

Enhancing Solar Radiation Forecasting with Machine Learning

Wee Fong Lee¹, Peter Ling¹, and Aaron Wilson^{2,3}

KEYWORDS. artificial intelligence, controlled environment agriculture, greenhouse, high tunnel, passive climate control, site-specific weather forecasting

ABSTRACT. This project aimed to enhance local weather forecasts by improving 1-hour, on-site predictions using the High-Resolution Rapid Refresh (HRRR) dataset. These forecasts can support the high tunnel weather forecast model, providing growers with critical insights to respond to high-temperature events. The project's objectives included developing a streamlined data preparation process and an on-site predictive model using machine learning (ML). After considering all potential weather variables, our analysis focused on solar radiation intensities exceeding 400 W/m² during the Northern Hemisphere's transition periods (March and October). The study used HRRR and observational data from three locations, including Wooster, OH, USA; West Lafayette, IN, USA; and Geneva, NY, USA for model training. Data preprocessing, including parsing, time synchronization, format unification, and missing data handling, was managed using Python. The complex meteorological HRRR data, originally in GRIB2 format, was transformed into a more accessible CSV format with selected variables and a significantly reduced file size, making it more usable for high tunnel producers. For the ML model, one neural network architecture effectively served all three locations, suggesting the potential for a generalized model that can be applied across sites at similar latitudes. Among the five input-feature designs, the HRRR forecast variables for the current time and next hour performed the best across all locations. The ML model outperformed HRRR, reducing root mean square error (RMSE) from 114 to 64 W/m² and mean error from 34 to 4 W/m² while improving R² from 0.47 to 0.67 for Wooster, OH. Similar performance gains were observed at the other locations. These findings support broader agricultural applications, including high tunnels, greenhouses, and outdoor farming.

High tunnels, as a low-tech form of controlled environment agriculture (CEA), have gained popularity among growers seeking cost-effective growing structures. These simple structures trap heat during the day, extending the growing season by 3 to 4 weeks in colder regions compared with in-field growing (Drost 2011). Unlike advanced greenhouses with active

climate control systems, high tunnels rely on passive methods, such as heat retention and roll-up sidewalls for temperature regulation, making them vulnerable to rapid heat surges that can stress plants within minutes on sunny days (Wahid et al. 2007). Given their dependence on external environmental conditions, short-term weather forecasts (1 or 2 h ahead) are crucial for providing growers with the necessary time to respond effectively.

High tunnel weather can be predicted using a local weather station, as site-specific weather data enable more precise agricultural management decisions, potentially enhancing crop yields and resource efficiency (Ahmad and Dar 2020). However, not all sites have a nearby weather station, making on-line weather forecasts the next best alternative. These forecasts rely on numerical weather prediction (NWP) models, such as the HRRR model (Dowell et al. 2022) developed by the National Oceanic and Atmospheric Administration. HRRR provides high spatial (3 km) and temporal (1 h)

resolution for key weather variables but lacks the site-specific accuracy, interpretability, and accessibility required by growers. These models are not designed to capture hyperlocal weather variability, affected by topography, land surface characteristics, and other site-specific factors. Hence, its predictions may fail to accurately reflect actual on-site conditions.

This project sought to improve HRRR forecasts of key variables impacting high tunnel conditions, including air temperature, relative humidity (RH), wind speed, and solar radiation. The focus was on enhancing the accuracy of the variable with the poorest forecast performance using ML. Because hyperlocal forecasts require site-specific weather data and HRRR data for model development, the study also focused on simplifying HRRR's complex dataset for ML compatibility, ensuring efficient training on common computing resources. A reliable data preparation framework was crucial to ensure high-quality inputs for model training. Leveraging HRRR as a foundation, this research aimed to refine weather forecasts at the local level, equipping growers with actionable insights for better climate control during extreme heat events. The project's objectives were to 1) develop a streamlined data preparation process for common weather prediction software tools and 2) develop an on-site predictive model for solar radiation using ML.

Background

Weather forecasting has traditionally relied on methods such as observations, NWP, and ensemble forecasts, which require substantial computational power and large computer systems (Bhawsar et al. 2021). However, with the rise of ML, forecasting has become more accessible, allowing predictions to be made on household computers. Although ML approaches are not new, their popularity has surged in recent years. ML, particularly deep learning (DL), offers significant advantages for weather forecasting. These include the ability to analyze vast datasets, detect intricate patterns, and predict outcomes based on those patterns. Unlike deterministic models, which can fall short of capturing the full complexity of weather systems, ML excels at identifying

Received for publication 13 Mar 2025. Accepted for publication 22 Apr 2025.

Published online 20 Jun 2025.

¹Food, Agricultural and Biological Engineering Department, Ohio State University, Wooster, OH 44691, USA

²Department of Extension, Ohio State University, Columbus, OH 43210, USA

³Byrd Polar and Climate Research Center, Ohio State University, Columbus, OH 43210, USA

This project is supported by the Specialty Crop Research Initiative (grant no. 2022-51181-38324, Project ADVANCEA) from the US Department of Agriculture National Institute of Food and Agriculture.

P.L. is the corresponding author. E-mail: ling.23@osu.edu.

This is an open access article distributed under the CC BY-NC license (<https://creativecommons.org/licenses/by-nc/4.0/>).

<https://doi.org/10.21273/HORTTECH05644-25>

nonlinear and continuous relationships within meteorological factors (Abdalla et al. 2021; Lam et al. 2022). By training on historical weather data, ML algorithms could discern the intricate interplay between weather variables, offering more accurate forecasts than the traditional NWP models alone (Lam et al. 2022). This approach is particularly valuable in harnessing big data and uncovering insights beyond the scope of traditional deterministic methods.

ML is used in various ways for weather forecasting. Some models focus on overall weather patterns across regions, whereas others specialize in predicting natural disasters like hurricanes, tornadoes, and floods. Some ML models target individual weather parameters such as wind speed (Du 2019), rainfall (Das et al. 2017; Rasel et al. 2018; Singh et al. 2019), or air temperature (Abrahamsen et al. 2018; Jakaria et al. 2020; Suleman and Shridevi 2022). Researchers use various ML techniques for weather forecasting, including support vector regression, artificial neural networks (ANNs), DL, such as recurrent neural networks and convolutional neural networks, decision trees, random forests, autoregression, and ensemble methods. The effectiveness of each

method varies from task to task. However, DL and ANN have emerged as the most used techniques (Bochenek and Ustrnul 2022). For instance, DL techniques are used to incorporate historical data from multiple nearby weather stations to forecast future conditions (Jakaria et al. 2020), predict forecast uncertainty (Scher and Messori 2018), or even simultaneously implement single-value forecasting and uncertainty quantification (Wang et al. 2019).

Although ML has demonstrated significant utility in weather forecasting, it is not intended to replace traditional numerical methods outright. Traditional NWP remains highly valuable and effective. Instead, ML serves as a potential complementary tool to existing methods, enhancing their capabilities (Lam et al. 2022). For instance, data assimilation from NWP can provide valuable input for ML algorithms, enriching their predictive power and improving forecast accuracy.

Methodology

HRRR PERFORMANCE EVALUATION.

Most growers rely on online weather forecasts or nearby local weather stations, when available, to predict conditions at their specific sites. Among

the commonly reported weather features, air temperature, RH, wind speed, and solar radiation are key attributes that significantly influence high tunnel conditions. To assess the HRRR model's suitability for site-specific forecasting, an evaluation of its performance in predicting these four attributes was conducted.

The evaluation compared HRRR's forecasts with ground truth data using Mar 2021 observations. While the coefficient of determination (R^2) was the primary metric for assessing forecast accuracy, the distribution of data points was analyzed to gauge the reliability of the forecasts during periods with a high probability of elevated temperatures. The decision on which attribute to improve was based not only on these evaluation metrics but also on the attribute's impact on high tunnel temperatures.

