Final Project Report
PH 582-001 Machine Learning

Submitted by- Pushpendra Raghav, Hitesh Prasad Thakur, Arun Niddish
Team: D1

Modeling Evapotranspiration Using Physics-Guided Machine Learning

1. Introduction

Accurate estimation of actual evapotranspiration (ETa) is critical in hydrological and climate modeling: for drought analysis, flood forecasting, irrigation scheduling, and for constraining ground water recharge (Katul *et al.*, 2012). ETa is controlled by various land and atmospheric states (of air temperature, land surface temperature, humidity, precipitation, soil characteristics, vegetation states, etc.) and thus, if we have measurements of these states then ETa can be modeled using machine learning algorithms. There is vast amount of the measurements of these states either from ground-based observatory systems or remote sensing. People have harnessed these databases of climate and hydrological variables extensively in varieties of machine learning algorithms to model ETa (Granata *et al.*, 2020; Yang *et al.*, 2006). Another way to model ETa is process-based modeling (i.e., explicit physical representation of water vapor exchange between land and atmosphere). Problem with process-based modeling is that it needs extensive model parameterization (e.g., resistance to water transport, vegetation phenology, etc.) that is a challenge especially modeling ETa at large scale. On the other hand, the issue with ML techniques-based models is that those models typically do not conserve the surface energy budget (i.e., not respecting the physics behind the process), which can be a major issue for model assessment of various components of surface energy budget (e.g., sensible and latent heat fluxes, net radiation, etc.). Recent studies have shown that process-based models and ML-based models should combine in order to have better generalization and interpretability and optimally extract information from the data (Reichstein *et al.*, 2019). A perspective paper by Reichstein et al. (2019) discusses that ***"Current ML approaches may not be optimal when system behavior is dominated by spatial or temporal context. Integration of domain knowledge and achievement of physical consistency by teaching ML models about the governing physical rules of the Earth system can provide very strong theoretical constraints on the top of the observational ones***." However, it is still unclear that whether such approaches have any advantages over process-based modeling and/or pure machine learning approaches.

Here, we developed a ML model (deep neural network) constrained by physics to model ETa. We combined energy-conserving Penman-Monteith like equation (Leuning *et al.*, 2008) and a ML model. We expect that this physics-guided ML model can benefit from both the strengths of physical modeling (theoretical foundations and interpretability) and ML (extracting information from data).

The ***overarching objective of this project is to develop a Physics-Guided Machine Learning (PGML) model for the estimation of actual evapotranspiration***. The specific objectives of this project include: (1) Developing a pure ML model and a physics-guided ML model for the prediction of ETa and comparing their performances; and (2) compare the capacity of pure ML model and physics-guided ML model to generalize.

2. Tasks and Methodology

2.1. Pure Machine Learning Model

A typical pure ML model predicts the target variable (LE in this case) based on N-sets of variables by minimizing/maximizing a loss function (Mean Squared Error of LE in this case). We used a feedforward Artificial Neural Network (ANN) to predict LE based on 13 sets of input observations including Fraction of Photosynthetically Active Radiation (fPAR), Soil Moisture (SM), Vapor Pressure Deficit (VPD), Plant Functional Type (PFT), Air Temperature (Ta), Carbon concentration (Ca), Wind Speed (WS), Air Pressure (Pa), Relative Humidity (RH), Net radiation (Rn), Soil Heat Flux (G), Photosynthetically Active Radiation (PAR) and, Canopy Height (h_canopy) (see Figure 1a) based on study by Alemohammad *et al.* (2017). The Mean Squared Error (MSE) of LE was used as a loss function in the training process of ANN. All the input variables were normalized to accelerate the learning pace except the PFTs for which we kept the original category values (1, 2, 3,……,9) as input. Rectified Linear Unit (ReLU) was used an activation function.

The very first step was to decide the optimal structure of the ANN model. For this, we shuffled the entire dataset in time and space and then splitted it into training (64%), validation (16%), and testing (20%) sets. We repeated the training of the ANN model with varying numbers of hidden layers and number of neurons for per layer. As a result, the ANN model with 5 hidden layers and 64 neurons each layer was used for further analysis (see Figure 2).

