

Article

A Physics-Informed Bayesian Framework for Calibrated, Multi-Horizon Forecasting of Solar, Wind, and Hybrid Renewable Generation

Assem Shayakhmetova ¹, Nurbolat Tasbolatuly ^{2,*}, Guldana Taganova ^{2,3,4,*}, Kamalbek Berkimbayev ⁵, Anar Sultangaziyeva ² , Gulnur Karakhanova ², Marat Shurenov ⁶ and Aigul Bissarinova ⁷

¹ Department of Artificial Intelligence and Big Data, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan

² Higher School of Information Technology and Engineering, Astana International University, Astana 010000, Kazakhstan

³ Faculty of Information Technology, L. N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan

⁴ Department of Software Engineering, International Engineering and Technological University, Almaty 050060, Kazakhstan

⁵ Department of Pedagogy and Psychology, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan 161200, Kazakhstan

⁶ Department of IT and Service, “Q” University, Almaty 050026, Kazakhstan

⁷ Department of Information Systems, International University of Information Technologies, Almaty 050040, Kazakhstan

* Correspondence: nurbolat.tasbolatuly@aiu.edu.kz (N.T.); guldana.taganova@aiu.edu.kz (G.T.)

Abstract

Renewable-generation forecasting is a probabilistic time-series problem in which point accuracy alone is insufficient for operational decision-making. This study proposes a physics-informed Bayesian framework for multi-horizon forecasting of solar, wind, and total renewable generation. The task is formulated as a 15-dimensional target-horizon problem covering three generation families and five forecast horizons: H1, H3, H6, H12, and H24. A leakage-safe data construction protocol generates causal predictors from meteorological observations, generation history, calendar cycles, lagged and rolling statistics, ramp descriptors, and physics-informed transformations. PI-BHTF partitions the 422-dimensional predictor space into solar, wind, temporal-calendar, and cross-context components, encodes them through parallel nonlinear branches, and combines the representations using cross-energy gated fusion. Its neural core uses a heteroscedastic predictive head and Monte Carlo dropout to distinguish input-dependent aleatoric uncertainty from epistemic variability, whereas the final hybrid forecasts are calibrated using residual quantiles computed from a chronologically held-out validation segment. Across three prespecified random seeds, the full PI-BHTF achieved a mean MAE of 0.0797 ± 0.0007 on the internal chronological test and 0.0686 ± 0.0006 on locked external SCADA validation. Performance and calibration varied substantially across target-horizon tasks, including marked short-horizon external undercoverage for wind generation. Component-wise ablation supported semantic feature partitioning and the heteroscedastic head, whereas the physics-informed features, cross-energy gate, and physics-consistency loss did not independently reduce aggregate MAE. PI-BHTF should therefore be interpreted as a reproducibly competitive framework that balances multi-horizon accuracy, structured representation, uncertainty estimation, and external transferability rather than as a universally dominant model.



Academic Editors: Carlos Llopis-Albert and Carlos Devece Carañana

Received: 22 July 2026

Revised: 21 August 2026

Accepted: 22 August 2026

Published: 29 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: multi-horizon forecasting; renewable generation; Bayesian forecasting; uncertainty quantification; conformal prediction; calibrated prediction intervals; physics-informed machine learning; solar power forecasting; wind power forecasting; probabilistic time series

MSC: 62M10; 62F15; 65C05; 68T07; 62P30

1. Introduction

The increasing penetration of renewable energy sources has made short-term and multi-horizon generation forecasting an important applied mathematical problem. Solar and wind power are weather-dependent, variable, and only partially controllable. Their uncertainty affects reserve planning, grid balancing, electricity-market operation, and the secure integration of renewable resources into modern power systems. Previous studies have shown that renewable-generation forecasting remains challenging because wind and solar power depend on nonlinear meteorological processes, temporal autocorrelation, seasonal effects, and site-specific operating conditions [1–3].

Solar and wind generation differ substantially in their physical mechanisms and temporal behavior. Photovoltaic generation is mainly driven by irradiance, daylight conditions, temperature, humidity, and seasonal cycles [2]. Its time series often contains strong diurnal regularity and many zero or near-zero values during night-time periods. Wind generation is less regular and is strongly affected by wind speed, air density, turbulence, and local atmospheric conditions [1,3]. The nonlinear relationship between wind speed and available kinetic energy makes wind forecasting particularly sensitive to regime changes and ramp events. When solar and wind generation are combined into total renewable generation, the modeling problem becomes more complex because the total output reflects two different physical processes with different sources of uncertainty.

Many existing forecasting approaches evaluate renewable-generation models mainly through averaged point-error metrics, such as MAE, RMSE, or R^2 . These metrics are useful, but they are not sufficient for a complete assessment of model behavior. A model may achieve a good average error while performing poorly for a specific energy source or forecast horizon. For example, a method may be accurate for one-hour-ahead solar prediction but much less reliable for wind predictions 12 or 24 h ahead. Similarly, a model may perform well on an internal test split but lose accuracy when applied to independent SCADA data from different stations, installed capacities, and meteorological conditions. Therefore, renewable-generation forecasting should be evaluated as a structured target-horizon problem rather than as a single aggregated regression task.

Another key limitation of point forecasting is the lack of calibrated uncertainty information. In operational settings, a single forecast value is rarely enough. System operators need to know how reliable the forecast is and how wide the plausible range of future generation may be. Probabilistic energy forecasting has therefore become an important research direction because it provides information about the distribution of future outcomes rather than only their conditional mean [4]. A useful probabilistic forecast should be both calibrated and sharp: the prediction intervals should cover the observed values with the intended frequency, while remaining as narrow as possible [5]. If the forecast interval is too narrow, the system may underestimate balancing risk; if it is too wide, the forecast becomes less useful for decision-making.

From a mathematical perspective, the problem studied in this paper is a probabilistic multi-output time-series forecasting problem. The objective is to estimate future solar generation, wind generation, and total renewable generation over several lead times. In

this study, the forecasting task is formulated as a 15-dimensional target–horizon problem: three target families—solar, wind, and total generation—are predicted at five forecast horizons, H1, H3, H6, H12, and H24, corresponding to forecasts 1, 3, 6, 12, and 24 h ahead. This formulation allows the model to be evaluated separately for each target and each horizon, making it possible to identify where the method is strong, where it is comparable to baseline models, and where limitations remain. Scale-independent evaluation is important because different target variables may have different magnitudes and naive forecasting difficulty [6]. In addition, performance should be reported separately across forecast horizons to avoid hiding horizon-specific weaknesses.

A major methodological issue in time-series forecasting is data leakage. Leakage may occur when future information enters the predictor space, or when preprocessing parameters are estimated using validation, test, or external data. This problem is particularly serious in multi-horizon forecasting because future targets are explicitly constructed by shifting the time series forward. To avoid optimistic evaluation, all predictors must be causal, all transformations must be fitted only on the training data, and external validation data must remain fully locked until the final assessment. In this work, the data construction procedure follows this principle: the main dataset is split chronologically into training, validation, and internal test subsets, while an independent Chinese State Grid SCADA dataset is used only for locked external validation.

Physics-informed modeling provides a natural way to improve the structure of renewable-generation forecasting models. Instead of treating the input variables as a purely generic tabular representation, physically meaningful transformations can be introduced into the predictor space. For solar generation, irradiance-based variables, daylight indicators, and irradiance–weather interaction terms can reflect the relationship between solar radiation and photovoltaic output. For wind generation, wind-speed transformations, air-density adjustment, and power-curve-inspired terms can represent the nonlinear dependence of wind power on atmospheric conditions. These transformations do not impose a complete physical model, but they provide a useful inductive bias for statistical learning.

Bayesian and probabilistic machine-learning methods are also relevant for this task because renewable-generation uncertainty is not constant across time, weather regimes, and horizons. Monte Carlo dropout has been widely used as a practical approximation to Bayesian inference in neural networks, allowing predictive uncertainty to be estimated through stochastic forward passes [7]. However, approximate Bayesian uncertainty alone does not guarantee calibrated prediction intervals, especially under model misspecification or distribution shift [8]. For this reason, conformal prediction has become a useful post-hoc calibration framework because it can provide finite-sample marginal coverage guarantees under exchangeability assumptions [9,10].

This paper proposes a Physics-Informed Bayesian Hybrid Temporal Forecaster, denoted as PI-BHTF, for multi-horizon forecasting of solar, wind, and total renewable generation. The methodological contribution is architectural and procedural rather than the isolated novelty of heteroscedastic regression, Monte Carlo dropout, or conformal calibration. In one reproducible framework, PI-BHTF combines leakage-safe causal feature construction, semantic partitioning of the input space, joint forecasting of solar, wind, and total generation over five horizons, validation-defined hybrid combination, separate characterization of uncertainty components in the neural core, and locked external SCADA validation. The full 422-feature input is divided into solar, wind, temporal-calendar, and cross-context groups, encoded by separate nonlinear branches, and combined through cross-energy gated fusion.

The proposed framework is evaluated using both internal and external validation. The internal test set is obtained from the main dataset through a chronological split, while

the external validation set is based on independent SCADA data. The model is compared with classical statistical models, regularized linear models, gradient-boosting methods, probabilistic regressors, temporal deep-learning models, and compact neural forecasting architectures inspired by recent multi-horizon time-series models. The inclusion of strong baseline models is important because modern time-series forecasting has advanced rapidly through both classical ensemble methods and neural architectures designed for multi-horizon prediction [11,12]. The evaluation includes point forecasting metrics, interval metrics, calibration metrics, ablation analysis, and paired statistical tests.

The experimental results indicate that PI-BHTF provides a reproducibly competitive balance between point forecasting accuracy, uncertainty characterization, and external transferability. The revised component-wise analysis does not support the claim that every physics-informed component independently reduces point error. Instead, the most consistent architectural evidence is obtained for semantic feature partitioning, while the heteroscedastic head has a distinct role in the uncertainty-aware neural core. Some linear, boosting, recurrent, and compact neural baselines remain stronger for specific targets or horizons, reinforcing the need for target-horizon-specific interpretation.

The main contributions of this study are as follows:

A probabilistic target-horizon formulation is introduced for renewable-generation forecasting, where solar, wind, and total generation are jointly represented as a 15-dimensional multi-output forecasting problem over five forecast horizons.

A leakage-safe data construction procedure is developed to ensure that all predictors are causal, all preprocessing operations are fitted only on the training split, and the external SCADA dataset remains locked until final evaluation.

A physics-informed predictor representation is constructed using generation memory, meteorological covariates, cyclic calendar terms, lagged and rolling statistics, A Bayesian hybrid forecasting architecture, PI-BHTF, is proposed. The model uses semantic feature partitioning, parallel branch encoders, cross-energy gated fusion, and a heteroscedastic neural core. The contribution is evaluated component-wise rather than assuming that every physics-informed block independently improves accuracy.

An uncertainty-estimation and calibration strategy is formulated in which the heteroscedastic head characterizes input-dependent aleatoric uncertainty, Monte Carlo dropout characterizes epistemic variability in the neural core, and the final hybrid prediction intervals are empirically adjusted using chronological validation residuals.

A strict empirical validation design is implemented using internal chronological testing, locked external SCADA validation, ablation analysis, and paired statistical comparison against strong classical, probabilistic, boosting, and neural baselines.

The remainder of this paper is organized as follows. Section 2 reviews related work on renewable-energy forecasting, probabilistic time-series modeling, Bayesian uncertainty estimation, and conformal prediction. Section 3 presents the mathematical formulation of the multi-horizon forecasting problem. Section 4 describes the datasets, leakage-safe data construction, feature engineering, PI-BHTF architecture, and experimental design. Section 5 reports the internal and external validation results. Section 6 discusses the findings, limitations, and implications of the study. Section 7 concludes the paper and outlines future research directions.

2. Related Work

2.1. Renewable-Generation Forecasting

Renewable-generation forecasting has been widely studied because solar and wind power are variable, weather-dependent, and difficult to dispatch directly. Wind power forecasting is especially challenging because wind speed, air density, turbulence, local

topography, and atmospheric regimes interact in a nonlinear way [1,3]. Solar power forecasting has a different structure: photovoltaic output is more strongly associated with irradiance, daylight conditions, temperature, cloud-related variability, and seasonal cycles [2]. These differences suggest that a single generic forecasting representation may be insufficient when solar generation, wind generation, and total renewable generation are predicted jointly.

Earlier studies and reviews have shown that the difficulty of renewable-energy forecasting depends strongly on the forecast horizon [1–4]. Very short-term forecasts often benefit from temporal persistence and lagged generation values, whereas longer horizons require stronger use of meteorological, seasonal, and structural information. This horizon dependence motivates the target–horizon formulation used in the present study. Instead of evaluating one aggregated forecasting task, solar, wind, and total generation are assessed separately at H1, H3, H6, H12, and H24.

The information available at the forecast origin is equally important when interpreting horizon-dependent accuracy. Reviews of wind, photovoltaic, and probabilistic energy forecasting emphasize that very short-term tasks can rely heavily on persistence and recent generation history, whereas the importance of meteorological information increases as lead time grows [1–4]. In operational H12–H24 forecasting, numerical or ensemble weather predictions are therefore often important. The present study deliberately does not replace unavailable future weather with future observations; instead, it evaluates a historical-observation-based forecasting setting. This distinction is revisited as a limitation in Section 6.

