eng

Article

A Digital Decision-Support Framework for Green Hydrogen-Based Steam Production in the Food Industry

Andreas Poyias [1], Panayiotis Mourtopallas [1], Diamanto Platanou [1], Chrysa Politi [2,*], Despoina Georgopoulou [2] and Antonis Peppas [2]

[1] MGD Energy, 10-12 Kifisias Avenue, 15125 Maroussi, Greece; poyias@mgd-energy.gr (A.P.); mourtopallas@mgd-energy.gr (P.M.); platanou@mgd-energy.gr (D.P.)
[2] School of Mining and Metallurgical Engineering, National Technical University of Athens (NTUA), 15780 Athens, Greece; desgeo@metal.ntua.gr (D.G.); peppas@metal.ntua.gr (A.P.)
* Correspondence: chrysapol@metal.ntua.gr

Abstract

The decarbonization of industrial steam production, representing up to 57% of energy use in the food industry, is critical for achieving EU climate neutrality goals. This study developed an integrated digital framework for the research project Hy4GreenSteam to optimize green-hydrogen integration through advanced predictive modeling. The employed LightGBM gradient-boosting algorithms were trained on 68,697 PV power measurements and 57,000 meteorological observations from 2020 to 2022. A "Production-Split" methodology was introduced for 24 h ahead forecasting, segmenting training into high (>2 kW) and low ($\leq$2 kW) production regimes to manage solar heteroscedasticity. Results show the 15 min model achieved an R^2 of 0.868 and the 1 h model an R^2 of 0.832, while the day-ahead model—trained exclusively on information available at forecast issue time—achieved an R^2 of 0.701, a 70% relative improvement over same-time-yesterday persistence. A complementary regime analysis shows that the production regime is predictable with 90.7% accuracy and quantifies the accuracy headroom of regime-specialized models (oracle R^2 0.794). These methods were integrated into a real-time React-based platform that calculates optimal H_2/CH_4 blending; for the reference pilot configuration, driven by measured on-site PV generation, the computed CO_2 emission reduction reaches 34% relative to natural-gas-only operation during high-solar operating intervals. Predictive modeling combined with a Digital Twin interface provides a TRL 6 decision-support solution, demonstrated in a relevant industrial environment, for managing renewable sources in industrial hydrogen applications.

Keywords: hydrogen blending; green hydrogen; steam generation; PEM electrolyzer; PV forecasting; natural gas; CO_2 reduction; decision support system; industrial decarbonization

Academic Editor: A. T. Alexandridis

Received: 24 June 2026
Revised: 4 August 2026
Accepted: 10 August 2026
Published: 15 August 2026

1. Introduction

The industrial sector is currently facing major challenges to reduce its carbon footprint while maintaining competitiveness. Global efforts are currently focused on limiting climate change in order to keep temperature increases within 1.5–2 °C [1]. Despite these goals, 2020 data reveals that reliance on fossil fuels cannot be easily discontinued, as out of a total global energy consumption exceeding 606 million TJ [1,2], only 14% is based on renewable sources. As a result, the European Union (EU) has introduced the "Green Deal", [3] a legislative framework aimed at achieving climate neutrality by 2050. The transition requires measures that should be followed across energy, transport, and agriculture fields. Regarding the industrial sector, this is governed by several key strategies, such as the Circular Economy Action Plan and the EU Chemicals Strategy, which are built upon the Energy Efficiency Directives [4]. These regulations introduce mandatory actions that should be taken in order to increase energy efficiency to at least 32.5% by 2030. The European Union's "Fit for 55" package makes it mandatory to reduce by 55% the greenhouse emissions by 2030, positioning hydrogen as a main method for achieving climate neutrality by 2050 [5].

While the past policies mainly focused on the infrastructure sector, it should be taken into consideration that the industrial sector accounted for 27% of the EU's final energy consumption in 2021. In some countries, such as Norway, the percentage is even higher. Currently, the sector remains fossil-fuel dependent, with natural gas providing 37% of its energy. While countries such as Sweden and Finland lead in renewable integration, countries with remarkable industrial activity like Germany and France remain the largest consumers, due to energy-intensive industries like steel and chemical production [6,7].

The food and beverage industry, specifically, accounts for nearly 12% of total industrial energy use, with a high demand for thermal energy still covered by fossil fuels. The main challenge the sector has to face is that, despite large industrial companies having the specialized staff and capital to adjust to the new measures, the EU's 290,000 food industry SMEs often lack the technical expertise and tools to transition [1]. As a result, new methods for optimizing energy systems should be introduced.

Steam production is a critical area for intervention, as it accounts for one of the biggest amounts of industrial energy use—for instance, 57% in food processing and up to 81% in paper manufacturing. Green hydrogen offers a solution, due to its high energy content per unit of weight (119.97 MJ/kg), which is relatively higher than other conventional fuels. Most significantly, its combustion only produces water and heat, and as a result it is considered as a zero-carbon energy source [6]. While creating hydrogen networks is a long-term goal, the near-term integration of hydrogen blends into existing gas infrastructure is a viable first stage of industrial decarbonization transition.

Recent literature emphasized the fact that, while full electrification is the goal, hydrogen blending serves as a critical transitional technology for existing industrial infrastructure. Studies on combustion dynamics indicate that hydrogen's high flame speed and low volumetric energy density need precise control systems in dual-fuel boilers to maintain thermal efficiency and safety [7–9]. Industrial steam generation has historically relied on the combustion of fossil fuels —especially natural gas, heavy fuel oil and coal—in fire-tube or water-tube boilers operating at temperatures between 150 and 250 °C, depending on process requirements. In industry, steam is widely used as a heat carrier for thermal processes in the 100–200 °C range and is conventionally produced in gas-fired steam boilers [10]. Traditional steam-generation methods are characterized by high carbon emissions and operational costs. Among the transitional strategies, hydrogen blending into the existing natural-gas boiler infrastructure has attracted attention. As the hydrogen fraction increases from 0% to 100% in a natural gas blend fired in an industrial boiler, CO_2 emissions decrease proportionally to zero, and heat recovery can improve thermal efficiency [11]. It should be noted that this proportionality holds when the hydrogen fraction is defined on an energy basis. Due to hydrogen's significantly lower volumetric energy content (10.6 MJ/Nm^3 vs. 35.3 MJ/Nm^3 for methane), a volumetric blend of 30–40% H_2 corresponds to a much smaller share of energy displacement. The platform accounts for this through the energy-equivalent methane displacement formula (Section 2.3.2), which uses LHV ratios to convert hydrogen flow to displaced methane.

