Professional Certificate in AI for Renewable Energy Forecasting (Thailand)

Model Evaluation and Selection for Renewable Energy Forecasting

Model evaluation and selection are the twin pillars that determine whether a renewable-energy forecasting system will succeed in real-world operations. In the context of the Professional Certificate in AI for Renewable Energy Forecasting, students must master a vocabulary that spans statistical theory, machine-learning practice, and domain-specific considerations for solar, wind, and hydro resources. The following exposition presents the essential terms, their definitions, practical examples, and common challenges, organized by thematic clusters. Each term is highlighted with bold or italic tags only when it represents a core concept; the tags are kept to a few words at a time to preserve readability.

Training set – The subset of historical data used to fit the parameters of a model. In solar-PV forecasting, the training set might consist of three years of hourly irradiance, temperature, and power output measurements. The quality of the training set directly influences the model's ability to capture seasonal patterns and diurnal cycles. A common pitfall is to include future information inadvertently, such as using a day-ahead forecast as an input for a same-day model, which creates data leakage and inflates performance metrics.

Validation set – A separate portion of the data, distinct from the training set, employed to tune hyper-parameters and to assess how well the model generalizes to unseen data. For wind-speed prediction, a typical validation strategy is to reserve the most recent six months of data, ensuring that temporal trends and evolving turbine dynamics are reflected. The validation set is not used for final performance reporting; its role is to guide model refinement.

Test set – The final hold-out data used to obtain an unbiased estimate of model performance after all training and validation steps are complete. In a day-ahead market scenario, the test set may comprise the last three months of actual market outcomes, allowing the practitioner to compute realistic error statistics that stakeholders will trust. The test set must remain untouched until the very end of the workflow.

Overfitting – A condition where a model captures noise or idiosyncrasies of the training data rather than the underlying signal. Overfitted wind-forecast models often display exceptionally low training error but dramatically higher error on validation and test sets. Symptoms include wildly fluctuating predictions during calm periods or unrealistic spikes during storm events. Regularization techniques, such as L1 or L2 penalties, are commonly used to mitigate overfitting.

Underfitting – The opposite extreme, where a model is too simple to represent the complexity of the data, leading to high error on both training and unseen data. A linear regression that only uses temperature as a

predictor for solar power will typically underfit because it ignores cloud cover, angle of incidence, and shading effects. Adding relevant features or increasing model capacity can remedy underfitting.

Bias–variance trade-off – The fundamental tension between model simplicity (bias) and flexibility (variance). High bias models, such as simple persistence forecasts, systematically miss patterns, while high variance models, such as deep neural networks with many layers, may overreact to random fluctuations. Achieving an optimal balance is the central goal of model selection. Graphical illustrations often plot bias and variance as functions of model complexity, showing a “U-shaped” total error curve.

Cross-validation – A systematic technique for estimating model performance by repeatedly training and validating on different data splits. In renewable forecasting, the most common form is k-fold cross-validation, where the dataset is divided into k equally sized folds; each fold serves once as a validation set while the remaining k-1 folds constitute the training set. For example, with k = 5, a solar-irradiance dataset of 5,000 hourly observations yields five training-validation cycles, each providing an error estimate that can be averaged for robustness.

Time-series cross-validation – A variant of cross-validation that respects temporal ordering, crucial for renewable energy where autocorrelation is strong. The rolling-origin or walk-forward approach expands the training window forward in time, using the most recent observations for validation. This mimics the operational setting where forecasts are generated with all data available up to the current hour. It helps detect performance degradation caused by concept drift, such as changes in turbine wake effects over years.

Leave-one-out cross-validation (LOOCV) – An extreme case where k equals the number of observations. While LOOCV provides an almost unbiased error estimate, it is computationally prohibitive for large renewable datasets, especially when each model training involves deep neural networks or ensemble methods. Practitioners often prefer k-fold with k = 5 or 10 for a reasonable trade-off between computational cost and estimate accuracy.

Performance metrics – Quantitative measures used to compare models. In renewable forecasting, both deterministic and probabilistic metrics are essential.

- Mean Absolute Error (MAE) – The average absolute difference between forecasted and observed values. MAE is intuitive: an MAE of 0.5 MW for a 10 MW wind farm indicates, on average, a 5 % error. It is less sensitive to outliers than squared-error metrics.

- Mean Squared Error (MSE) – The average of squared errors. MSE penalizes larger deviations more heavily, making it useful when extreme under- or over-predictions are costly, such as in balancing market penalties.