2.2. Physics-Guided Machine Learning (PGML)

PGML was developed by coupling ANN (same as described in section 2.1) with modified Penman-Monteith equation to predict LE. The ANN was set to predict logarithmic of surface resistance (r_s) instead of LE as it is more normally distributed (see Figure 1b). Mathematical details are as follows:

The regular Penman-Monteith equation has been widely used for the estimation of ETa or LE (Monteith, 1965; Penman, 1948):

$$LE = \frac{\Delta(R_n - G) + \frac{\rho C_p (e_s - e_a)}{r_a}}{\Delta + \gamma\left(1 + \frac{r_s}{r_a}\right)} \qquad (1)$$

where LE (=λETa) is the latent heat flux [W m^{-2}], λ is the latent heat of water vaporization, Δ is the slope of the saturated vapor pressure curve [kPa K^{-1}], R_n is the net radiation [W m^{-2}], G is the ground heat flux [W m^{-2}], ρ is the air density [kg m^{-3}], C_p is the specific heat of air [=1012 J kg^{-1} K^{-1}], e_s is the saturation vapor pressure [kPa], e_a is the actual vapor pressure [kPa], r_a is the aerodynamic resistance to heat transfer [s m^{-1}], γ is the psychrometric constant [kPa K^{-1}], and r_s is the bulk surface resistance to water transport [s m^{-1}].

Equation 1 is based on the surface energy balance equation:

$$R_n - G = LE + H \qquad (2a)$$

$$with\ H = \frac{\rho C_p (T_s - T_a)}{r_a} \qquad (2b)$$

$$and\ LE = \frac{\rho C_p (e_s - e_a)}{\gamma (r_a + r_s)} \qquad (2c)$$

and a first-order Taylor approximation for the saturation vapor pressure function (e_s) of the air temperature (T_a) which avoids the need of surface temperature (T_s) (which is difficult to obtain in many cases):

$$e_s(T_s) = e_s(T_a + dT) \; with \; dT = T_s - Ta$$

$$\text{or } e_s(T_s) \approx e_s(T_a) + \left.\frac{de_s}{dT}\right|_{Ta} (T_s - T_a) \qquad (3)$$

where $e_s(T_s)$ is saturated vapor pressure at the surface temperature (T_s).

The first-order approximation however may lead to underestimation of LE (Gao, 1988). 2nd order approximation may better capture the curvature of e_s:

$$e_s(T_s) = e_s(T_a) + \left.\frac{de_s}{dT}\right|_{Ta} (T_s - T_a) \; + \frac{1}{2}\left.\frac{d^2e_s}{dT^2}\right|_{Ta} (T_s - T_a)^2 \; + \cdots$$

or

$$e_s(T_s) \approx e_s(T_a) + \Delta(T_s - T_a) \; + \frac{1}{2}\left.\frac{d\Delta}{dT}\right|_{Ta} (T_s - T_a)^2 \quad (4)$$

Combining equations (2) and (3) and eliminating T_s, we lead to following quadratic equation for LE (see Gao (1988) for more details):

$$aLE^2 + bLE + c = 0 \qquad (5a)$$

with

$$a = \frac{1}{2}\beta\left(\frac{r_a}{\rho C_p}\right)^2 \qquad (5b)$$

$$b = -\left[\beta\left(\frac{r_a}{\rho C_p}\right)^2 (R_n - G) + \frac{\Delta r_a}{\rho C_p} + \frac{\gamma(r_a + r_s)}{\rho C_p}\right] \qquad (5c)$$

$$c = \frac{1}{2}\,\beta\left(\frac{r_a}{\rho C_p}\right)^2 (R_n - G)^2 + (e_s - e_a) + \frac{\Delta r_a (R_n - G)}{\rho C_p} \qquad (5d)$$

and

$$\beta = -\frac{5006.12(T_a - 1811.79)e^{\frac{17.27T_a}{237.3+T_a}}}{(237.3+T_a)^4} \qquad (5e)$$