The operational complexity of hydrogen blending systems proves the need for advanced digital monitoring and control tools. Advanced industrial technologies, including cloud computing and data analytics platforms, have been increasingly applied to hydrogen energy plants to enhance efficiency, safety and regulatory compliance through real-time data acquisition and predictive control [12]. Supervisory Control and Data Acquisition systems that are integrated with Internet of Things (IoT) sensors and artificial intelligence modules enable dynamic monitoring of electrolyzer performance, renewable energy availability and fuel-blending ratios, thereby reducing operational risk and improving response times [13]. The International Renewable Energy Agency (IRENA) has further highlighted the fact that data-driven operation is essential, given that electricity accounts for approximately 70% of total hydrogen production costs, making accurate forecasting and real-time optimization vital for cost-competitive use [14].

Digital Twin (DT) Technology represents a further evolution in this direction, and enables virtual replication of physical energy systems for continuous simulation and monitoring, as well as optimization. Systematic reviews confirm that DT implementation can yield energy savings of up to 30% and support predictive maintenance strategies that extend equipment lifespan, while contributing to measurable reductions in carbon emissions in industrial environments [15]. Applied to hydrogen production, DTs integrate real-time sensor data with process models to simulate electrolyzer degradation, PV variability and hydrogen storage dynamics, providing operators with decision-relevant insights that conventional engineering systems cannot offer [16].

The intermittent and non-linear nature of renewable energy sources introduces significant forecasting challenges when these are coupled with green-hydrogen production systems. Previous research has explored a wide spectrum of predictive technologies, ranging from classical statistical models to complex deep-learning architectures. The foundational ARIMA framework of Box et al. [17] relies on assumptions of linearity and stationarity, modeling future values as linear combinations of past observations [18]. However, as demonstrated by Erbay [19] and Antonanzas et al. [7], both solar irradiance data and electrochemical processes such as electrolysis efficiency exhibit pronounced non-linearity and are subject to exogenous disturbances including cloud cover, ambient temperature and thermal fluctuations in the stack. These characteristics fundamentally limit the applicability of linear-series models, which systematically underperform under the high-frequency variability and regime-switching behavior characteristic of photovoltaic power output. As established by Erbay [19], inaccurate forecasts lead to electrolyzer underutilization or the need for a costly back-up power system, which increases the Levelized Cost of Hydrogen. According to Antonanzas et al. [7], ensemble learning methods like XGBoost and LightGBM offer superior performance in handling the non-linearities of photovoltaic power output. ML algorithms use gradient boosting to perform internal feature selection. They can map multi-dimensional equations and interactions between variables that are statistically not considered in traditional linear methods. Gradient-Boosting Regression (GBR) has emerged as an alternative, refining predictions by fitting subsequent models to residual errors. However, the main disadvantage that characterizes GBR is the fact that it is often offset by high computational demands and a high sensitivity to hyperparameter tuning, such as learning rates and estimator counts [20]. In contrast, Random Forest (RF) uses an ensemble of decision trees trained on bootstrapped data subsets, offering significant resilience against overfitting and facilitating parallel computation. While RF is highly effective at reducing variance in datasets, its interpretability decreases as the complexity of the forest grows [18].

More advanced frameworks of boosting such as XGBoost and LightGBM have introduced critical optimization of industrial-scale data. XGBoost enhances traditional GBR through the integration of tree pruning and distributing computing capabilities [18]. LightGBM further advances this by employing a leaf-wise tree growth strategy and the histogram-based feature called data binning. These innovations result in faster training cycles and lower memory consumption, making LightGBM particularly suitable for the real-time or day-ahead operational requirements of projects like Hy4GreenSteam [21]. Finally, Support Vector Regression (SVR) remains a prominent choice for handling high-dimensional data through kernel functions [22]. Unlike tree-based ensembles, SVR focuses on defining an optimal hyperplane in a high-dimensional space to minimize generalization error, providing a different but effective approach to mapping the complex non-linear relationships inherent in solar hydrogen systems [23].

This work addressed these challenges by developing an integrated methodology specifically for the food industry. The study demonstrates the transition from fossil fuels to green hydrogen with the use of a system that features a PEM electrolysis unit and a 400 kWth dual-fuel boiler capable of burning natural gas and hydrogen blends. A main outcome of this work is the development of a specialized digital decision-support system (DSS). The platform integrates real-time data acquisition with advanced predictive modeling to optimize hydrogen production and use. This paper presents the technical and experimental validation of the pilot case, which specifically focused on the implementation of a machine-learning forecasting layer achieving leakage-free day-ahead accuracy of $R^2 = 0.701$ (0.868 at 15 min, 0.832 at 1 h), and, finally, the evaluation of a web-based management platform designed to bridge the gap between energy data and industrial decision making.

2. System Overview

2.1. Framework Concept

The framework targets food-industry SMEs that have already deployed on-site renewable generation but lack the technical capacity to translate intermittent renewable availability into a coherent operational strategy for their thermal infrastructure. Within this context, the role of the proposed DSS is to bridge two decision horizons that are conventionally handled by separate tools: the operational horizon, where the recommended H_2/CH_4 blending ratio must respond to the next 15 min of measured photovoltaic output, and the planning horizon, where the production schedule must align with the 24 h ahead availability of green hydrogen. By coupling a probabilistic representation of renewable supply with a deterministic representation of the downstream conversion chain, the framework produces a single, internally consistent set of decision variables—recommended blending ratio, methane displacement, avoided CO_2 emissions and operating cost—that the operator can act on without further analytical effort.

The DSS for industrial steam production (Figure 1) through the blending of hydrogen (H_2) with natural gas, primarily methane (CH_4), integrates the following:

- PV Power Forecasting using Machine Learning (LightGBM).
- Hydrogen Production Modeling using a PEM electrolyzer.
- Fuel-Blending Optimization for steam generation.
- Environmental Impact Assessment, including CO_2 emissions reduction, according to standard inventory methodologies and emission factor approaches [24,25].

2.2. Digital System Architecture

The Hy4GreenSteam decision-support platform was designed as a modular, web-based application that integrates real-time renewable-energy data, machine-learning-based forecasting and process-level blending optimization into a single interface for industrial operators. The platform follows a three-layer architecture: a data acquisition layer that connects to live energy and meteorological sources at the pilot site, a data-driven Digital Twin layer that hosts the predictive models and the energy-flow equations of the PV–electrolzer–boiler chain, and a Digital Twin Dashboard that renders live measurements and twin predictions side-by-side for the operator. This decomposition is in line with the architectural patterns for data-driven Digital Twins surveyed by Ba et al. [15] and Kim et al. [16].