Another important issue is model transferability. A model that performs well on one dataset may not retain the same behavior when applied to data from different sites, capacities, climates, or measurement systems. This is particularly relevant for SCADA-based renewable-generation datasets, where station composition and local weather distributions may differ from the training data. Therefore, external validation is an important requirement for assessing whether a forecasting model is robust beyond the original data source.

2.2. Point Forecasting, Probabilistic Forecasting, and Calibration

Many renewable-energy forecasting studies focus primarily on point predictions and compare models using MAE, RMSE, MAPE, or R^2 . These metrics are useful for measuring central forecast accuracy, but they do not describe the reliability of forecast uncertainty. In practical power-system operation, uncertainty is not a secondary issue. A forecast with a low average error may still be operationally risky if its prediction intervals are poorly calibrated or too narrow.

Probabilistic energy forecasting addresses this limitation by estimating a predictive distribution or prediction interval rather than only a conditional mean [4]. A good probabilistic forecast should be calibrated and sharp: it should cover future observations with the intended empirical frequency while avoiding unnecessarily wide intervals [5]. This calibration–sharpness trade-off is important for renewable generation because forecast errors are heteroscedastic. Error variance can change with time of day, season, weather conditions, generation level, and forecast horizon.

Scale-independent evaluation is also important when several target variables are analyzed together. Solar, wind, and total generation may have different magnitudes, distributions, and naive forecast difficulty. For this reason, metrics such as MASE are useful because they allow forecasting performance to be compared relative to a simple benchmark rather than only in the raw scale of the target variable [6]. In the present study, point metrics are therefore interpreted together with interval and calibration metrics, rather than as a single isolated ranking.

2.3. Machine-Learning and Deep-Learning Approaches

Classical machine-learning and gradient-boosting methods remain strong baselines for renewable-generation forecasting, particularly when the input space contains engineered meteorological, lagged, rolling, and interaction features. Tree-based boosting methods such as XGBoost, LightGBM, and CatBoost are widely used because they can model nonlinear relationships, feature interactions, and heterogeneous tabular predictors effectively [13–15]. In renewable-energy applications, such models are often competitive because much of the temporal structure can be represented through carefully designed lagged and rolling features.

Deep-learning models provide a different modeling strategy by learning temporal representations from sequential inputs. Recurrent architectures such as LSTM and GRU were developed to capture temporal dependencies and have been widely applied to time-series prediction tasks [16,17]. Temporal convolutional networks provide another approach by using causal and dilated convolutions to model sequential dependencies efficiently [18]. Transformer architectures introduced a self-attention-based alternative to recurrent and convolutional sequence modeling, enabling long-range dependencies to be represented through attention mechanisms [19]. More recent multi-horizon architectures, including the Temporal Fusion Transformer and N-BEATS, have also influenced forecasting research by introducing attention-based, residual, and interpretable sequence-modeling ideas [11,12].

Recent studies have also explored hybrid spatio-temporal architectures for photovoltaic power forecasting. A multi-branch CNN/RNN–Transformer architecture was developed for rare-event-aware PV power prediction [20], whereas a HybridCNN–STTransformer was proposed for sub-hourly multi-horizon quantile forecasting [21]. Together, these studies demonstrate the potential of combining local feature extraction and temporal representation learning for deterministic, rare-event-aware, and probabilistic PV forecasting.

2.4. Bayesian Uncertainty and Conformal Prediction

Bayesian uncertainty estimation is relevant for renewable-generation forecasting because the uncertainty of solar and wind output is not constant. Monte Carlo dropout provides a practical approximation to Bayesian inference in neural networks by retaining dropout during inference and generating stochastic forward passes [7]. This allows the model to estimate predictive variability without training a full Bayesian neural network. Nevertheless, approximate Bayesian uncertainty can be miscalibrated, especially under distribution shift, limited training data, or model misspecification [8].

Conformal prediction provides a complementary calibration mechanism. Instead of relying only on the internal uncertainty estimates of a model, conformal methods use validation residuals to adjust prediction intervals and improve empirical coverage [9,10]. This is useful in renewable-generation forecasting because errors can differ across targets and horizons. For example, wind generation may require wider intervals at medium or long horizons than solar generation, while total generation may exhibit a different uncertainty structure due to the combination of two energy sources.

The present study combines these ideas while keeping their roles distinct. Bayesian heteroscedastic modeling and Monte Carlo dropout characterize aleatoric and epistemic uncertainty in the neural core, respectively. After the final hybrid point forecast is formed, conformal residual correction is applied as a separate validation-based empirical calibration step because the full predictive distributions of the base models are not propagated through the stack.

2.5. Positioning of the Present Study

The reviewed literature shows four important gaps. First, many studies still emphasize point forecasting more than calibrated uncertainty. Second, solar and wind generation are often modeled separately, while total renewable generation is less frequently treated as a structured joint forecasting problem. Third, model quality is often summarized by averaged metrics, which can hide target-specific and horizon-specific weaknesses. Fourth, external validation on independent SCADA data is still less common than internal testing.

This paper addresses these gaps by formulating renewable-generation forecasting as a probabilistic 15-dimensional target-horizon problem. The proposed PI-BHTF model combines physics-informed feature representation, Bayesian heteroscedastic predictive heads, Monte Carlo dropout, and conformal residual calibration. The experimental design evaluates the model on internal chronological testing and locked external validation, with comparisons against strong classical, boosting, probabilistic, and neural forecasting baselines.

3. Mathematical Problem Formulation

This section mathematically formalizes the implemented forecasting pipeline and the PI-BHTF model structure. The forecasting task is defined as a probabilistic multi-output time-series problem with causal predictors, multi-horizon targets, heteroscedastic predictive uncertainty, and empirically adjusted prediction intervals. The detailed data construction procedure, PI-BHTF architecture, and experimental design are described in Section 4.

Let $t \in T$ denote an hourly timestamp. At each time t , the observed renewable-generation and meteorological state is represented as

$$z_t = [s_t, w_t, g_t, m_t, c_t] \quad (1)$$

where s_t , w_t , and g_t denote normalized solar, wind, and total renewable generation, respectively. The total generation variable is constructed from raw solar and wind generation before the final scaling step and is treated as a separate target family. The vector m_t contains meteorological variables, including wind speed, air density, temperature, humidity, and irradiance-related variables, while c_t contains calendar and cyclic time indicators.

The forecasting problem is defined for three target families,

$$\mathcal{M} = \{\text{solar}, \text{wind}, \text{total}\}, \quad (2)$$

and five forecast horizons,

$$\mathcal{H} = \{1, 3, 6, 12, 24\}, \quad (3)$$

corresponding to forecasts 1, 3, 6, 12, and 24 h ahead. For each timestamp t , the multi-horizon target vector is defined as

$$y_t^{\mathcal{H}} = \left[y_{t+h}^{(m)} \right]_{m \in \mathcal{M}, h \in \mathcal{H}} \in \mathbb{R}^{15}. \quad (4)$$

Thus, each supervised instance has a 15-dimensional output formed by three target families and five forecast horizons. This formulation follows the prepared forecasting pipeline, where the output task matrix contains solar, wind, and total generation targets for H1, H3, H6, H12, and H24.

A central requirement of the formulation is causality. The predictor vector at time t must be constructed only from information available at or before t . Let L denote the

maximum lag or rolling-window length used in feature construction. The causal feature operator is defined as

$$x_t = \Phi(z_t, z_{t-1}, \dots, z_{t-L}) \in \mathbb{R}^d, \quad (5)$$

where $\Phi(\cdot)$ includes lagged variables, rolling statistics, ramp descriptors, calendar encodings, missingness indicators, and physics-informed transformations. In the implemented pipeline, the final engineered predictor dimension is $d = 422$, and future targets are shifted only after causal feature construction, which prevents future information from entering the predictor vector.

The supervised dataset is written as

$$\mathcal{D} = \left\{ (x_t | y_t^{\mathcal{H}}) : t \in \mathcal{T}^* \right\}, \quad (6)$$

where $\mathcal{T}^*$ is the set of timestamps for which all forecast horizons in $\mathcal{H}$ are available. Rows with unavailable future targets are excluded from the model-ready matrix.

The main dataset is split chronologically into training, validation, and internal test subsets:

$$\mathcal{D}_{\text{main}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}}, \quad \mathcal{D}_{\text{train}} \prec \mathcal{D}_{\text{val}} \prec \mathcal{D}_{\text{test}}, \quad (7)$$

where $\prec$ denotes temporal precedence. An additional external dataset is used only for locked external validation:

$$\mathcal{D}_{\text{ext}} \cap (\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}}) = \emptyset. \quad (8)$$

No preprocessing parameters, feature filters, model-selection decisions, or calibration parameters are estimated from $\mathcal{D}_{\text{test}}$ or $\mathcal{D}_{\text{ext}}$.

This restriction is necessary to prevent information leakage and optimistic performance estimation.

The forecasting model is defined as a probabilistic mapping

$$f_{\theta} : \mathbb{R}^d \rightarrow \mathbb{R}^{15} \times \mathbb{R}^{15}, \quad (9)$$

where θ denotes the trainable parameters. For each input vector x_t , the model outputs a conditional mean vector and a log-variance vector:

$$f_{\theta}(x_t) = (\mu_{\theta}(x_t), \eta_{\theta}(x_t)) \quad (10)$$

where the two 15-dimensional outputs are the conditional mean and log-variance vectors:

$$\mu_{\theta}(x_t) \in \mathbb{R}^{15}, \quad \eta_{\theta}(x_t) \in \mathbb{R}^{15}. \quad (11)$$

The conditional variance vector is obtained as

$$\sigma_{\theta}^2(x_t) = \exp(\eta_{\theta}(x_t)). \quad (12)$$

The predictive distribution is approximated by a diagonal heteroscedastic Gaussian model:

$$p_{\theta}(y_t^{\mathcal{H}} | x_t) = N\left(\mu_{\theta}(x_t), \text{diag}(\sigma_{\theta}^2(x_t))\right). \quad (13)$$

The diagonal covariance assumption is used as a tractable approximation. Cross-target and cross-horizon dependencies are not modeled through a full covariance matrix; instead, they are represented indirectly through shared feature construction, branch encoders, cross-

energy fusion, and joint multi-output training. This is consistent with the implemented PI-BHTF architecture, which uses a shared representation and two 15-dimensional predictive heads: one for the mean and one for the log-variance.

For each target-horizon component $j \in \{1, \dots, 15\}$, the point prediction is defined as

$$\hat{y}_{t,j} = \mu_{\theta,j}(x_t). \quad (14)$$

The corresponding $(1 - \alpha)$ -level prediction interval is

$$\mathcal{I}_{t,j}^{(1-\alpha)} = [L_{t,j}^{(1-\alpha)}, U_{t,j}^{(1-\alpha)}], \quad (15)$$

where $\alpha = 0.10$ for 90% prediction intervals. The desired calibration property is

$$\mathbb{P}(y_{t,j} \in \mathcal{I}_{t,j}^{(1-\alpha)}) = 1 - \alpha. \quad (16)$$

Equation (16) expresses the nominal calibration target. In the temporally dependent and externally shifted settings considered here, it is not interpreted as an unconditional finite-sample coverage guarantee; empirical coverage is reported separately on held-out chronological and external data.

At the neural-core level, heteroscedastic prediction and Monte Carlo dropout characterize aleatoric and epistemic uncertainty, respectively. The final hybrid PI-BHTF interval does not directly propagate the component predictive distributions through the stack; instead, it is adjusted using residual quantiles from the chronological calibration segment.

The full model-ready input matrix is

$$X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times 422}, \quad (17)$$

and the target matrix is

$$Y = [y_1^{\mathcal{H}}, \dots, y_n^{\mathcal{H}}]^\top \in \mathbb{R}^{n \times 15}. \quad (18)$$

The overall objective is to learn a model that minimizes point prediction error, characterizes input-dependent uncertainty in the neural core, supports empirically adjusted prediction intervals for the final hybrid forecast, and preserves generalization under internal chronological testing and locked external validation.

4. Materials and Methods

4.1. Datasets and Experimental Roles

This study uses two publicly available renewable-generation datasets with different experimental roles. The first dataset is the Wind and Solar Power Generation Dataset from Mendeley Data [22]. It is used as the main modeling source because it contains hourly solar and wind generation records together with relevant meteorological variables, including wind speed, air density, temperature, humidity, ground irradiance, and atmospheric irradiance. This dataset provides the basis for training, validation-based model selection, and internal chronological testing.

The second dataset is the Chinese State Grid Renewable Energy Generation Forecasting Competition dataset [23]. It contains SCADA-based solar and wind generation observations from multiple renewable-generation sites. In this study, this dataset is used only as a locked external validation source. It is not used for training, hyperparameter tuning, feature selection, calibration, or model selection. This separation is necessary because the purpose of the external dataset is to test transferability under a different data source, station composition, generation scale, and meteorological distribution.