In equation (5a), coefficient a is positive and b is negative so equation (5a) will have two roots. To avoid this issue, equation can be converted to H as:

$$AH^2 + BH + C = 0 \qquad (6a)$$

with

$$A = \frac{1}{2}\beta\left(\frac{r_a}{\rho C_p}\right)^2 \qquad (6b)$$

$$B = \frac{\Delta r_a}{\rho C_p} + \frac{\gamma(r_a + r_s)}{\rho C_p} \qquad (6c)$$

and

$$C = (e_s - e_a) - \frac{\gamma(r_a + r_s)(R_n - G)}{\rho C_p} \qquad (6d)$$

In equation (6a), both the coefficients A and B are positive, so there can be only one solution to the equation (6a) and H should be positive, so:

$$H = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \qquad (7)$$

In the PGML, H is constrained for the loss function, but the value of A is generally small (~ 10^{-10}), which can lead to instabilities with vanishing or infinite gradients. To avoid this issue, coefficients B and C can be directly targeted instead of H in the loss function:

$$Loss(\hat{H}, H) = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(\hat{B}_i - B_i)^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{N}(\hat{C}_i - C_i)^2} \qquad (8)$$

Finally, the LE can be obtained as the residual of the surface energy budget:

$$LE = R_n - G - H \qquad (9)$$

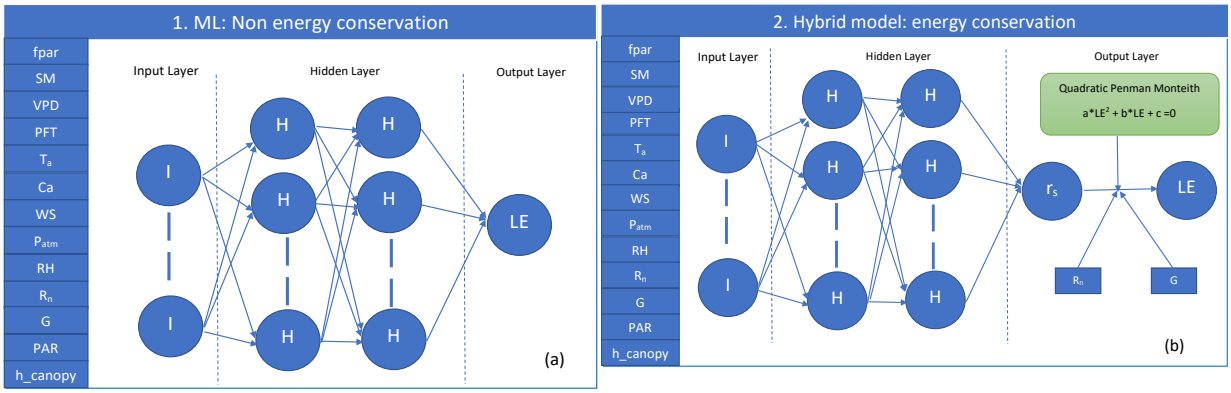

Figure 1. Model architectures of (1) Pure machine learning model; (2) physics-guided machine learning model (PGML).

3. Data

We obtained data from 82 eddy covariance sites from FLUXNET 2015 Tier 1 data set (https://fluxnet.fluxdata.org/data/download-data/) which provides measurements of various climate and hydrological variables under different landcover settings including croplands, grasslands, evergreen broadleaf forests, evergreen needleleaf forests, deciduous broadleaf forests, mixed forests, closed shrublands, woody savannas and, savannas. The data is available at every 30 minutes interval. We followed the data filtering procedure provided in Li *et al.* (2019) as:

- Selected only measured and good quality gap-filled data.
- Only daytime data was used to avoid stable boundary layer conditions.
- All the rainy days were removed to avoid measurements errors.
- Negative LE, GPP, and VPD were filtered out.
- The extremes (below 5th percentile and above 95th percentile) of net radiation, carbon concentration, and ground heat flux were filtered out.
- Assuming the EC towers measure Bowen ratio ($B = H/LE$) more accurately, the fluxes were corrected to ensure energy balance closure (EBC; $R_n - G = LE + H$) at each site.

