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 #2:

Comment 1. This paper is not well motivated and lacks focus.

Response:

1.1 'not well motivated'

Without any specific reasoning, it is unclear to us why the Referee considered our work to be 'not well-motivated'.

However, as already articulated in the Abstract, Introduction, main body (e.g., line # 282-287 on page 13), and in the Conclusions section of the original manuscript, the main motivations for the work and development of a new probabilistic Machine Learning (ML) model are to

(i) seek an alternative method to the Penman-Monteith equation to avoid computationally involved net solar radiation computations in ET_o calculations,

(ii) overcome uncertainties associated with the pan coefficients and pan evaporation measurements,

(iii) offset high capital & maintenance costs of EC towers used for ET_a measurements, and

(iv) assess uncertainties associated with the ML predictions

The first three motivation and the associated findings have been articulated throughout the manuscript (e.g., the last paragraph on page 13). Motivation (iv) was associated with the challenge brought up by Tang et al. (2018), which was already cited and discussed in line # 52 of the original manuscript.

Although we believe that the main motivation of the work was clear in the original manuscript, we plan to include a stand-alone 'motivation' discussion, as discussed above, in the Introduction section of the revised manuscript to enhance the clarity.

1.2 'lacks focus'.

Again, without any specific reasoning, it is unclear to us why the Referee considered the manuscript 'lacks focus'.

However, as already highlighted in the Title, noted in the Abstract, and discussed throughout the manuscript, the manuscript **focuses** on the use of a novel probabilistic ML model (a hybrid XGBoost-NGBoost model) to calculate the daily and monthly ET_o, E_{sw}, and ET_a using local hydroclimatic data, and to identify & quantify the importance and interactions of features (hydroclimatic variables) in predicting targeted evapotranspiration measures using Shapley values - a method from the coalition game theory.

This information was already provided in **'Objectives 1−3'** in line # 39−44 (on page 2) of the Introduction section of the manuscript, which was the **main focus** of the subject manuscript and associated with the main conclusions elaborated in line # 410−428 of the Conclusion sections (on pages 21−22) of the manuscript. This is also consistent with the highlights we provided in line # 4−14 of the Abstract.

Briefly, the main focus of the manuscript was consistently articulated in the Abstract, Introduction, throughout the text of the manuscript, and in the Conclusion section. So, we believe the main focus of the work has

been carried out consistently throughout the manuscript.

As we highlighted in the original manuscript, the use of the hybrid XGBoost-NGBoost in ET_o, E_{sw}, and ET_a prediction, and the use of Shapley values (computed based on the coalition game theory) to determine the importance (considering their interactions explicitly) of the features (i.e., hydroclimatic variables) on the targeted evapotranspiration are unprecedented. Although we discussed the rationale and motivation for the use of these new methods in evapotranspiration prediction in the manuscript, we provided further justification for the use of these novel modeling features in our responses to the Referee's comments below.

Comment 2. The point of E0 modeling by XGBoost was to saved computation. The P-M equation is very simple and the computational time is negligible.

Response:

The main motivation of the ET_o modeling using the hybrid XGBoost-NGBoost model was to reduce (or eliminate) the number of equations (and associated parameters) implemented with the time-series of climate or derived data in ET_o calculations. Eq. (1) may look simple, yet it is computationally involved (see Fig. 1 for the main calculation steps to generate Fig. 5d in the manuscript). Its full-scale implementation is not 'very simple' as the Referee suggested, but requires very diligent coding, critical implementation steps (e.g., sunset time in calculating extra terrestrial solar radiation, comparison of cloudiness computed by the PME to the measured cloudiness) in calculations, and thorough analysis.

Using the hybrid XGBoost-NGBoost model in the testing phase of the ML modeling, all the equations listed below (in reference to equation numbers in the FAO report by Allen et al. (1998, cited in the manuscript)) and their calculations in Fig. 1 below were replaced by 'Fig. 9a' in the manuscript (i.e., considerable reductions in computational steps). On the other hand, 'Fig. 9a' in the manuscript was generated using only daily R_s, T_a, RH, P, and u_2 without any other intermediate calculations and net solar radiation calculations. The equations eliminated using the ML model to calculate ET_o involved:

