egusphere-2026-2300 & Comments

NeuralFAO56 v1.0: A Scalable Physics-Informed Deep Learning Framework for Real-Time Evapotranspiration Estimation Across CONUS

The manuscript presents an operationally useful, well-designed Python framework of NeuralFAO56 v1.0 that integrates FAO-56 Penman–Monteith reference-ET calculation with LSTM- and Transformer-based ET forecasting, using NWS real-time and forecast meteorology and NCEI historical archives, across a continental network of NWS–NCEI matched stations. The scope is suitable for a GMD Model description. The supporting evidence for the current publication, however, is not sufficient. The manuscript should be rethought after the authors make a response to the following substantive concerns. I would suggest the following comments, and hope that it will be helpful for the authors. I would suggest another resubmission:

1- Line 1: “Real-Time Evapotranspiration Estimation”, is an overstatement of what is demonstrated. The training evaluation is done based on NCEI data in a backward look, with the actual deployment being conducted for just one week. Take into account "Near-Real-Time Reference Evapotranspiration Forecasting". In addition, Physics-Informed is a good term in the DL literature. In the current manuscript the physics only appears as an offline FAO-56 label generator, not as a loss-function constraint; it would be better to call it “Physics-Guided” or “Hybrid Physics/Data-Driven”.
2- Line 10-32: Methods and the figure captions report 845 stations although there are 867 reported by the Abstract. Please check everywhere and explain the QA/QC step that decreases the number.
3- Line 10-32: The sentence “… support broader adoption of physics-informed machine learning in hydrologic prediction” is not supported by the evidence in the manuscript and should be toned down. The term “irrigation planning” also goes too far because ETo is a boundary condition, and scheduling needs crop coefficients (Kc) and/or ETa.
4- Line 52: Please remove “(4/30/2026 12:05:00 PMAllen et al., 1998” and replace with a clean citation.
5- Line 134: Figure 1 caption: replace “utilis” with “utils”. Please verify all module names against the released code.
6- Line 137: replace “componenet” with “component”.
7- Please start the introduction by explicitly stating (1) the maximum number of stations that previous DL-ETo studies have used, (2) that there are no existing packages that combine NWS real-time data for ETo with historical data from NCEI, and (3) that no previous continental-scale evaluation has been performed for all nine NOAA climate regions. NeuralFAO56 should be compared respect to Valipour et al. (2023), Ahmadi et al. (2023), Medina et al. (2018), Thorp (2022), and Vremec et al. (2024) that would help sharpen the contribution provided in table 1.
8- There is no independent evidence of the validity of ETo. The forecasts from LSTM/Transformer and NWS are compared to the FAO-56 ETo based on observed meteorology. The reference suffers itself from a systematic, regionally variable error, as is its solar radiation (Rs) which is derived from Hargreaves–Samani with a globally fixed kRS = 0.16, and no independent measurement of ETo has been used to validate either the reference

or the DL forecasts. The manuscript should be benchmarked with the AmeriFlux/FLUXNET/SURFRAD/USCRN eddy-covariance ETo/ETref, and an independent ET product from remote sensing in a subset of stations co-located with the manuscript to support the claim of "real-time ETo forecasting".

9- The DL forecasts are only compared to the physics-based FAO-56 pipeline. There are no naive baselines reported. Each defensible GMD-level evaluation should contain, at each lead time and for each NOAA region. These baselines should be beaten by a statistically significant margin for any DL claim before using the additional complexity of LSTM/Transformer.

10- Only deterministic skill of point forecast (NSE, KGE, MAE) is reported in the manuscript. Prediction intervals are needed for a GMD Model description, for an operational forecasting system. Two lightweight approaches are Monte-Carlo dropout during inference and multiple models using random seeds (deep ensembles of 5 models). The interval coverage should be provided in terms of Prediction interval coverage probability and Mean interval width at nominal 90% coverage.

11- In module FAO-56, Section 2, please specify if NeuralFAO56 imports the package pyfao56 (Thorp, 2022), imports and wraps the package, or re-implements FAO-56 equations. When depending on pyfao56 specify the exact version. In one paragraph, discuss the relation with PyEt (Vremec et al., 2024).