Table 1 lists all the relevant information on the data used in this study.

Table 1. The variables to be used for training neural network.

Number	Variables	Full Name	Units	Data source	Frequency
1	fPAR	Fraction of photosynthetically active radiation	-	MCD15A3H (https://modis.ornl.gov/globalsubset/)	4 days
2	SM	Soil water content	$m^3 m^{-3}$	Fluxnet 2015 Tier 1 dataset	1 hour
3	VPD	Vapor pressure deficit	kPa	Fluxnet 2015 Tier 1 dataset	1 hour
4	PFT	Plant function type	-	Fluxnet 2015 Tier 1 dataset	
5	TA	Air temperature	℃	Fluxnet 2015 Tier 1 dataset	1 hour
6	Ca	CO_2 concentration	umol mol^{-1}	Fluxnet 2015 Tier 1 dataset	1 hour
7	WS	Wind speed	m s^{-1}	Fluxnet 2015 Tier 1 dataset	1 hour
8	PA	Atmospheric pressure	kPa	Fluxnet 2015 Tier 1 dataset	1 hour
9	RH	Relative humidity	%	Fluxnet 2015 Tier 1 dataset	1 hour
10	R_n	Net radiation	W m^{-2}	Fluxnet 2015 Tier 1 dataset	1 hour
11	G	Soil heat flux	W m^{-2}	Fluxnet 2015 Tier 1 dataset	1 hour
12	PAR	Photosynthetically active radiation	µmolPhoton m^{-2} s^{-1}	Fluxnet 2015 Tier 1 dataset	1 hour
13	h_canopy	Canopy height	m	(Pennypacker and Baldocchi, 2016)	constant over time
14	r_s	surface resistance	s m^{-1}	Inverted from Penman-Monteith's equation	1 hour
15	LE	Latent Heat flux	W m^{-2}	Fluxnet 2015 Tier 1 datasset	1 hour

Note: Variables from 1-13 will be used as input variables and from 14-15 as target variables.

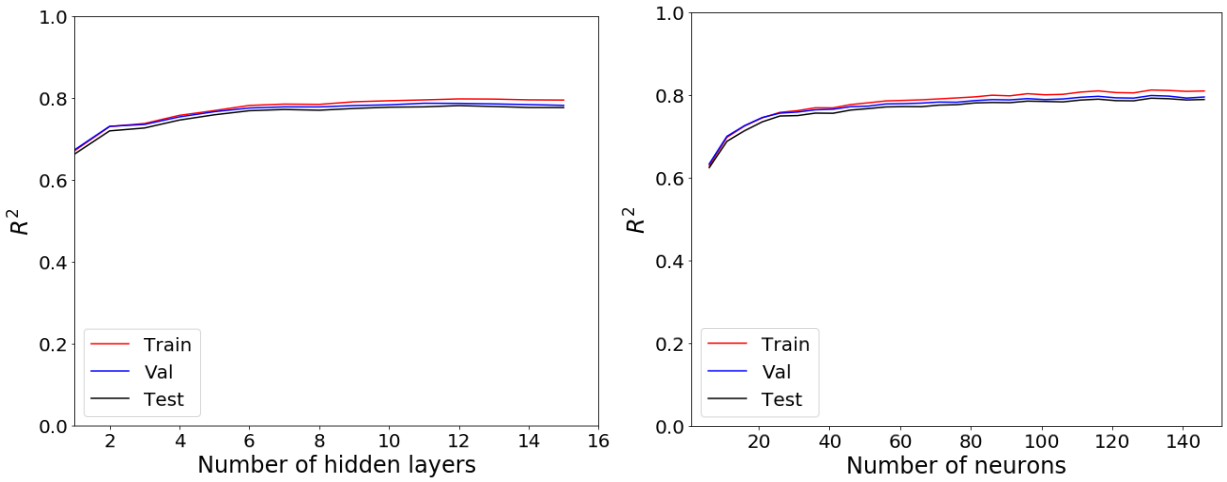

Figure 2. Demonstration for the optimal ANN structure. (a) Coefficients of determination (R^2) between ANN retrievals of *LE* and target data as function of number of neurons in hidden layer, the ANN has 5 hidden layers. (b) R^2 between ANN retrievals of *LE* and target data as function of number of hidden layers, the number of neurons in each hidden layer is set as 64.