The experimental roles of the two datasets are summarized in Table 1.

Table 1. Dataset roles in the experimental design.

Dataset	Temporal Resolution	Role in This Study	Use in Model Development
Mendeley wind–solar dataset [22]	Hourly	Main modeling dataset	Training, validation, internal chronological testing
Chinese State Grid SCADA dataset [23]	15 min, resampled to hourly	Locked external validation	Final external evaluation only
Processed model-ready data	Hourly	Harmonized feature and target matrices	Generated after leakage-safe preprocessing

Before future-target construction, the harmonized Mendeley dataset contains 8752 hourly observations. The external Chinese State Grid source contains 931,695 native 15-min SCADA records: 510,816 solar observations from eight solar sites and 420,879 wind observations from six wind sites. Four 15-min observations are first averaged within each hour at the site level; generation and installed capacity are then summed across solar or wind sites, while meteorological variables are averaged. This procedure yields 17,544 hourly timestamps, of which 17,520 remain after removing rows without a complete H24 target. The detailed harmonization protocol, site capacities, and capacity-state audit are provided in Supplementary Tables S1, S1a, and S1b.

The use of these two datasets follows the leakage-safe design shown in Figure 1. The main dataset is split chronologically into training, validation, and internal test subsets, whereas the external SCADA dataset remains locked until the final evaluation stage. This protocol follows the prepared pipeline, where the Mendeley data are used for chronological train–validation–internal test splitting and the Figshare/Chinese State Grid SCADA data are reserved only for locked external validation.

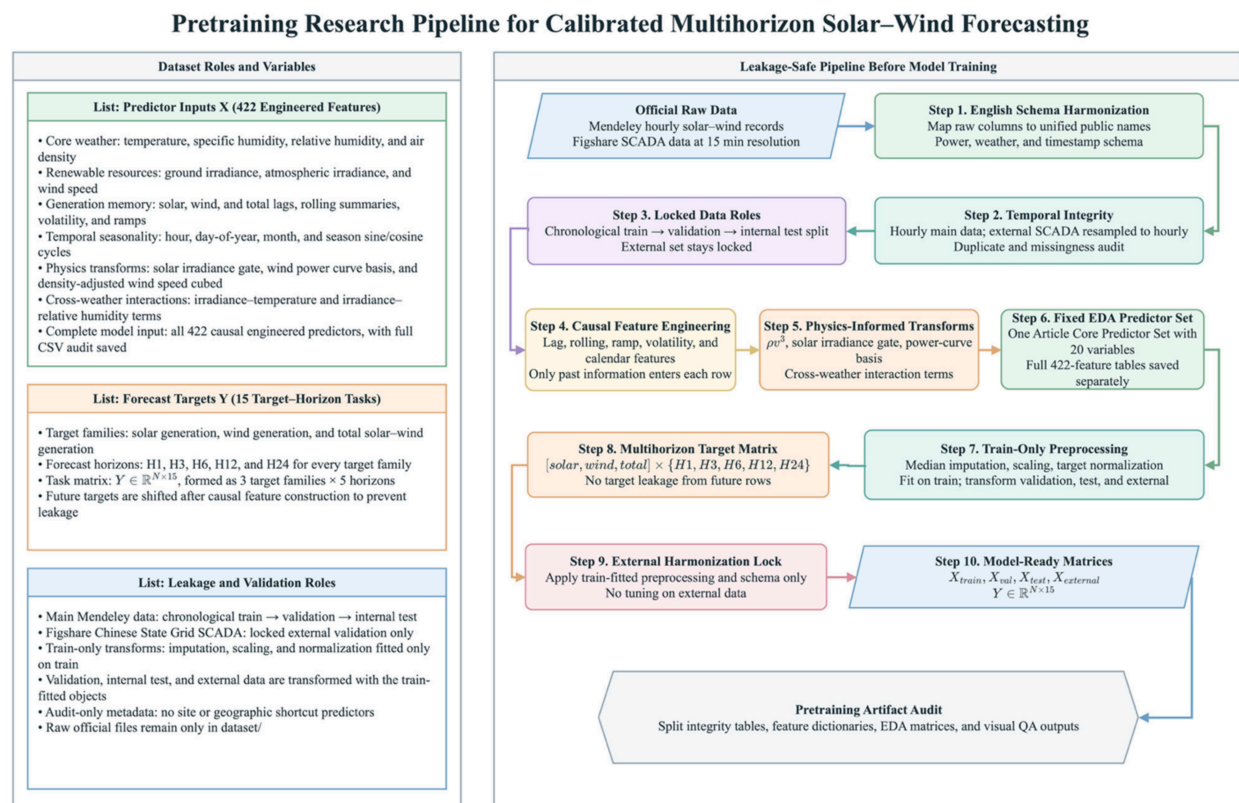


Figure 1. Leakage-safe data construction pipeline for calibrated multi-horizon solar–wind forecasting.

The pipeline includes schema harmonization, temporal integrity checks, locked data roles, causal feature engineering, physics-informed transformations, train-only preprocessing, multi-horizon target construction, external harmonization lock, and model-ready matrix generation.

4.2. Leakage-Safe Data Construction

A central methodological requirement of the study is to prevent information leakage. The data preparation procedure was therefore designed so that no future observation, internal test information, or external validation information could affect training or model selection.

The raw data were first harmonized into a unified English schema. The working variables included timestamp, solar power, wind power, total power, wind speed, air density, temperature, humidity, ground irradiance, atmospheric irradiance, and derived calendar indicators. The external SCADA data were resampled from 15-min frequency to hourly resolution in order to match the temporal frequency of the main dataset.

After harmonization, the main dataset was split chronologically into training, validation, and internal test subsets. No random shuffling was used. This is important because random splitting in time-series forecasting may mix past and future observations and lead to overly optimistic performance estimates. The validation subset was used within the predefined optimization, checkpoint-selection, hybrid-selection, and calibration protocol. Prespecified architecture and fixed hyperparameter settings were not selected using the internal test or locked external validation data.

All preprocessing operations that may influence model performance were fitted only on the training data. These operations included train-defined clipping thresholds, median imputation, scaling, normalization, and feature filtering. The fitted preprocessing objects were then applied without refitting to the validation, internal test, and external validation subsets. The 1309 validation observations were also kept chronological: the first 785 observations were used for optimization and hybrid combination, whereas the final 524 observations were used for final hybrid-configuration selection and residual-quantile calibration. The internal test and locked external SCADA data were not used to compute calibration quantiles or to refit preprocessing parameters.

The model information set contains only values available no later than the forecast origin t : current observations, historical lags, shifted rolling statistics, ramp descriptors, and calendar features. Neither future observed meteorological variables at $t + h$ nor numerical weather-prediction trajectories for H1–H24 are used. The experiment therefore represents multi-horizon forecasting from historically available information, not a complete operational day-ahead system driven by future numerical weather forecasts. This restriction is particularly important for interpreting H12 and H24 results.

4.3. Feature Engineering and Target Construction

This subsection defines the complete causally admissible predictor set and the construction of the 15 target-horizon combinations. All feature transformations are formed before future-target shifting and use only information available no later than the forecast origin.

The physics-informed representation is formed directly from meteorological variables observed at forecast origin t and does not use information from future time points.

For ground solar irradiance G_t , temperature T_t , specific humidity q_t , relative humidity RH_t , wind speed v_t , and air density ρ_t , the implemented physics-informed transformations are defined as follows:

$$x_t^{(IT)} = G_t T_t, \quad (19)$$

$$x_t^{(Iq)} = G_t q_t, \quad x_t^{(IRH)} = G_t RH_t, \quad (20)$$

$$x_t^{(v\rho)} = v_t \rho_t, \quad x_t^{(v^2)} = v_t^2, \quad x_t^{(v^3)} = v_t^3, \quad (21)$$

$$x_t^{(\rho v^3)} = \rho_t v_t^3. \quad (22)$$

A smooth solar-irradiance gate is defined as

$$g_t^{(s)} = \frac{1}{1 + \exp[-0.02(G_t - 20)]}, \quad (23)$$

and the nonlinear wind power-curve basis as

$$g_t^{(w)} = \frac{1}{1 + \exp[-(v_t - 7)]}. \quad (24)$$

The following regime indicators are additionally used:

$$r_t = \mathbb{I}(G_t \leq 20), \quad \mathbb{I}(v_t \leq 3), \quad \mathbb{I}(v_t \geq 12), \quad \mathbb{I}(RH_t \geq 0.80). \quad (25)$$

These transformations do not replace a deterministic photovoltaic-panel or wind-turbine model. Their role is to introduce physically motivated structure into the statistical predictor space.

The final predictor space contains 422 engineered features. These features were constructed from several information groups:

- meteorological variables;
- renewable-resource variables;
- historical generation values;
- lagged and rolling statistics;
- ramp and volatility descriptors;
- cyclic calendar indicators;
- physics-informed transformations;
- cross-weather interaction terms.

The meteorological group includes temperature, humidity-related variables, air density, wind speed, and irradiance-related variables. The generation-memory group includes lagged solar, wind, and total generation values, together with rolling summaries and ramp indicators. The temporal group represents hour, day-of-year, month, and seasonal cycles using cyclic encodings. The physics-informed group includes irradiance-based solar transformations and wind-speed/air-density transformations inspired by the nonlinear structure of wind power. The cross-weather group includes irradiance–temperature and irradiance–humidity interaction terms.

The feature engineering stage is not intended to impose a complete physical model of renewable generation. Rather, it introduces physically meaningful inductive bias into the statistical learning problem. Solar-related features emphasize irradiance and daylight behavior, while wind-related features emphasize wind speed, air density, and power-curve-like nonlinearities. This is important because the two energy sources have different physical mechanisms and different uncertainty structures.

For the main article, the full 422-dimensional predictor space is used in modeling. A smaller article-core subset may be used only for visualization and interpretability. It does not replace the complete model input. The pipeline explicitly records the full engineered feature space and the feature dictionaries as part of the pretraining artifact audit.

The forecasting task is defined as a multi-output target–horizon problem. Three target families are considered: solar generation, wind generation, and total renewable generation. For each target family, five forecast horizons are constructed: H1, H3, H6, H12, and H24,

corresponding to forecasts 1, 3, 6, 12, and 24 h ahead. This gives 15 target–horizon tasks in total.

The target matrix is constructed after causal feature engineering. This order is important: if target shifting were performed before feature construction without proper safeguards, future information could enter lagged or rolling predictors. In the implemented pipeline, future targets are shifted only after the causal predictor rows have been created, which prevents target leakage.

The final model-ready data dimensions and chronological split boundaries are summarized in Table 2.

Table 2. Model-ready split structure and target–horizon design.

Split	Rows	Start	End	Role
Train	6109	1 January 2019 08:00:00	12 September 2019 20:00:00	Model fitting
Validation	1309	12 September 2019 21:00:00	6 November 2019 09:00:00	Chronological optimization/hybrid selection and calibration (785 + 524 observations)
Internal test	1310	6 November 2019 10:00:00	30 December 2019 23:00:00	Final internal evaluation
External locked	17,520	1 January 2019 00:00:00	30 December 2020 23:00:00	Final external validation

The final predictor matrix contains 422 engineered features, and the target matrix contains 15 target–horizon variables. The chronological split boundaries and matrix dimensions follow the model-ready data summary in the technical document.

4.4. Proposed PI-BHTF Model

4.4.1. Model Overview

The proposed model, denoted as PI-BHTF, is a Physics-Informed Bayesian Hybrid Temporal Forecaster designed for calibrated multi-horizon renewable-generation forecasting. The model receives a leakage-safe engineered predictor vector $x_t \in \mathbb{R}^{422}$ and produces a 15-dimensional output corresponding to solar, wind, and total renewable generation at horizons H1, H3, H6, H12, and H24.

The architecture is based on four principles. First, the input space is divided into physically meaningful feature groups instead of being processed as a single homogeneous vector. Second, each feature group is encoded by a separate nonlinear branch. Third, solar and wind representations are combined through a cross-energy gate to represent interactions between the two renewable sources. Fourth, the neural core outputs conditional means and heteroscedastic variances for uncertainty characterization; after hybrid point-forecast formation, the final interval is constructed separately using chronological residual calibration.

The internal structure of PI-BHTF is summarized in Figure 2.

The model partitions the 422-dimensional engineered input vector into solar, wind, temporal-calendar, and cross-context branches; applies parallel encoders; uses cross-energy gated fusion; and produces 15-dimensional mean and log-variance outputs. Monte Carlo dropout characterizes epistemic variability in the neural core, whereas the final hybrid 90% interval is obtained separately through chronological residual correction.