Hourly mean saturation vapor pressure (Eq.11, p.36, main document, FAO56, Allen et al., 1998)
Hourly mean actual vapor pressure (Eq.54, p.74, main document, FAO56, Allen et al., 1998)
Slope of the saturation pressure curve (Eq.13, p.37, main document, FAO56, Allen et al., 1998)
Psychrometric constant (Eq.8, p.32, main document, FAO56, Allen et al., 1998)
Longwave outgoing radiation (Eq.39, p.52, Box 9, FAO56, Allen et al., 1998)
Solar time angle midpoint of hourly period (Eq.31, p.48, Box 9, FAO56, Allen et al., 1998)
Seasonal correction for solar time (Eqs.32-33, p.48, Box 9, FAO56, Allen et al., 1998)
Inverse relative distance Earth-Sun (Eq. 23, p.46, Box 9, FAO56, Allen et al., 1998)
Solar declination (Eq. 24, p.46, Box 9, FAO56, Allen et al., 1998)
Sunset hour angle (Eq. 25, p.46, Box 9, FAO56, Allen et al., 1998)
Extra-terrestrial radiation for hourly period (Eq. 28, p.47, main document, FAO56, Allen et al., 1998)
Relative solar radiation (cloudiness) (Eq. 3-15, p.226, main document, FAO56, Allen et al., 1998)
Net radiation (Eq. 40, p.53 main document, FAO56, Allen et al., 1998)

In brief, the hybrid XGBoost-NGBoost model effectively and significantly reduced the computational steps in ET_o while eliminating computationally-involved net solar radiation calculations, including also extra-terrestrial radiation and longwave radiation calculations.