DATA PREPARATION. Once the most impactful weather attribute was identified, it became the predictor of our ML model. Data preparation was crucial for ML modeling, involving not only sourcing available data but also organizing it into accessible formats to ensure that researchers without in-depth knowledge of HRRR could effectively analyze and prepare the datasets. For this project, hourly

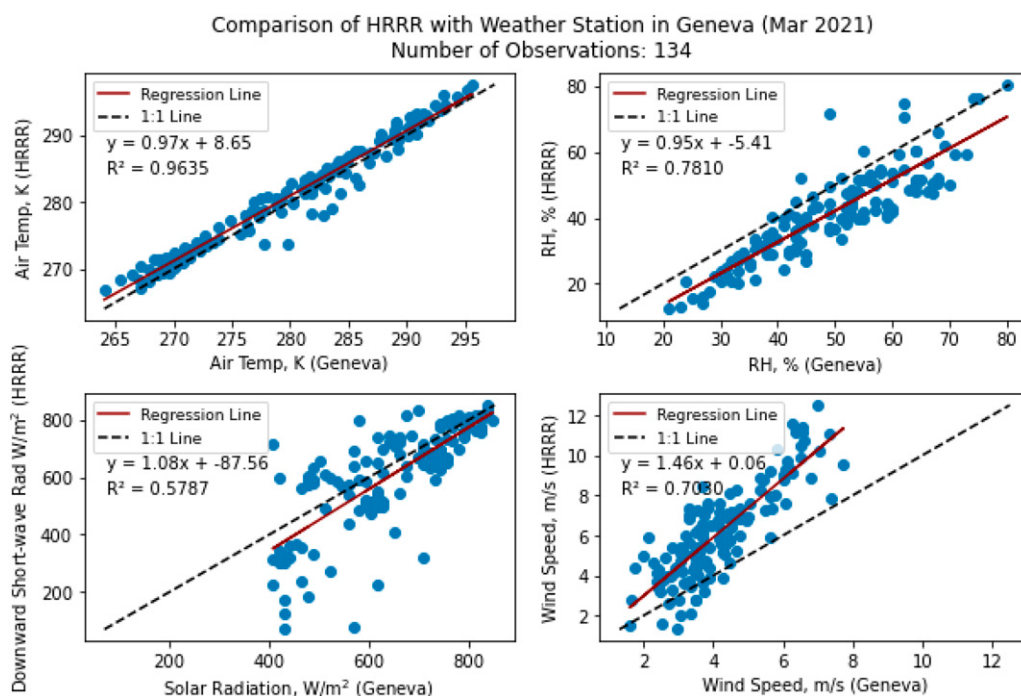


Fig. 1. Regression plot of HRRR forecast vs. observed data from a local weather station in Geneva, NY, Mar 2021, focusing on data points with solar radiation levels above $400 W/m^2$.

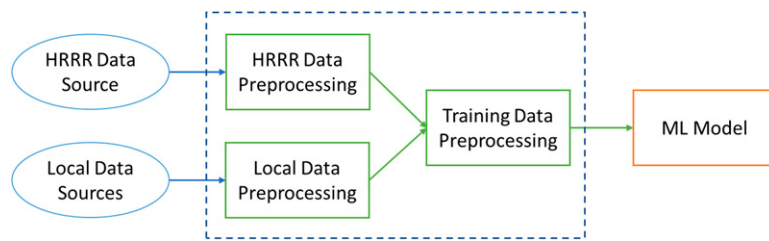


Fig. 2. Schematic of the training data preparation process before model development. Green boxes represent individual programs developed to automate data into the ML-ready dataset.

data were essential to align with the HRRR forecast intervals. Several data preparation programs were developed to download and transform data into an ML-ready training dataset.

HRRR DATA. The HRRR data are available on cloud services and provide extensive meteorological information (Dowell et al. 2022). It includes more than 170 atmospheric variables across various levels (e.g., surface, upper level) and different temporal resolutions (e.g., hourly, sub-hourly). The data are stored in GRIB2 (gridded binary) format, which is not commonly used outside the meteorological community and requires significant storage capacity. Each file represents weather data for a single hour across the entire Continental United States (CONUS) and is ~150 MB, resulting in more than 3 GB of data per day and more than 90 GB per month. For broader applications, it is crucial to convert the data into a more consumer-accessible format, such as CSV for Microsoft Excel, and to reduce its size and dimensionality for effective scope-specific modeling.

For this project, data from March and October over 3 years (2019–21) was selected for the model training. The Herbie Python package (Blaylock 2023) was used to download the raw hourly HRRR data in GRIB2 format. Herbie facilitates the retrieval of specific data points by specifying the NWP model, field, date/time, and forecast time. Our program extended its capabilities to select data based on site coordinates, a date range, and variables of interest.

To process the GRIB2 data, the program used PyGrib for data extraction and employed the k-d tree nearest neighbor search method from the SciPy library to find the HRRR grid point closest to the target location. Information was extracted from only 21 variables on the surface level deemed relevant due to their potential relationship or influence on solar radiation forecasts. Missing data were addressed, and hourly data were aggregated into daily and monthly summaries. The processed data were then output in CSV format, significantly reducing the file size and making it more accessible for analysis.

GROUND TRUTH DATA. In addition to the HRRR forecast data, observational data are also necessary for ML model development. To explore the potential for ML in enhancing weather forecasting on a broader scale, three local weather datasets from different locations: Wooster, OH (40.78, −81.93), West Lafayette, IN (40.47, −86.99), and Geneva, NY (42.88, −77.03), were used. Notably, West Lafayette sits directly on an HRRR grid point, whereas Wooster and Geneva are off-grid locations, allowing for a comprehensive assessment of the ML model’s performance across both scenarios. Weather data for Wooster was obtained from The Ohio State University College of Food, Agricultural, and Environmental Sciences (CFAES) Wooster weather station (Ohio State University 2023). Data for West Lafayette was sourced from the Purdue Mesonet Agronomy Center for Research and Education (ACRE) station (Purdue University 2023). Data for Geneva were collected from the Network for Environment and Weather Applications (NEWA) Geneva station (Cornell University 2023).

Retrieving observational data required manual execution due to differences in website interfaces. Although this process might seem straightforward, it required domain knowledge to evaluate the accuracy of the data instead of blindly trusting what was published, because mistakes can occur. In addition, overcoming the challenges posed by varying data handling practices at each station was crucial to creating interpretable data accessible to non-meteorologists. These challenges

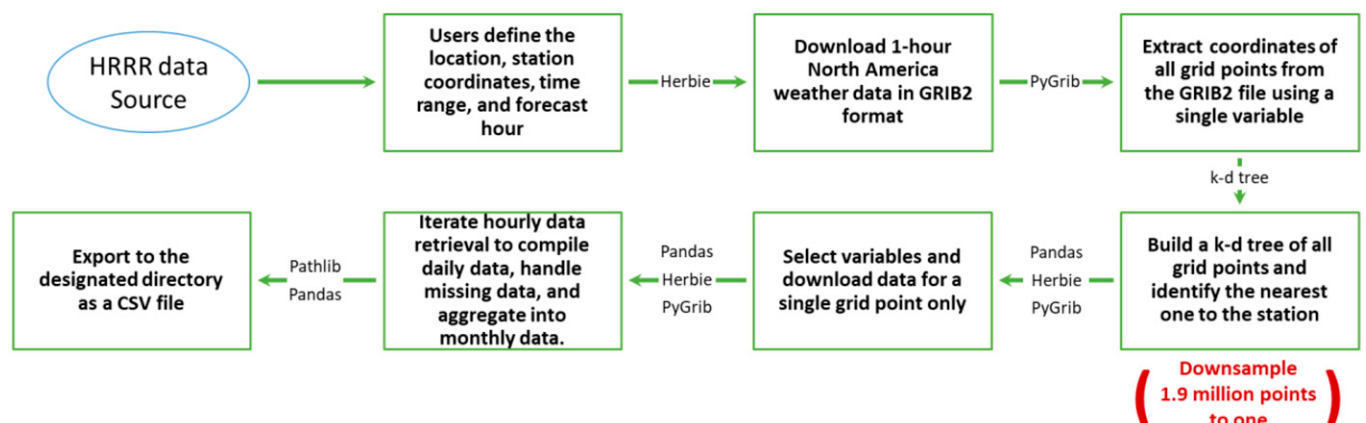


Fig. 3. Diagram of the “HRRR Data Preprocessing” program, illustrating the methods used to efficiently process large volumes of HRRR data, select relevant variables, and transform the data into CSV format for analysis.



