We thank Referee #1 for the questions & comments. We have interspersed our responses between the questions and comments. New texts, figures, Tables planned to be included in the manuscript in response to the comments are marked in blue in our responses below.

Referee #1:

In this work,the authors analyzed the relationship between potential evapotranspiration (ETo), actual evapotranspiration (ETa), and surface water evaporation, using data from multiple sources.

Major comments: 1. Study objective #1 is not clear. Relationship between PET and Actual ET has been well studied in the literature. The authors should mention the previous works in this area. I found the literature survey is too cursory. Suggest that the authors move some materials from Section 2 to the Introduction. Even so, it is not clear to me from the Intro why the existing results are not sufficient, such that the authors need a sophisticated ML to revisit the problem. The motivation needs to be elaborated further in the Intro.

Response:

1.1 'objective #1 is not clear. Relationship between PET and Actual ET has been well studied in the literature'

To be more accurate with the definition of different evapotranspiration processes in our revised manuscript in accordance with the description by Allen(1998) and McMohan et al. (2013) (both references were cited in the original manuscript), we have replaced 'PET' (referring to ET_o in the original manuscript) with the 'reference crop evapotranspiration', and also provided the description of 'PET' and differentiated it from the reference crop evapotranspiration, ET_o. To that end, we plan to include the following description in the end of the first paragraph of the Introduction section:

Evapotranspiration can be computed as reference crop evapotranspiration (ET_o), actual evapotranspiration (ET_a), or as potential evapotranspiration from wet surfaces (ET_p) with a specific crop type or surfaces covered by large volume of water, such as wetlands or lakes (Stagnitti et al., 1989). E_{sw} is used to represent ET_p from a lake in this paper. E_{sw} from a free water surface has been commonly estimated using the Penman equation (Penman, 1948) that combines the energy budget and mass transfer approaches.

After all, the correct statement in the revised manuscript will read as

'...The deterministic analysis reveals that reference crop evapotranspiration, ET_o, set the upper bound for ET_a, but the lower bound for E_{sw} in the study area...'

For the sake of clarity, in Objective #1, we look into the relationship not only between reference crop evapotranspiration (ET_o) and Actual ET, but also their relations with lake evaporation.

As per the referee's comment, however, we plan to include the following discussion (and additional reference) on the relationships between reference crop evapotranspiration, actual ET, and lake evaporation in the Introduction section:

From a practical standpoint, ET_p has been applied mostly in hydrology, meteorology and climatology; whereas, ET_o has been applied mostly in agronomy, agriculture, irrigation and ecology (Xiang et al., 2020). In particular, ET_p rather than ET_a is a common input for hydrological models, such as HYDRUS, SWAP, SWAT, and MODFLOW-2000 (Li et al., 2016). In drought characterization, ET_p, approximated by the ET_o, has been used to calculate the aridity index (Kingston et al., 2009, Greve et al., 2019). Although Kristensen and Jensen (1975) reported that ET_p may not be the upper limit of ET_a for all crops or development stages, typically ET_p sets the upper bound for ET_a due to limited water availability for evapotranspiration (Lascano and Bavel 2007, Li et al., 2016). When $ET_a < ET_p$, moisture becomes limited, the air becomes drier and the excess energy heats up the atmosphere, which subsequently increases ET_p (Wang and Zlotkonik, 2012). However, $ET_p \cong ET_a \cong E_{sw}$ holds for wet surface evaporation (Mortan, 1965, Milly and Dunne, 2016). (ET_a/ET_p) represents the evaporative stress index, (ESI), in which ET_p was approximated by Liu et al. (2019) using the PME-computed PME. The ESI was used to study short term droughts (Choi et al., 2013) and evaluate the irrigation need for crop growth and land classification (Yao, 1974) and water stress using remotely sensed hydrological and ecological properties (Anderson, 2016). If soil moisture data are available, ET_a can be computed by multiplying ET_p by the soil moisture extraction function, defined as the ratio of the measured soil moisture to the field capacity (Lingling et al., 2013). For more comprehensive discussion on different evapostranspiration measures, the readers may refer to the paper by McMahon et al. (2013).

