Dear Reviewer,

We sincerely thank you for reviewing our manuscript. We appreciate your positive comments regarding the clarity and significance of our work, as well as the detailed and constructive feedback you provided. We have carefully considered each point and revised the manuscript accordingly. We believe these revisions can strengthen the paper's reproducibility and theoretical depth. Below, please find our point-by-point response to your comments.

1. Reviewer comment 1: "A) Why 'The initial 12 hours of each simulation run were discarded'. Add more context here on why only 12 and what happens when you discard this initial forecast?"

Response: We appreciate the reviewer's request for clarification on this critical detail. We have implemented the following revisions in the manuscript to clarify this question:

In the WRF model, the atmospheric state at the initial time is typically imbalanced. The 'spin-up' period is employed to allow the model's internal physical variables (such as soil moisture and cloud microphysics) to reach thermodynamic equilibrium. If this period is not discarded, the pronounced fluctuations present at the beginning may contaminate the subsequent weather prediction results (Mallard et al., 2026). We have added the revised text to **Page 5, Lines 113-116**.

2. Reviewer comment 2: B) Clarity is required on what is meant by - itransformer, physics-itransformer, PC-itransformer, PINN-itransformer. These terminologies are used further in the results section. What I find is that the writeup for these models in data and section seems to look OK, but it is confusing due to cross referencing with previous published work. I suggest that more details and fundamental information be added in the 4 models to make it accessible for readers.

Response: We appreciate the comment regarding the clarity of our model terminology. To address this, we have expanded Section 2 to provide a more self-contained explanation of each model variant. We have also clarified the hierarchy of these models to show how each adds a specific layer of physical constraint or architectural refinement. To further improve accessibility, we have added a Summary Table (Table 1) in Section 2 that compares these four models. This allows readers to quickly identify the novelty of the PINNs-iTransformer relative to the other variants. All changes are highlighted in blue in the revised manuscript (**Page 10, Lines 241-264**) for the reviewer's convenience.

3. Reviewer comment 3: "C) GPOA is used here, but defined only in the results section. I suggest rewriting the description of models from scratch, do not assume that the readers would have read the referenced work by Fan, Sun et al. etc. before. I would suggest that the authors follow this pattern: § Describe the model in plain English by adding what the model can do and give an application example § Describe the core functioning of the model, its equations and layer description § Describe in detail how this model will be used in the subsequent steps/validation studies. eg. $dP(t)/dt = -k\cdot(P(t)-Peq(t))$ this equation just comes from nowhere and we do not get a context on why this is important. The only pointer is that some components eg.

equilibrium is defined from Fan et al. paper."

Response: We are grateful for the comments on improving the clarity and organization of our methodological description. In line with the reviewer's suggested presentation, we have revised the descriptions of the $\boldsymbol{G_{POA}}$ calculation and the PINNs-iTransformer framework as follows:

1.The definition and physical significance of $\boldsymbol{G_{POA}}$ are now introduced at its first occurrence in Section 2.4, ensuring readers understand its role as the primary driver for PV power before reaching the results.

2. In Section 2.5.2, we have restructured the description of the PINNs component to improve readability and physical interpretability:

(1). We state explicitly, in plain language, that the PINNs term serves as a mathematical regulator that discourages physically inconsistent predictions.

(2). We introduce and define the equilibrium power ($\boldsymbol{P_{eq}}$) and the learnable relaxation coefficient ($\boldsymbol{k}$), and clarify that $\boldsymbol{P_{eq}}$ denotes the theoretical steady-state power implied by the instantaneous meteorological conditions.

(3). We describe the discretization procedure (forward difference) and explain how the physics-informed loss ($\boldsymbol{\mathcal{L}_{pinn}}$) is incorporated into the total loss to guide the optimization.

(4). We add an explicit rationale for adopting the relaxation-ODE formulation: it provides a restoration mechanism that drives the predictions back toward physical consistency when they deviate from the theoretical equilibrium.

All changes are highlighted in blue in the revised manuscript (**Pages 6-7, Lines 143-180** and **Page 9, Lines 214-240**) for the reviewer's convenience. We believe these revisions have substantially improved the clarity of our description.

4. Reviewer comment 4: A) GHI correction: the 400/100/400 split is counterintuitive, Here you are loosing nearly 50% of your data points. I would suggest using stratified sampling and doing 10 fold cross validation technique to use all data for training and prediction. e.g. in the 10 fold cross validation 90% of the samples are used for training and 10% for validation, this is then used in conjunction with rolling window approach to use all data for training. If this is unsuitable then do 700/100/100 split.