PI-BHTF Internal Architecture: Physics-Informed Bayesian Hybrid Temporal Forecaster

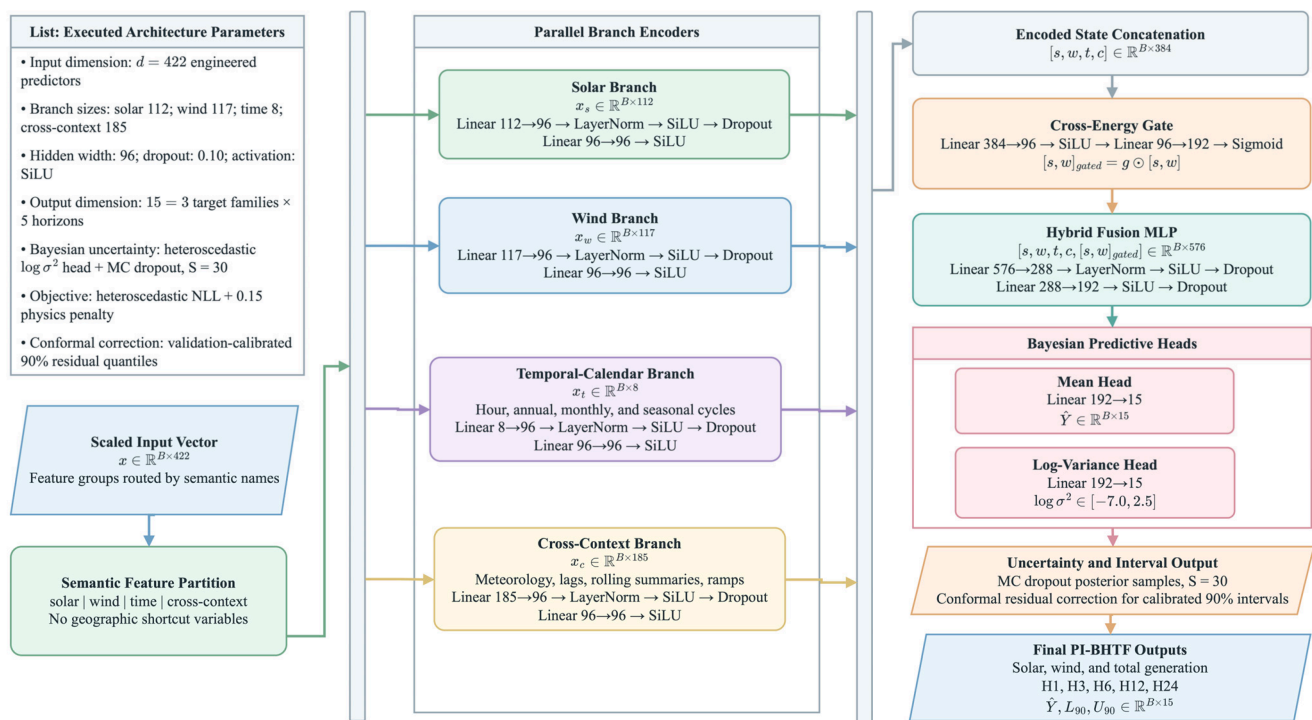


Figure 2. Internal architecture of the proposed PI-BHTF model.

4.4.2. Semantic Feature Partition

The input vector is divided into four disjoint semantic groups:

$$x_t = [x_t^s, x_t^w, x_t^t, x_t^c], \quad (26)$$

The implemented branch dimensions are as follows:

$$x_t^s \in \mathbb{R}^{112}, \quad x_t^w \in \mathbb{R}^{117}, \quad x_t^t \in \mathbb{R}^8, \quad x_t^c \in \mathbb{R}^{185}. \quad (27)$$

The four branch dimensions sum to the complete 422-dimensional input space:

$$112 + 117 + 8 + 185 = 422. \quad (28)$$

Feature routing is deterministic after construction of the complete predictor set. The solar branch contains solar generation, irradiance, daylight-regime variables, and their corresponding lagged and rolling characteristics (112 features). The wind branch contains wind speed, air density, wind generation, and their corresponding lagged and rolling characteristics (117 features). The temporal-calendar branch contains hourly, annual, monthly, and seasonal cyclic features (8 features). The cross-context branch contains the remaining meteorological variables, total-generation lags, rolling summaries, ramp descriptors, and interaction features (185 features). Thus, source-specific solar and wind lagged features are routed to the corresponding specialized branches, whereas general and total temporal descriptors remain in the cross-context branch.

4.4.3. Branch Encoders

Each feature group is passed through a separate nonlinear encoder. For a branch $k \in \{s, w, \tau, c\}$, the encoded representation is defined as

$$h_t^k = E_k(x_t^k), \quad (29)$$

where $E_k(\cdot)$ denotes a branch-specific feed-forward encoder. In the implemented model, each encoder maps its input to a hidden representation of width 96 using linear layers, layer normalization, SiLU activation, and dropout:

$$E_k(x_t^k) = \psi_k(x_t^k; \theta_k), \quad h_t^k \in \mathbb{R}^{96}. \quad (30)$$

The four encoded branch states are concatenated as follows:

$$h_t = [h_t^s, h_t^w, h_t^\tau, h_t^c] \in \mathbb{R}^{384}. \quad (31)$$

4.4.4. Cross-Energy Gated Fusion

To represent interactions between solar and wind components, PI-BHTF uses a cross-energy gate. The gate is computed from the concatenated encoded state:

$$q_t = G(h_t) = \sigma(W_{g,2}\phi(W_{g,1}h_t + b_{g,1}) + b_{g,2}), \quad (32)$$

where $\phi(\cdot)$ is the SiLU activation, $\sigma(\cdot)$ is the sigmoid function.

$$\text{The gate has dimension } q_t \in \mathbb{R}^{192}. \quad (33)$$

The gate is applied to the concatenated solar–wind representation:

$$\tilde{h}_t^{sw} = q_t \odot [h_t^s, h_t^w]. \quad (34)$$

The final fusion vector is constructed as

$$r_t = \begin{bmatrix} h_t, \tilde{h}_t^{sw} \end{bmatrix} \in \mathbb{R}^{576}. \quad (35)$$

The hybrid fusion multilayer perceptron produces

$$u_t = F(r_t) \in \mathbb{R}^{192}. \quad (36)$$

4.4.5. Bayesian Predictive Heads

The fused representation is passed to two predictive heads, one for the conditional mean and one for the log-variance.

The first predictive head estimates the conditional mean vector:

$$\mu_\theta(x_t) = W_\mu u_t + b_\mu \in \mathbb{R}^{15}. \quad (37)$$

The second predictive head estimates the log-variance vector:

$$\eta_\theta(x_t) = W_\eta u_t + b_\eta \in \mathbb{R}^{15}. \quad (38)$$

The conditional variance is obtained by exponentiation:

$$\sigma_\theta^2(x_t) = \exp(\eta_\theta(x_t)). \quad (39)$$

It is important to clarify that PI-BHTF is not a full Bayesian neural network in the strict posterior-inference sense. Rather, it uses a Bayesian approximation based on heteroscedastic predictive heads and Monte Carlo dropout. This is a practical uncertainty-estimation strategy suitable for multi-output forecasting under moderate data size and high-dimensional engineered predictors.

4.4.6. Objective Function

The model is trained by minimizing the heteroscedastic negative log-likelihood over all target-horizon components. For one sample t and one target-horizon component j , the heteroscedastic negative log-likelihood is

$$\ell_{\text{NLL}}(y_{t,j}, \mu_{t,j}, \sigma_{t,j}^2) = \frac{1}{2} \left[\log \sigma_{t,j}^2 + \frac{(y_{t,j} - \mu_{t,j})^2}{\sigma_{t,j}^2} \right]. \quad (40)$$

For the full training set and all 15 target-horizon components, the average heteroscedastic loss is

$$\mathcal{L}_{\text{NLL}} = \frac{1}{15 |\mathcal{D}_{\text{train}}|} \sum_{t \in \mathcal{D}_{\text{train}}} \sum_{j=1}^{15} \ell_{\text{NLL}}(y_{t,j}, \mu_{t,j}, \sigma_{t,j}^2). \quad (41)$$

A physics-consistency penalty is used as a soft structural regularizer between the separately predicted solar, wind, and total targets. The scaling weights are computed exclusively from the power-normalization constants obtained on the training set; they are not trainable parameters.

$$\begin{aligned} \omega_s &= \frac{C_s}{C_s + C_w} = 0.4733720678, \\ \omega_w &= \frac{C_w}{C_s + C_w} = 0.5266279322, \end{aligned} \quad (42)$$

Here the training-derived normalization constants are $C_s = 4342.518250$ and $C_w = 4831.065375$, which yield $\omega_s = 0.473372$ and $\omega_w = 0.526628$, with $\omega_s + \omega_w = 1$. These weights are not learned and are not determined from the internal test or external validation data.

For the horizon set $H = \{1, 3, 6, 12, 24\}$, the physics-consistency term is

$$\mathcal{L}_{\text{phys}} = \frac{1}{B |\mathcal{H}|} \sum_{i=1}^B \sum_{h \in \mathcal{H}} \left[\hat{y}_{i,h}^{\text{total}} - (\omega_s \hat{y}_{i,h}^{\text{solar}} + \omega_w \hat{y}_{i,h}^{\text{wind}}) \right]^2 + \frac{0.10}{BK} \sum_{i=1}^B \sum_{k=1}^K [\max(0, -\hat{y}_{i,k})]^2, \quad (43)$$

where $\hat{s}_{t+h}$, $\hat{w}_{t+h}$, and $\hat{g}_{t+h}$ denote the predicted solar, wind, and total renewable generation at horizon h . The coefficients ω_s and ω_w are used to account for the scale relationship between the component targets and the total-generation target after normalization.

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \lambda_{\text{phys}} \mathcal{L}_{\text{phys}}, \quad \lambda_{\text{phys}} = 0.15. \quad (44)$$

where $\lambda_{\text{phys}} = 0.15$ in the implemented model.

The coefficient $\lambda_{\text{phys}} = 0.15$ was fixed a priori in the main experimental configuration before model-quality evaluation. No grid search or other selection of this coefficient was performed on the basis of validation MAE, the internal test, or locked external validation. With λ_{phys} kept unchanged at 0.15, the chronological validation segment was used only to monitor training, select the saved checkpoint, and apply early stopping. The $\lambda_{\text{phys}} = 0$ configuration was introduced later solely as a component-wise ablation and did not participate in selecting the parameters of the main PI-BHTF model. A hard equality is not imposed because solar, wind, and total generation are modeled as separate normalized target tasks, and the main and external datasets use different power-scaling schemes.

The role of the physics penalty is not to impose a rigid physical equality after normalization. Instead, it acts as a soft regularization term that discourages internally inconsistent forecasts across solar, wind, and total generation outputs. This is important because total generation is treated as a separate target family, but it remains physically related to the component sources.

4.4.7. Monte Carlo Dropout, Neural-Core Uncertainty, and Conformal Calibration

Dropout remains active during inference, and $S = 30$ stochastic forward passes are generated for each input.

For stochastic forward pass r , the neural core produces

$$\left\{ \mu_{\theta}^{(r)}(x_t) \right\}_{r=1}^S, \quad S = 30. \quad (45)$$

For the neural core, two uncertainty sources are distinguished.

The Monte Carlo predictive mean is

$$\bar{\mu}_t = \frac{1}{S} \sum_{r=1}^S \mu_t^{(r)}. \quad (46)$$

Epistemic variance is estimated from the variability of the conditional means:

$$\sigma_{\text{epi},t}^2 = \text{Var}_{r=1,\dots,S} \left(\mu_t^{(r)} \right). \quad (47)$$

Input-dependent aleatoric variance is obtained from the heteroscedastic head:

$$\sigma_{\text{ale},t}^2 = \frac{1}{S} \sum_{r=1}^S \exp \left(\eta_t^{(r)} \right). \quad (48)$$

The total predictive uncertainty of the neural core is summarized as

$$\sigma_{\text{pred},t} = \sqrt{\sigma_{\text{epi},t}^2 + \sigma_{\text{ale},t}^2}. \quad (49)$$

These quantities characterize the uncertainty-aware neural core. They should not be confused with the final hybrid PI-BHTF interval because the full probabilistic distributions of the individual base models are not propagated through the final stack.

Calibration is performed on a chronologically held-out validation segment.

The 1309 validation observations were kept in chronological order. The first 785 observations were used for model optimization and hybrid combination, whereas the final 524 observations formed the calibration subset $\mathcal{D}_{\text{cal}}$, which was used for final hybrid-configuration selection and residual-quantile calibration. Neither the internal test set nor the locked external SCADA dataset was used to compute these quantiles or for recalibration.

$$\mathcal{D}_{\text{val}} = \mathcal{D}_{\text{opt}} \cup \mathcal{D}_{\text{cal}}$$

where $|\mathcal{D}_{\text{opt}}| = 785$, $|\mathcal{D}_{\text{cal}}| = 524$.

For each target-horizon component j , absolute validation residuals are computed on the chronological calibration segment as

$$e_{t,j}^{\text{cal}} = |y_{t,j}^{\text{cal}} - \bar{\mu}_{t,j}^{\text{cal}}|. \quad (50)$$

The conformal correction is the empirical $(1 - \alpha)$ -quantile of the chronological calibration residuals:

$$q_j^{(1-\alpha)} = Q_{1-\alpha} \left(\left\{ e_{t,j}^{\text{val}} : t \in \mathcal{D}_{\text{val}} \right\} \right). \quad (51)$$

For nominal 90% intervals, $\alpha = 0.10$.