Fig. 4. Diagram of daily HRRR data transformation using the “HRRR Data Preprocessing” program, illustrating the results of dimensionality reduction, storage size optimization, and file formation conversion.

included evaluating data accuracy, handling missing data, accounting for daylight saving time, adjusting for time zones, harmonizing sampling frequencies, managing unit differences, and reconciling variable name discrepancies.

To address these issues, three custom programs were developed for the respective locations. The program for Wooster resampled 5-min interval data to hourly increments to align with HRRR’s format. The program for West Lafayette managed data stored in different encodings and handled duplicated times on daylight saving dates. The program for Geneva addressed both duplicated time entries and time skips due to daylight saving time and converted missing data labels to empty cells, ensuring compatibility with the dataset processing pipeline. In addition, all programs standardized variable names, converted units to match those of HRRR, filled missing data with empty rows, selected the necessary variables, and outputted the data in CSV format.

TRAINING DATA. The final step in data preparation involved merging HRRR and local data into a unified training dataset. A key challenge was synchronizing time across different sources: HRRR data are in Coordinated Universal Time (UTC), whereas Wooster uses Eastern daylight saving time (EDT), West Lafayette follows Eastern Standard Time (EST), and Geneva observes daylight saving time. In Geneva, daylight saving time shifts the clock from 1:00 AM to 3:00 AM in March, whereas standard time in November repeats 1:00 AM twice. To align all data to UTC, local data from the previous month was included to cover missing hours caused by time differences and daylight saving adjustments.

A program was developed to manage these time adjustments and realign all datasets to UTC. This program also introduced new features such as previous-hour downward short-wave radiation (DSW-1), previous-hour downward longwave radiation (DLW-1), previous-hour upward shortwave radiation (USW-1), and previous-hour upward longwave radiation (ULW-1),

that was derived from the HRRR data. In addition, the date/time value was converted into a numerical format. Rows with missing values and the days where daylight saving time transition occurred were removed. Finally, the data were aggregated into monthly summaries and saved in CSV format.

ML MODEL DEVELOPMENT. This project leveraged DL for its ability to manage the complexities of weather data, particularly its strength in handling high-dimensional datasets and using extensive historical HRRR data. However, instead of using a highly complex DL model with hundreds of parameters, this study developed an ANN that retained the benefits of DL while remaining computationally efficient. By using fewer parameters, the model ensured accessibility for general use without sacrificing predictive capability. Its simplified design also allowed training on a laptop (Dell Latitude 5530) equipped with a 12th-generation Intel Core i5-1245U CPU, integrated Intel UHD graphics GPU, 8 GB RAM, and Microsoft Windows 11 Enterprise. Furthermore, the proven success of DL in previous weather forecasting studies (Abrahamsen et al. 2018; Du 2019; Lam et al. 2022; Rasel et al. 2018; Scher and Messori 2018; Suleman and Shridevi 2022; Wang et al. 2019) reinforced its selection for this project.

With data preparation procedures established, the next step was to

```
122:Upward long-wave radiation flux:W m**-2 (instant):lambert:nominalTop:level 0:fcst time 1 hrs:from 202103011000
123:Downward short-wave radiation flux:W m**-2 (instant):lambert:surface:level 0:fcst time 1 hrs:from 202103011000
124:Downward long-wave radiation flux:W m**-2 (instant):lambert:surface:level 0:fcst time 1 hrs:from 202103011000
125:Upward short-wave radiation flux:W m**-2 (instant):lambert:surface:level 0:fcst time 1 hrs:from 202103011000
126:Upward long-wave radiation flux:W m**-2 (instant):lambert:surface:level 0:fcst time 1 hrs:from 202103011000
127:Cloud Forcing Net Solar Flux:W m**-2 (instant):lambert:surface:level 0:fcst time 1 hrs:from 202103011000
128:Visible Beam Downward Solar Flux:W m**-2 (instant):lambert:surface:level 0:fcst time 1 hrs:from 202103011000
129:Visible Diffuse Downward Solar Flux:W m**-2 (instant):lambert:surface:level 0:fcst time 1 hrs:from 202103011000
130:Upward short-wave radiation flux:W m**-2 (instant):lambert:nominalTop:level 0:fcst time 1 hrs:from 202103011000
```

Raw HRRR Data



Processed HRRR Data

C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Date_Time	Surface pr	2 metre ter	2 metre sp	2 metre de	2 metre re	10 metre L	10 metre V	10 metre v	Total Prec	Sensible F	Latent hea	Ground he	Precipitat	Low cloud
Hour (UTC)	Pa	K	kg kg**-1	K	%	m s**-1	m s**-1	m s**-1	kg m**-2	W m**-2	W m**-2	W m**-2	kg m**-2	%
3/1/2021 1:00	97450	285.8029	0.0091	285.355	91.5	1.483303	2.731548	3.093012	0.025	-27	-14	-37	20.05909	100
3/1/2021 2:00	97540	286.2531	0.00923	285.5623	90.7	3.807292	2.878998	4.934376	0.061	-40	4	-33	23.7538	100
3/1/2021 3:00	97450	285.7726	0.00901	285.1585	91.2	2.309099	2.636093	4.18424	0	-30	-2	-16	24.98477	37
3/1/2021 4:00	97460	285.6869	0.009	285.1866	91.2	2.674988	3.05698	4.024266	0.015	-37	-9	-32	26.46475	100
3/1/2021 5:00	97560	284.9401	0.00808	283.5622	88.4	6.504423	-0.45437	6.478207	0.114	-29	44	11	24.71027	62

Fig. 5. Product of the “HRRR Data Preprocessing” program, demonstrating the transformation of complex raw HRRR data into simplified data with selected variables displayed in a spreadsheet format.