Mean Square Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2) were used as performance metrices in this study.

4. Results and Discussion

4.1. Pure ML model vs PGML for the prediction of LE

Figure 3 compares the performances of pure ML model and PGML model for the prediction of LE. Both the models show the similar performances in predicting the LE (see the performances statistics in Figure 3). LE predicted by PGML has slightly lower Mean Square Absolute Error (MAPE) (19.59% vs 21.22%) on testing set.

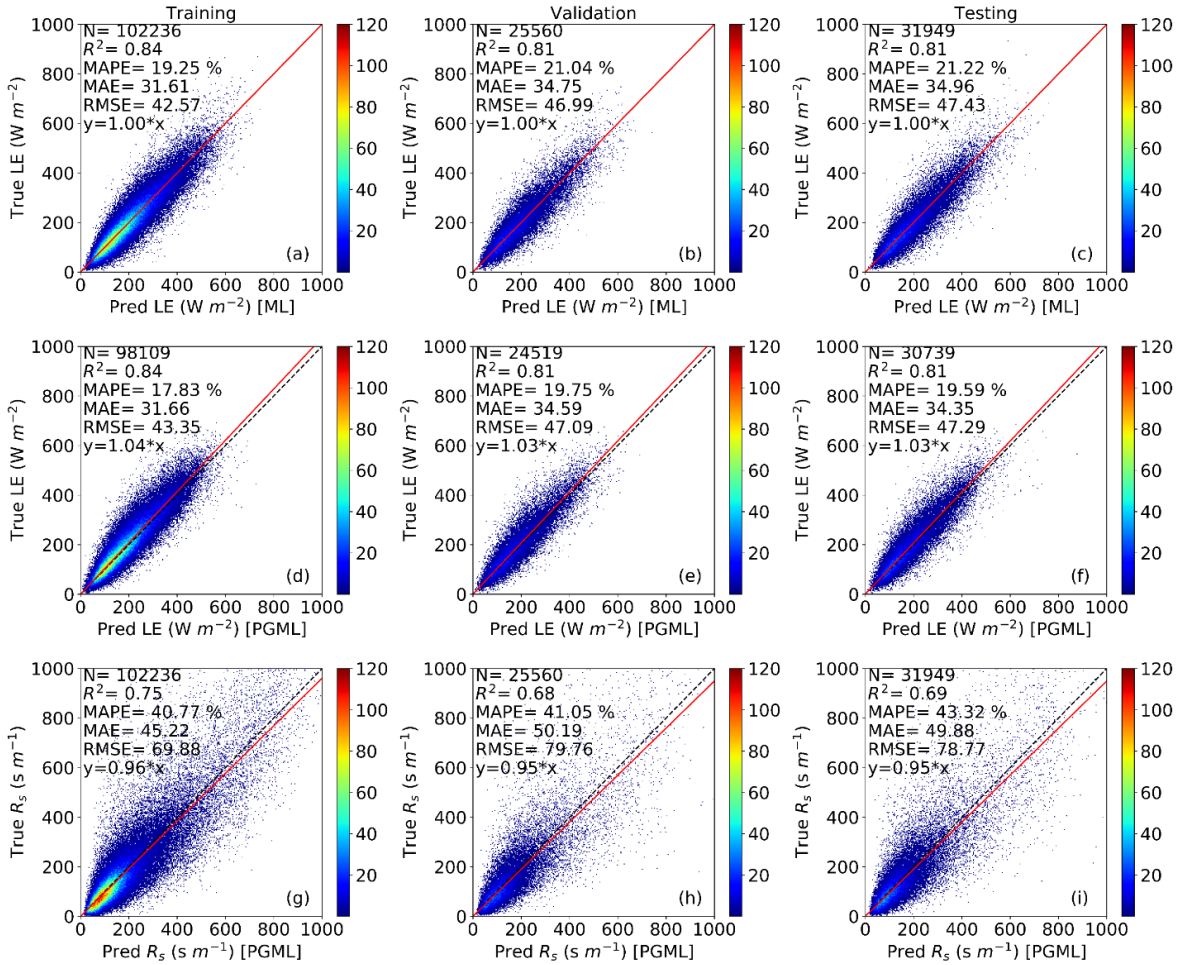

Figure 3. Comparing the performance of pure machine learning model and Physics-guided machine learning (PGML) for the prediction of LE. (a)-(c) LE predictions by pure ML model on training (left panel), validation (middle panel), and testing (right panel) sets. (d)-(f) LE predictions by PGML on training (left panel), validation (middle panel), and testing (right panel) sets. (g)-(i) surface resistance (R_s) predictions by PGML on training (left panel), validation (middle panel), and testing (right panel) sets. The red solid line is nonbiased linear regression line (i.e., $y = mx$) and, the diagonal black dashed line is 1:1 line. The scalebar represents the density of the plotted scattered points. N, R^2, MAPE, MAE, and RMSE represent the number of data points, coefficient of determination, mean absolute percent error, mean absolute error, and root-mean-square error, respectively.

Although, the pure ML model and PGML model showed the same skills in predicting the LE but they differ in their skills to conserve the surface energy budget (see Equation 2a). From figure 4, it is clear that PGML respects the surface energy budget (i.e., $R_n - G = H + LE$) but pure ML model does not conserve the surface energy budget over the land surface. The surface energy imbalance was 78 $W\ m^{-2}$ on test set from pure ML model (see Figure 4d).