At the neural-core level, the residual-adjusted interval is

$$\mathcal{I}_{t,j}^{0.90} = [\bar{\mu}_{t,j} - q_j^{0.90}, \bar{\mu}_{t,j} + q_j^{0.90}]. \quad (52)$$

The neural-core output consists of the point forecast and interval bounds:

$$(\hat{y}_{t,j}, L_{t,j}^{0.90}, U_{t,j}^{0.90}), \quad j = 1, \dots, 15. \quad (53)$$

The residual correction is fitted only on the final chronological calibration segment and is not refitted on the internal test or external validation data. PICP90 is therefore interpreted as empirical held-out coverage rather than as an unconditional finite-sample guarantee under temporal dependence or cross-domain distribution shift.

4.4.8. Final Hybrid Forecast and Residual Calibration

After the hybrid point forecast has been formed, the complete predictive distributions of the individual base models are not propagated through the stack. The final PI-BHTF interval is therefore constructed as a residual-conformal adjustment around the final hybrid forecast, using the chronological calibration residuals:

Here, $q_{0.90,j}^{\text{gate}}$ denotes the empirical 0.90 quantile of the absolute residuals of the final hybrid point forecast computed on the calibration subset $\mathcal{D}_{\text{cal}}$.

The final hybrid PI-BHTF interval is

$$\mathcal{I}_{t,j}^{\text{PI-BHTF}} = [\hat{y}_{t,j}^{\text{hyb}}, -q_{0.90,j}^{\text{gate}} \hat{y}_{t,j}^{\text{hyb}} + q_{0.90,j}^{\text{gate}}]. \quad (54)$$

This distinction separates the roles of the uncertainty components: the heteroscedastic head represents input-dependent aleatoric uncertainty in the neural core, Monte Carlo dropout represents epistemic variability, and the final conformal layer performs an empirical coverage correction around the hybrid point forecast. Component-wise results for these blocks are reported in Section 5.6.

4.5. Experimental Design

The experimental design compares PI-BHTF with statistical, boosting, probabilistic, and neural models under common chronological splits and a common target-horizon definition. Validation data are used only within the predefined optimization, checkpoint-selection, hybrid-selection, and calibration protocol; the internal test and locked external datasets are not used to tune hyperparameters.

4.5.1. Baseline Models

The proposed PI-BHTF model was evaluated against several groups of baseline models. The comparison was designed to avoid weak-baseline evaluation and to determine whether the proposed framework remains competitive against classical, probabilistic, boosting, and neural forecasting approaches.

The first group included simple forecasting references, such as persistence, seasonal persistence, and rolling-mean baselines. These models served as basic reference points for interpreting the difficulty of the forecasting task.

The second group consisted of linear and regularized statistical models, including Ridge regression, ElasticNet, and Bayesian Ridge. These models were included because renewable-generation time series often contain strong autoregressive and exogenous structure, which can sometimes be captured effectively by regularized linear predictors.

The third group included tree-based ensemble and gradient-boosting models, such as Random Forest, ExtraTrees, HistGradientBoosting, XGBoost, LightGBM, and CatBoost. These methods are strong tabular baselines because they can model nonlinear feature interactions and heterogeneous engineered predictors.

The fourth group consisted of probabilistic and quantile-oriented models, including quantile gradient boosting and related median or interval forecasting variants. These models were used to compare PI-BHTF with uncertainty-aware alternatives.

The fifth group included neural forecasting models such as ResidualMLP, GRU, NBEATStyleCompact, SAINTStyleCompact, and a compact TFT-inspired attention model. The labels ‘NBEATStyleCompact’, ‘SAINTStyleCompact’, and ‘TFTInspiredAttention’ denote controlled compact implementations of architectural ideas rather than claims of full industrial-scale reproduction of the original architectures [11,12].

All main comparative models used the same causally admissible matrices, with the specific exception that QuantileGradientBoostingMedian used 80 predictors selected using training data only. Hyperparameters were fixed before internal-test and external evaluation. LightGBM used 220 trees with learning rate 0.045; CatBoost used 220 iterations, depth 5, and learning rate 0.045; HistGradientBoosting used 160 iterations, learning rate 0.06, and 31 leaf nodes; and XGBoost used 180 trees, depth 4, and learning rate 0.045. Exact architecture and reproducibility details are given in Supplementary Table S2.

For ResidualMLP, GRU, NBEATStyleCompact, and SAINTStyleCompact, the common neural protocol used AdamW with learning rate 10^{-3} , weight decay 10^{-4} , batch size 128, a maximum of 18 epochs, and early stopping with patience 3. PI-BHTF used its prespecified probabilistic training regime with learning rate 0.0015, weight decay 10^{-4} , batch size 256, a maximum of 140 epochs, patience 18, dropout 0.10, and $\lambda_{\text{phys}} = 0.15$. Repeated-seed evaluation used the three fixed seeds 20,260,621, 20,260,622, and 20,260,623; no seed was discarded or selected after inspecting the internal-test or external results. The PI-BHTF value $\lambda_{\text{phys}} = 0.15$ was fixed a priori and was not tuned using validation results; validation was used at this fixed value for training monitoring, checkpoint selection, and early stopping.

4.5.2. Point Forecasting Metrics

Point forecasting performance was evaluated separately for each of the 15 target-horizon components. The main point forecasting metrics were MAE, RMSE, nRMSE, MASE, and R^2 . MAE and RMSE were used to measure absolute and squared-error accuracy. nRMSE was used to compare errors across target variables with different empirical scales. MASE was included because it evaluates model error relative to a naive forecasting reference and is therefore useful when several target variables and horizons are compared jointly. R^2 was used to measure the proportion of explained variance in each target-horizon task.

The metrics were computed first for each individual target-horizon component and then summarized across target families and horizons. This procedure prevents strong performance on one target or horizon from hiding weak performance on another.

4.5.3. Interval and Calibration Metrics

Because the proposed framework produces prediction intervals, interval quality was evaluated together with point accuracy. The main interval and calibration metrics were PICP90, Mean Interval Width, and Winkler score [24].

PICP90 measures the empirical coverage of the nominal 90% prediction intervals. Mean Interval Width measures the average width of the predicted intervals and is used to assess sharpness. The Winkler score combines coverage and sharpness by penalizing intervals that are too wide or fail to contain the observed value.

These metrics are important because high coverage alone is not sufficient. A model can achieve high coverage by producing very wide intervals, but such intervals may be operationally uninformative. Therefore, interval quality was interpreted as a trade-off between empirical coverage and interval sharpness.

4.5.4. Statistical Testing Procedure

Statistical testing was used to assess whether observed differences in forecasting error were stable under paired comparison. For each model pair, absolute errors were compared at the same timestamps and for the same target–horizon components.

The statistical analysis included paired *t*-tests, Wilcoxon signed-rank tests, Diebold–Mariano tests with Newey–West correction [25,26], bootstrap confidence intervals [27], and Holm correction for multiple comparisons [28]. Comparisons were performed on errors matched by timestamp and target–horizon task. The paired mean MAE difference is reported as an effect magnitude together with its bootstrap confidence interval. Newey–West correction addresses temporal dependence in the Diebold–Mariano comparison, while Holm adjustment controls the stated family of multiple pairwise significance tests.

The statistical tests were used to support cautious interpretation of model differences. They were not used to claim universal superiority of one model across all targets, horizons, and validation settings.

4.5.5. Ablation Study

The revised ablation study evaluates PI-BHTF components separately. For the final hybrid model, the tested removals are physics-informed features, semantic branch separation, the cross-energy gate, and the physics-consistency loss. For the neural core, the tested removals are the heteroscedastic head, Monte Carlo stochastic inference, and conformal correction. This design separates architectural and uncertainty-estimation effects that were combined in the earlier ClassicalStack, NeuralAugmented, and NeuralOnly variants.

The ablation results are interpreted using internal MAE, external MAE, external R^2 , external PICP90, and mean interval width. The analysis does not assume that every physics-informed component must independently reduce point error; negative ablation results are retained and interpreted as evidence about the actual role of each component.

The ablation analysis was treated as a robustness check rather than as a standalone ranking exercise.

4.5.6. Evaluation Protocol

The full evaluation protocol was organized as follows. First, all models were trained using the same training split and the same leakage-safe feature matrix. Second, the validation set was used for model selection, hyperparameter decisions, and conformal calibration. Third, final point and interval metrics were computed on the internal chronological test set. Fourth, the locked external SCADA dataset was evaluated only after model selection and calibration had been completed. Fifth, statistical tests were applied to paired error sequences to support cautious interpretation of performance differences.

This protocol ensures that model comparison is based on the same data construction process, the same target–horizon structure, and the same external validation restrictions.

5. Results

5.1. Internal Chronological Test Results

Table 3 summarizes the internal chronological test results across the evaluated target–horizon tasks. The comparison shows that the proposed PI-BHTF family achieved the strongest average performance among the evaluated models on the internal test set. The

best average MAE values were obtained by PI-BHTF ClassicalStack, PI-BHTF NeuralAugmented, and the full PI-BHTF model, all of which remained close to 0.080. The nearest boosting competitors were LightGBM and HistGradientBoosting, with average MAE values of approximately 0.082 and 0.083, respectively.

Table 3. Summary of internal chronological test performance by model family.

Model	Model Family	Mean MAE	Main Interpretation
PI-BHTF ClassicalStack	PI-BHTF variant	0.080	Best internal average performance
PI-BHTF NeuralAugmented	PI-BHTF variant	0.080	Comparable to the best hybrid configuration
PI-BHTF	Proposed model	0.080	Strong overall balance across targets and horizons
LightGBM	Boosting	0.082	Strongest boosting competitor
HistGradientBoosting	Boosting	0.083	Competitive classical ensemble
XGBoost	Boosting	0.084	Strong but slightly weaker than LightGBM
QuantileGradientBoostingMedian	Probabilistic/quantile	0.086	Competitive interval-oriented baseline
ResidualMLP	Neural baseline	0.088	Strong neural baseline
PI-BHTF NeuralOnly	PI-BHTF variant	0.089	Weaker than hybrid PI-BHTF variants

The internal results indicate that the hybrid PI-BHTF configurations are competitive with the strongest internal baselines. The NeuralOnly variant remains competitive but does not match the hybrid configurations. This comparison should not be interpreted as evidence that every constituent component independently reduces error; the separate component-wise ablation in Section 5.6 evaluates those effects directly.

At the horizon level, PI-BHTF variants were strongest for H1–H6 and H24. At H1, PI-BHTF achieved MAE close to 0.022 and R^2 close to 0.965, indicating very strong short-term accuracy. At H3 and H6, PI-BHTF ClassicalStack achieved the lowest MAE values among the compared models. At H12, HistGradientBoosting produced the lowest MAE, but PI-BHTF variants remained competitive and showed favorable RMSE and R^2 behavior. At H24, PI-BHTF ClassicalStack again achieved the lowest MAE, followed closely by the full PI-BHTF model.

These results show that the proposed framework is not uniformly dominant at every horizon, but it provides the strongest overall internal balance across the 15 target–horizon tasks. This is important because the study is designed to evaluate structured target–horizon performance rather than a single aggregated forecasting task.

5.2. Locked External Validation Results

Table 4 reports the locked external validation results. The external validation set represents a more difficult evaluation scenario because it comes from an independent SCADA source and was not used for training, model selection, preprocessing adaptation, or calibration. Therefore, external performance is interpreted as evidence of transferability rather than only in-sample forecasting accuracy.

Table 4. Summary of locked external validation performance by model family.

Model	Model Family	Mean MAE	Main Interpretation
PI-BHTF NeuralAugmented	PI-BHTF variant	0.069	Among the best external models
PI-BHTF ClassicalStack	PI-BHTF variant	0.069	Strong external transferability
SAINTStyleCompact	Compact neural/tabular transformer	0.069	Strongest external competitor
PI-BHTF	Proposed model	0.070	Competitive, but not sole external leader
GRU	Temporal deep learning	0.071	Strong medium-horizon behavior
CatBoost	Boosting	0.072	Strong external boosting baseline
NBEATStyleCompact	Compact neural forecasting	0.074	Competitive neural baseline
QuantileGradientBoostingMedian	Probabilistic/quantile	0.075	Stable interval-oriented baseline
LightGBM	Boosting	0.080	Weaker external transfer than internal test

The external validation results show a more complex pattern than the internal test results. The proposed PI-BHTF model remains among the best-performing approaches, but it is not the unique best model on the external dataset. PI-BHTF NeuralAugmented, PI-BHTF ClassicalStack, and SAINTStyleCompact achieved very similar average MAE values, all close to 0.069. The full PI-BHTF model achieved an average MAE close to 0.070, which remains competitive and clearly better than LightGBM on the external dataset.

The horizon-wise external results show that the best model depends on the forecast horizon. At H1, ElasticNet achieved the lowest external MAE, indicating that short-term external forecasting contains a strong linear autoregressive-exogenous component. At H3, PI-BHTF ClassicalStack achieved the best MAE. At H6 and H12, compact neural and recurrent models, including SAINTStyleCompact and GRU, were stronger. At H24, PI-BHTF ClassicalStack achieved the lowest MAE, followed closely by the full PI-BHTF model.