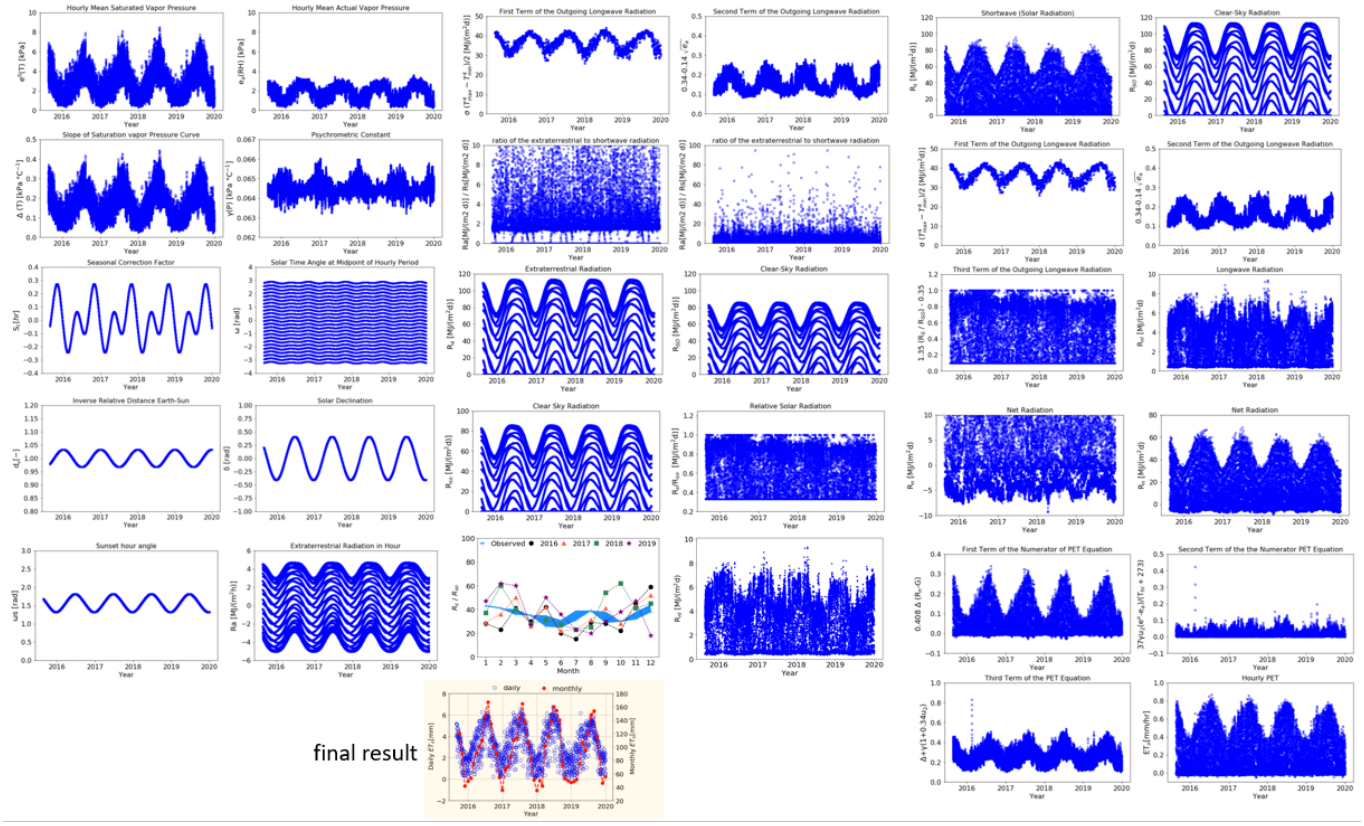

Figure 1: The main steps in ET_o calculations using the PME.

Comment 3. Ea and Esw might have some value, but there lacked comparisons with other methods.

Response:

It is unclear what the Referee meant by 'other methods' in the comment above, as no specifics have been provided for 'other methods'.

However, daily and monthly E_{sw} 'measurements'(upscaled from pan evaporation data) were already compared against the computed E_{sw} via Meyer's formula ('another method', if we adopt the Referee's terminology), using Eqs. 6 and 7 in the manuscript in Fig. 6 of the manuscript, and against ET_o (using Eq. 1) in Fig. 7 of the manuscript. As already mentioned in line #149 of the manuscript, Eq. 6 is the best form of Meyer Formula to predict daily E_{sw} from free water surface – the associated references were already provided just before Eq. 6 in the original manuscript.

It is also unclear from the comment above why we need another model to predict E_a, as the daily E_a (and hence, monthly aggregated E_a) in Fig. 8 are the 'measured data' acquired from the EC tower. We used these data to test the predictive performance of our new hybrid XGBoost-NGBoost model.

As for the use of probabilistic ML model for E_{sw} and ET_a prediction, as discussed in the Abstract, in the main body of the manuscript, and in the Conclusion section, the use of a hybrid XGBoost-NGBoost model would eliminate high capital and maintenance costs of the EC towers (as mentioned in line #276 of the manuscript, the capital cost for the EC tower was $40,000 and required frequent maintenance) and overcome uncertainties associated with pan evaporation and pan coefficient in E_{sw}, which are critical issues in practice.

Comment 4. Why can't you estimate a relationship between E0 and Ea? Instead of XGBoost, why can't we use linear regression or autoregressive function?

Response: Fig.11c of the manuscript reveals that the statistical correlation between the daily ET_o and daily ET_a is not high ($R^2 = 0.74$). The underlying reasoning was already elaborated in line #369−391 (on page 19) of the original manuscript. Moreover, the relation between ET_o and ET_a is clearly nonlinear in Fig. 8. Therefore, the linear regression method that the Referee suggested is not expected to perform well here. To demonstrate this, we set another ML model based on Linear Regression model (labeled as 'Baseline') and compared its performance against the hybrid NGBoost-XGBoost model (labeled as 'Hybrid') in predicting ET_o, E_{sw}, and ET_a in Table 1.

As expected, Table 1 below shows that the hybrid NGBoost-XGBoost model clearly and consistently outperformed the baseline linear regression model in predicting ET_o, E_{sw}, and ET_a in terms of the 'statistical metrics' for point predictions (e.g., in terms of RMSE, MSE, R^2). But, more importantly, the hybrid NGBoost-XGBoost model provided uncertainty estimates through 'probabilistic predictions' (unlike the Linear Regression model and/or other previously developed ML models discussed in line # 45−57 (on page 2) of the original manuscript), which is imperative to practically deploy such models with confidence.

As for the Referee's suggestion on autoregressive models, autoregressive models are beyond the scope of this work, because (i) the prediction of ET_o, E_{sw}, and ET_a is a 'multivariate problem' (please see Eqs. 1, 6, and 7) and autoregressive models cannot unveil the relative importance and interactions of each features (hydroclimatic variables) on the targeted evapotranspiration measure as in Figs. 10 and 12, which are critical in practice, as evident from a number of studies we reviewed, discussed, and compared against our findings on pages 16−19 of the manuscript; (ii) the presence of lagged variables (t-1, t-2, etc.) in autoregressive models leads to accumulation of errors, which is unsuitable for long-term predictions, and also, the lagged variables take away the focus from the climatic variables, due to their high correlation with the target variable, thereby not allowing us to correctly identify and detect the importance and interactions between the hydroclimatic variables using Shapley values; and (iii) the standalone autoregressive models cannot produce probabilistic predictions, which were made possible with the proposed hybrid model as in Fig. 9, which is imperative for uncertainty analysis, and hence, for the use of such models with confidence in practice.