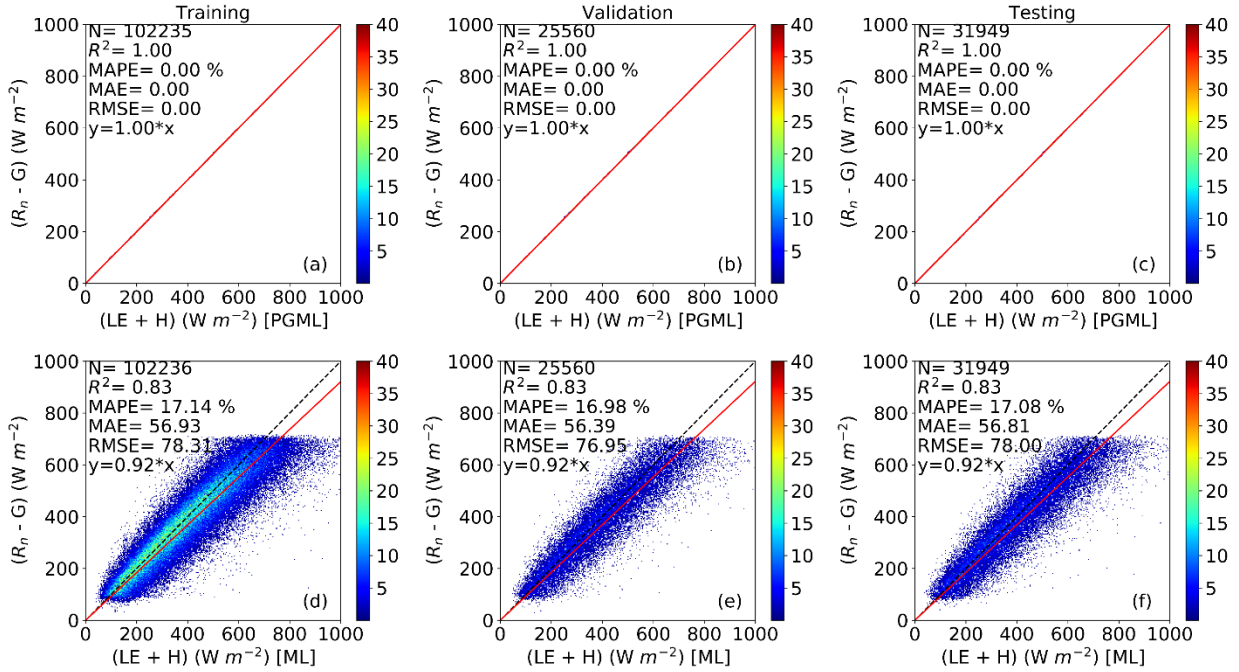

Figure 4. Comparing the capacity of pure machine learning model and Physics-guided machine learning (PGML) in conserving the surface energy budget (i.e., $R_n - G = LE + H$). (a)-(c) Energy conservation by PGML model. (d)-(f) Energy conservation by pure ML model.

4.2. Variables controlling LE and/or R_s

We further performed sensitivity analysis to find how different variables affect the LE and R_s. For this, we added perturbations (10-90% standard deviation increase) to each variable. From PGML, we found that soil moisture (SM) and height of the canopy (h_canopy) are the most critical variables for both LE (see Figure 5b) and R_s (see Figure 5a). From pure ML model, we found that relative humidity (RH) and fraction of Photosynthetically Activation Ratio are even more important than SM and h_canopy in controlling the LE (see Figure 5c). There are clear evidences in the literatures that SM is first order controlling variable to the LE. So, PGML assigns the role of different variables more accurately than pure ML model. Interestingly, h_canopy plays very important role in controlling the LE or r_s. This is very exciting because role of h_canopy has been underestimated in the past. h_canopy controls the transpiration (and thus LE) through plant hydraulic traits.