To summarize the external transferability of the evaluated models, Figure 3 shows the ranking of model families by mean MAE on the locked external validation set. This comparison is important because the external SCADA data were not used for training, preprocessing adaptation, model selection, or calibration. Therefore, the ranking reflects how well each model transfers to an independent renewable-generation data source.

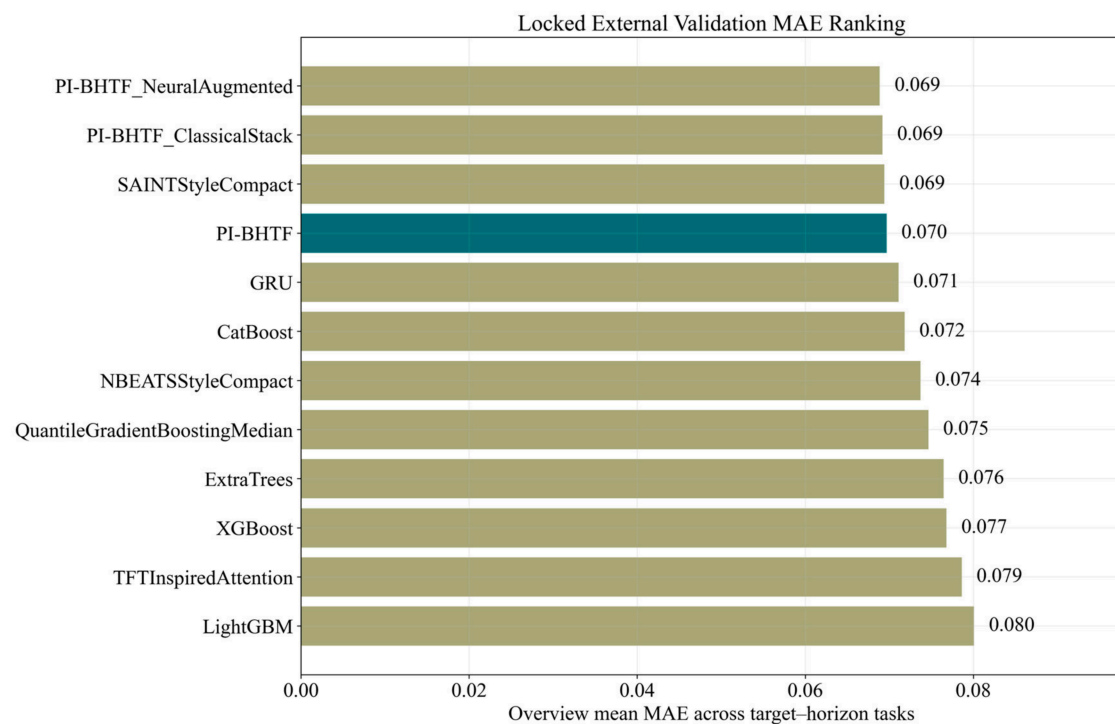


Figure 3. Model ranking by mean MAE on locked external validation.

Figure 3 presents the ranking of evaluated models by mean MAE on the locked external validation set. The results show that PI-BHTF variants remain among the strongest models under external validation. PI-BHTF NeuralAugmented, PI-BHTF ClassicalStack, and SAINTStyleCompact achieved the lowest average external MAE, while the full PI-BHTF model remained very close to the leading group.

This horizon-dependent behavior is important. It shows that external forecasting should not be interpreted through one average metric alone. The proposed model is especially valuable in longer-horizon and total-generation settings, but some baseline models remain stronger for selected short- and medium-horizon tasks.

For the locked external SCADA data, normalized errors were also converted to physical units using the installed capacity associated with the future target timestamp $t + h$. The conversion is

$$e_{t,h}^{MW} = |y_{t,h}^{norm} - \hat{y}_{t,h}^{norm}| C_{t+h}. \quad (55)$$

and the capacity-normalized percentage error is

$$e_{t,h}^{\%C} = 100 \frac{e_{t,h}^{MW}}{C_{t+h}}. \quad (56)$$

Capacity at $t + h$, rather than at the forecast origin, was used for all 15 target–horizon tasks. This conversion is applied only to the external SCADA dataset, which contains physical installed-capacity metadata. For PI-BHTF, external MAE corresponds to 3.9% of installed solar capacity at H1, reaches 11.7% of installed wind capacity at H12, and is 2.3% of total installed capacity at H1. Full target-wise physical-unit results and model averages are reported in Supplementary Tables S5 and S5a and Figures S1–S3.

5.3. Target–Horizon Behavior of PI-BHTF

The target–horizon analysis of PI-BHTF shows different behavior for solar, wind, and total renewable generation. On the internal chronological test set, the model achieved strong H1 performance for all three target families. The R^2 values remained high at H1, reaching approximately 0.959 for solar generation, 0.970 for wind generation, and 0.966 for total generation. At longer horizons, performance decreased, especially for wind generation, which is expected because wind dynamics are more irregular and more sensitive to regime changes.

On the locked external validation set, PI-BHTF achieved the most stable behavior for total renewable generation. External MAE for total generation was approximately 0.023 at H1, increased to approximately 0.064 at H6, and then remained around 0.059–0.061 at H12–H24. Solar MAE was approximately 0.039 at H1, reached approximately 0.086 at H6, and decreased to approximately 0.051 at H24. Wind generation remained the most difficult target family, with MAE values increasing from approximately 0.069 at H1 to approximately 0.117 at H12, before decreasing slightly at H24.

To provide a more detailed view of the proposed model across all target–horizon tasks, Figure 4 presents the internal chronological test performance of PI-BHTF for solar, wind, and total renewable generation. The figure reports the main point and interval metrics across H1, H3, H6, H12, and H24. This visualization makes it possible to identify not only the average behavior of the model, but also the specific targets and horizons where forecasting becomes more difficult.

Figure 4 shows the target–horizon performance profile of PI-BHTF on the internal chronological test set. The heatmaps report MAE, RMSE, nRMSE, MASE, R^2 , and PICP90 for solar, wind, and total generation across H1, H3, H6, H12, and H24. The results show that short-horizon performance is strongest at H1, while errors increase at longer horizons, especially for wind generation.

These results suggest that the proposed architecture is more stable for total renewable generation than for individual wind generation. This can be explained by the fact that total generation combines two energy sources, and the cross-energy representation may help the model learn compensating patterns between solar and wind components.

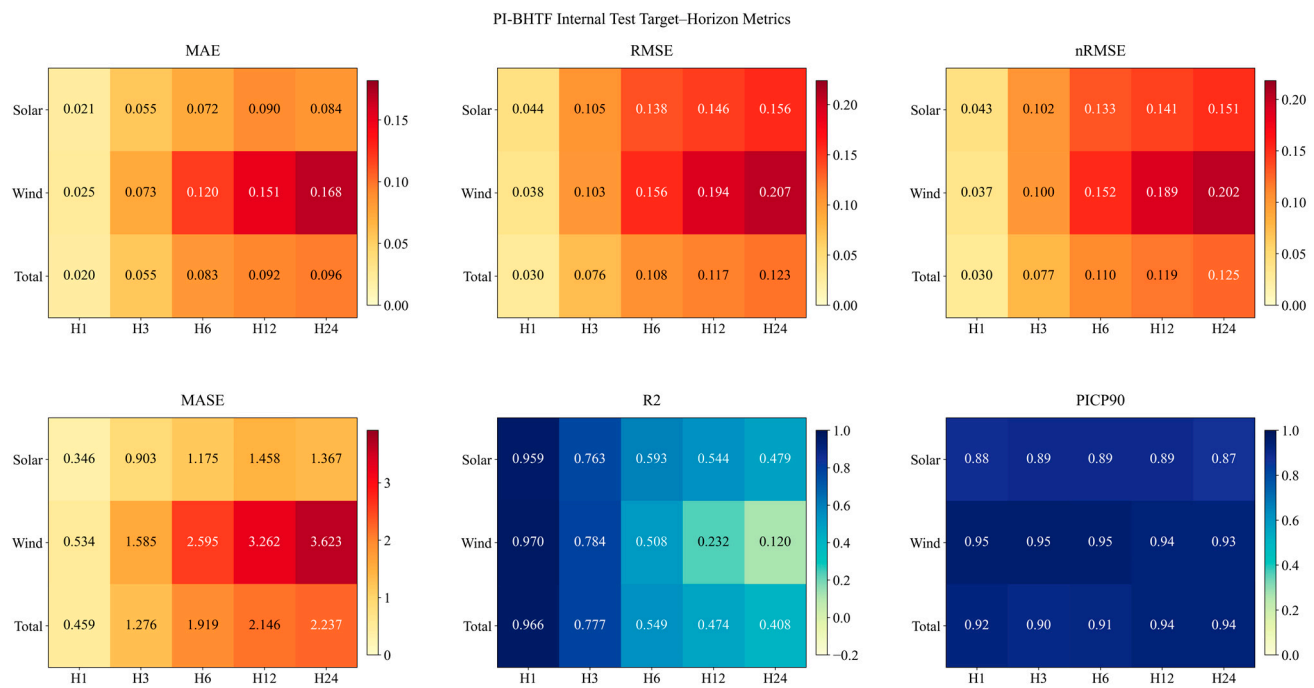


Figure 4. Target–horizon performance of PI-BHTF on the internal chronological test set.

5.4. Interval Coverage and Calibration

The interval results show heterogeneous empirical calibration rather than uniformly calibrated uncertainty. On the internal chronological test, PICP90 is generally between approximately 0.87 and 0.95. On locked external validation, the mean PICP90 of the final PI-BHTF is 0.931648, but coverage differs substantially across target–horizon tasks. These values are interpreted as empirical held-out coverage, not as unconditional conformal guarantees under temporal dependence and domain shift.

The most pronounced external calibration deficit occurs for Wind H1, where PICP90 is approximately 0.65, about 25 percentage points below the nominal 0.90 level. Solar H1 is also undercovered at approximately 0.77. At longer horizons the pattern reverses: Wind H12–H24 reaches approximately 0.99, and Total H12–H24 reaches approximately 1.00, indicating conservative rather than optimally sharp intervals.

The external PICP90 results by target family and forecast horizon are summarized in Table 5.

Table 5. External PICP90 behavior of PI-BHTF by target family and horizon.

Target Family	H1	H3	H6	H12	H24	Interpretation
Solar	0.77	0.94	0.89	0.97	0.97	Undercoverage at H1, conservative at long horizons Strong undercoverage at H1, high coverage later High and conservative coverage
Wind	0.65	0.95	0.97	0.99	0.99	
Total	0.93	0.97	0.98	1.00	1.00	

This pattern shows that fixed residual calibration improves coverage but does not remove temporal dependence or external distribution shift. In the neural-core ablation, removing conformal correction changes external PICP90 from 0.996465 to 0.923059 while reducing mean interval width from 0.656041 to 0.376986; the corresponding Winkler score changes from 0.659115 to 0.454233. Thus, increased empirical coverage is obtained at the cost of wider intervals. Conformalized quantile regression [10], block-based calibration, rolling calibration, and adaptive conformal methods are therefore considered as future alternatives.

5.5. Statistical Testing

Statistical tests were applied to errors matched by timestamp. On the internal chronological test, the mean paired MAE improvement of PI-BHTF relative to LightGBM is 0.001769095462. The paired t -test gives $p = 0.000000330356$, and the Wilcoxon signed-rank test gives $p = 0.000000342344$; after Holm correction, the adjusted value for these comparisons is $p_{\text{Holm}} = 0.000000991069$. The Diebold–Mariano test with Newey–West correction gives $p = 0.049696875681$. The 95% block-bootstrap confidence interval for the mean paired difference is $[-0.000045701421, 0.003578161917]$ and includes zero. The aggregate internal effect is therefore characterized as small rather than as a large practical advantage.

On locked external validation, the mean paired difference relative to SAINTStyleCompact in the saved comparison is -0.000259171734 . The paired t -test gives $p = 0.080818669291$, the Wilcoxon test gives $p = 0.063541241550$, and the Holm-adjusted value is $p_{\text{Holm}} = 0.190623724651$. The Diebold–Mariano test with Newey–West correction gives $p = 0.537440659450$. The 95% block-bootstrap confidence interval $[-0.001072744012, 0.000584046637]$ includes zero. Thus, the saved aggregate comparison does not support a stable external superiority of either model. Full statistical evidence, including test statistics and adjusted p -values, is provided in Supplementary Table S4a.

The component-wise ablation and repeated-seed analyses are reported separately in Sections 5.6 and 5.7 so that architectural effects are not conflated with one-off model rankings.

Together, the statistical results support a cautious interpretation of the relatively small aggregate differences among leading configurations and motivate the component-wise and repeated-seed analyses that follow.

5.6. Component-Wise Ablation Analysis

Table 6 reports separate removals of the requested architectural, physics-informed, and uncertainty components. The results do not support the assumption that every physics-informed component independently improves point accuracy. The clearest positive architectural result is semantic feature partitioning: removing the semantic branches increases external MAE from 0.069217 to 0.071166 and reduces external R^2 from 0.444705 to 0.394838.