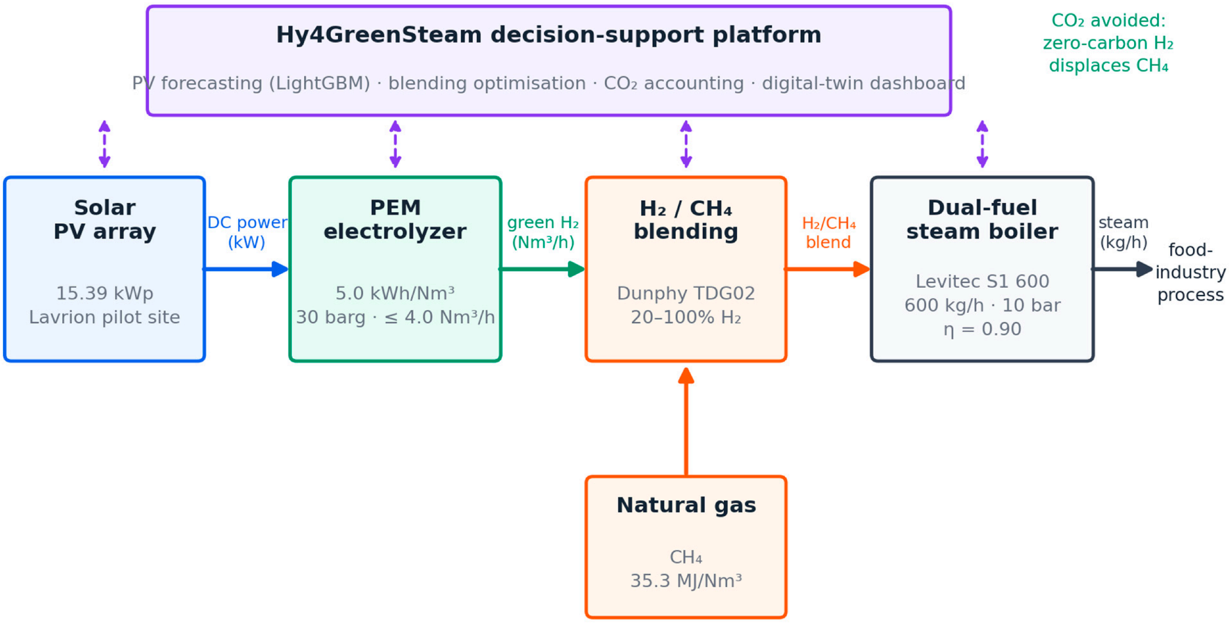

Figure 1. Energy Flow Diagram.

2.2.1. Data Acquisition Layer

Live operational data is ingested from independent sources connected to the pilot installation at Lavrion (Attica, Greece). The Lavrion PV API provides real-time photovoltaic power measurements from the 15.39 kWp reference array. The Lavrion Weather API delivers meteorological observations—global horizontal irradiance, ambient temperature and wind speed—that act as exogenous variables for the forecasting models. Day-ahead weather forecasts are retrieved from the Open-Meteo API for display in the deployed platform; the forecasting models evaluated in this study do not consume them, relying exclusively on measurements available at forecast issue time (Section 3.1.3). The platform also exposes a Lavrion Wind API connected to six wind-turbines (6 × 6 kW PROVEN wind-turbine); this channel allows photovoltaic and wind generation to be aggregated into a single renewable input, although the present study restricts the analysis to the photovoltaic component, for clarity. All incoming streams are timestamp-aligned to Europe/Athens local solar time and are structured in a unified time-series schema. Figure 2 depicts the Pilot Installation System diagram at Lavrion.

2.2.2. Data-Driven Digital Twin

Together, the forecasting models and the energy-flow equations constitute a data-driven Digital Twin of the PV–electrolyer–boiler chain: a virtual replica that is continuously synchronized with live sensor measurements and that can be queried for short-, medium- and long-term predictions of system state [14,16]. Unlike physics-based Digital Twins that simulate component-level dynamics from first principles, the proposed twin learns the input–output relationship of the photovoltaic plant directly from operational data. This approach avoids the need for site-specific calibration of an electrochemical or thermal model, and accelerates deployment in food-industry SMEs, in line with the digital-backbone vision articulated by IRENA [14].

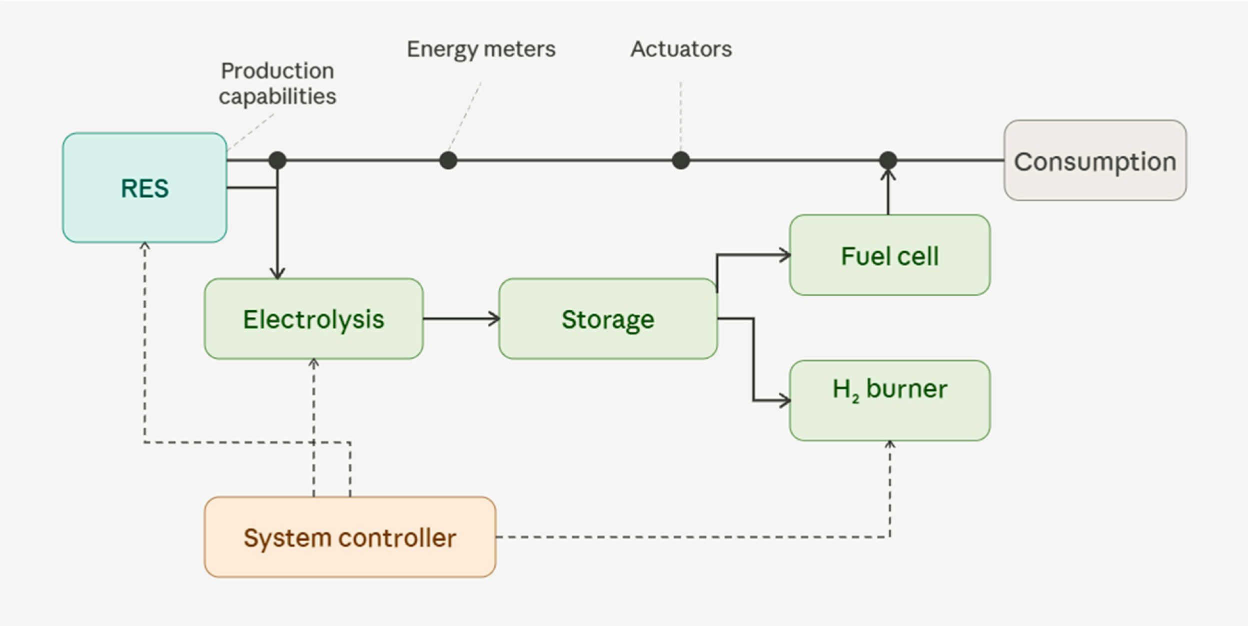

Figure 2. Pilot installation system at Lavrion.