Applications discussed above show that different, yet interrelated, evapotranspiration measures have been used in practice, although they may be converted from one into the other using empirical relations and/or additional hydroclimatic variables. Additional complexities in evapotranspiration calculations and projection are introduced by changes in climate (Milly and Dunne 2016) and land use (Ozdogan and Salvucci, 2004), which alter land surface and lower atmosphere energy budget, and hence, evapotranspiration rates. For example, expansion of irrigated areas in the southern parts of Turkey resulted in $\sim 50\%$ reduction in ET_p and E_p in 23 years due to decreases in wind speed and increases in humidity. Similarly, ET_o exhibited a decreasing trend with an average value of 3 mm/year in the northwest China over 50 years due to decreasing wind speed and radiation and increasing humidity and temperature (Huo et al., 2013).

1.2 'literature survey is too cursory'

In the original manuscript, we provided extensive literature review with 50+ references cited in the Introduction and Methods sections in describing different evaporation measures (ET_o, ET_a, E_{sw}), related measurement and calculation techniques, the rationale for their inclusion in our analysis, previously used ML techniques, and the main advantages of our ML model for evapotranspration predictions.

As per the referee's comment, however, we have included additional references (previously reported relations in the literature among the evapotranspiration measures) in the revised manuscript to further improve the discussion and clarity (please see our response 1.1 above for further details).

1.3 'moving some materials from Section 2 to the Introduction'

Agree. As per the referee's comment, the Introduction section and Section 2 will be merged, restructured, and revised to enhance the clarity. The discussion on the Edwards aquifer system (between line $\#$ 24-30 in the Intro section of the original manuscript) will move to Section 2.1, which will be named 'Study Area and Data Availability' in the revised manuscript.

1.4 'why existing results are insufficient; why a sophisticated ML to revisit the problem'

To our knowledge, this is the first ML model proposed & tested to *simultaneously* predict ET_o (while avoiding computationally involved net solar radiation calculations), E_{sw} (while suppressing uncertainties associated with pan evaporation coefficient, measurements, and upscaling), and ET_a (while offsetting high capital & maintenance costs of Eddy Covariance towers) using the standard sets of climatic data obtained from local weather stations, with the exception that water temperature is also required in E_{sw} calculations. From a practical standpoint, it is a cost-effective and efficient computational method that can be used to predict different evaporation measures *simultaneously with high accuracy*, explicitly addressing prediction *'uncertainties'*, different from traditional ML models.

To further provide insights into why we want to use a more sophisticated ML model, we plan to include the following 'research question' in the Introduction section.

Considering the presence of different models for representing evapotranspiration processes as discussed above, we raise the following research question: Can we have a computationally-efficient and unified data-driven machine learning (ML) model to (i) avoid calibration parameters and empirical relations, (ii) calculate different evapotranspiration measures using the standard hydroclimatic data sets, (iii) analyze and report the order of importance of hydroclimatic variables, *while explicitly accommodating their interactions with each other* to identify the most crucial data that need to be acquired for particular evapotranspiration process, (iv) seek new knowledge that may not be readily available from non-probabilistic ML, numerical, or empirical models, and (v) perform probabilistic predictions over the entire solution space for more accurate assessment of uncertainties related to hydrological predictions?

Moreover, in reference to other ML algorithms listed in the Intro section of the original manuscript, the rationale for the use of a sophisticated ML (hybrid XGBoost-NGBoost) was already provided in line # 51 and 159-165 of the original manuscript. The new ML model is capable of taking 'uncertainties' into consideration. As emphasized in the Intro section, this challenge was brought up by Tang et al. (2018). As illustrated in Fig. 1 and discussed throughout the manuscript, the proposed hybrid ML model is capable of producing not only point predictions (as can be done by traditional ML models), but also a *probability distribution* over the entire solution space (new modeling capability by our hybrid ML model) required for quantifying the *uncertainties* related to hydrological predictions.

More precisely, the hybrid model can produce prediction intervals in addition to point predictions that can effectively inform the user about the model's confidence by quantifying the total number of target datapoints that fall within the specified bounds (e.g. 95%) of the prediction interval. The statistical metrics calculated on the point predictions - such as R^2 and RMSE - may not be conclusive about the predictive uncertainties of the resulting models. For example, the statistical metrics for the ET_a model (in Table 1) suggest that the model performs reasonably or fairly, whereas the ML-evaluated 'probabilistic distribution' reveals that 91% of the predicted values are within the 95% prediction interval of the target variable, which provides additional confidence in the model's performance from a practical point-of-view.

1.5 'motivation needs to be elaborated further in the Intro'

Agree. Please see our response to Major comment 2 (our reponse 2.1) below.

Major comment 2: Similarly, in Section 2 the motivation of using ML is not clear. What regression methods have been used before? Why existing methods are insufficient in terms of model performance? The authors need to touch on these aspects. Otherwise, the work seems to focus on a new ML algorithm without justification and no baseline results (e.g., multivariate linear regression) were provided.