Table 6. Component-wise ablation analysis of PI-BHTF.

Scope	Configuration	Internal MAE	External MAE	External R^2	External PICIP90	Mean Width
Hybrid model	Full PI-BHTF	0.080465	0.069217	0.444705	0.931648	0.364367
Hybrid model	Without physics-informed features	0.079579	0.068304	0.462938	0.933916	0.365384
Hybrid model	Without semantic branches	0.080080	0.071166	0.394838	0.926511	0.365539
Hybrid model	Without cross-energy gate	0.079612	0.068861	0.456007	0.933033	0.366095
Hybrid model	Without physics-consistency loss	0.079325	0.067269	0.468789	0.934958	0.364618
Neural core	Full PI-BHTF neural core	0.089000	0.084737	0.112028	0.996465	0.656041
Neural core	Without heteroscedastic head	0.086561	0.094385	−0.081103	0.952538	0.478857
Neural core	Without Monte Carlo	0.087353	0.084137	0.114139	0.994140	0.640413
Neural core	Without conformal correction	0.089000	0.084737	0.112028	0.923059	0.376986

Removing physics-informed features, the cross-energy gate, or the physics-consistency loss does not increase aggregate external MAE; the lowest external MAE among these ablations is 0.067269 without the physics-consistency loss. These components are therefore interpreted as physically motivated structural priors and consistency mechanisms rather than as independently demonstrated sources of point-error reduction. In the uncertainty block, removing the heteroscedastic head substantially worsens external neural-core quality, while conformal correction primarily changes coverage and interval width.

Figure 5 compares mean MAE across the 15 target–horizon tasks for the full PI-BHTF and four architectural ablations. Internal differences remain small, whereas the clearest external degradation occurs when semantic branches are removed. The figure also confirms that removing the physics-informed features, cross-energy gate, or physics-consistency loss does not worsen aggregate external MAE.

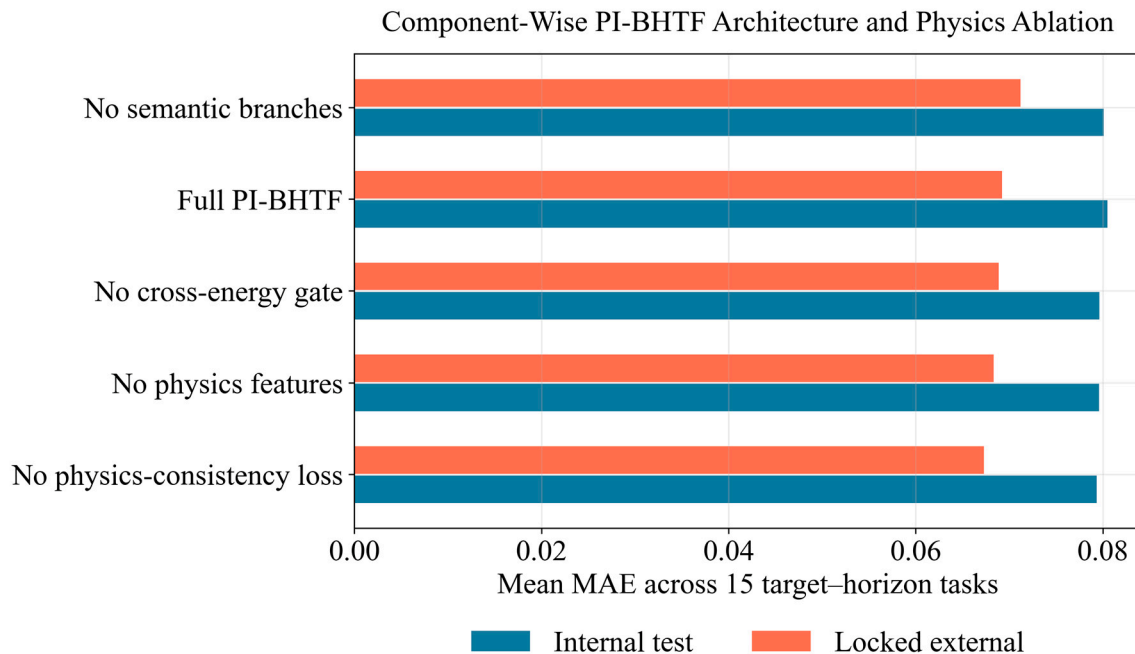


Figure 5. Component-wise analysis of PI-BHTF architectural and physics-informed components.

5.7. Repeated-Seed Stability and Statistical Uncertainty

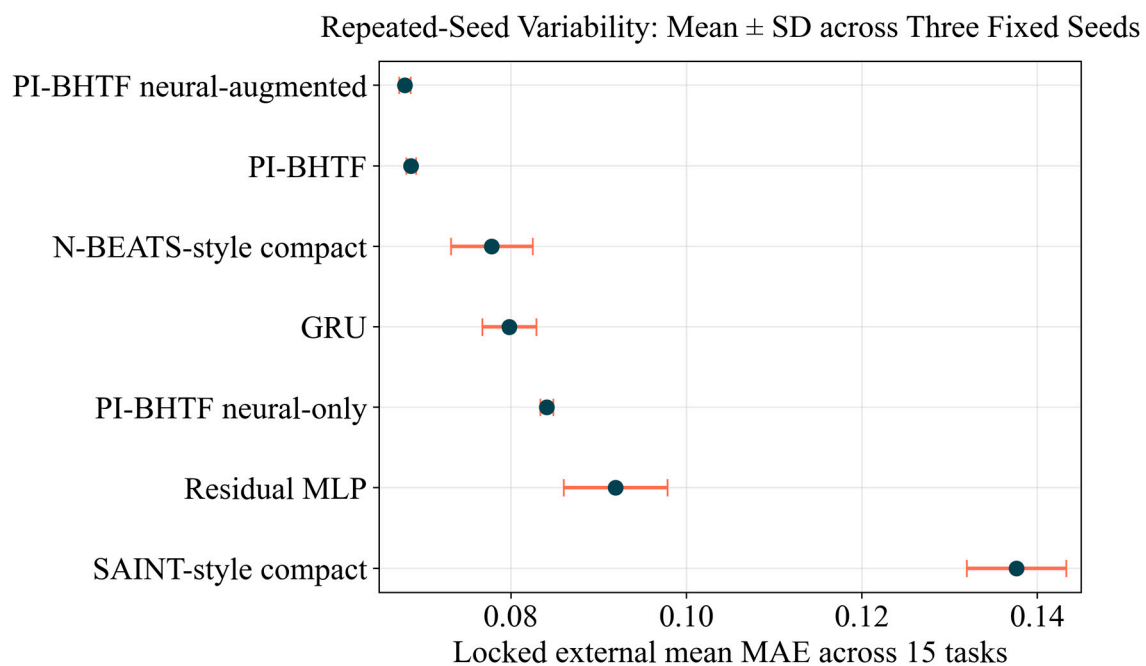
The main neural and hybrid models were re-evaluated using the three prespecified seeds 20,260,621, 20,260,622, and 20,260,623, with no post hoc selection of a favorable run. Table 7 reports the mean $\pm$ standard deviation across these fixed initializations.

Table 7 shows that the full PI-BHTF has low sensitivity to random initialization. The MAE standard deviation is 0.000683 on the internal test and 0.000571 on locked external validation; the corresponding values for PI-BHTF NeuralAugmented are also small (0.000725 and 0.000651). The repeated-seed analysis also changes the interpretation of the original single-seed PI-BHTF NeuralOnly–ResidualMLP comparison: the mean internal MAE is 0.085874 for NeuralOnly and 0.089298 for ResidualMLP, so the small one-run advantage of ResidualMLP does not reproduce as a stable effect. High PICP90 should not be interpreted independently of point error: for example, SAINTStyleCompact has external PICP90 close to 1.00 together with MAE 0.137596 and negative R^2 , whereas PI-BHTF NeuralOnly also has very high coverage but substantially weaker R^2 than the final hybrid PI-BHTF. These results support joint assessment of point accuracy, calibration, and interval sharpness.

Figure 6 shows the mean locked-external MAE and standard deviation across the three fixed seeds. The hybrid PI-BHTF configurations lie in the lowest-error region and have short standard-deviation bars, whereas several comparative neural architectures show higher mean errors and/or larger between-run variability.

Table 7. Repeated-seed variability of the main neural models across three prespecified initializations.

Panel A. Internal Test				
Model	MAE, Mean $\pm$ SD	RMSE, Mean $\pm$ SD	R ² , Mean $\pm$ SD	PICP90, Mean $\pm$ SD
PI-BHTF	0.079674 $\pm$ 0.000683	0.115492 $\pm$ 0.000509	0.611816 $\pm$ 0.003014	0.918524 $\pm$ 0.000982
PI-BHTF NeuralAugmented	0.079660 $\pm$ 0.000725	0.115662 $\pm$ 0.000696	0.610517 $\pm$ 0.004516	0.919559 $\pm$ 0.001838
PI-BHTF NeuralOnly	0.085874 $\pm$ 0.002668	0.123110 $\pm$ 0.001394	0.570143 $\pm$ 0.010287	0.998236 $\pm$ 0.000427
ResidualMLP	0.089298 $\pm$ 0.001499	0.128360 $\pm$ 0.002078	0.546926 $\pm$ 0.009036	0.922375 $\pm$ 0.004813
GRU	0.098225 $\pm$ 0.003015	0.132592 $\pm$ 0.002654	0.520520 $\pm$ 0.020732	0.924003 $\pm$ 0.006359
NBEATSSStyleCompact	0.090003 $\pm$ 0.003113	0.127287 $\pm$ 0.002901	0.556079 $\pm$ 0.017720	0.926497 $\pm$ 0.008795
SAINTStyleCompact	0.197235 $\pm$ 0.009002	0.225034 $\pm$ 0.007217	−0.290157 $\pm$ 0.086028	0.972180 $\pm$ 0.000885
Panel B. Locked External Validation				
Model	MAE, Mean $\pm$ SD	RMSE, Mean $\pm$ SD	R ² , Mean $\pm$ SD	PICP90, Mean $\pm$ SD
PI-BHTF	0.068614 $\pm$ 0.000571	0.093526 $\pm$ 0.000762	0.455494 $\pm$ 0.010536	0.933144 $\pm$ 0.001696
PI-BHTF NeuralAugmented	0.067910 $\pm$ 0.000651	0.092264 $\pm$ 0.000998	0.461956 $\pm$ 0.011198	0.936175 $\pm$ 0.000660
PI-BHTF NeuralOnly	0.084071 $\pm$ 0.000729	0.114064 $\pm$ 0.001840	0.112643 $\pm$ 0.007672	0.995758 $\pm$ 0.001027
ResidualMLP	0.091917 $\pm$ 0.005919	0.120715 $\pm$ 0.006425	−0.045433 $\pm$ 0.123714	0.893865 $\pm$ 0.016569
GRU	0.079798 $\pm$ 0.003093	0.104638 $\pm$ 0.006135	0.267128 $\pm$ 0.063544	0.937457 $\pm$ 0.014091
NBEATSSStyleCompact	0.077805 $\pm$ 0.004647	0.101435 $\pm$ 0.007259	0.300867 $\pm$ 0.114287	0.949830 $\pm$ 0.022263
SAINTStyleCompact	0.137596 $\pm$ 0.005668	0.158391 $\pm$ 0.005728	−0.258401 $\pm$ 0.072153	0.999980 $\pm$ 0.000014

**Figure 6.** Repeated-seed variability: mean $\pm$ standard deviation across three fixed seeds. Dots indicate the mean MAE values, and the horizontal red error bars represent ± 1 standard deviation across the three fixed seeds

The ratio of 422 predictors to 6109 training observations creates a potential overfitting risk. The evaluation therefore does not rely on training error. Train-defined clipping, median imputation, scaling, dropout 0.10, AdamW weight decay 10^{-4} , gradient-norm clipping, chronological validation monitoring, and early stopping are used within the stated protocol. Independent assessment is then performed on 1310 internal-test observations and

17,520 locked external SCADA observations. The low repeated-seed variability of the full PI-BHTF provides an additional stability check.

Computational cost was measured in the saved experimental protocol. Table 8 separates training time from inference time because the latter is directly relevant to periodic forecast updates. The absolute times should not be interpreted as a hardware-normalized benchmark because LightGBM and CatBoost were run on CPU, whereas the neural models were run on MPS.

Table 8. Computational cost in the saved experimental protocol.

Model	Device	Training Time (s)	Passes	Inference Time (s)	Rows
LightGBM	CPU	29.242	1	1.335	20,139
CatBoost	CPU	13.199	1	0.038	20,139
ResidualMLP	MPS	6.285	1	0.730	20,139
GRU	MPS	11.393	1	1.408	20,139
NBEATStyleCompact	MPS	4.435	1	0.821	20,139
SAINTStyleCompact	MPS	173.992	1	8.729	20,139
PI-BHTF, deterministic	MPS	19.056	1	0.181	17,520
PI-BHTF, Monte Carlo	MPS	19.056	30	9.217	17,520

Deterministic PI-BHTF inference requires 0.0103 ms per observation. Thirty stochastic passes increase latency to 0.5261 ms per observation, approximately 51 times the deterministic mode, with 9.217 s required for all 17,520 external observations. The main additional cost of PI-BHTF is therefore associated with uncertainty estimation rather than deterministic point forecasting, and the measured latency remains small relative to the hourly data frequency.