The forecasting component comprises three LightGBM regressors operating at 15 min, 1 h and 24 h horizons (Section 3.1). The day-ahead horizon uses a single LightGBM model trained exclusively on information available at forecast issue time; a complementary Production-Split regime analysis (Section 3.2) quantifies the accuracy headroom available to regime-specialized variants. The forecasted renewable power is forwarded to a deterministic energy-flow model that computes the hydrogen production rate of the PEM electrolyzer, the H_2/CH_4 blending ratio for the 400 kWth dual-fuel boiler, and the resulting CO_2 emissions, using the equations presented in Section 2.3. The Digital Twin, therefore, couples a statistical model of the renewable input with a mass–energy balance of the downstream conversion chain, producing operationally relevant decision variables at every refresh interval.

2.2.3. Digital-Twin Dashboard

The DT state is exposed to the operator through a single-page web dashboard implemented in React (version 18.3.1). The dashboard renders live sensor measurements and twin predictions side-by-side, allowing the operator to verify the alignment between physical state and virtual replica continuously; sustained divergence between the two is itself a diagnostic signal, indicating either sensor anomalies or model drift requiring retraining [15].

The interface is organized around a row of real-time key performance indicator (KPI) cards (instantaneous PV power, hydrogen production rate, steam demand and CO_2 savings) and a coordinated set of operational charts covering the renewable-energy profile, hydrogen production, recommended H_2/CH_4 blending ratio, fuel cost relative to a natural-gas-only baseline, CO_2 emissions and boiler efficiency. A dedicated cases module supports scenario analysis: the operator specifies a steam-demand profile, electrolyzer efficiency, fuel- and electricity-cost assumptions, and emission factors, and the platform evaluates the resulting blending strategy against both the live forecast and the natural-gas baseline, as shown in Figure 3. This functionality is intended to support informed operational and planning decisions by SME operators who lack the in-house engineering capacity to perform such analyses manually [9]. The application is deployed as a containerized service and is accessed through a standard web browser, requiring no client-side installation and no dedicated information-technology infrastructure on the operator's side.

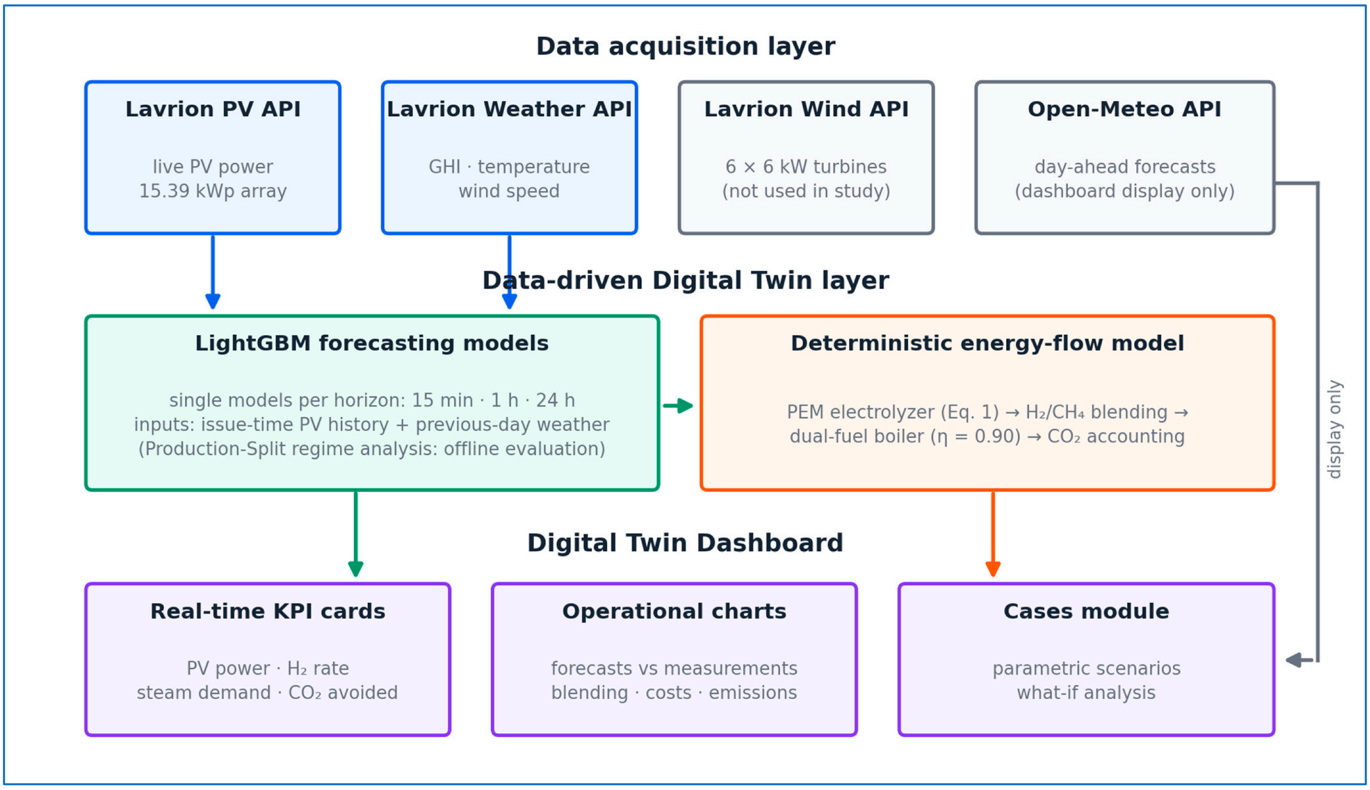

Figure 3. Three-layer architecture of the Hy4GreenSteam platform: live data acquisition, a data-driven Digital Twin combining LightGBM forecasting with a deterministic energy-flow model, and a dashboard exposing the live and predicted state to the operator.

2.3. Energy-Flow Model Overview

The energy flow within this system follows a chain that links electricity generated through renewable sources to industrial steam production. The energy conversion chain is composed of five stages: photovoltaic power generation, water electrolysis, hydrogen production and storage, fuel blending, and steam generation. At each stage, the output of each process constitutes the primary input to the following one, forming a closed mass–energy balance that the Digital Twin evaluates at every interval. The power generated by PV installations is supplied directly to the PEM electrolyzer, which converts it into green hydrogen at a rate governed by the electrolysis efficiency and the specific energy consumption. The hydrogen flow rate thus obtained is forwarded to the fuel-blending model, where it determines the volumetric hydrogen fraction in the H_2/CH_4 mixture delivered to the 400 kWth dual-fuel boiler (Section 2.3.2). The blending ratio is computed to fulfil the steam demand while maximizing the substitution of natural gas; the resulting thermal power balance and the corresponding CO_2 avoided are quantified by the steam-generation model (Section 2.3.3). Because the hydrogen fraction carries a zero CO_2 emission factor, each unit increase in the blending ratio yields a proportional reduction in the carbon intensity of the steam produced, providing the environmental performance indicator reported in Section 4.