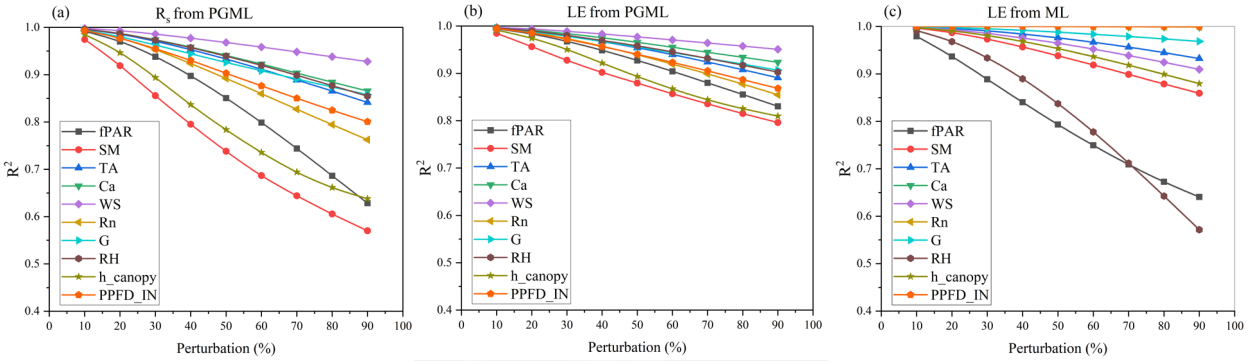

Figure 5. (a) Sensitivity analysis for surface resistance (R_s) predictions by PGML for each input variable over the test data set by giving perturbations of 10, 20, 30, 40, 50, 60, 70, 80, and 90% standard deviation increase in each input variable. R^2 is the coefficient of determination between R_s predictions with and without perturbation. (b) Sensitivity analysis for latent heat flux (LE) predictions by PGML for each input variable over the test data set by giving perturbations of 10, 20, 30, 40, 50, 60, 70, 80, and 90% standard deviation increase in each input variable. R^2 is the coefficient of determination between LE predictions by PGML with and without perturbation. (c) Sensitivity analysis for latent heat flux (LE) predictions by pure ML for each input variable over the test data set by giving perturbations of 10, 20, 30, 40, 50, 60, 70, 80, and 90% standard deviation increase in each input variable. R^2 is the coefficient of determination between LE predictions by pure ML with and without perturbation.