Response:

2.1 'motivation'

Although the main motivation for the use of ML was hinted in the Abstract (line # 8-13), throughout the manuscript (e.g., line # 283-287, 300-302, 309-311, 320-322), and in the Conclusion section (line $\#400 - 412$) of the original manuscript, as per the referee's comment, we plan to include the following info in the Intro section of the revised manuscript to enhance the clarity:

The main motivation for development of such a model is to (i) seek an alternative method to the Penman-Monteith equation to avoid computationally intensive net solar radiation computations in ET_o calculations; (ii) overcome uncertainties associated with the pan coefficients, pan evaporation measurements, and upscaling methods for E_{sw} estimates; (iii) offset high capital & maintenance costs of EC towers used for ET_a measurements, and (iv) assessing uncertainties associated with the ML predictions.

2.2 What regression methods have been used before?

As discussed in line # 45 - 57 of the original manuscript, previous regression-based methods that were utilized include neural networks, clustering, tree-based ensembles, fuzzy models, multivariate adaptive regression splines, and extreme learning machines. However, the crux of the problem is that none of these ML models are designed to produce prediction intervals for continuous target variables such as ET_o, E_{sw}, and ET_a, which is critical to account for the inherent uncertainties in predictions. Therefore, we applied NGBoost to solve this problem, which is designed to produce prediction intervals for continuous target data. However, as mentioned by Duan et al. [1], "NGBoost is not specifically designed for point estimation" and "better tree-based base learners and regularization (such as XGBoost by Chen and Guestrin [2]) are more likely to improve performance". These future work indications in their paper led us to develop the hybrid NGBoost-XGBoost model that provides the best of both models in terms of their ability to accurately generate both the prediction intervals and point predictions, respectively (this information and references will also be included in the revised manuscript).

2.3. ‘why existing methods are insufficient in terms of model performance?’

As elaborated in the original manuscript (e.g., line # 159-165), the existing ML models provide only the point predictions, **but not the probability distributions** over the entire outcome space of continuous target variables. The latter, however, is critical for enhanced uncertainty assessments and building confidence in model predictions.

For example, unlike the other ML methods listed in line #51 of the original manuscript, ‘uncertainties’ in predictions are accommodated by the NGBoost-XGBoost model, which provided the confidence that at least 90% of the predicted PET, actual ET, and lake evaporation in our ML-based calculations were within 95% of the prediction interval of the target variables. Please see our responses in 1.4 and 2.2 for additional discussion.

2.4 ‘without justification and no baseline results?’

‘justification’

For justification for the use of the XGBoost-NGBoost model, pleased see (i) through (v) under ‘research question’ in our responses 1.4, and (i) through (iv) under ‘motivation’ in our response 2.1 from the application standpoint, in addition to ‘built-in uncertainty calculation’ for the predicted target variables from the ML modeling standpoint. This will be highlighted in the Intro section of the revised manuscript to enhance the clarity.

‘baseline results’

As per the referee’s comment, we have compared the performance of the hybrid model with respect to a baseline linear regression model and included the results in Table 1 below (new analyses and results are marked in blue), which will be included the revised manuscript. The comparison revealed that our proposed hybrid model performs better than the baseline in terms of the statistical metrics for point predictions. But, more importantly, it provides uncertainty estimates through ‘probabilistic predictions’ (as we further elaborated in our response to Comment # 1.4), which is imperative to practically deploy such models with confidence.

Table 1: Hybrid NGBoost-XGBoost ML model accuracy test with statistical measures and comparison with a baseline linear regression model.

	Model	Data	RMSE*(mm)	MAE†(mm)	R^2 ‡	$C_f^{\S}$ (%)
ET_o	Baseline	Training data only	0.205	1.364	0.984	-
		Testing data only	0.191	1.374	0.986	-
	Hybrid	Training data only	0.099	0.074	0.996	100
		Testing data only	0.139	0.102	0.992	99.4
E_{sw}	Baseline	Training data only	0.953	1.493	0.711	-
		Testing data only	1.015	1.504	0.695	-
	Hybrid	Training data only	0.703	0.545	0.843	99.1
		Testing data only	0.918	0.736	0.750	89.9
ET_a	Baseline	Training data only	0.647	1.003	0.698	-
		Testing data only	0.719	1.011	0.643	-
	Hybrid	Training data only	0.388	0.291	0.891	99.9
		Testing data only	0.533	0.411	0.804	91

(*) Root mean square error; † Mean absolute error; ‡ Correlation Coefficient; § Percentage of datapoint within the model’s 95% prediction interval.