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.

As per this comment, Table 1 will be included in the revised manuscript to demonstrate that the hybrid

NGBoost-XGBoost model outperformed the traditional Linear Regression models, which was suggested by the Referee, in analyzing nonlinear relations between ET_o and ET_a.

Comment 5. What is it in here that we cannot get elsewhere?

Response:

As elaborated in the original manuscript (e.g., line # 45−57 and 159−165) and in our response to the previous comment, 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 of ET_o, ET_a, and E_{sw} in practice.

For example, unlike the other ML methods listed in line #51 of the original manuscript, 'uncertainties' in predictions are accommodated by the hybrid NGBoost-XGBoost model, which provided the confidence that at least 90% of the predicted ET_o, ET_a, and E_{sw} in our ML-based calculations were within 95% of the prediction interval of the target variables. This conclusion **cannot be obtained** from the previously developed ML model, or from the Linear Regression and Autoregressive models that are suggested by the Referee.

Moreover, typical features (independent variables) importance calculations are equivalent to a sensitivity analysis, measuring *relative* contribution of a specific predictor variable to the target observation, but *without* accommodating the dynamic interaction of that specific predictor variable 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. Using the Shapley values, 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 and new information **cannot be** obtained from the traditional features importance analysis and plots.

Comment 6. The author then introduced Shapley values and claimed it a game-theory-based feature importance ranking. If the purpose was to introduce a new feature ranking scheme, it should be compared to other methods to show its validity? This dilutes the focus and looks like a demonstration of software capabilities.

Response:

6.1 '...and claimed it a game-theory-based '

We did not 'claim' that Shapley values are a game theory based importance ranking. It is a fact. Please refer to the excellent study by Lundberg et al. (2020; Nature Machine Intelligence), which was already cited in the original manuscript.

6.2 'it should be compared to other methods to show its validity '

Again, it is unclear which other methods the Referee is referring to in this comment. Without any specifics in this comment, here we assume that the Referee might be referring to tree-based features importance analysis or sensitivity or correlation based feature importance analysis.

To the best of our knowledge, although the Shapley values have not been implemented in hydrological problems in the literature to date, its importance has already been highlighted in the recent literature. For example, Schmidth et al. (2020, Water Resources Research, 10.1029/2019WR025924) noted that :

'*Therefore, other model-agnostic methods that have been made available in the recent past like **Shapley values** or Local Interpretable Model-agnostic Explanations should be applied additionally to **rule out any bias** that is specific to the respective method*.'

Moreover, as already elaborated in the original manuscript, 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 variable 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.

Unlike the tree-based feature importance methods, the Shapley values method enables:

(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), and

(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.

On the other hand, if the Referee is referring to traditional sensitivity analysis or correlation maps, as already shown in Fig. 11 of the manuscript, such analysis cannot accommodate the dynamic interaction of a specific predictor variable with the other predictor variables in calculating the contribution of a specific predictor variable to the target evaporation measures (ET_o, ET_a, E_{sw}).

In brief, as elaborated in the original manuscript, the Shapley values presented in Fig. 10 captured the underlying physics of the evapotranspiration process reasonably well. Because the tree-based methods, traditional sensitivity, and correlation analysis are incapable of accounting for such dynamic interactions among the predictor variables simultaneously (which is made possible with the cooperative game theory-based Shapley values) in calculating the contribution of a specific predictor variable to the target output, a comparison of the results from the Shapley values to the results from the other traditional methods would be uninformative and inconclusive. For example, neither the sensitivity and correlation analysis nor tree-based feature importance methods is capable of producing the results and capturing the conclusions in Fig. 12, and hence, making such comparison is impossible.

6.3 'This dilutes the focus and looks like a demonstration of software capabilities'

With all due respect, we believe that it does not dilute the focus, as the Shapley values-based feature importance analysis constitutes 'Objective 3' (one one the **main focuses**) of the manuscript, as described in line $43-44$ of the manuscript. Moreover, this analysis is important to (i) enhance the transparency of the ML model by identifying & quantifying dynamic interactions between the predictor (hydroclimatic) variables and the target evapotranspiration measures; and (ii) assess the acceptability of applicability of the simplified evapotranspiration models implemented at basins with scarce data (see the discussion in the last paragraph of page 16 of the manuscript). Thus, the Shapley-value based feature importance analysis is an essential component of the hybrid XGBoost-NGBoost modeling, and has nothing to do with demonstration of software capabilities (we even never used the term 'software' in our manuscript, but we focused on a new ML model/method), as the Referee suggested in the comment above.