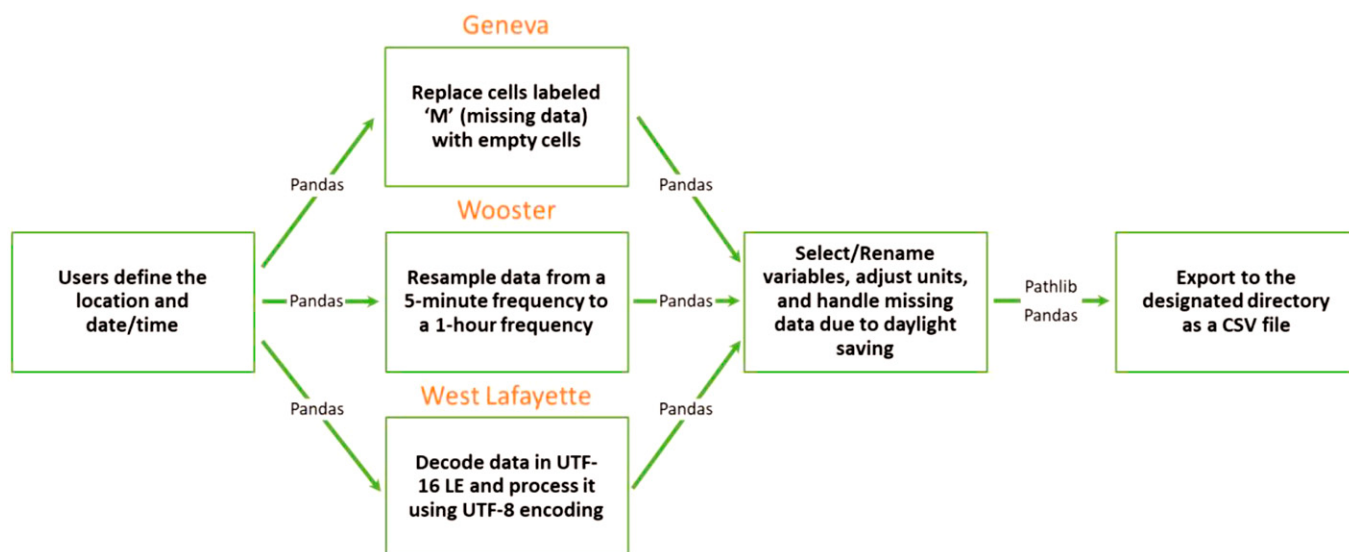


Fig. 6. Diagram of the “Local Data Preprocessing” program, illustrating the methods used to transform weather stations’ data into an accessible format for analysis.

develop an ML model to forecast 1-h on-site weather conditions, focusing specifically on solar radiation, the attribute that needed the most improvement. The model targeted daytime data, with an emphasis on instances in which solar radiation exceeded 400 W/m^2 , as high temperatures were most likely to occur inside high tunnels under these conditions. To account for the influence of prior environmental conditions on future solar radiation trends, the model incorporated HRRR data from three timesteps: previous-hour, current time, and a 1-h forecast for selected features. This approach aimed to capture the temporal effects that might be missed when treating data points in isolation.

MODEL ARCHITECTURE. The ML model was implemented using Google’s TensorFlow library (Abadi

et al. 2016), incorporating ELU (Exponential Linear Unit) activation, Adam optimizer (Kingma and Ba 2014), Huber loss function (Huber 1964), Early Stopping, and hyperparameter tuning. The architecture consisted of one input dense layer containing 25 neurons, three hidden dense layers with 25 neurons each, and one output dense layer with a single neuron. The number of input features depended on the five feature designs that are discussed in the “Feature Selection” section. The activation functions applied to the input and hidden layers were ELU, whereas the output layer used a linear activation function.

ReLU (Rectified Linear Unit) is a popular activation function for regression problems due to its simplicity and computational efficiency. However, it can encounter the vanishing

gradient issue, where the gradients become very small during training, causing the weights to stop updating and hindering convergence. To address this, ELU was chosen because it can alleviate the vanishing gradient issue. Although ELU introduces a modest computational overhead due to its exponential function, it generally improves training stability and convergence (Clevert et al. 2015).

The optimal model parameters were trained using the Adam optimizer (Adaptive Moment Estimation), an adaptive optimization algorithm based on stochastic gradient descent. Adam dynamically adjusts learning rates for each parameter using first- and second-moment estimations of past gradients. This adaptive learning rate adjustment makes it particularly efficient when working with large

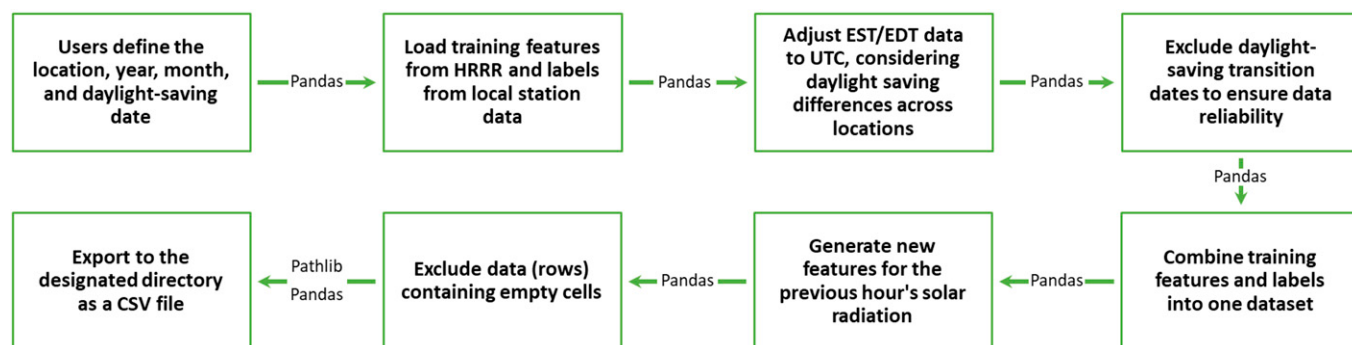


Fig. 7. Diagram of the “Training Data Preprocessing” program, illustrating the methods used to unify training features and labels into one dataset.

Table 1. Summary of solar radiation forecasting performance by High-Resolution Rapid Refresh (HRRR) and machine learning (ML) models, comparing the results across five feature designs, including mean, standard deviation (*SD*), and root mean square error (RMSE) of prediction errors. Positive values indicate overprediction and negative values indicate underprediction. Dark orange, orange, and light orange boxes highlight the best-performing ML feature design at each location for mean error, RMSE, and R^2 , respectively, compared to HRRR.

Location	Prediction error								
	HRRR				Feature design	ML			
	Mean (W/m ²)	SD (W/m ²)	RMSE (W/m ²)	R^2		Mean (W/m ²)	SD (W/m ²)	RMSE (W/m ²)	R^2
Geneva	-16.43	93.56	94.18	0.62	F01	0.85	75.97	75.30	0.69
					F00	-10.58	92.20	92.00	0.55
					F00 F01	-3.75	79.44	78.83	0.68
W. Lafayette	118.71	99.03	153.71	0.34	F01	-3.19	70.49	69.57	0.54
					F00	4.85	63.27	62.58	0.63
					F00 F01	3.60	50.85	50.27	0.76
Wooster	33.90	109.54	113.78	0.47	F01	14.48	61.03	62.22	0.69
					F00	2.69	81.98	81.33	0.46
					F00 F01	3.85	64.69	64.26	0.67
					F00 F01 Mix	4.53	66.39	65.99	0.65
					F-1 F00 F01 Mix	-1.71	58.25	57.76	0.62
	27.37	100.70	103.49	0.43					

datasets, allowing for faster convergence during training.

For the loss function, this project chose to use the Huber loss during model compilation, balancing the advantages of mean squared error (MSE) and mean absolute error (MAE) for the task at hand. Although MSE is effective for penalizing large errors, MAE is less sensitive to outliers. Huber loss blends these attributes by employing MSE for small errors and MAE for larger errors, offering a smoother transition between the two. This approach reduces the impact of outliers on the overall loss calculation, making it well-suited for handling real-world data that typically contains noise and outliers.

Hyperparameter tuning was performed manually to optimize the model's learning pace by testing various combinations of learning rates, batch sizes, number of layers, neurons per layer, and activation functions. This approach helped reduce the risk of overfitting.

FEATURE SELECTION. Feature selection is crucial in developing an effective ML system, enhancing efficiency, performance, and interpretability by reducing dimensionality, minimizing overfitting, addressing multicollinearity, and simplifying models. It also improves training and inference time.