Major comment 3: The ML pipeline is not clear. A diagram is needed to show inputs and output to the ML model. Around L185, the authors simply spelled out the inputs, without much reasoning. Why these features are selected? What is the lead time of prediction? The promise of ML is not so much for well gauged sites, but for sites with a lot of missing data.

Response:

3.1 'diagram'

As per the referee's comment, we have revised the flow chart in Fig. 1 of the original manuscript, as shown below, which we plan to include in the revised manuscript.

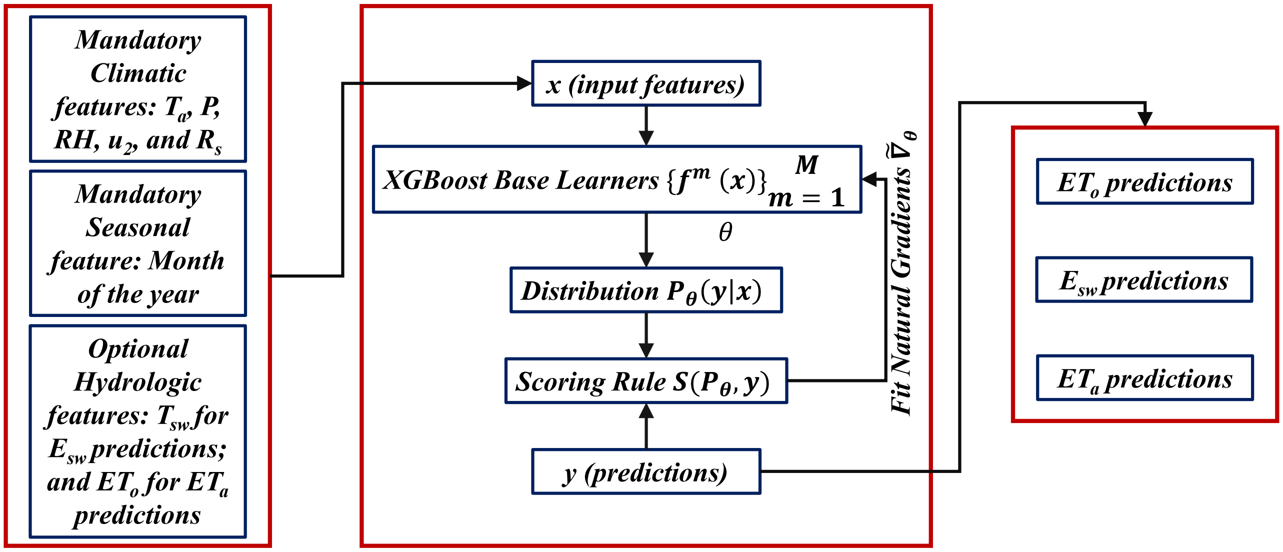
Mandatory Climatic features: T_a, P, RH, u_2, and R_s
Mandatory Seasonal feature: Month of the year
Optional Hydrologic features: T_{sw} for E_{sw} predictions; and ET_o for ET_a predictions
x (input features)
XGBoost Base Learners $\{f^m(x)\}_{m=1}^{M}$
θ
Distribution $P_\theta(y|x)$
Scoring Rule $S(P_\theta, y)$
y (predictions)
Fit Natural Gradients $\tilde{\nabla}_\theta$
ET_o predictions
E_{sw} predictions
ET_a predictions

Figure 1: Conceptual representation of the hybrid NGBoost-XGBoost model for ET_o, E_{sw}, and ET_a prediction.

3.2 'why these features are selected?'

As per the referee's comment, we will include this information in the manuscript. Briefly,

The climatic variables (T_a, P, RH, u_2, R_s) in Eq. 1 (PME) were chosen as the features for the ET_o predictive model. The same climatic variables were used as the features for the ET_a predictive model, in addition to ET_o to quantify its contribution to ET_a. T_{sw} in Eqs. 6 and 7 (Meyer's formula) was added as a new feature to the climatic variables in the E_{sw} predictive model. Moreover, 'month' was chosen as an additional feature in all predictive models based on the observed seasonality in T_{sw} data, ET_a measurements from the EC tower, and expected seasonality in soil moisture content at the site where the the EC tower is located.

3.3 what is the lead time in predictions?