- Root Mean Squared Error (RMSE) – The square root of MSE, expressed in the same units as the original variable, facilitating direct interpretation. An RMSE of 2 % of installed capacity is often considered a benchmark for short-term wind forecasts.

- Mean Absolute Percentage Error (MAPE) – The average absolute error expressed as a percentage of the observed value. MAPE can be misleading when observed values approach zero, a common situation for solar output during night hours; therefore, practitioners sometimes use symmetric MAPE (sMAPE).

- Coefficient of Determination (R^2) – The proportion of variance explained by the model. While R^2 is widely reported, it can be inflated by autocorrelation and does not reflect calibration of probabilistic forecasts.

- Skill Score – A relative metric comparing a model against a baseline, often the persistence or climatology forecast. The skill score is defined as 1 – (RMSE_model / RMSE_baseline). Positive skill indicates improvement over the baseline, while negative skill warns of a deteriorating model.

Probabilistic metrics – When forecasts provide a full distribution or prediction intervals, additional evaluation criteria are needed.

- Continuous Ranked Probability Score (CRPS) – A generalization of MAE for probability distributions. Lower CRPS values indicate sharper and more reliable forecasts. In wind-power forecasting, CRPS is frequently used to assess ensemble outputs.

- Reliability diagram – A graphical tool that plots observed frequencies against forecasted probabilities. A well-calibrated solar-irradiance forecast will have points lying close to the diagonal, indicating that a 70 % confidence interval indeed contains the true value about 70 % of the time.

- Sharpness – The concentration of the predictive distribution, independent of calibration. Sharpness is desirable but must be balanced against reliability; overly narrow intervals can lead to frequent miss-coverage.

- Quantile loss (Pinball loss) – Used when models predict specific quantiles (e.g., the 10th and 90th percentiles). The loss penalizes deviations asymmetrically, reflecting the direction of error. This metric is popular for generating prediction intervals in solar-PV power forecasting.

Model selection criteria – Formal rules that help choose among competing models, especially when they differ in complexity.

- Akaike Information Criterion (AIC) – Combines goodness-of-fit with a penalty for the number of parameters. Lower AIC values indicate a better trade-off. AIC is applicable to statistical models such as ARIMA or linear regression, where the likelihood can be computed.

- Bayesian Information Criterion (BIC) – Similar to AIC but imposes a stronger penalty for model complexity, favoring simpler models when the sample size is large. BIC is useful when selecting among multiple weather-regression models for wind-speed prediction.

- Adjusted R^2 – An extension of R^2 that accounts for the number of predictors. It prevents spurious inflation of R^2 when irrelevant features are added. In solar-forecasting, adding redundant satellite channels may

increase raw R^2 but lower adjusted R^2, signaling over-parameterization.

Hyper-parameter tuning – The process of optimizing settings that control model structure rather than model parameters learned from data. Examples include the number of trees in a Random Forest, the learning rate of a gradient-boosting machine, or the number of hidden layers in a neural network. Proper tuning can dramatically improve forecast skill.

- Grid search – Exhaustively evaluates a predefined set of hyper-parameter combinations. For a wind-power model, a grid search might explore tree depths from 3 to 10 and learning rates from 0.01 to 0.1. The exhaustive nature guarantees coverage but can be computationally expensive.

- Random search – Samples hyper-parameter combinations randomly from a defined distribution. Empirical studies show that random search often finds good configurations more efficiently than grid search, especially when only a few hyper-parameters strongly influence performance.

- Bayesian optimization – Builds a probabilistic surrogate model of the hyper-parameter space and selects new points to evaluate based on expected improvement. Tools such as Gaussian-process-based optimizers can reduce the number of model trainings required, which is valuable when each training run involves a large neural network for solar-forecasting.

Early stopping – A regularization technique that halts training when validation error ceases to improve, preventing overfitting. In deep learning for wind-turbine power prediction, early stopping is typically combined with a patience parameter (e.g., stop after 10 epochs without improvement). The saved model at the best validation epoch is then used for final testing.

Regularization – Adding a penalty term to the loss function to discourage overly complex models.

- L1 regularization (Lasso) – Encourages sparsity by shrinking some coefficients exactly to zero. In a linear regression linking solar output to dozens of satellite channels, L1 can automatically select the most informative channels.

- L2 regularization (Ridge) – Penalizes the squared magnitude of coefficients, leading to smaller but non-zero weights. L2 is often used in ridge regression for wind-speed forecasting when all predictors are believed to contain some signal.

- Elastic Net – A combination of L1 and L2 penalties, balancing sparsity and shrinkage. Elastic Net is advantageous when predictors are correlated, a common situation with meteorological variables.

