data of the seven environmental variables and NEE in 1999 to 2014 where the vertical line separates the training period (1999-2009) and the testing period (2010-2014).

B MODEL DETAILS AND EVALUATION METRICS

In all the numerical experiments, we used a single layered LSTM network with 50 hidden neurons. The Adam optimizer [19] was used to minimize the mean squared error loss function with a learning rate of 0.001. The look-back window size was set to 60 days according to [20] and after hyperparameter tuning.

We consider two evaluation metrics, the coefficient of determination (R^2) and root mean squared error (RMSE).

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(Y_i^{obs} - Y_i^{sim}\right)^2}{\sum_{i=1}^{N} \left(Y_i^{obs} - \overline{Y}^{obs}\right)^2} \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(Y_i^{obs} - Y_i^{sim}\right)^2}{N}}. \tag{2}$$

where Y_i^{obs} is the ith observation, Y_i^{sim} is the ith simulated value, $\overline{Y}^{obs}$ is the mean of observation, and N is the total number of observations. R^2 is a normalized statistic that determines the relative magnitude of the residual variance compared to the observation variance. It indicates how well the plot of observed versus simulated data fits the 1:1 consistency line. R^2 value ranges from negative infinity to 1. A value of 1 corresponds to a perfect match of model simulations to observations, and a negative R^2 indicates that the model performs worse than the observed mean. RMSE measures the estimation errors in squared sense; its value varies from the optimal 0 to a large positive number. The lower RMSE and the better the model simulation performance.

C PREDICTION PERFORMANCE OF DIFFERENT MODELS

This section presents visualizations of the prediction performance of the models II, III, IV, and V in Table 1.

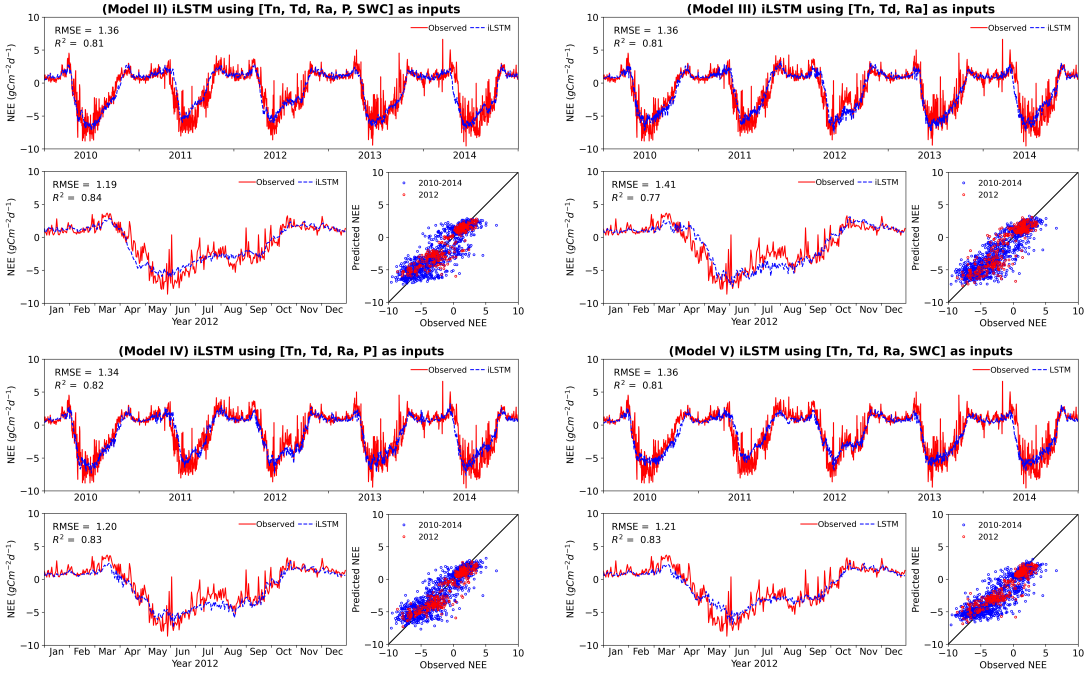

Figure 4: Prediction performance of NEE between different iLSTM models in Table 1.