Response: We fully agree with the reviewer's assertion that the data should be used in a reasonable and comprehensive manner. We recognize that our original description was insufficiently clear and may have led to a misunderstanding. We would like to clarify the methodology's actual meaning and provide our rationale below.

1. Clarification of Data Usage and the Rolling Prediction Mechanism

The "400/100/400" split denotes three consecutive temporal phases within a continuous time series, rather than a random partition of the dataset. In fact, no data are discarded throughout the evaluation process. Specifically, the training phase (Steps 1–400) is used to initially optimize model parameters. The validation phase (Steps 401–500) is used for hyperparameter tuning and for implementing early stopping strategies to prevent overfitting. Finally, the test phase (Steps 501–900) is reserved for an independent and unbiased assessment of the model's predictive performance.

Crucially, we employ a non-overlapping rolling-window mechanism during the test phase. For a forecast horizon ***H***, the model uses the preceding historical window to predict the next ***H*** steps, then advances to the next time step. This process iterates until the entire test period (400 steps) is covered. This ensures that 100% of the test data is used for evaluation, effectively addressing concerns about data loss.

2. Rationale for Adopting 'Walk-Forward Validation' over Cross-Validation

We agree that k-fold cross-validation is robust for static datasets. However, it is generally unsuitable for operational time-series forecasting because it violates temporal causality.

(1). Prevention of Look-Ahead Bias: Standard or stratified cross-validation involves random shuffling, which would allow the model to train on future data to predict past events (data leakage). This yields overly optimistic error metrics that do not reflect real-world performance.

(2). Simulation of Operational Conditions: Our study aims to simulate the real-world workflow of the WRF operational forecasting system. In this context, predictions must be made strictly sequentially, using only information that is historically available.

(3). Adherence to Best Practices: Our approach aligns with the 'Walk-Forward Validation' (or rolling-origin) method, which is the standard for time-series evaluation (Bergmeir and Benítez, 2012; Tashman, 2000). This ensures our results represent a realistic assessment of the model's operational capability.

All changes are highlighted in blue in the revised manuscript (**Pages 12-13, Lines 300-319**) for the reviewer's convenience. We believe these revisions have substantially improved the clarity and rigor of our methodology description.

Reviewer comment 5: "B) Why use only RMSE/MAE as the metric for evaluation, why not other error metics. What I mean is that define the rational for RMSE/MAE."

Response: We thank the reviewer for this pertinent question regarding our selection of evaluation metrics. RMSE and MAE are widely used in forecasting studies and offer complementary perspectives: MAE provides an easily interpretable measure of the average error magnitude, treating all deviations equally and being robust to outliers; RMSE gives greater weight to larger errors, which is critical in solar PV forecasting where substantial deviations can impact grid stability and operational planning. Together, RMSE and MAE offer a balanced assessment of forecast performance, which is essential for our application. According to the reviewer's suggestion, we have added a subsection that clarifies the rationale for our choice of RMSE and MAE as primary evaluation metrics in the revised manuscript.

Revised Text (Pages 10-11, Lines 265-271): To quantify the predictive accuracy of the models, this study employs two widely recognized error metrics: the RMSE and the MAE. RMSE is particularly sensitive to large errors, effectively penalizing significant deviations or peak-prediction inaccuracies. In contrast, MAE provides a more straightforward measure of the average magnitude of prediction errors, reflecting the overall level of predictive performance (Jannah et al., 2024). The prevalence of these metrics in the field is underscored by a comprehensive review (Al-Dahidi et al., 2024; Pandžić and Capuder, 2024), which indicates that over half of the surveyed literature utilizes both RMSE and MAE for model evaluation.

6. Reviewer comment 6: "C) Why is this logical? "the inherent challenge for neural networks to model long-range temporal dependencies". what are the inherent challenges? why not use RNN/LSTM for long term forecast? How long is the long term forecast? I am assuming here that the forecast cycle can have seasonal and repetitive behaviour after 365 days?"

Response: We sincerely thank the reviewer for raising this question regarding the challenges of long-range temporal dependency modeling. We clarify the specific definitions and constraints within the context of our study.