Ensemble methods – Techniques that combine multiple base learners to improve robustness and accuracy.

- Bagging (Bootstrap Aggregating) – Trains each base learner on a bootstrap sample of the data and averages predictions. Random Forests are a classic bagging approach that work well for wind-speed classification and regression because they reduce variance without increasing bias.

- Boosting – Sequentially adds base learners that correct the errors of previous learners. Gradient Boosting Machines (GBM) and XGBoost have become popular for solar-power forecasting due to their ability to capture nonlinear relationships and interactions among weather variables.

- Stacking – Learns a meta-model that combines predictions from heterogeneous base models (e.g., a neural network, a support vector machine, and a decision tree). In a multi-source wind-forecasting system, stacking can fuse outputs from a physics-based NWP model, a statistical ARIMA model, and a machine-learning model, often achieving higher skill than any single component.

Model interpretability – The degree to which a model's internal logic can be understood by humans. In renewable energy, interpretability is important for regulatory compliance and for gaining stakeholder confidence.

- Feature importance – Quantifies the contribution of each predictor to model performance. Tree-based models provide built-in importance scores, while permutation importance can be applied to any model. For example, a solar-forecast model may reveal that cloud-cover index and surface temperature dominate the prediction, guiding data-collection priorities.

- Partial dependence plots (PDP) – Visualize the marginal effect of a single feature while averaging out others. A PDP for wind-direction in a turbine power model can illustrate how output drops when wind comes from the side, reflecting turbine yaw limitations.

- SHAP values – A unified framework for attributing predictions to individual features based on game-theoretic concepts. SHAP is increasingly used to explain deep-learning forecasts of photovoltaic output, helping operators understand why a model predicts a sudden dip in generation.

Probabilistic forecasting – Instead of a single point estimate, the model outputs a full probability distribution or a set of quantiles. Probabilistic forecasts are essential for markets where penalties depend on confidence intervals, such as the Australian National Electricity Market's "reliability-margin" product.

- Quantile regression – Directly predicts specific percentiles of the target distribution. A wind-farm operator may request the 10th and 90th percentile forecasts to plan reserve procurement. Quantile regression loss functions penalize under- and over-predictions asymmetrically, aligning model training with business objectives.

- Ensemble prediction systems (EPS) – Generate multiple forecasts by perturbing initial conditions or model physics, commonly used in Numerical Weather Prediction (NWP). The resulting ensemble spread provides a natural measure of forecast uncertainty, which can be calibrated using statistical post-processing methods such as Bayesian Model Averaging.

Calibration – Adjusting the raw probabilistic forecasts so that predicted probabilities match observed frequencies. For solar-irradiance EPS, raw ensemble spreads are often under-dispersive; calibration methods

like reliability-based scaling or Gaussian copula can correct this, improving CRPS scores.

Sharpness vs. reliability trade-off – Sharp forecasts are narrow but may be unreliable; reliable forecasts are well-calibrated but may be overly broad. The ideal forecast balances both, often evaluated with a "sharpness-reliability diagram" that plots CRPS against ensemble spread.

Domain-specific variables – Renewable-energy forecasting relies on a set of specialized inputs.

- Solar irradiance – Global horizontal irradiance (GHI), direct normal irradiance (DNI), and diffuse horizontal irradiance (DHI). Accurate measurement or satellite-derived estimates are critical for PV output models.

- Wind speed and direction – Typically measured at hub height or extrapolated using the power-law profile. Wind direction is often encoded as sine and cosine components to preserve circular continuity.

- Temperature, humidity, and pressure – Influence air density and turbine performance; they are also used to correct NWP wind fields.

- Capacity factor – The ratio of actual output to rated capacity, used as a normalized target in many forecasting studies.

- Load demand – For integrated forecasting, the net demand after accounting for renewable generation is a key variable, especially in regions with high solar penetration.

Feature engineering – The process of creating informative variables from raw data.

- Lag features – Historical values of the target (e.g., previous hour's wind speed) used as inputs to capture autocorrelation.

- Rolling statistics – Moving averages, variances, or min/max over a window (e.g., 3-hour rolling mean of cloud cover) to smooth noisy observations.

- Interaction terms – Products of variables, such as temperature × wind speed, to capture synergistic effects on turbine power.

- Fourier or wavelet transforms – Decompose time series into frequency components, useful for capturing diurnal cycles in solar generation.

- Spatial aggregation – Averaging or weighted combining of measurements from multiple nearby weather stations to reduce local measurement error.