Comment 7. The organization and writing of the paper are poor.?

Response:

We separated this part of the comment, as it refers to the overall 'manuscript', different from the critiques on the Intro section in the following sentence of the comment. However, it is unclear to us what parts of the writing in the overall 'manuscript' (other than the critiques on the Intro section below) are poor and why? Similarly, what parts of the structure of the overall manuscript is poor and why? Without specific reasoning, it is not

possible to respond to this comment.

Having said this, we followed the 'guideline for manuscript preparation' when we prepared (and organized) the manuscript. The Introduction section provides the background information, main objectives, motivation for the study, rationale for the chosen methods, and briefly explains critical findings. Next, we provide more detailed information about the Methods and Available Data used in the ML analysis. These sections are followed by the Results and Discussion, and the main findings and contributions are elaborated in the Conclusion section. It is unclear to us why the Referee considered this flow of information and manuscript organization to be poor.

More importantly, because our manuscript was accepted for the peer review after the initial screening by the Editors, we believe the overall structure and the writing style of the manuscript should be acceptable and in compliance with the Journal's guideline. We feel like such comments are probably not applicable at this phase of the review process, but certainly without specific reasoning, they are not constructive at all.

Comment 8. There needs to be a clear positioning of the paper in the context of what has been known. The introduction is supposed to position the work in the context of the literature.

Response:

8.1 'clear positioning of the paper in the context of what has been known '

This has been addressed in the original manuscript, but we agree with the comment that it may still require further improvements. In the first two paragraphs below, we elaborate how this has already been addressed in the manuscript, and in the subsequent paragraphs we explain how to implement further improvements.

We noted in line #45−57 on page 2) that various ML models/methods have been recently developed and used for ET prediction **without accounting for inherit uncertainties**. This issue was brought up by Tang et al (2018), as we stated in the Introduction section, and we confronted this challenge in our manuscript for the first time in predicting ETo, ET_a, E_{sw} prediction (as described in the Introduction section). This laid the foundation for motivation (iv), as further elaborated in our response 1.1 above, and Objective 2 (line # 2) in the Introduction section of the original manuscript.

Moreover, one of the main objectives is to develop a probabilistic ML model to predict ET_o, ET_a, and E_{sw} simultaneously from standard hydroclimatic data sets. To our best of knowledge, there is 'no' such a probabilistic ML model in the literature to predict all these evapotranspiration measures; therefore, we used the term 'unprecedented', clearly positioning the paper in the context of what has been known in the literature, as the Referee suggested in the comment above. In brief, this thought process has already been embedded in the manuscript. But, we agree with the Referee's comment that it needs further improvements and we elaborate below how we plan to incorporate such improvements.

As part of further improvements, we plan to include the following discussion and additional references in the Introduction section, as we elaborated in our response to Referee 1's comments, which is also applicable to this comment.

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.

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).

8.2 'introduction is supposed to position the work in the context of the literature '

Response:

We plan to include the following statement in the revised Introduction after the extended literature review provided in our response above to further strengthen the position of the work in the context of the literature:

Considering the presence of different models for evapotranspiration processes as discussed above, we raise the following research questions: Can we have a computationally-efficient and unified data-driven machine learning (ML) model to (i) avoid calibration parameters (e.g., pan coefficients) and empirical relations (e.g.,, Meyer formula), (ii) calculate different evapotranspiration measures (ET_a, E_{sw}, and E_a) 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 datasets 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, a stand alone 'motivation' paragraph, as discussed in our Response 1.1 will also be included in the Introduction section to highlight the main contributions of this work beyond what is already available in the existing literature.

Comment 9. For ET, that is a lot of literature. Although the authors cited some ML for ET papers, they were discussed very surficially. Instead the introduction was extremely surficial and entirely omitted all the literature in ET in hydrology. ?

Response: As we discussed in our response above, the additional literature review on the different evapotranspiration measures and their interrelations (as per Referee 1's comment) is planned to be included in the revised manuscript. If the Referee 2 has any additional suggested references relevant to our work, we are open to his/her suggestions and we will address them in the manuscript as appropriate.