12- Repeat each experiment, reported in Section 2, at least 10 times across ≥ 10 random seeds, with the exact hyperparameter search space reported; report the hyperparameter sweep if the reported configuration is a default, or if the reported configuration was the result of tuning. Report averages ± SD in all tables and all figures.

13- For gaps greater than about 2 days, it is not defensible to use Section 2, preprocessing: forward-fill and backward-fill on daily meteorological time series. Use linear interpolation for small gaps (≤ 3 days) and a Kalman smoother or regression imputation for larger gaps or remove larger gaps. Justify the 5% threshold for missingness; Zhang & Post (2018) recommend that gap-filled series are still useful for hydrometeorological data at 10-20% missingness.

14- Section 2, train and test protocol: the 60/40 split by year is acceptable in principle, but the manuscript should confirm: 1- that only the train years are used for the fit of the Min–Max scaling parameters, 2- that there is no 7-day window of years extending across the split point, and 3- that the model selection uses a validation subset of years, not the test years.

15- Please provide an extra figure of station level NSE against lag-1 autocorrelation of daily ETo in Section 2.4.2 in Fig. 6. This will provide a measure of the informal saying that skill is primarily constrained by data characteristics.

16- The reported architectures are at the bottom of the possible design space. This observed equivalence may be due to either true data-limited saturation or under-parameterization. A short scaling study needs to be conducted consisting of 32, 64, 128 and 256 widths and 1, 2, 4 depths, evaluated on at least one representative station for each NOAA region, and reported best-of-search NSE, with 5-fold or block cross-validation. The only way the equivalence claim can be supported or if it is not, the paper becomes stronger as a result.

17- The same value of kRS is used for all stations and all nine NOAA climate regions ,0.16. This is contrary to FAO-56 recommendation which suggests that interior and coastal sites should be separated (≈0.16 vs. ≈0.19). It is also in conflict with the calibration literature. The authors should either follow the regional lookup or utilize observed Rs at co-located

SURFRAD/USCRN sites, if available, or include a sensitivity study for kRS ± 25% and show that the conclusions in the "results" section are valid.

18- Three studies are cited and not engaged with which are closely related. 1- The current Python implementation of FAO-56, pyfao56 (Thorp, 2022, SoftwareX 19, 101208), is the 'legacy' Python implementation of FAO-56 and should be mentioned as an explicit dependency or comparison. 2- For a GMD reader, the primary reference is the current GMD paper by Vremec et al. (2024) which uses a Python PET package. 3- The improvement that NeuralFAO56 can offer over the work of Valipour et al. (2023) should be measured, preferably in a Table 1, for example, by calculating the RMSE for the 1-day forecasts for all the 30 CONUS sites with hybrid ML/DL at the 10-day horizon in order to quantify this contribution.

19- Look-back sensitivity is not mentioned in the text in section 2.4.2, but there is no table or figure provided in this regard. Include a sensitivity table (rows = look-back; columns = NSE/KGE/MAE by NOAA region), or be clear that only 7 days were tested and provide justification.

20- In addition to the training cost, report also inference-time cost per station per day, as specified in Section 2.4.6 in Table 3.

21- As of yet, the current “Code and Data Availability” section does not satisfy Copernicus GMD policy requirements. All requirements must be completed prior to publication.

22- The near-real-time evaluation of the forecast allows a window of 7 days. This is not enough to cover the overall “operational readiness” terminology in the Abstract and Conclusion. The evaluation must be made at least once every four seasons and provide MAE distributions per window for all 845 stations. These expanded results may be presented in the Abstract and Conclusion.

23- In the conclusion section, the sentence “reinforces its reliability, scalability, and operational readiness of environmental prediction systems” is not backed up without independent validation, uncertainty quantification and multi-window real-time evaluation. Revise it and provide a supported claim.

24- While making Code and Data available, split them into two DOIs, attach a tagged release from GitHub, provide a license, specify the environment and provide a user manual.

25- I recommend removing section 2.4. Results and Discussion should be separated into two different sections. This would enhance readability of the manuscript and help to differentiate the findings of the study from their interpretation.

26- Use consistent units throughout (mm day-1 not “mm day-1”).

27- In the reference list, Vremec: the correct one is https://doi:10.5194/gmd-17-7083-2024.