Data preprocessing – Steps required before model training.

- Missing-value imputation – Techniques such as forward-fill, interpolation, or model-based imputation (e.g., using k-nearest neighbours) to handle gaps in sensor data. In wind farms, missing anemometer data for a

few minutes can be recovered using neighboring turbines' measurements.

- Outlier detection – Identifying and optionally removing anomalous observations that can distort model learning. Statistical tests (e.g., Z-score) or robust methods (e.g., median absolute deviation) are commonly applied.

- Normalization and scaling – Transforming variables to a common range (e.g., min-max scaling) or standardizing to zero mean and unit variance. Neural networks are particularly sensitive to feature scaling, especially when using activation functions like ReLU.

Concept drift – The phenomenon where the statistical relationship between inputs and outputs changes over time. In renewable forecasting, drift can arise from turbine aging, changes in land-use affecting wind flow, or upgrades to satellite sensors. Detecting drift often involves monitoring validation error over successive periods and triggering model retraining when a threshold is exceeded.

Model retraining strategies – Approaches to keep forecasts up-to-date.

- Periodic retraining – Retraining the model on a fixed schedule (e.g., monthly). This is straightforward but may miss sudden drifts.

- Incremental learning – Updating model parameters continuously as new data arrive, common for algorithms like online gradient descent or adaptive filters. Incremental learning is suitable for high-frequency PV forecasting where data streams are continuous.

- Transfer learning – Adapting a model trained on one site or climate regime to another, using a smaller amount of target-site data. For example, a deep-learning model trained on a large dataset from the United States can be fine-tuned with a few months of Thai solar data, reducing the need for extensive local data collection.

Computational considerations – Forecasting pipelines must balance accuracy with runtime and resource constraints.

- Training time – Deep neural networks for multi-hour solar forecasts can require hours of GPU time, whereas a Random Forest may train in minutes on a CPU. Practitioners often benchmark training time alongside accuracy to select a feasible solution.

- Inference latency – The time required to produce a forecast after receiving the latest inputs. Real-time market participation demands low latency; models that need extensive feature extraction (e.g., satellite image processing) may require pre-computed features or model simplification.

- Memory footprint – Large ensembles or deep networks can exceed the memory capacity of edge devices, prompting the use of model compression techniques such as pruning or quantization.

Evaluation workflow – A typical end-to-end process for renewable forecasting projects.

1. Data acquisition – Gather historical weather observations, NWP outputs, satellite imagery, and power measurements.
2. Preprocessing – Clean, impute, and scale the data; engineer lag and statistical features.
3. Dataset split – Partition into training, validation, and test sets using time-aware strategies.
4. Model development – Choose a family of models (e.g., gradient-boosted trees) and define hyper-parameter ranges.
5. Hyper-parameter optimization – Apply random search or Bayesian optimization, evaluating each trial with k-fold time-series cross-validation.
6. Model selection – Compare candidates using a combination of deterministic metrics (RMSE, MAE) and probabilistic scores (CRPS). Use AIC or BIC for statistical models when appropriate.
7. Final training – Retrain the selected model on the full training + validation data.
8. Testing – Compute performance on the untouched test set; report skill scores relative to persistence.
9. Post-processing – Calibrate probabilistic outputs, generate prediction intervals, and assess reliability.
10. Deployment – Package the model, monitor online performance, and schedule periodic retraining.

Practical example: wind-speed forecast

A wind-farm operator wishes to predict 6-hour ahead wind speed at hub height. The data set contains 5 years of hourly NWP wind vectors, surface observations, and turbine-level power output. The workflow proceeds as follows:

- The dataset is split using a rolling-origin strategy, with the most recent 6 months held out as the test set.
- Lag features (t-1, t-2, t-3) and moving-average wind speed over the past 12 hours are created.
- A Gradient Boosting Machine is selected; hyper-parameters (learning rate, max depth, number of estimators) are tuned via Bayesian optimization, each trial evaluated with a 5-fold time-series cross-validation.
- The best model achieves an RMSE of 1.2 $m\,s^{-1}$ on validation folds, a skill score of 0.18 over persistence, and a CRPS of 0.85 $m\,s^{-1}$ for probabilistic forecasts.
- The model is retrained on the full training + validation data, then tested on the hold-out set, yielding an RMSE of 1.3 $m\,s^{-1}$ and a calibrated 90 % prediction interval coverage of 92 %.
- SHAP analysis reveals that NWP wind speed at 850 hPa, surface temperature, and the 12-hour rolling mean are the top contributors, guiding future data-collection efforts.