If the Referee refers to the 'training time' of the ET_o, E_{sw}, and ET_a predictive models by the 'lead time in predictions', this information was already included in the 'Predictive ML Models' section of the original manuscript. Briefly, the total training time of the ET_o, E_{sw}, and ET_a predictive models are $\sim$ 30 min, $\sim$ 6 min, and $\sim$ 9 min. As mentioned in the manuscript, the models were developed using an Intel Core i9-9980XE processor and 64 GB RAM computer. Otherwise, please provide clarification on what the Referee means by the 'lead time in predictions'.

3.4 'The promise of ML is not so much for well gauged sites, but for sites with a lot of missing data.'

There is a growing number of studies in the literature, in which ML models have been primarily used for assessment, simulation, estimation, solving and capturing nonlinear complexity, and projection of various hydrologic processes, including for example, precipitation, evapotranspiration, droughts, floods, groundwater levels ([3]–[15] are just a few examples from a long list of recent articles), different from other types of studies, in which the ML methods are primarily used for imputing missing data, as the Referee mentioned about. These two types applications of the ML models (simulation & prediction vs. data imputation) were emphasized in Refs. [7] and [16], as an example, without favoring one over another, as the ML methods have emerged as promising modeling tools in both application fields, but not as a promising tool only for 'ungauged' systems with lots of missing data, as the Referee suggested.

In the first type of applications discussed above (which is well-aligned with the the main scope of our current manuscript), data imputation are often taken care of outside the ML modeling (e.g., by removing missing or suspicious data as in Refs. [17], [18] or using statistical packages as in Ref. [3]). Thus, the main goal of the ML modeling in this case is to decipher the nonlinear dynamics between the input (predictive variables) and output (target variables) and use such information for prediction & projection of target variables without requiring detailed physical information (e.g., constitutive equations reflecting physical laws) on the investigated system, which would otherwise require a large volume of data, non-unique calibration processes, and high computational costs, as emphasized by [7] and [10], as an example.

After all, we are well aware that ML modeling for hydrological simulations and predictions and ML modeling for data imputations are two different (but, sometimes integrated) applications. We have another manuscript currently in review at one of the Artificial Intelligence journals, focusing on a new ML model -based on the transfer learning approach- to impute long-stretches (a few months to a year) of missing data. This is another application of the ML model the referee is referring to. In brief, in light of current literature and our experience, we believe ML modeling is as important to analyze, predict, and project nonlinear dynamics between predictor and target variables as to impute missing data. And, our current manuscript focused on the former application.

Major comment 4: Variable importance calculation is well established for tree-based method, which entails finding whether a variable is selected to split on during the tree building process, and how much the squared error (over all tress) us improved or reduced as a result. Why a new variable importance method is needed?

Response: The typical variable importance calculations are equivalent to a sensitivity analysis, measuring *relative* contribution of a specific predictor variable to the target observation *without* accommodating the dynamic interaction of that specific predictor variables with the other predictor variables. In contrast, as mentioned in the original manuscript, the Shapley value is the average marginal contribution of each feature value across all possible combinations of features. Thus enabling (i) global interpretability: the collective SHAP values can show how much each feature contributes to the target, which is similar to the traditional tree-based permuted feature importance, however, the SHAP plots can additionally explain the positive or negative relationship between each feature value and the target (see Fig. 10 in the manuscript); (ii) local interpretability: traditional variable importance plots only show the results across the entire dataset, but not on each individual datapoints. In contrast, with the new SHAP-based technique each observation gets its own set of SHAP values (see Fig. 12 in the manuscript). This greatly increases the transparency of the ML models and reveals new insights. For example, Fig. 12 reveals new knowledge on low ET_a predictions despite high ET_o & low RH measures in hot and dry summer, which will be critical in future climate scenarios in Texas or elsewhere. Such insights are not available from the traditional tree-based variable importance plots.

Minor comment: In Abstract, the authors concluded "the deterministic analysis reveals that ETo set the upper bound for Eta", isn't this expected?

Response

As we discussed in our response to the Major Comment $\#1$, ET_o represents the reference crop evapotranspiration, not the potential evapotransporation (this was a typo in the original manuscript and we corrected in the revised version).

Fig. 8b show that $ET_a < ET_o$ for most of the time, except for October through December of 2019, in which $ET_a \cong ET_o$, which we did not expect initially for a semi-arid region in Texas.