2.3.1. Hydrogen Production Modeling

Hydrogen production is calculated from the available photovoltaic power and the specific energy consumption, based on Equation (1) and Table 1:

$$q_{H2} = \frac{P_{pv}}{E_{specific}} \quad (1)$$

Table 1. Input parameters and their values for the PEM electrolyzer model.

Symbol	Description	Value	Unit
q_{H2}	Hydrogen flow rate	Calculated	Nm^3/h
P_{pv}	Available PV power	Variable	kW
$E_{specific}$	Specific energy consumption	5.0	kWh/Nm^3

The value of the specific energy consumption (Especific = 5.0 kWh/Nm^3) corresponds to typical operating parameters of commercial PEM electrolyzers under nominal load. The specific energy consumption $E_{specific}$ = 5.0 kWh/Nm^3 represents the gross electrical input per Nm^3 of H_2 at nominal load, as reported in the manufacturer specifications for the PEM electrolyzer used in the pilot installation. This value reflects the conversion performance of the electrolyzer, and a separate efficiency multiplier is therefore not applied. The available photovoltaic power Ppv is supplied at every time step by the forecasting layer described in Section 3.1, while qH_2 is computed deterministically from Equation 1 and fed forward to the fuel-blending model presented in the next subsection.

2.3.2. Fuel-Blending Model

The lower heating value of the hydrogen–methane fuel mixture is calculated using the volumetric hydrogen fraction:

$$LHV_{mix} = x \cdot LHV_{H2} + (1 - x) \cdot LHV_{CH4} \quad (2)$$

The reference values reported in Table 2 are adopted from standard thermophysical handbooks and IPCC inventory guidelines for natural gas combustion. The volumetric energy density of hydrogen (10.6 MJ/Nm^3) is approximately one-third that of methane (35.3 MJ/Nm^3), which implies that maintaining a constant thermal output as the hydrogen fraction in the blend increases requires a proportional adjustment of the total volumetric fuel flow—a relationship that is captured by the steam-generation model presented in Section 2.3.3. Conversely, the zero CO_2 emission factor of hydrogen translates the blending ratio directly into a CO_2-displacement quantity, which forms the basis of the environmental assessment performed by the platform.

Table 2. Thermophysical and combustion properties of hydrogen and methane used in the fuel-blending model.

Property	H_2	CH_4	Unit
$LHV_{mass\ basis}$	120.3	50.0	MJ/kg
$LHV_{volume\ basis}$	10.6	35.3	MJ/Nm^3
Density	0.0899	0.646	kg/m^3 (standard conditions)
CO_2 emission factor	0	1.96	kg CO_2/Nm^3

Four hydrogen quantities are distinguished in the analysis: hydrogen produced (Equation (1)); hydrogen consumed in the blend; hydrogen stored in the 30-barg buffer of the electrolyzer skid, which decouples production from consumption within the day;

and hydrogen curtailed when production exceeds both the contemporaneous fuel demand and the buffer capacity. The dual-fuel burner fires either pure natural gas or H_2/CH_4 blends within its certified 20–100 vol.% H_2 envelope; when the instantaneously available hydrogen corresponds to a volumetric fraction below 20%, the burner operates on natural gas, and hydrogen accumulates in the buffer until a compliant blend can be sustained. Time-averaged hydrogen fractions below 20 vol.%—as reported for the solar-poor scenarios in Section 4.3.5—therefore reflect averaging over alternating pure-natural-gas and blended intervals, not sub-envelope burner operation.

2.3.3. Steam Generation Model

Steam demand is converted into required thermal power using

$$P_{thermal} = \frac{m_{steam} \times \left(h_{steam} - h_{feedwater}\right)}{3600} \tag{3}$$

The total fuel-mixture flow required to satisfy steam demand is

$$q_{mix} = \frac{P_{thermal} \times 3.6}{\eta_{boiler} \cdot LHV_{mix}} \tag{4}$$

The hydrogen and methane flow rates are calculated as

$$\begin{gathered} q_{H2} = x \times q_{mix} \\ q_{CH4} = q_{mix} - q_{H2} \end{gathered} \tag{5}$$

The methane displaced by hydrogen is calculated on an energy-equivalent basis:

$$q_{CH4,saved} = q_{H2} \times \frac{LHV_{H2}}{LHV_{CH4}} \tag{6}$$

The steam generation model is parameterized using the actual operating specifications of the Levitec Steam Way S1 600 boiler (Levitec, Aspropyrgos, Greece) installed at the Lavrion pilot site. The boiler operates at a working pressure of 10 bar, at which pressure saturated steam has a specific enthalpy of approximately h_{steam} = 2778 kJ/kg. The feedwater enters the boiler at a temperature of 80–100 °C (as specified by the manufacturer), corresponding to a feedwater enthalpy of $h_{feedwater} \approx$ 335–419 kJ/kg; a representative value of 90 °C ($h_{feedwater}$ = 377 kJ/kg) is used in the model, giving an enthalpy rise of $\Delta h \approx$ 2401 kJ/kg. Boiler thermal efficiency is taken as η_{boiler} = 0.90, as specified in the manufacturer's datasheet (±1%). The rated steam output of 600 kg/h at 10 bar corresponds to a thermal capacity of approximately 400 kWth.

3. Methods

3.1. Photovoltaic Power Forecasting

Photovoltaic power output is predicted by a gradient-boosting regression model based on the LightGBM framework (version 4) [26]. LightGBM was selected after a systematic comparison against XGBoost [27], Random Forest [18] and persistence baselines [17], the results of which are presented in Section 4. Among the candidates evaluated, the gradient-boosting estimators offered the best combination of accuracy and computational cost across all horizons, and simple persistence proved insufficient beyond the shortest horizon, in agreement with the analyses of Antonanzas et al. [7] and Erbay [19]. LightGBM was retained as the operational model on the basis of its leaf-wise tree growth and histogram-based feature binning [26], which yielded approximately 40% faster training cycles than

XGBoost on the same dataset while producing comparable predictive accuracy, consistent with the findings of Hokmabad et al. [22] for day-ahead solar forecasting.

3.1.1. Dataset

The model was trained on operational data from the 15.39 kWp photovoltaic installation at the Lavrion pilot site (Attica, Greece), covering the period from January 2020 to January 2022. The dataset comprises 68,697 PV power measurements at 15 min resolution (98% temporal coverage) paired with more than 57,000 meteorological observations (82% coverage). Weather variables include outside temperature (1.17–38.56 °C), Global Horizontal Irradiance (0–1234 W/m^2) and wind speed (up to 12.54 m/s). The two-year horizon ensures that the model is exposed to the full seasonal envelope of Mediterranean solar conditions, which is essential for the construction of a data-driven DT that must remain valid across all calendar regimes.