This project initially selected features based on domain knowledge and further refined them using the

Genetic Algorithm (GA) (Siedlecki and Sklansky 1989). In addition, the study examined how different input features at different timesteps influenced model performance. Because solar radiation exhibits temporal dependencies, feature designs incorporating past, current, and forecasted data were tested to assess their impact on predictions. As a result, five feature designs were evaluated: F01, F00, F00 F01, F00 F01 Mix, and F-1 F00 F01 Mix.

1. **F01** included only 1-h forecast features selected by GA from domain-informed HRRR variables. This design tested whether HRRR forecast data alone could serve as

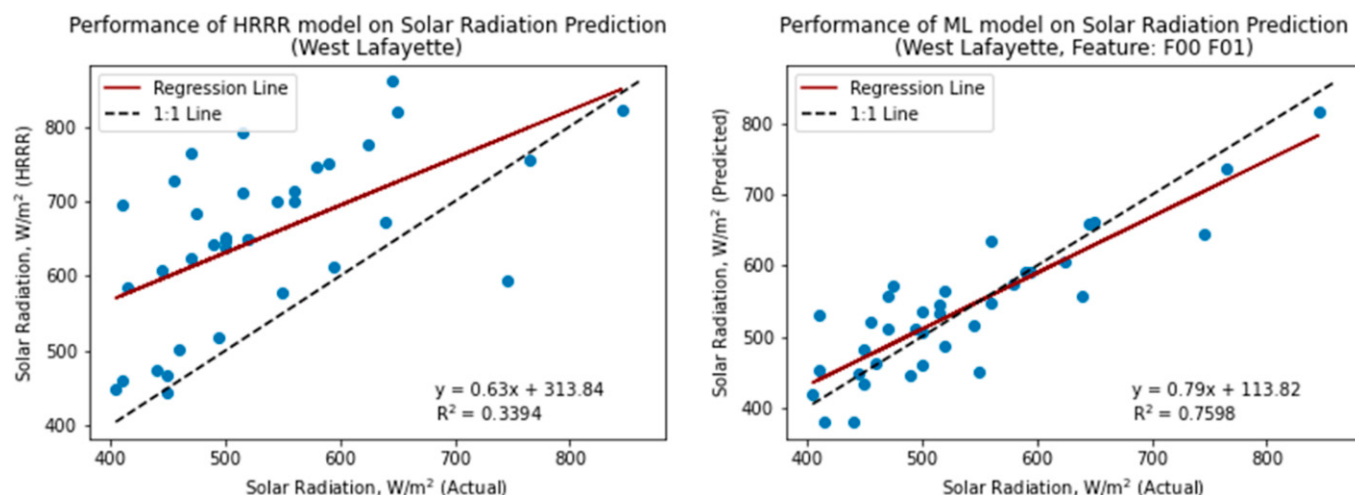


Fig. 8. Regression plot of HRRR and ML (F00 F01 feature) solar radiation forecasts vs. observations in West Lafayette, IN, illustrating the models' performance in terms of R^2 values and HRRR bias based on data deviation from the 1:1 line.

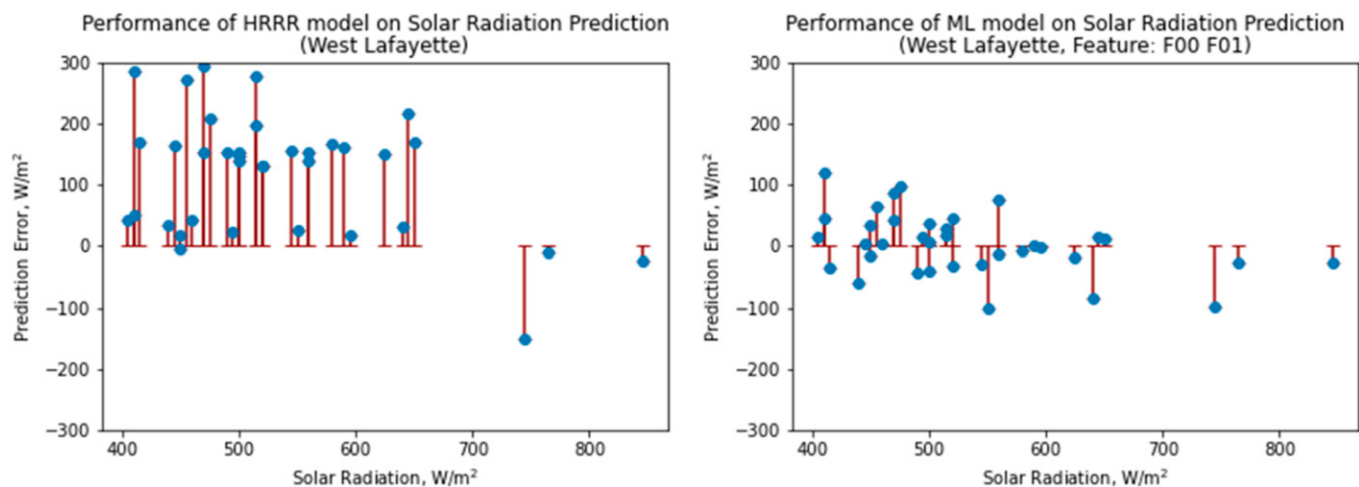


Fig. 9. Solar radiation forecast error plot for HRRR and ML (F00 F01 feature) models of West Lafayette, IN, illustrating the models' forecasting performance in terms of bias.

reliable inputs for a site-specific ML model to improve HRRR's global-scale predictions.

2. **F00** used the same feature set as F01 but with current-time data, assessing whether real-time atmospheric conditions better captured the immediate influences on solar radiation.
3. **F00 F01** combined current-time and 1-h forecast features to determine if leveraging both observed conditions and short-term predictions could enhance accuracy by capturing subtle temporal patterns and interactions.
4. **F00 F01 Mix** was a refined version of F00 F01, in which GA was applied again to select only the most relevant features from the combined pool of domain-informed current-time and 1-h forecast features. This design aimed to reduce dimensionality while preserving key predictive variables.
5. **F-1 F00 F01 Mix** extended F00 F01 Mix by incorporating previous-hour radiation features to capture temporal dependencies across three timesteps, addressing lag effects in solar radiation response.

These five designs were tested to determine the optimal balance between information richness and model efficiency.

DATA PARTITION, OUTLIERS, AND NORMALIZATION. Apart from timestep considerations, maintaining a balanced

dataset for model training, validation, and testing was essential. Equal representation from March and October ensured that the model learned from varying seasonal conditions. The first four feature designs combined all March and October data from 2019, 2020, and 2021 into a single dataset before being split into 80% training, 10% validation, and 10% test sets. In contrast, the fifth design split March and October data into separate groups before dividing each group into 80% training, 10% validation, and 10% test sets. The corresponding March and October subsets were then merged to maintain seasonal balance. Because a model trained exclusively on summer data, for instance, could struggle with winter predictions, this partitioning method ensured better generalization across seasonal variations.

In addition, the Isolation Forest technique (Liu et al. 2008) was used in the fifth feature design to remove 5% outliers based on the downward shortwave radiation of the HRRR 1-h forecast. Isolation Forest is an anomaly detection algorithm that identifies outliers by constructing an ensemble of binary decision trees. It randomly partitions the data, and anomalies are identified as instances with fewer splits in the trees. Although the Huber loss function used in model training can handle outliers, actively removing outliers ensured they did not adversely impact the training process.

Because the model included many features that varied in scale, normalization was essential to ensure they contributed equally. This step

also helped achieve faster convergence during training, thereby improving training efficiency. Z-score standardization was used for normalization, transforming the data to have a mean of zero and a standard deviation of one. To prevent data leakage, in which the information from outside the training dataset is accidentally used to create the model, the training, validation, and test sets were normalized separately.