Having said this, $ET_a < ET_o$ confirms the reliability and confidence in measured local climatic data from the local weather stations, our ET_o calculations using the PME, and the actual ET measurements from the EC tower. The validity of this relation confirms that the ML models were operated on ‘physically reasonable’ input data sets, eliminating concerns on potential uncertainties or inconsistencies in the input data used in ML analysis.

References

[1] T. Duan, A. Avati, D. Y. Ding, S. Basu, A. Y. Ng, and A. Schuler, “Ngboost: Natural gradient boosting for probabilistic prediction”, *ArXiv preprint arXiv:1910.03225*, 2019.

[2] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system”, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[3] P. Malakar, A. Mukherjee, S. Bhanja, D. Saha, R. K. Ray, S. Sarkar, and A. Zahid, “Importance of spatial and depth-dependent drivers in groundwater level modeling through machine learning”, *Hydrol. Earth Syst. Sci. Discuss.*, 2020, in review. DOI: `https://doi.org/10.5194/hess-2020-208`.

[4] M. K. Nema, D. Khare, and S. K. Chandniha, “Application of artificial intelligence to estimate the reference evapotranspiration in sub-humid doon valley”, *App. Water Sci.*, vol. 7, no. 7, pp. 3903–3910, 2017.

[5] S. Pan, N. Pan, H. Tian, P. Friedlingstein, S. Sitch, H. Shi, V. Arora, V. Haverd, A. Jain, E. Kato, S. Lienert, D. Lombardozzi, J. Nabel, C. Ottlé, B. Poulter, S. Zaehle, and S. Running, “Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches in remote sensing, machine learning and land surface modeling”, *Hydrol. Earth Syst. Sci.*, vol. 24, pp. 1485–1509, 2020.

[6] C. Chong, H. Wei, Z. Han, X. Yaru, and Z. Mingda, “A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China”, *Sci. Rep.*, vol. 10, p. 3904, 2020.

[7] R. Taormina, K.-W. Chau, and R. Sethi, “Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon”, *Eng. Appl. Artif. Intell.*, vol. 25, no. 8, pp. 1670–1676, 2012.

[8] A. Zhang, J. Winterle, and C. Yang, “Performance comparison of physical process-based and data-driven models: A case study on the Edwards Aquifer, USA”, *Hydrogeol. J.*, 2020.

[9] Z. A. Al-Sudani, S. Q. Salih, A. Sharafati, and Z. M. Yaseen, “Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation”, *J. Hydrol.*, vol. 573, pp. 1–12, 2019.

[10] T. Rajaeea, H. Ebrahimia, and V. Nouranib, “A review of the artificial intelligence methods in groundwater level modeling”, *J. Hydrol.*, vol. 572, pp. 336–351, 2019.

[11] S. Jovic, B. Nedeljkovic, Z. Golubovic, and N. Kostic, “Evolutionary algorithm for reference evapotranspiration analysis”, *Comput. Electron. Agric.*, vol. 150, pp. 1–4, 2018.

[12] L. Schmidt, F. Heße, S. Attinger, and R. Kumar, “Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany”, *Water Resour. Res.*, vol. 56, e2019WR025924, 2020.

[13] O. Kisi and M. Alizamir, “Modelling reference evapotranspiration using a new wavelet conjunction heuristic method: Wavelet extreme learning machine vs wavelet neural networks”, *Agr. Forest Meteorol.*, vol. 263, pp. 41–48, 2018.

[14] N. Khan, D. A. Sachindra, S. Shahid, K. Ahmed, M. S. Shiru, and N. Nawaz, “Prediction of droughts over Pakistan using machine learning algorithms”, *Adv. Water Resour.*, vol. 139, p. 103 562, 2020.

[15] K. Ahmed, D. A. Sachindra, S. Shahid, Z. Iqbala, N. Nawaz, and K. Najeebullah, “Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms”, *Atmos. Res.*, vol. 236, p. 104 806, 2020.

[16] X. Dou and Y. Yang, “Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems”, *Comput. Electron. Agric.*, vol. 148, pp. 95–106, 2018.

[17] J. Fan, W. Yue, L. Wu, F. Zhang, H. Cai, X. Wang, X. Lu, and Y. Xiang, “Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China”, *Agr. Forest Meteorol.*, vol. 263, pp. 225–241, 2018.

[18] X. Lu, Y. Ju, L. Wu, J. Fan, F. Zhang, and Z. Li, “Daily pan evaporation modeling from local and cross-station data using three tree-based machine learning models”, *J. Hydrol.*, vol. 566, pp. 668–684, 2018.