1. Clarification of 'Inherent Challenges'

The term 'inherent challenges' refers to well-documented limitations of recurrent neural networks in capturing long-range dependencies: (1). Gradient propagation constraints: Although LSTM's gating mechanisms alleviate the vanishing gradient problem, gradient signals still attenuate over extended sequences during backpropagation through time (Hochreiter and Schmidhuber, 1997). In our study, the 75-step forecast horizon requires 225 steps of historical input, which places substantial demands on the network's memory capacity. (2). Information bottleneck: The fixed-dimensional hidden state must compress all historical information, inevitably leading to information loss as sequence length increases. (3). Attention dilution: In our TPA-LSTM model, attention weights become increasingly dispersed over longer sequences, reducing the model's ability to focus on critical historical patterns.

2. Response to ' How long is the long term forecast? Why Not Use RNN/LSTM for Long-Term Forecasting?'

Following the widely-accepted taxonomy from (Zhou et al., 2020), long-term forecasting is defined as prediction horizons of 48 steps or longer for hourly data. Our maximum forecast horizon of 75 hours clearly satisfies this criterion. We acknowledge that this extended horizon poses significant challenges for solar irradiance prediction, primarily due to cumulative atmospheric uncertainties and the progressive propagation of model errors over time.

We would like to clarify that our model is based on LSTM architecture (TPA-LSTM). The reviewer's question addresses an important point: even LSTM, which was specifically designed for long-term dependency modeling, exhibits performance degradation with extended sequences. This is a fundamental characteristic of sequence-to-sequence modeling, not a limitation unique to our implementation.

3. Discussion on Seasonal and Cyclical Behavior

We acknowledge that solar irradiance has a strong annual periodicity. However, our study focuses on bias correction in the WRF rather than stand-alone long-range climate forecasting. Due to the limited dataset duration (60 days), our model is trained to capture synoptic-scale variability and diurnal patterns rather than full seasonal cycles.

Revised Text (Page 14, Lines 333-340): The 6-step forecast horizon (6-hour lead time) yielded optimal correction performance, with RMSE and MAE reductions of 33% and 36%, respectively. However, correction performance declined as the forecast horizon extended to 75 steps (5-day lead time). This degradation is attributable to: (1) the inherent difficulty of LSTM networks in maintaining effective memory over extended sequences (Hochreiter and Schmidhuber, 1997), as the 225-step historical input required for 75-step predictions approaches practical memory limits; (2) the chaotic nature of atmospheric dynamics, which causes forecast errors to accumulate nonlinearly with increasing lead time; and (3) the limited temporal coverage of the current dataset (60 days), which constrains the model's capacity to learn high-frequency meteorological fluctuations. Future work will address these limitations by expanding dataset coverage and incorporating explicit periodic encoding mechanisms.

7. Reviewer comment 7: "D) Please define in detail what the residuals are in different models. I assume that in the physics-itransformer, the physics model is predicting the PV output and you compare that with actual observations to generate the residuals?"

Response: The reviewer's statement is correct. The Physics-iTransformer utilizes a physical model to generate an initial theoretical PV power estimate, with the residual defined as the difference between the observed ground truth and this physical baseline.

We have also introduced a new subsection (Section 2.5.3) to rigorously define all baseline models and their integration strategies. The detailed revised text for this section is presented in our response to Comment 2. Please check the revised text to **Page 10, Lines 241-265**.

8. Reviewer comment 8: "3) For figure number 15 comparisons, I think it would be beneficial if the authors can provide the ranking of the different comparative models to understand how accurate the PINN-itransformer is from the current state of art."

Response: We sincerely thank the reviewer for this valuable comment. Accordingly, we have revised Figures 13 and 14 (formerly Figure 15) to include new subplots at the bottom. These subplots show the rankings of all models based on MAE and RMSE. This quantitative ranking clearly demonstrates the performance advantage of PINNs-iTransformer over baseline models. You can check the revised figure below.