Formal multiplicity adjustment is reported for the saved family of pairwise significance tests in Table S4a; the complete target–horizon tables are interpreted descriptively rather than as 120 separate multiplicity-adjusted hypothesis tests.

6. Discussion

The results show that renewable-generation forecasting does not have a single architecture that is uniformly optimal across all targets and horizons. At H1, much of the predictable signal is carried by current generation, short lags, and the observed meteorological state. Regularized linear models can therefore remain competitive in selected external short-horizon tasks because a more complex nonlinear representation does not necessarily add useful information at very short lead times. At H6–H12, temporal context becomes more important, which is consistent with the stronger performance of some recurrent and compact neural models. At H24, daily periodicity, calendar variables, and longer memory become more influential.

This horizon dependence is consistent with the renewable-forecasting literature, where short-horizon tasks are more strongly associated with persistence and recent generation history, whereas increasing lead time places more weight on meteorological and structural information [1–4]. It is also compatible with multi-horizon modeling approaches such as the Temporal Fusion Transformer and N-BEATS [11,12] and with recent hybrid CNN/RNN–Transformer and HybridCNN–STTransformer photovoltaic forecasting studies [20,21]. PI-BHTF differs in using explicit semantic partitioning of solar, wind, temporal, and cross-context predictors and in evaluating transfer to an independent SCADA domain.

The absence of future numerical weather-prediction inputs is an important limitation of the present setup. Neither future observed meteorological variables nor NWP trajectories are used at H1–H24. The current results therefore characterize forecasting from historically available information rather than a complete operational day-ahead dispatch forecast. This

limitation is especially important for H12 and H24, where future irradiance, wind speed, and atmospheric conditions can materially affect forecast quality. Incorporating numerical or ensemble weather forecasts is a direct direction for future work and may change the relative ranking of models.

PI-BHTF should not be interpreted as a universally error-minimizing model. Across three fixed seeds, the full model has an external MAE of 0.068614 ± 0.000571 , but individual competing configurations remain stronger on specific tasks. The component-wise ablation also narrows the claim about the source of model performance. Semantic feature partitioning has a positive external result: its removal raises external MAE to 0.071166 and reduces R^2 to 0.394838. In contrast, removing the physics-informed features, cross-energy gate, or physics-consistency loss does not worsen aggregate MAE. Accordingly, the term “physics-informed” in PI-BHTF refers to the way physical knowledge is embedded in the model representation and cross-target structure, rather than to a claim that each physics-oriented component independently improves point accuracy. The designation is supported by the explicit irradiance- and wind-based transformations, source-specific semantic routing, cross-energy coupling, and soft consistency regularization between solar, wind, and total outputs. The component-wise ablation shows that these mechanisms should be interpreted as physically motivated structural priors and consistency constraints; their independent contribution to reducing external MAE is not supported by the present results. The uncertainty results likewise require component-specific interpretation. The heteroscedastic head represents input-dependent aleatoric uncertainty in the neural core, while Monte Carlo dropout represents epistemic variability. The final hybrid interval is instead obtained by chronological residual calibration around the hybrid point forecast because full base-model predictive distributions are not propagated through the stack. The neural-core ablation shows that removing the heteroscedastic head worsens external MAE and R^2 , while removing conformal correction primarily changes coverage and width. These results clarify that the three uncertainty-related blocks perform different functions.

Conformal coverage should also be interpreted cautiously for temporally dependent and externally shifted time series. The calibration residuals are taken from the final 524 observations of the chronological validation set, and neither the internal test nor the external SCADA set is recalibrated. Accordingly, PICP90 is reported as empirical held-out coverage rather than as a distribution-free guarantee. Wind H1 external PICP90 of approximately 0.65 is a substantive calibration failure, not a mild deviation; Solar H1 is also undercovered at approximately 0.77. Conversely, Wind H12–H24 approaches 0.99, and Total H12–H24 approaches 1.00, indicating conservative intervals. Conformalized quantile regression [10] and block-based, rolling, or adaptive conformal schemes are relevant alternatives for future work.

The external SCADA evaluation remains important because its station composition, installed capacities, meteorological distribution, and measurement system differ from the main modeling dataset, and no external samples are used for hyperparameter selection, checkpoint selection, preprocessing refitting, or calibration. At the same time, a single external source cannot represent all climates, plant configurations, or grid regimes. The diagonal Gaussian approximation in the neural core also does not explicitly model full cross-target and cross-horizon covariance. These limitations define the scope of the current evidence and motivate broader cross-region validation and multivariate uncertainty models.

Overall, the supported contribution of PI-BHTF is a reproducibly competitive combination of structured multi-horizon forecasting, semantic feature partitioning, uncertainty-aware neural modeling, chronological residual calibration, and independent external validation. The revised analysis deliberately retains the negative ablation findings and the

external calibration deficits so that the claimed contribution matches the evidence rather than implying universal superiority.

7. Conclusions

This study presented PI-BHTF, a physics-informed Bayesian hybrid framework for a 15-dimensional solar–wind–total target–horizon forecasting problem. The revised experiments support a specific rather than universal conclusion: PI-BHTF provides a reproducibly competitive balance among multi-horizon accuracy, structured feature representation, uncertainty estimation, and transfer to independent SCADA data.

Component-wise ablation supports the contribution of semantic feature partitioning and the heteroscedastic head, whereas the physics-informed features, cross-energy gate, and physics-consistency loss do not show independent reductions in aggregate MAE. These negative results are retained because they clarify the demonstrated role of the individual components. Repeated-seed evaluation further shows that the main PI-BHTF result is stable across the three prespecified initializations.

Uncertainty calibration remains heterogeneous. Chronological residual correction improves empirical coverage but does not remove external distribution shift, with a particularly notable deficit for Wind H1 and conservative coverage at several longer horizons. The present H1–H24 setup also uses historically available information only and does not include numerical weather forecasts, which limits its interpretation as a complete operational H12–H24 dispatch forecast.

Future work should therefore prioritize numerical or ensemble weather-prediction inputs, adaptive or block-based conformal calibration, conformalized quantile regression, broader cross-region validation, and multivariate uncertainty models with explicit cross-target dependence. These findings define both the demonstrated scientific contribution of PI-BHTF and the limits of the current implementation.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/math14173104/s1>, Table S1: Harmonization protocol for the main and external datasets; Table S1a: Nominal installed capacities of the external SCADA sites; Table S1b: Available-capacity states in the external model-ready sample; Table S2: Architecture, hyperparameters, and reproducibility protocol for the main comparative models; Table S3: Complete PI-BHTF target–horizon results on internal and locked external validation; Table S3a: Completeness check for the main model target–horizon results; Table S3b: Internal test: solar generation; Table S3c: Internal test: wind generation; Table S3d: Internal test: total generation; Table S3e: Locked external validation: solar generation; Table S3f: Locked external validation: wind generation; Table S3g: Locked external validation: total generation; Table S4: Structure of the complete repeated-seed target–horizon result set; Table S4a: Statistical evidence for the PI-BHTF comparative analysis; Table S5: PI-BHTF error on locked external validation as a percentage of installed capacity; Table S5a: Mean physical error of the main models across the five external horizons; Figure S1: Physical-unit error of solar generation on locked external validation by forecast horizon; Figure S2: Physical-unit error of wind generation on locked external validation by forecast horizon; Figure S3: Physical-unit error of total generation on locked external validation by forecast horizon; Figure S4: Computational cost in the saved experimental protocol.

Author Contributions: Conceptualization, A.S. (Assem Shayakhmetova) and N.T.; methodology, G.T.; software, G.T.; validation, K.B., A.S. (Anar Sultangazyeva) and G.K.; formal analysis, M.S.; investigation, A.B.; resources, A.S. (Anar Sultangazyeva); data curation, M.S.; writing—original draft preparation, N.T.; writing—review and editing, G.T.; visualization, G.T.; supervision, A.S. (Assem Shayakhmetova). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in (1) Mendeley Data at <https://doi.org/10.17632/gxc6j5btrx.1>; (2) Scientific Data at <https://doi.org/10.1038/s41597-022-01696-6>; and (3) Github at <https://github.com/Guldek1987/pi-bhtf-renewable-forecasting> (accessed on 16 July 2026).

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.5 Thinking) and Grammarly (version v.1.2.256.1883) for language editing, grammar checking, and readability improvement. The authors reviewed, verified, and edited all AI-assisted content and take full responsibility for the final version of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PI-BHTF	Physics-Informed Bayesian Hybrid Temporal Forecaster
H1	One-hour-ahead forecast horizon
H3	Three-hour-ahead forecast horizon
H6	Six-hour-ahead forecast horizon
H12	Twelve-hour-ahead forecast horizon
H24	Twenty-four-hour-ahead forecast horizon
PICP90	Prediction Interval Coverage Probability for nominal 90% intervals

References

1. Foley, A.M.; Leahy, P.G.; Marvuglia, A.; McKeogh, E.J. Current methods and advances in forecasting of wind power generation. *Renew. Energy* **2012**, *37*, 1–8. [\[CrossRef\]](#)
2. Antonanzas, J.; Osorio, N.; Escobar, R.; Urraca, R.; Martinez-de-Pison, F.J.; Antonanzas-Torres, F. Review of photovoltaic power forecasting. *Sol. Energy* **2016**, *136*, 78–111. [\[CrossRef\]](#)
3. Zhang, Y.; Wang, J.; Wang, X. Review on probabilistic forecasting of wind power generation. *Renew. Sustain. Energy Rev.* **2014**, *32*, 255–270. [\[CrossRef\]](#)
4. Hong, T.; Pinson, P.; Fan, S.; Zareipour, H.; Troccoli, A.; Hyndman, R.J. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *Int. J. Forecast.* **2016**, *32*, 896–913. [\[CrossRef\]](#)
5. Gneiting, T.; Balabdaoui, F.; Raftery, A.E. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2007**, *69*, 243–268. [\[CrossRef\]](#)
6. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [\[CrossRef\]](#)
7. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; Volume 48, pp. 1050–1059. Available online: <https://proceedings.mlr.press/v48/gal16.html> (accessed on 16 July 2026).
8. Kuleshov, V.; Fenner, N.; Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 2796–2804. Available online: <https://proceedings.mlr.press/v80/kuleshov18a.html> (accessed on 16 July 2026).
9. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, NY, USA, 2005. [\[CrossRef\]](#)
10. Romano, Y.; Patterson, E.; Candès, E.J. Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019. Available online: <https://papers.nips.cc/paper/8613-conformalized-quantile-regression> (accessed on 16 July 2026).
11. Lim, B.; Arik, S.O.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [\[CrossRef\]](#)
12. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020; Available online: <https://openreview.net/forum?id=r1ecqn4YwB> (accessed on 16 July 2026).
13. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
14. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 3146–3154.

- Available online: <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree> (accessed on 16 July 2026).
15. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; pp. 6639–6649. Available online: <https://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features> (accessed on 16 July 2026).
 16. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
 17. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 25–29 October 2014; pp. 1724–1734. [[CrossRef](#)]
 18. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271. [[CrossRef](#)]
 19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008. Available online: <https://papers.neurips.cc/paper/7181-attention-is-all-you-need> (accessed on 22 July 2026).
 20. Taganova, G.; Tussupov, J.; Abdildayeva, A.; Kaldarova, M.; Kazi, A.; Simpson, R.C.; Zakirova, A.; Nurbekov, B. HST–MB–CREH: A Hybrid Spatio-Temporal Transformer with Multi-Branch CNN/RNN for Rare-Event-Aware PV Power Forecasting. *Algorithms* **2026**, *19*, 94. [[CrossRef](#)]
 21. Taganova, G.; Zakirova, A.; Abdildayeva, A.; Nurbekov, B.; Akhayeveva, Z.; Azykanov, T. Sub-Hourly Multi-Horizon Quantile Forecasting of Photovoltaic Power Using Meteorological Data and a HybridCNN–STTransformer. *Algorithms* **2026**, *19*, 123. [[CrossRef](#)]
 22. Liu, Y. Wind and Solar Power Generation Dataset. Mendeley Data. 2024, V1. Available online: <https://data.mendeley.com/datasets/gxc6j5btrx/1> (accessed on 10 July 2026).
 23. Chen, Y.; Xu, J. Solar and wind power data from the Chinese State Grid Renewable Energy Generation Forecasting Competition. *Sci. Data* **2022**, *9*, 577. [[CrossRef](#)] [[PubMed](#)]
 24. Winkler, R.L. A decision-theoretic approach to interval estimation. *J. Am. Stat. Assoc.* **1972**, *67*, 187–191. [[CrossRef](#)]
 25. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [[CrossRef](#)]
 26. Newey, W.K.; West, K.D. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **1987**, *55*, 703–708. [[CrossRef](#)]
 27. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [[CrossRef](#)]
 28. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.