However, although some literature on the evapotranspiration measures might be missing in the original

manuscript, we do not agree with the statement in the comment above that 'the authors entirely omitted 'All' the literature in ET in hydrology'. With all due respect, that's incorrect and 'in contradiction' with the Referee's own words (see the underlined first part of the second sentence of the Referee's comment above). In contrast to the Referee's incorrect statement, we had many key references on the evapotranspiration in the Introduction and also in pages 16 through 18 (under the 'Future Importance' section) of the original manuscript.

Please note that this is not a review manuscript on evapotranspiration, yet we will be citing over 80 references in the revised manuscript, including the new references in blue-colored new paragraphs in our Response to Comment #8. Besides, we refereed to McMahon et al. (2013)'s paper in our manuscript for more comprehensive discussion on different evapotranspiration measures. Our manuscript rather aims at introducing and briefly describing different evapotranspiration measures, prediction using a novel probabilistic ML framework and standardized climate data, and assess the accuracy, robustness, and prediction uncertainties using the PME-computed ET_o, upscaled E_{sw} data, and measured ET_a data from the EC tower.

So, the main objectives (as already stated in the Abstract, Introduction, main body, and conclusions sections of the manuscript) is development of a new data-driven, probabilistic **ML method** to predict ET_o, E_{sw}, and ET_a by avoiding computationally involved net solar radiation calculations, overcoming uncertainties in pan evaporation and pan coefficients, and offsetting high capital and maintenance costs in ET_a measured from an EC tower (already described in the Introduction, main body, and Conclusion sections of the manuscript).

Having said this, we are still open to any references (not included in the original manuscript and the blue-marked new sections) the Referee may suggest.

9.1 'Although the authors cited some ML for ET papers, they were discussed very surficially '

We respectfully disagree with this comment. The purpose of citations for ML-based ET prediction (in line #45−57 on page 2) to show that various ML models/methods have been developed and used for ET prediction, **but without accounting for inherit prediction uncertainties**. This issue was brought up recently by Tang et al. (2018), as we stated in the Introduction section, and we confronted this challenge in our manuscript for the first time in predicting ETo, ET_a, and E_{sw} prediction. This laid the foundation for motivation (iv), as further elaborated in our response 1.1 above and Objective 2 in the Introduction section of the original manuscript.

Because this is not a review paper, we needed to extract the **key points** from the references to lay out the main foundation for the main objectives and motivation for this work, and our review of the ML-based ET papers accomplished this; therefore, we do not consider the discussion 'surficial', as the Referee suggested.

Comment 10. Then the authors spent 8 figures presenting stuff that has no modeling components. ?

Response:

'8 figures with no modeling components' in the comment above is an inaccurate statement.

We had 12 figures in the original manuscript, excluding the figures in Appendix). Among them, Fig. 5 (modeled by the PME), Fig. 6 (modeled by MF), Fig. 7 (modeled by the PME), Fig. 8 (modeled by the PME), Fig. 9 (modeled by hybrid XGBoost-NGBoost), Fig. 10 (modeled by hybrid XGBoost-NGBoost and Shapley Values method) and Fig. 12 (modeled by hybrid XGBoost-NGBoost and Shapley Values method).

Therefore, the number of figures with 'no direct modeling component' is 5 out of 12 (40% of the total figures). The remaining figures display the location map, input data for the models, and correlation maps to interpret data, which are still related to the model setups.

Having said this, we did not see any requirement in the manuscript preparation guideline for authors about the required number of figures that need to allocated to the 'modeling component'. We believe that the figures in our manuscript enhances its transparency by introducing the readers to the data that we used for modeling

and the modeling results.

Comment 11. It was very hard to understand what were the inputs and what were the outputs from Methods. This was instead discussed in the Results section.

Response:

Input and output data were described in the original manuscript far before the 'Results' section. For example, as described in line #198−199 of the manuscript (section 2.1 on page 7), T_a, P, RH, u_2, and R_s are the **input** data for the PME (and hence, for the probabilistic ML modeling used for ET_o prediction). As described in line #215, T_a (input data) was replaced by T_{SW} (another input) in E_{sw} calculations. In the original manuscript, the section entitle 'Data Availability' (along with the Appendix) is dedicated for the input data. Moreover, Shapley plots in Fig. 10 list all the input data for the probabilistic ML modeling along the y-axis, as the method quantifies the importance and interaction of each input feature on ET_o, E_{sw}, and E_a.