3.1.2. Pre-Processing

All UTC timestamps were converted to Europe/Athens local solar time, since photovoltaic generation is determined by local solar position rather than universal time. Power values were standardized to kilowatts and irradiance to W/m^2. Night-time records, defined as samples with Global Horizontal Irradiance below 20 W/m^2, were excluded from training, in line with the recommendations of Antonanzas et al. [7]. This decision concentrates the learning signal on productive daylight hours and prevents the long zero-power tail of nocturnal records from biasing the regression toward trivial solutions.

The two data streams are ingested independently, timestamp-aligned (UTC → Europe/Athens) and resampled to a complete 15 min grid using within-interval means; the grid is kept complete so that all lag, rolling, and horizon-label computations correspond to exact time offsets. Records missing either of the two mandatory meteorological variables (GHI and ambient temperature) are excluded from training rather than imputed, to avoid fabricating irradiance values; this, together with sensor outages, accounts for the 82% meteorological coverage. Wind speed, an auxiliary variable, is imputed with its median where missing, and physically implausible records (negative GHI; temperature outside $-40\ldots+70$ °C) are treated as missing. Feature rows with undefined lag or rolling features at series boundaries or across gaps are dropped.

3.1.3. Feature Engineering

A compact set of engineered features was constructed to capture the temporal and meteorological structure of solar generation. Cyclical encodings (sine and cosine transformations of hour-of-day and day-of-year) preserve the circular topology of time without introducing artificial discontinuities at midnight or year change. Two autoregressive lag features (15 and 30 min) exploit the short-term autocorrelation induced by cloud-movement patterns [20]. Rolling statistics—45 min and 90 min moving averages and a 90 min rolling maximum—provide a smoothed representation of the underlying signal that distinguishes transient fluctuations from sustained trends. Lag features at 15, 30 and 45 min were also generated for ambient temperature, irradiance and wind speed, capturing the temporal persistence of the atmospheric state [7].

Table 3 summarizes the complete feature set of each horizon, together with its availability at prediction time; the day-ahead models use only issue-time PV history and previous-day observed weather, so no target-day meteorological information enters any model.

Table 3. Model inputs per forecasting horizon and their availability at prediction time.

Feature Group	Horizons	Known at Prediction Time
Cyclical time encodings (hour, day-of-year); weekend indicator	All	Yes (calendar)
PV power lags (15, 30 min); rolling statistics (45/90 min mean, 90 min max)	All	Yes (measured at issue time)
GHI, ambient temperature, wind speed at issue time	15 min, 1 h	Yes (measured at issue time)
Previous-day weather: same-time-yesterday, morning/noon/evening samples, daily max/mean/min, two-day trends (GHI, temperature, wind)	24 h	Yes (measured $\geq$ 24 h before target)
Previous-day PV persistence (same-time-yesterday, daily max/mean)	24 h	Yes (measured $\geq$ 24 h before target)
Season index	24 h	Yes (calendar)

3.1.4. Model Configuration and Validation

Hyperparameters of every model were selected by grid search over n_estimators (50–250), max_depth (3–10), learning_rate (0.05–0.2), num_leaves (20–100) and subsample/colsample_bytree (0.8–0.9), using TimeSeriesSplit cross-validation with three folds and negative mean absolute error as the scoring metric; the selected values of all models are listed in Table 4. The dataset was partitioned chronologically into 80% training and 20% test sets; this temporal split preserves causality and prevents look-ahead leakage, which is essential when the resulting model is to be deployed as part of a Digital Twin operating on streaming data.

Table 4. Selected LightGBM hyperparameters per model (random seed 42; remaining parameters at library defaults).

Parameter	15 min	1 h	24 h Single	24 h Split High	24 h Split Low
n_estimators	50	100	50	150	50
max_depth	6	9	9	10	4
learning_rate	0.1	0.05	0.1	0.05	0.1
num_leaves	31	31	70	70	20
subsample	0.8	0.8	0.8	0.8	0.8
colsample_bytree	0.9	0.8	0.8	0.9	0.9

3.2. Production-Split Methodology

Standard regression approaches achieve excellent accuracy at the 15 min horizon but degrade sharply at the 24 h horizon, where weather-forecast uncertainty dominates and the response variable becomes strongly heteroscedastic. This behavior is widely reported in the photovoltaic forecasting literature [7,19] and reflects a fundamental property of solar generation: the conditional variance of the output is itself a function of the magnitude of the predicted signal. Residual analysis of the baseline 24 h LightGBM model confirmed this pattern—error variance scaled with predicted power level, indicating that low- and high-production periods follow fundamentally different predictability regimes.

To address this limitation, we introduce a Production-Split methodology that partitions the training set into two regimes on the basis of the observed PV power: a low-production regime ($\leq$2 kW), dominated by morning, evening and overcast periods, and a high-production regime (>2 kW), corresponding to mid-day clear-sky conditions. Two independent LightGBM models are trained on these subsets, each specializing in the statistical structure of its regime. In Section 4.2 the split is analyzed both under ideal gating (actual regime known—an oracle upper bound) and under operational gating by a regime classifier restricted to issue-time information. The 2 kW threshold corresponds to the boundary between low-irradiance and clear-sky regimes for the reference installation, and was identified as the level at which the residual variance of the baseline single-model formulation begins to grow rapidly. Specifically, residuals of the baseline day-ahead model were binned by observed power level, and 2 kW (approximately 13% of rated capacity) marks the onset of rapid growth in binned residual variance; the sensitivity of the results to this choice is examined in Section 4.2.

This formulation respects the heteroscedastic nature of the underlying physical process. The two regimes differ not only in the magnitude of the response, but also in the dominant exogenous drivers: low-production periods are governed primarily by cloud cover and time-of-day, whereas high-production periods are strongly modulated by panel temperature and angle of incidence. A single global model averages over these regimes and necessarily produces a compromise estimator. The split formulation, by contrast, allows each sub-model to allocate capacity to the features that are most informative for its regime, in the spirit of mixture-of-experts approaches but with a deterministic, physically motivated gating function.

3.3. Performance Metrics

Model performance is evaluated using four complementary metrics. The Mean Absolute Error (MAE) provides an interpretable measure of the average absolute deviation between predicted and observed power, expressed in kilowatts. The Root Mean Square Error (RMSE) penalizes larger deviations more strongly than MAE and is sensitive to outliers. The coefficient of determination (R^2) quantifies the proportion of variance in the observed power explained by the model and is reported as the primary metric for cross-horizon comparison. The Mean Absolute Percentage Error (MAPE) provides a scale-independent view of relative error; it is computed only on samples with non-trivial production (>0.5 kW) to avoid division by near-zero values, following common practice in the photovoltaic forecasting literature [7].

Cross-validation was performed using a TimeSeriesSplit scheme with three folds, which preserves temporal ordering and prevents leakage from future to past—a precaution that is essential whenever the resulting model is to be deployed within a Digital Twin operating on streaming data. Residual diagnostics, including the empirical distribution of errors, their temporal autocorrelation and their dependence on predicted power, were used to verify model validity at every horizon and to motivate the Production-Split formulation introduced in Section 3.2.