4.3. Performance of the Pure ML and PGML for the predictions of extremes

Here, we tested our hypothesis that PGML can better reproduce extremes and thus can better generalize outside of the range of the training data set compared to the regular ML algorithm. For this, we tested the performance of both the models on extremes (0^{th}-1^{st} percentile or 99^{th} -100^{th} percentile) of the input variables. It should be noted here that these extremes were exclude from the model training process. Here, the aim is to assess the performance of PGML and pure ML for the predictions of LE in extreme conditions: for instance, 0^{th}-1^{st} percentile of soil moisture means drought conditions, 99^{th}-100^{th} percentile of temperature means heat waves, etc. From Table 1, it is clear that PGML performed consistently better than pure ML even under extreme conditions. This means that PGML ET model performs better in extrapolation and for out-of-sample generalization than the pure ML model which is very encouraging.

Table 1. Performance of extrapolation for the PGML and pure ML models over the extreme data set for all sites. R^2 is the coefficient of determination between LE predictions and observations using the extreme data set for all sites.

Variables	**R^2 (Pure ML)**		**R^2 (PGML)**	
	0^{th}-1^{st} percentile	99^{th}-100^{th} percentile	0^{th}-1^{st} percentile	99^{th}-100^{th} percentile
Soil moisture	0.62	0.59	0.71	0.69
Temperature	0.38	0.69	0.47	0.76
PAR	0.68	0.72	0.74	0.71

5. Inferences/Conclusions

We developed a Physics-guided ML model to predict ET. Pure ML and PGML showed the similar skills in predicting ET, but PGML

- Respects the surface energy balance.
- Respects the physics of evapotranspiration.
- Can be used to find the ET controlling factors and their relative contribution.
- Better generalizes during extremes.

References

Alemohammad, S.H., Fang, B., Konings, A.G., Aires, F., Green, J.K., Kolassa, J., Miralles, D., Prigent, C. and Gentine, P. 2017. Water, Energy, and Carbon with Artificial Neural Networks (WECANN): a statistically based estimate of global surface turbulent fluxes and gross primary productivity using solar-induced fluorescence. Biogeosciences 14(18), 4101-4124.

Gao, W. 1988. Applications of solutions to non-linear energy budget equations. Agricultural and Forest Meteorology 43(2), 121-145.

Granata, F., Gargano, R. and de Marinis, G. 2020. Artificial intelligence based approaches to evaluate actual evapotranspiration in wetlands. Science of The Total Environment 703, 135653.

Katul, G.G., Oren, R., Manzoni, S., Higgins, C. and Parlange, M.B. 2012. Evapotranspiration: a process driving mass transport and energy exchange in the soil-plant-atmosphere-climate system. Reviews of Geophysics 50(3).

Leuning, R., Zhang, Y., Rajaud, A., Cleugh, H. and Tu, K. 2008. A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman-Monteith equation. Water Resources Research 44(10).

Li, X., Gentine, P., Lin, C., Zhou, S., Sun, Z., Zheng, Y., Liu, J. and Zheng, C. 2019. A simple and objective method to partition evapotranspiration into transpiration and evaporation at eddy-covariance sites. Agricultural and Forest Meteorology 265, 171-182.

Monteith, J.L. 1965 Evaporation and environment, pp. 205-234, Cambridge University Press (CUP) Cambridge.

Penman, H.L. 1948. Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences 193(1032), 120-145.

Pennypacker, S. and Baldocchi, D. 2016. Seeing the fields and forests: Application of surface-layer theory and flux-tower data to calculating vegetation canopy height. Boundary-Layer Meteorology 158(2), 165-182.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J. and Carvalhais, N. 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566(7743), 195-204.

Yang, F., White, M.A., Michaelis, A.R., Ichii, K., Hashimoto, H., Votava, P., Zhu, A.-X. and Nemani, R.R. 2006. Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine. IEEE Transactions on Geoscience and Remote Sensing 44(11), 3452-3461.