However, we see the concern raised by the Referee. Therefore, to enhance the clarity about the inputs and outputs, we will replace Fig. 1 in the original manuscript with the new figure in Fig. 2 below.

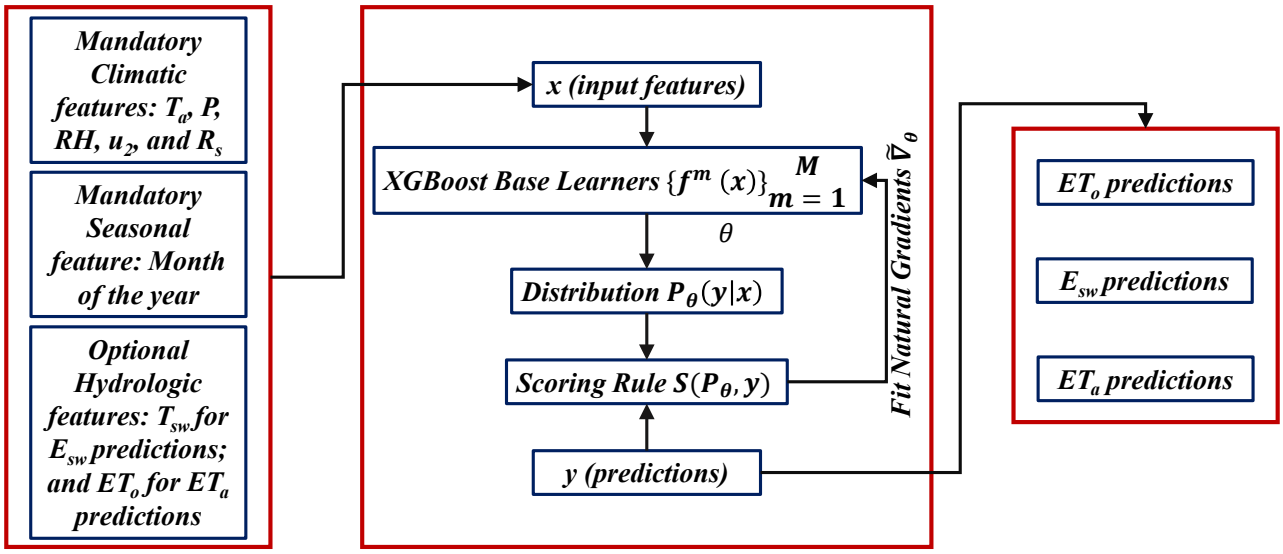

Figure 2: Conceptual representation of the hybrid NGBoost-XGBoost model for ET_o, E_{sw}, and ET_a prediction.

Comment 11. The authors used 'unprecedented' and 'for the first time' very casually. There are things we do and things we don't do and there must be a clear reason why we do them.

Response:

Based on our current literature review, we use the terms 'unprecedented' and 'for the first time' very meticulously (not causally at all) throughout the manuscript. However, if the Referee has any compelling and convincing references and 'reason' (not included in the Referee's comment above) that may invalidate our use of these terms in the sentences below, we kindly ask the Referee to share these references and information with us for our review.

... Unprecedentedly, we demonstrate that a newly developed probabilistic machine learning (ML) model, using a hybridized NGBoost-XGBoost framework, can accurately predict the daily ET_o, E_{sw}, ET_a from local climate data. The probabilistic approach exhibits great potential in overcoming data uncertainties, in which 99% of the ETo, 90% of the E_{sw}, and 91% of the ET_a test data at three watersheds were within the model's

95% prediction interval....

Question to Referee: If the Referee thinks that we use the term 'unprecedented' causally in the sentence above, we would like to know which journal article has already reported development and use of a probabilistic ML model to predict ET_o, E_{sw}, ET_a from local climate data within $\geq 90\%$ within the model's 95% prediction interval?

also

... Finally, we demonstrate, for the first time, a coalition game theory approach to identify the order of importance, dependencies & interactions of climatic variables on the ML-based ET_o, E_{sw}, ET_a predictions....

Question to Referee: If the Referee thinks that we use the term 'for the first time' causally in the sentence above, we would like to know which journal article reports the use of the coalition game theory approach to determine the order of importance, dependencies & interactions of climatic variables on the ML-based ET_o, E_{sw}, and ET_a predictions?