4. Results

4.1. Short-Term Forecasting Performance

At the operational horizons of 15 min and 1 h, the LightGBM model achieved high predictive accuracy (Figure 4). At 15 min the model attained an R^2 of 0.868 with a mean absolute error of 0.79 kW, corresponding to approximately 5% of the system rated capacity. At the 1 h horizon, performance degraded moderately to an R^2 of 0.832 and an MAE of 1.03 kW, reflecting the increasing influence of stochastic cloud dynamics over longer prediction windows. XGBoost [27] produced comparable accuracy at both horizons, confirming the robustness of the gradient-boosting paradigm for short-term solar forecasting; LightGBM [26] was retained as the operational model on the basis of its lower training cost—approximately 40% faster than XGBoost on the same dataset—and faster inference latency, in agreement with the findings of Hokmabad et al. [22].

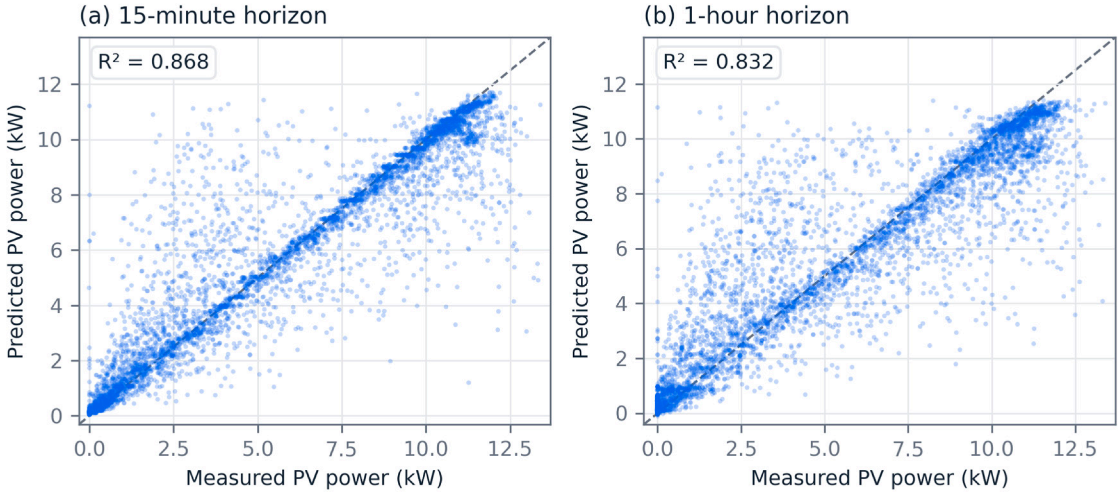

Figure 4. Predicted vs. actual PV power for the LightGBM model at (**a**) 15 min and (**b**) 1 h horizons on the held-out chronological test set. Both horizons exhibit tight clustering around the identity line, with R^2 of 0.868 and 0.832, respectively.

As shown in Table 5. persistence baselines confirm that the models capture structure beyond diurnal autocorrelation: last-observed-value persistence achieves an R^2 of 0.741 at 15 min and 0.489 at 1 h, and same-time-yesterday persistence reaches 0.413 at the 24 h horizon (principal day-ahead test set)—all well below the corresponding LightGBM models, with the margin widening as the horizon lengthens. A Random Forest regressor trained on the same day-ahead feature set attains an R^2 of 0.695, statistically comparable to the LightGBM day-ahead model (R^2 = 0.701), which is retained for its lower training and inference cost and its consistency across horizons. These results reinforce the conclusion of Antonanzas et al. [7] and Erbay [19] that exogenous weather information and non-linear estimators are essential for operationally useful forecasts.

Table 5. Comparative performance of all forecasting models evaluated. Day-ahead rows are evaluated on the principal held-out test set (n = 3816); the 15 min and 1 h rows use the corresponding chronological test sets. The Production-Split row reports the ideal-gating (oracle) configuration (Section 4.2).

Approach	Model	Horizon	R^2	MAE (kW)
Gradient boosting	XGBoost	15 min	0.865	0.82
Gradient boosting	XGBoost	1 h	0.827	1.07
Gradient boosting	LightGBM	15 min	0.868	0.79
Gradient boosting	LightGBM	1 h	0.832	1.03
Persistence	Last value	15 min	0.741	1.30
Persistence	Last value	1 h	0.489	2.27
Persistence	Same time yesterday	24 h	0.413	1.86
Ensemble	Random Forest (day-ahead features)	24 h	0.695	1.56
Gradient boosting	LightGBM (day-ahead features)	24 h	0.701	1.58
Production-Split	LightGBM (split, ideal gating)	24 h	0.794	1.20

Feature-importance analysis was broadly consistent across LightGBM and XGBoost, as shown in Table 6. The rolling PV statistics dominate the predictor set (33–78% relative importance), followed by the cyclical temporal encodings (12–33%) and the autoregressive PV lag features (7–29%), with Global Horizontal Irradiance (1–11%) and ambient temperature (0–9%) contributing smaller shares. This ranking indicates that short-horizon forecasting is driven primarily by recent production history—which itself embeds the current irradiance state—with calendar structure acquiring weight at longer horizons [20].

Table 6. Feature-importance analysis across both LightGBM and XGBoost.

Feature Group	Relative Importance (%)	Dominant Horizon
Global Horizontal Irradiance	1–11	15 min, 1 h
PV power lag (15–30 min)	7–29	15 min, 1 h
Cyclical temporal encodings	12–33	all
Ambient temperature	0–9	15 min, 1 h
Rolling statistics (45–90 min)	33–78	all

4.2. Comparative Analysis: Standard vs. Production-Split Modeling

At the day-ahead horizon, a single LightGBM regressor trained on the day-ahead feature set of Section 3.1.3—using only information available at forecast issue time—achieves an R^2 of 0.701 with a mean absolute error of 1.58 kW on the principal held-out test set (n = 3816, the union of the two regime test segments; all day-ahead results in this paper are evaluated on these identical samples). Residual analysis, nevertheless, shows regime-dependent behavior: error variance grows with predicted power, motivating the regime analysis that follows.

The Production-Split formulation of Section 3.2 is evaluated under three gating assumptions on identical test samples (Table 7). Under ideal gating—each sample dispatched by its actual regime, an oracle upper bound—the dual-regime formulation achieves an R^2 of 0.794 and a mean absolute error of 1.20 kW. Under operational gating, a LightGBM regime classifier restricted to issue-time information reaches 90.7% gating accuracy (against a 74% trivial baseline), yet the resulting split forecast (R^2 of 0.685 with probability-weighted soft gating) does not outperform the single model: the residual gating errors concentrate at high-power mid-day samples, where dispatching to the wrong specialist is costly. Gating on the target day's actual meteorological observations—a proxy for an ideal weather forecast—gives the same outcome (87.8% accuracy; R^2 0.692 vs. 0.712 for the single model on the same rows). The regime analysis, therefore, quantifies the headroom available to regime-specialized models (R^2 0.794 vs. 0.701; MAE −24%) and identifies gating robustness, rather than regressor capacity, as the binding constraint on day-ahead accuracy; the deployed platform accordingly uses the single day-ahead model.