TRAINING PROCESS AND EVALUATION. The project aimed to maintain a consistent neural network architecture and hyperparameters for training models across three locations in the Midwest region to simplify the training process. The models were trained with a learning rate of 0.001, a batch size of 32, and 500 epochs. The 500 epochs were chosen because the models tended to converge around this range for our dataset. The first three feature designs (F01, F00, and F00 F01) were evaluated at all three locations to determine whether one model could be effective across similar regions. The fourth and fifth designs were tested exclusively at Wooster to assess whether these concepts could benefit model training or performance.

During training, each model was trained 10 times to test the stability of the model, and the best-performing model was selected based on the lowest RMSE obtained from the test set. To evaluate the performance of the best models for each feature design and location, mean prediction error, standard deviation, RMSE, and R^2

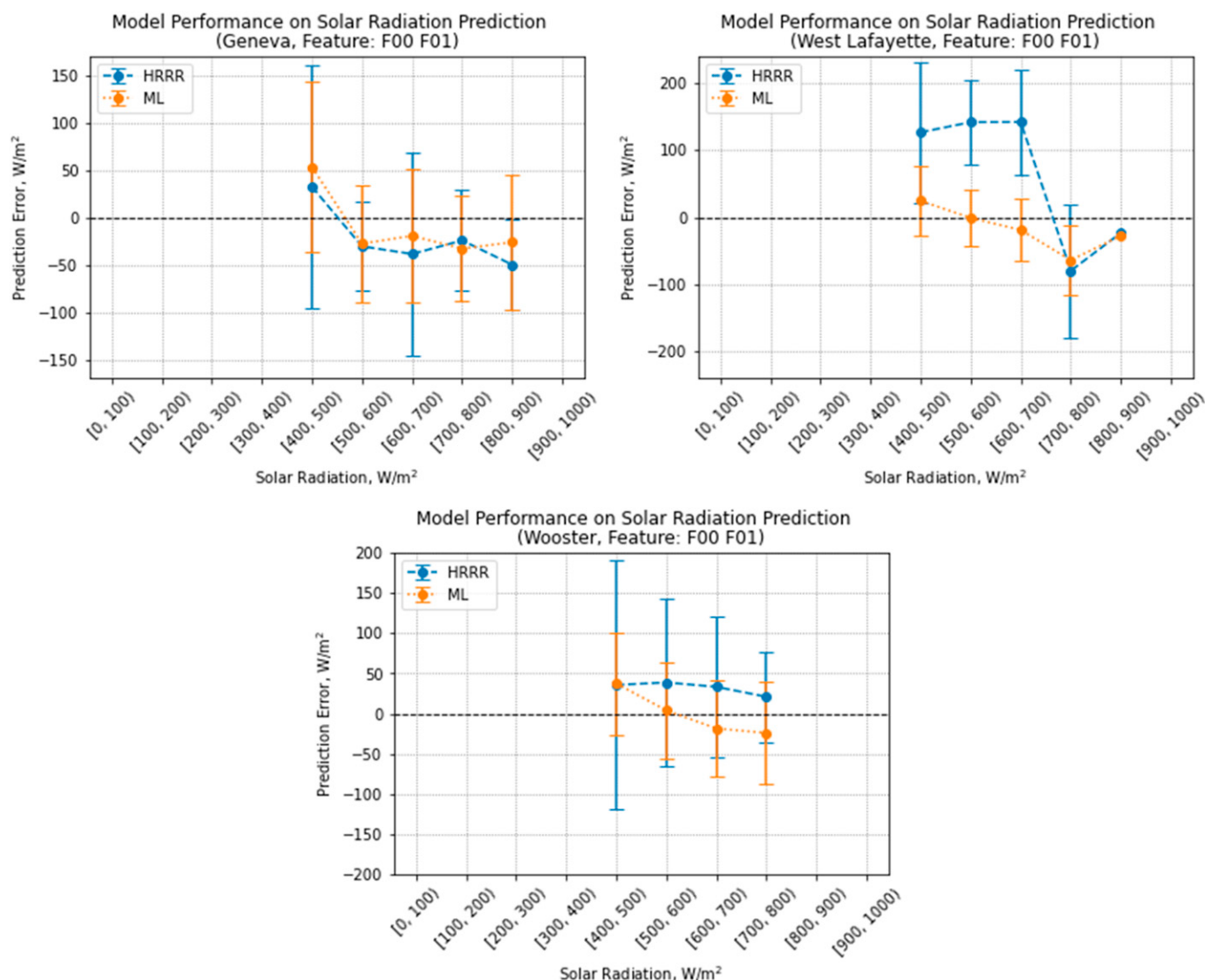


Fig. 10. Diagram of solar radiation forecast error in Geneva, NY, West Lafayette, IN, and Wooster, OH, grouped by solar radiation levels for HRRR and ML (F00 F01 feature) models, illustrating model performance through mean error (dot) and standard deviation (error bar).

were calculated. RMSE is sensitive to outliers, making it useful for assessing whether the ML model has effectively handled them and produced a more generalized result. In addition, the model performance was evaluated across different solar radiation levels to determine if it performed particularly well under specific conditions. The solar radiation levels were grouped in increments of $100 W/m^2$, allowing for a detailed analysis of how the model's accuracy varied with changes in solar radiation intensity.

Results

HRRR PERFORMANCE EVALUATION. The R^2 of solar radiation dropped to the worst performance compared with

the other three variables (Fig. 1). Given the significant impact of these variables on high tunnel temperature changes, solar radiation emerged as the most crucial factor. Hence, we identified solar radiation as the factor requiring the most improvement, making it the ML model's target variable.

DATA PREPARATION. To support model development, several Python programs were created for training data preparation. The data processing pipeline (Fig. 2) consisted of three stages: HRRR Data Preprocessing, Local Data Preprocessing, and Training Data Preprocessing, which handled HRRR data, local weather station data, and their integration.

During the "HRRR Data Preprocessing" stage (Fig. 3), the Herbie package downloaded comprehensive HRRR data for the CONUS region, consisting of 1.9 million grid points, whereas PyGrib read and selected 21 relevant variables from the GRIB2 format. These variables included surface pressure, air temperature, specific humidity, dewpoint temperature, RH, u-component wind, v-component wind, wind speed, sensible heat net flux, latent heat net flux, ground heat flux, total precipitation (liquid equivalents that reached the surface), precipitable water (total water vapor in a column of the atmosphere), low cloud cover, medium cloud cover, high cloud cover, total cloud cover,

Table 2. Feature design ranking (dark orange highlight) based on root mean square error (RMSE) improvement from High-Resolution Rapid Refresh to machine learning. Lower numbers indicate better designs. Improvements at 1% or below were considered negligible and did not affect rankings, allowing ties. The total ranking sum (orange highlight) is shown only for the first three designs, which are available across all three locations.

Feature design ranking			
Location	Feature design	RMSE improvement (%)	Ranking
Geneva	F01	20	1
	F00	2	3
	F00 F01	16	2
W. Lafayette	F01	55	3
	F00	59	2
	F00 F01	67	1
Wooster	F01	45	1
	F00	29	3
	F00 F01	44	1
	F00 F01 Mix	42	2
	F-1 F00 F01 Mix	44	1
Ranking sum			
		F01	5
		F00	8
		F00 F01	4

downward shortwave radiation, downward longwave radiation, upward shortwave radiation, and upward longwave radiation. The k-d tree method only selected the nearest grid point to the target site from all the grid points, whereas the Pandas library inserted empty rows for missing data, aggregated the data into monthly intervals, and exported the processed data as CSV files. This preprocessing step significantly reduced the data size from ~3.6 GB per day to just 6 KB by narrowing location grid points from 1.9 million to 1 and reducing the number of variables from 170 to 21 (Fig. 4). The raw HRRR data, initially too complex for direct use, was transformed into a spreadsheet-friendly format with selected variables, making it easier to interpret and integrate into the ML model (Fig. 5).