Practical example: solar-PV power forecast

A utility company needs day-ahead PV generation estimates for a 200 MW solar park in Thailand.

- Historical GHI, DNI, DHI, temperature, and power output (15-minute resolution) are aggregated to hourly intervals.

- Cloud-cover indices are derived from geostationary satellite imagery; these are combined with NWP forecasts to form a hybrid input set.
- A Long Short-Term Memory (LSTM) network with two layers (64 and 32 units) is chosen. Early stopping with a patience of 8 epochs prevents overfitting.
- The hyper-parameters (learning rate, dropout rate) are explored using random search; each configuration is evaluated with a 4-fold rolling-origin cross-validation.
- The final model yields a MAE of 0.45 MW (0.23 % of capacity) on the validation folds and a sMAPE of 4.2 % on the test set.
- Post-processing includes quantile regression to produce the 10th and 90th percentile forecasts; the resulting prediction intervals have a coverage of 88 % (target 90 %).
- Feature importance analysis via permutation shows that satellite-derived cloud-cover contributes 45 % of explanatory power, while GHI contributes 30 %, confirming the value of high-resolution satellite data in tropical climates.

Challenges specific to renewable forecasting

1. Non-stationarity – Weather patterns evolve with climate change, and turbine performance degrades over time. Models must be periodically updated, and evaluation pipelines should incorporate drift detection mechanisms.

2. Extreme events – Storms, fog, or sudden cloud shadows can cause large forecast errors. Traditional error metrics may under-represent these tail events; practitioners therefore supplement RMSE with metrics that emphasize extremes, such as the 95th-percentile absolute error.

3. Spatial correlation – Solar farms often span several kilometers, and wind farms may consist of multiple clusters. Ignoring spatial dependence can lead to underestimation of uncertainty. Kriging, Gaussian processes, or convolutional neural networks can model spatial structure, but they increase computational load.

4. Data quality and availability – In emerging markets, sensor networks may be sparse, and satellite data may have gaps due to cloud cover. Robust imputation and uncertainty quantification become essential to avoid biased forecasts.

5. Interpretability vs. performance – Deep learning models often outperform simpler methods in terms of raw accuracy, yet they are less transparent. Regulatory bodies and grid operators may require explainable models for market settlement; thus, a hybrid approach that blends interpretable statistical components with high-capacity black-box models is sometimes adopted.

6. Regulatory and market constraints – Forecasts must align with market timelines (e.g., 15-minute market intervals) and comply with standards such as the International Energy Agency's "Renewable Energy Forecasting Guidelines." Evaluation protocols therefore need to mimic market submission windows and

account for penalties associated with forecast error.

7. Computational resources – Real-time forecasting for large interconnections may demand parallel processing across multiple nodes. Model selection must therefore consider not only statistical performance but also scalability and ease of deployment on cloud or edge platforms.

8. Integration with dispatch models – Forecasts are often inputs to unit-commitment or economic-dispatch optimizers. Errors in forecasts propagate through these downstream models, affecting system cost and reliability. Sensitivity analysis can quantify the impact of forecast error on dispatch decisions, informing the choice of a more robust forecasting model.

Best practices for evaluation and selection

- Use time-aware cross-validation to respect the temporal ordering of renewable data.
- Report a suite of metrics, including deterministic (MAE, RMSE) and probabilistic (CRPS, reliability) scores, to capture different aspects of forecast quality.
- Compare models against meaningful baselines, such as persistence or climatology, and express improvements as skill scores.
- Perform hyper-parameter optimization with a limited budget, favoring random or Bayesian search over exhaustive grid search for high-dimensional spaces.
- Apply regularization and early stopping to control overfitting, especially when using deep neural networks.
- Conduct post-processing calibration of probabilistic forecasts to ensure reliable prediction intervals.
- Use interpretability tools (feature importance, SHAP) to validate that the model is leveraging physically plausible information.
- Monitor concept drift and schedule retraining or incremental updates as needed.
- Consider computational constraints early in the model selection phase to avoid deploying models that cannot meet real-time latency requirements.
- Document the entire evaluation pipeline, including data splits, random seeds, and software versions, to ensure reproducibility and facilitate peer review.

By mastering the terminology and associated practices described above, learners will be equipped to design, evaluate, and select forecasting models that meet the stringent accuracy, reliability, and operational demands of Thailand’s rapidly expanding renewable-energy sector. The vocabulary provided serves as a reference framework for communicating results, diagnosing issues, and advancing the state of AI-driven renewable forecasting.