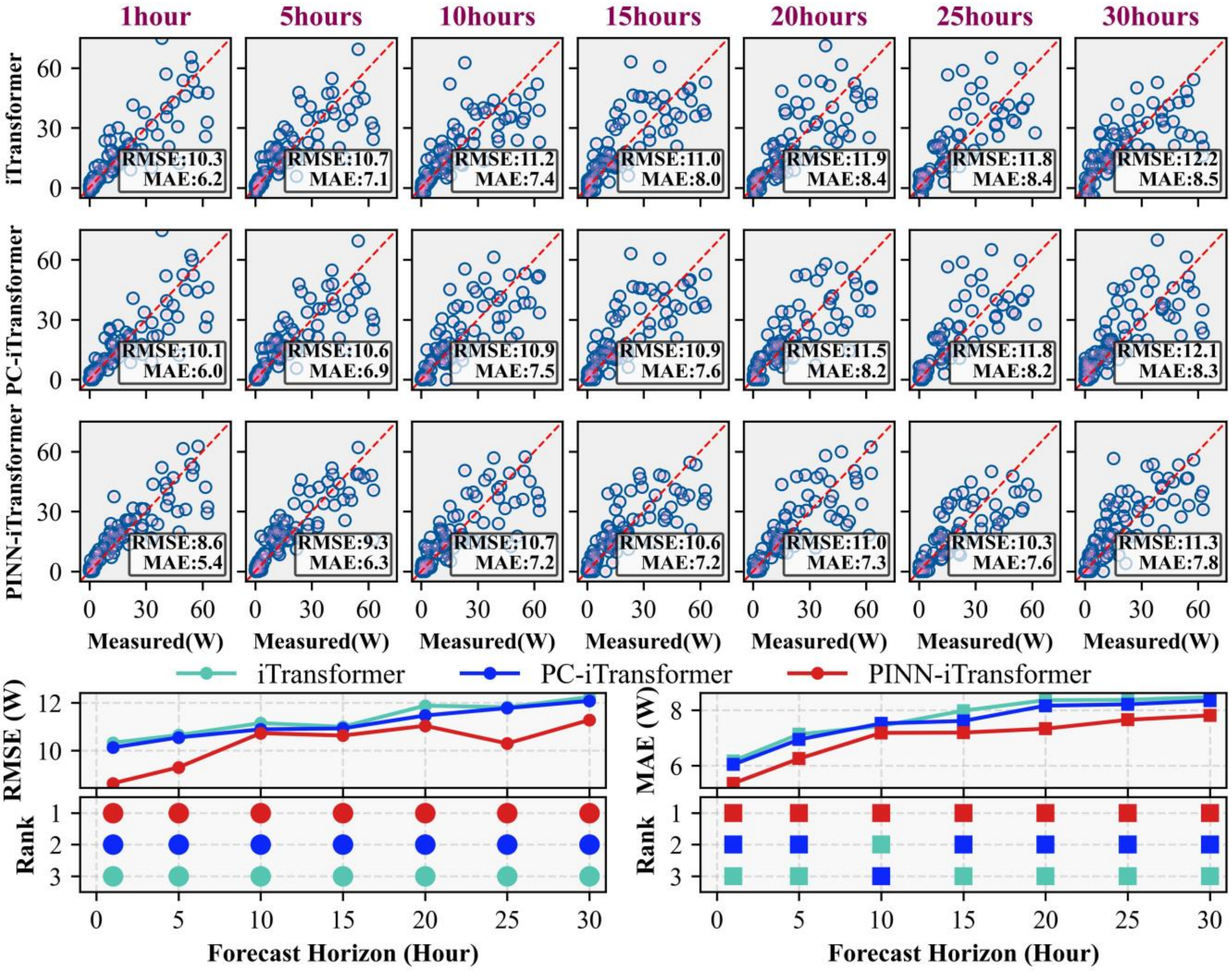

Figure 13: Multi-step PV power forecasting performance using a 5-day WRF forecast.