For the local weather data, domain knowledge was applied to address specific challenges and correct inaccuracies. Geneva's data required addressing letter-labeled missing cells, whereas Wooster's data needed re-sampling from 5-min to 1-h intervals and converting solar radiation measurements from Langley's per 5-min interval to W/m². West Lafayette's data required decoding UTF-16 LE encoding, and the 1-h solar radiation

data had been incorrectly published as the sum of 30-min intervals. Thus, the "Local Data Preprocessing" program was customized for each location to address these issues and standardize the datasets by renaming variables, adjusting units, and managing missing data due to daylight saving, ensuring all local data aligned with the HRRR format (Fig. 6).

In the "Training Data Preprocessing" stage (Fig. 7), HRRR and ground truth observations were merged into a single dataset. Because the three locations handled daylight saving differently, the program adjusted EST/EDT times to UTC, explicitly accounting for these variations rather than merely applying static time zone offsets. It excluded transition dates to ensure data reliability. It also generated new features, such as previous-hour solar radiation, and removed records with missing values. The finalized dataset was then ready for model training. This three-step data preparation process streamlined the transformation of HRRR's large, complex meteorological data into a simplified format while standardizing ground truth weather station data for ML model development.

ML MODEL DEVELOPMENT. Feature selection played a critical role in model development. GA reduced the

21 initial features to 14 key 1-h forecast variables for the F01 design, including surface pressure, air temperature, dewpoint temperature, RH, u-component wind, wind speed, sensible heat net flux, latent heat net flux, precipitable water, cloud cover (low, medium, total), and both downward shortwave and longwave radiations. F00 selected the same 14 features but used current-time data instead. F00 F01 combined all 14 features from F00 and F01, totaling 28.

F00 F01 Mix refined the selection further by applying GA to the 42 combined domain-informed features (21 current-time and 21 1-h forecast), selecting 16 features based on a frequency threshold of 20 out of 30 generations. The final selection included eight current-time features (surface pressure, air temperature, RH, wind speed, latent heat net flux, medium cloud cover, downward longwave radiation, and upward shortwave radiation) and eight forecast features (u-component wind, sensible heat net flux, latent heat net flux, ground heat flux, medium cloud cover, total cloud cover, downward shortwave radiation, and downward longwave radiation). Although GA initially selected the 1-h forecast total precipitation, it was excluded due to unreliability. In addition, a date/time feature was manually added to capture diurnal solar radiation trends, bringing the total to 17 features.

F-1 F00 F01 Mix expanded on this by incorporating three previous-hour radiation features (downward shortwave, downward longwave, and upward shortwave) and two additional radiation-related features (current-time downward shortwave radiation and 1-h forecast upward shortwave radiation) that GA had previously excluded. This resulted in a final set of 22 features.

The model performance evaluation metrics included R^2 , mean error, RMSE, and standard deviation of the prediction errors. RMSE measured accuracy, whereas the standard deviation indicated the variability of errors. If the model was unbiased (with a mean error close to 0), RMSE and standard deviation would be similar. Table 1 summarizes the performance of all feature designs. For Geneva, the F01 design performed the best, followed by the F00 F01, whereas for West Lafayette,

the F00 F01 design was the top choice, followed by the F00. In Wooster, the F-1 F00 F01 Mix was the most effective, followed by the F01.

For example, in West Lafayette, IN, the HRRR model exhibited a high mean error (119 W/m²), standard deviation (99 W/m²), and RMSE (154 W/m²), along with a low R^2 value (0.34). The ML model using the F00 F01 feature design significantly improved these metrics with the mean error dropping to 3.6 W/m², standard deviation to 51 W/m², RMSE to 50 W/m², and R^2 rising to 0.76. The large discrepancy between HRRR's standard deviation and RMSE suggests that the model exhibited bias, further supported by the regression plot (Fig. 8), where HRRR predictions cluster above the 1:1 line. In contrast, the ML model's data aligns more closely with the 1:1 line. Figure 9 highlights that the HRRR model consistently overpredicted solar radiation below 650 W/m², whereas the ML model produced smaller, more balanced errors centered around zero. Despite improvements, the ML model still had errors in underpredicting by 100 W/m² and overpredicting by 120 W/m². The HRRR model, comparatively, had underprediction and overprediction errors of as much as 150 W/m² and 294 W/m², respectively.

When evaluating performance across different solar radiation levels, the ML model (F00 F01 feature) generally outperformed the HRRR model (Fig. 10), although results varied by location. In Geneva, ML performed best at 600 to 700 W/m² but only moderately at 400 to 500 and 700 to 800 W/m². In West Lafayette, it performed well from 400 to 700 W/m², with weaker performance at 700 to 800 W/m². The 800 to 900 W/m² range had only one test data point. In Wooster, the ML model performed well at 500 to 700 W/m² but slightly worse at 400 to 500 and 700 to 800 W/m². The limited number of data points at higher flux levels likely impacted its accuracy. For example, West Lafayette had 177 points in the 400 to 500 W/m² range, but only 10 and 9 in the 700 to 800 and 800 to 900 W/m² ranges, respectively, before splitting for training. Similarly, Wooster had 185 points at 500 to 600 W/m² but only 14 at 800 to

900 W/m². This suggests performance at higher solar radiation levels could improve with more data.

To streamline training across locations, a ranking system based on RMSE improvement was employed. Improvements below 1% were deemed insignificant and did not affect rankings, allowing the same rank to be assigned in such cases. The final rankings were summed to determine the overall best design. Table 2 shows that the F00 F01 ranked the highest overall. However, for specific needs such as reducing input dimensionality to lower the risk of overfitting, the F-1 F00 F01 Mix may be better suited for certain locations.

The minimal improvement between the fourth and fifth designs suggests that the March and October data may share similar characteristics, making the additional partitioning step unnecessary, along with the inclusion of previous-hour features. The outlier removal step in the fifth design enhanced both the HRRR and ML models, reducing HRRR's RMSE from 114 to 103 W/m² and improving the ML's RMSE to 58 W/m². However, this step had a limited effect on the overall improvement, contributing only a 2% gain over the fourth design.

Conclusions

Although the HRRR model offers useful weather forecasts across CONUS, especially for those lacking local weather stations, a more targeted site-specific forecast with improved accuracy is desirable for high tunnel producers. This project identified that solar radiation forecast in HRRR requires the most enhancement compared with other key attributes like temperature, humidity, and wind for CEA facility temperature forecasts. The focus on a 1-h forecast period aimed to optimize accuracy and user response time.

This project successfully transformed complex HRRR GRIB2 data into a more accessible CSV format with selected grid points and variables, reducing dimensionality and file size. Challenges associated with multiple local weather stations, such as differing sampling rates and missing data, were overcome through various programs. Ensuring high-quality data were crucial for accurate modeling,

with domain expertise playing a key role in evaluating and converting data into reliable inputs to produce reliable forecasts.

The developed ML model significantly enhanced the accuracy and stability of 1-h solar radiation forecasts compared with the HRRR model, benefiting high tunnel users and other CEA facilities like greenhouses. Despite HRRR's high spatial resolution, the ML model outperformed it for both on-grid and off-grid locations. Although tailored to specific sites and timeframes, the methodology is adaptable to other locations.

This study demonstrated the potential to simplify modeling for locations at similar latitudes. A simple ANN effectively served Wooster, OH (40.78, -81.93), West Lafayette, IN (40.47, -86.99), and Geneva, NY (42.88, -77.03). The model development used HRRR forecast features for the current time and the next hour (the F00 F01 design), highlighting the robustness of the selected features across these locations. This approach showed that it was not necessary to use complex DL models, which typically require hundreds of parameters, because a simpler ANN model was sufficient to predict solar radiation.