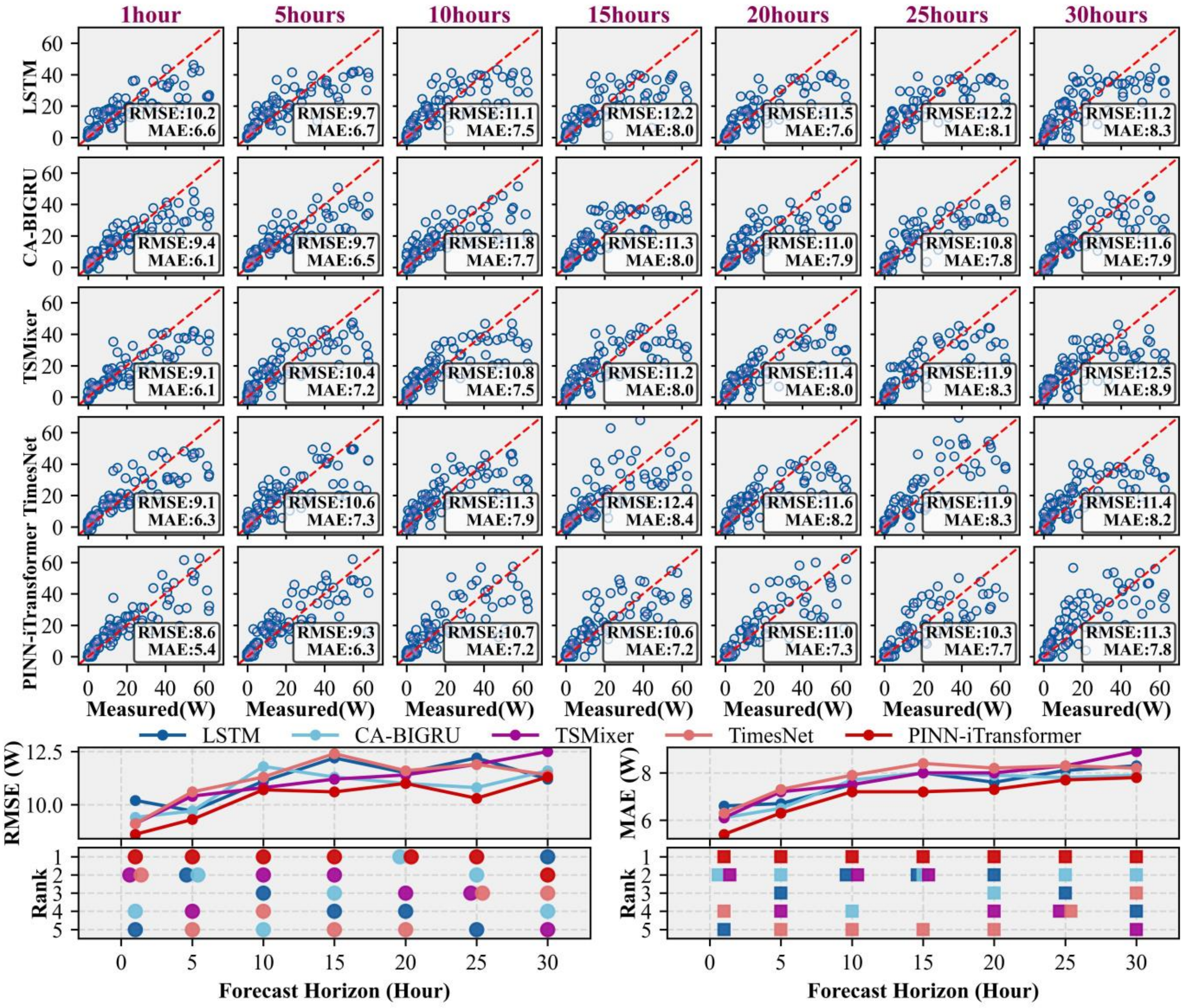

Figure 14: Performance comparison of different models for multi-step forecasting.

9.Reviewer comment 9: 4) Code-> I went through the Zenodo repository, here my recommendation would be to revoke hardcoded file paths from the python files. Additionally, please add some user manual here e.g. steps/order in which the files would be run.

Response: We thank the reviewer for their careful examination of our code repository and for their constructive suggestions to improve reproducibility. We have addressed both points as follows: (1) Hardcoded file paths: All hardcoded absolute paths have been removed from the Python scripts. The code now employs relative paths and configurable variables, ensuring portability across different computing environments. (2) User documentation: A comprehensive README.pdf has been added to the repository, which includes: (a) a description of the dataset structure, (b) step-by-step instructions for script execution order, and (c) configuration details for the WRF model setup. The revised code and documentation are available at the updated Zenodo repository: New DOI: https://doi.org/10.5281/zenodo.17850993

10. Reviewer comment 10: 5) Limitations -> Add limitations of the model eg. this model is only training on the historical data or very short term real data. Would this model break if we do the forecasting for 1 year/5 year/20 years? What is the result of validation with SolarGIS data at monthly level? What will happen when we forecast for regions that are in gobi desert etc. where there is not much cloud based variability. Have you tested for 2025 time series as the model was trained for 2020 timer series. Add additional limitations that the authors seems necessary.

Response: We sincerely thank the reviewer for these insightful questions regarding the model's robustness and generalizability. We have addressed these points by adding a dedicated '**Key factors and uncertainties in forecasting**' section (**Section 5, Pages 20-22, Lines 421-468**).