Improved solar radiation prediction can be used to enhance the accuracy of high tunnel weather models. For example, the solar radiation forecast from this model, combined with other HRRR parameters, can be used in a high tunnel temperature predictive model to better anticipate in-tunnel conditions. This enables growers to manage heat stress more effectively, prevent crop loss, and ultimately support healthier plants and higher yields. As the model expands to cover a broader range of solar radiation, it could also support daily light integral prediction, optimize plant scheduling, and assist with supplemental lighting planning.

In addition, as the model is applied to more locations and variables, further adjustments will be necessary. Although it currently addresses data source variation challenges, its capacity for year-round predictions, broader radiation ranges, and additional variables is promising. However, one limitation is its potential to miss abnormal events, risking under- or overprediction of

solar radiation, which could impact high-temperature events or trigger false alarms. Striking a balance between consistency and anomaly detection will be essential moving forward.

References cited

- Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X. 2016. TensorFlow: A system for large-scale machine learning, p 265–283. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16). <https://dl.acm.org/doi/10.5555/3026877.3026899>.
- Abdalla AM, Ghaith IH, Tamimi AA. 2021. Deep learning weather forecasting techniques: Literature survey, p 622–626. 2021 International Conference on Information Technology (ICIT). IEEE. <https://doi.org/10.1109/ICIT52682.2021.9491774>.
- Abrahamsen EB, Brastein OM, Lie B. (2018). Machine learning in Python for weather forecast based on freely available weather data, p 169–176. Proceedings of the 59th Conference on Simulation and Modelling (SIMS 59). Linköping Electronic Conference Proceedings 153. <http://dx.doi.org/10.3384/ecp18153169>.
- Ahmad SF, Dar AH. 2020. Precision farming for resource use efficiency, p 109–135. In: Kumar S, Meena RS, Jhariya MK (eds). *Resources Use Efficiency in Agriculture*. Springer, Singapore. https://doi.org/10.1007/978-981-15-6953-1_4.
- Bhawsar M, Tewari V, Khare P. 2021. A survey of weather forecasting based on machine learning and deep learning techniques. *International Journal of Emerging Trends in Engineering Research*. 9(7):988–993. <https://doi.org/10.30534/ijeter/2021/24972021>.
- Blaylock BK. 2023. Herbie: Retrieve numerical weather prediction model data (Version 2023.3.0) [Computer software]. <https://doi.org/10.5281/zenodo.4567540>.
- Bochenek B, Ustrnul Z. 2022. Machine learning in weather prediction and climate analyses—Applications and perspectives. *Atmosphere*. 13(2):180. <https://doi.org/10.3390/atmos13020180>.
- Cornell University. 2023. Network for environment and weather applications (NEWA). Station name: Geneva, NY. <https://newa.cornell.edu/all-weather-data-query/>. [accessed 24 May 2023].
- Das S, Chakraborty R, Maitra A. 2017. A random forest algorithm for nowcasting of intense precipitation events. *Adv Space Res*. 60(6):1271–1282. <https://doi.org/10.1016/j.asr.2017.03.026>.
- Drost D. 2011. High tunnel summer squash production. Utah State Univ. Ext. Publ. No. 237. https://digitalcommons.usu.edu/extension_curall/237.
- Dowell DC, Alexander CR, James EP, Weygandt SS, Benjamin SG, Manikin GS, Blake BT, Brown JM, Olson JB, Hu M, Smirnova TG, Ladwig T, Kenyon JS, Ahmadov R, Turner DD, Duda JD, Alcott TI. 2022. The high-resolution rapid refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Weather and Forecasting*. 37(8):1371–1395. <https://doi.org/10.1175/WAF-D-21-0151.1>.
- Clevert DA, Unterthiner T, Hochreiter S. 2015. Fast and accurate deep network learning by Exponential Linear Units (ELUs). arXiv. 1511.07289. <https://doi.org/10.48550/arXiv.1511.07289.1511.07289>.
- Du P. 2019. Ensemble machine learning-based wind forecasting to combine NWP output with data from weather station. *IEEE Trans Sustain Energy*. 10(4):2133–2141. <https://doi.org/10.1109/TSTE.2018.2880615>.
- Huber PJ. 1964. Robust estimation of a location parameter. *Ann Math Statist*. 35(1):73–101. <https://doi.org/10.1214/aoms/1177703732>.
- Jakaria AHM, Hossain MM, Rahman MA. 2020. Smart weather forecasting using machine learning: A case study in Tennessee. arXiv. 2008.10789. <https://doi.org/10.48550/arXiv.2008.10789.2008.10789>.
- Kingma DP, Ba J. 2014. Adam: A method for stochastic optimization. arXiv. 1412.6980. <https://doi.org/10.48550/arXiv.1412.6980.1412.6980>.
- Lam R, Sanchez-Gonzalez A, Willson M, Wirnsberger P, Fortunato M, Alet F, Ravuri S, Ewalds T, Eaton-Rosen Z, Hu W, Merose A, Hoyer S, Holland G, Vinyals O, Stott J, Pritzel A, Mohamed S, Battaglia P. 2022. GraphCast: Learning skillful medium-range global weather forecasting. arXiv. 2212.12794. <https://doi.org/10.48550/arXiv.2212.12794>.
- Liu FT, Ting KM, Zhou Z-H. 2008. Isolation forest. 2008 Eighth IEEE International Conference on Data Mining, 413–422. IEEE. <https://doi.org/10.1109/ICDM.2008.17>.
- Ohio State University. 2023. CFAES weather system. Station name: Wooster Station. <https://weather.cfaes.osu.edu/>. [accessed 27 Mar 2023].
- Purdue University. 2023. Purdue Mesonet Data Hub. Station name: ACRE. <https://ag.purdue.edu/indiana-state-climate/purdue-mesonet/purdue-mesonet-data-hub/>. [accessed 14 Apr 2023].
- Rasel RI, Sultana N, Meesad P. 2018. An application of data mining and machine learning for weather forecasting. Recent Advances in Information and Communication Technology 2017: Proceedings of the 13th International Conference on Computing and Information Technology (IC2IT), 169–178. https://doi.org/10.1007/978-3-319-60663-7_16.
- Scher S, Messori G. 2018. Predicting weather forecast uncertainty with machine learning. *Quart J Royal Meteor Soc*. 144(717):2830–2841. <https://doi.org/10.1002/qj.3410>.
- Siedlecki W, Sklansky J. 1989. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*. 10(5):335–347. [https://doi.org/10.1016/0167-8655\(89\)90037-8](https://doi.org/10.1016/0167-8655(89)90037-8).
- Singh N, Chaturvedi S, Akhter S. 2019. Weather forecasting using machine learning algorithm. In 2019 International Conference on Signal Processing and Communication (ICSC), 171–174. IEEE. <https://doi.org/10.1109/ICSC45622.2019.8938211>.
- Suleman MAR, Shridevi S. 2022. Short-term weather forecasting using spatial feature attention-based LSTM model. IEEE Access. 10:82456–82468. <https://doi.org/10.1109/ACCESS.2022.3196381>.
- Wahid A, Gelani S, Ashraf M, Foolad MR. 2007. Heat tolerance in plants: An overview. *Environmental and Experimental Botany*. 61(3):199–223. <https://doi.org/10.1016/j.envexpbot.2007.05.011>.
- Wang B, Lu J, Yan Z, Luo H, Li T, Zheng Y, Zhang G. 2019. Deep uncertainty quantification: A machine learning approach for weather forecasting. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2087–2095. <https://doi.org/10.1145/3292500.3330704>.