Regarding the specific concerns:

1.Temporal Robustness (1/5/20 years): Our model is designed for short-to-medium-term forecasting (up to 5 days) to support grid dispatch. For decadal forecasting (1-20 years), the model would indeed face challenges due to climate drift and sensor degradation. We have clarified that the model's current scope is operational forecasting rather than long-term climate projection.

2.Geographical and Climatic Generalizability: The study site is located in a semi-arid region, which shares characteristics with Gobi-like environments (e.g., high solar resource, low cloud frequency). To address the 'breakdown' concern, we conducted a SHAP analysis. The results (Figure 15) demonstrate that the model strategically shifts its reliance from temporal persistence (dominant at 1h) to physical meteorological signals (increasingly important at 5 days). This transition suggests that the model captures intrinsic physical couplings rather than just 'memorizing' historical patterns, supporting its potential for cross-regional application.

3.Data Constraints: We acknowledge the limitations of using a single-site dataset. Due to data privacy and the operational status of the sensors, 2025 time-series and SolarGIS monthly validation were not available during this study. However, the high stability of the SHAP values (Bootstrap R = 0.9992) confirms that the learned logic is statistically robust.

4.Although observational photovoltaic power data were unavailable due to experimental constraints, we validated our approach using NASA POWER data to correct solar irradiance forecasts at multiple stations for 2025. The proposed model can slso significantly reduce both RMSE and MAE for GHI predictions (Figure S3-S9). We believe these additions directly address the reviewer's concerns regarding robustness, generalizability, and model limitations.

We hope that the revisions and explanations provided adequately address the reviewer's concerns. We believe that these improvements have significantly strengthened the manuscript, making it suitable for publication.

Thank you once again for your time and effort in reviewing our work.

Sincerely,

The Authors

Reference

Al-Dahidi, S., Madhiarasan, M., Al-Ghussain, L., Abubaker, A. M., Ahmad, A. D., Alrbai, M., Aghaei, M., Alahmer, H., Alahmer, A., Baraldi, P., and Zio, E.: Forecasting Solar Photovoltaic Power Production: A Comprehensive Review and Innovative Data-Driven Modeling Framework, 10.3390/en17164145, 2024.

Anderson, K., Hansen, C., Holmgren, W., Jensen, A., Mikofski, M., and Driesse, A.: pvlib python: 2023 project update, Journal of Open Source Software, 8, 5994, 10.21105/joss.05994, 2023.

Bergmeir, C. and Benítez, J. M.: On the use of cross-validation for time series predictor evaluation, Information Sciences, 191, 192-213, https://doi.org/10.1016/j.ins.2011.12.028, 2012.

Erbs, D. G., Klein, S. A., and Duffie, J. A.: Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation, Solar Energy, 28, 293-302, https://doi.org/10.1016/0038-092X(82)90302-4, 1982.

Hay, J. E.: Calculation of monthly mean solar radiation for horizontal and inclined surfaces, Solar Energy, 23, 301-307, https://doi.org/10.1016/0038-092X(79)90123-3, 1979.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735-1780, 10.1162/neco.1997.9.8.1735, 1997.

Jannah, N., Gunawan, T. S., Yusoff, S. H., Hanifah, M. S. A., and Sapihie, S. N. M.: Recent Advances and Future Challenges of Solar Power Generation Forecasting, IEEE Access, 12, 168904-168924, 10.1109/ACCESS.2024.3496120, 2024.

Mahmoudi, E., de Souza Silva, J. L., and dos Santos Barros, T. A.: Assessing the performance of physical transposition models in photovoltaic power forecasting: A comprehensive micro and macro accuracy analysis, Energy Conversion and Management: X, 24, 100792, https://doi.org/10.1016/j.ecmx.2024.100792, 2024.

Mallard, M. S., Spero, T. L., Bowden, J. H., Willison, J., Nolte, C. G., and Jalowska, A. M.: Examining spin-up behaviour within WRF dynamical downscaling applications, Geosci. Model Dev., 19, 579-594, 10.5194/gmd-19-579-2026, 2026.

Pandžić, F. and Capuder, T.: Advances in Short-Term Solar Forecasting: A Review and Benchmark of Machine Learning Methods and Relevant Data Sources, 10.3390/en17010097, 2024.

Tashman, L. J. J. I. J. o. F.: Out-of-sample tests of forecasting accuracy: an analysis and review, 2000.

Zhou, H., Zhang, S., Peng, J., Zhang, S., and Zhang, W.: Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting, 2020.