Table 7. Day-ahead performance under the three gating assumptions, on identical held-out test samples (n = 3816).

Configuration	Gating Accuracy	R^2	MAE (kW)
Single day-ahead model	—	0.701	1.58
Operational split—hard gating (issue-time regime classifier)	90.7%	0.653	1.56
Operational split—probability-weighted soft gating	—	0.685	1.55
Ideal split (actual regime known)	100%	0.794	1.20

https://doi.org/10.3390/eng7080414

The sensitivity of the split formulation to the regime threshold was examined by retraining both sub-models at alternative thresholds under identical hyperparameters and protocol (Table 8): performance varies smoothly across the 1.5–3.0 kW range, with no instability around the chosen 2 kW value; the monotonic increase with threshold reflects the growing share of samples handled by the low-production specialist—a compositional effect, rather than a better operating point.

Table 8. Sensitivity of the Production-Split result (ideal gating) to the regime threshold.

Threshold	R^2 (Combined, Ideal Gating)	MAE (kW)
1.5 kW	0.760	1.30
2.0 kW	0.794	1.20
2.5 kW	0.823	1.10
3.0 kW	0.854	1.00

Uncertainty of the model comparison was quantified by case-resampling bootstrap (2000 resamples) over the held-out test set: single-model MAE 1.58 kW with 95% CI [1.53, 1.63] and R^2 0.701 [0.680, 0.722]; ideal-gating split MAE 1.20 kW [1.15, 1.24] and R^2 0.794 [0.777, 0.812]; paired MAE difference 0.38 kW [0.34, 0.42] (paired t-test on per-sample absolute errors: $t = 17.6$, $p < 10^{-65}$). The ideal-split and single-model intervals do not overlap, while the operational-gating configuration lies within the single model's uncertainty band.

As a probabilistic extension, LightGBM quantile regression (pinball loss, $\alpha = 0.1/0.5/0.9$) was trained under the same protocol: the nominal 80% prediction interval achieves an empirical coverage of 62.2% with a mean width of 2.83 kW, indicating that uncalibrated quantile regression underestimates day-ahead uncertainty for this installation. Calibrated (conformalized) quantile regression is identified as future work.

Table 9 summarizes the cross-horizon performance. The operational forecasting configuration of the Hy4GreenSteam DT combines the standard LightGBM models at all three horizons; the Production-Split formulation is retained as an analysis of regime-driven error structure. As illustrated in Figure 5, the Production-Split formulation under ideal gating more closely captures the timing and magnitude of peak PV production than the standard LightGBM model during the representative test week.

Table 9. Cross-horizon performance on the held-out chronological test set. The Production-Split row reports the ideal-gating (oracle) configuration; operational gating is compared in Table 7.

Model	Horizon	R^2	MAE (kW)	ΔR^2 vs. Standard 24 h
LightGBM (standard)	15 min	0.868	0.79	—
LightGBM (standard)	1 h	0.832	1.03	—
LightGBM (standard)	24 h	0.701	1.58	(reference)
LightGBM (Production-Split, ideal gating)	24 h	0.794	1.20	+13%

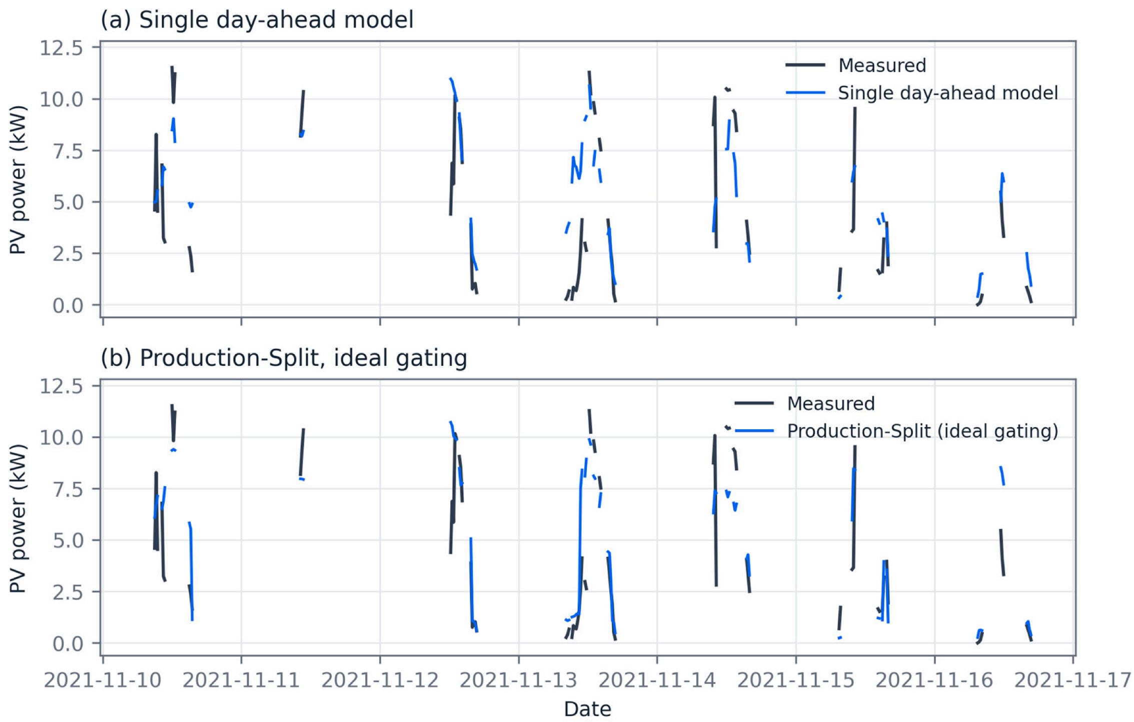

Figure 5. Day-ahead PV forecasting over a representative test week. (**a**) Standard LightGBM under-predicts peaks and fails to track the diurnal envelope. (**b**) Production-Split under ideal gating closely tracks both the timing and magnitude of peak production.

4.3. Digital-Twin Visualization and Blending Optimization

The forecasting models, the energy-flow equations and the blending optimization logic are integrated into a unified Digital-Twin Dashboard that mirrors the state of the physical hydrogen-blending system in real time [15,16]. The dashboard renders live sensor measurements and twin predictions side-by-side, allowing the operator to verify the alignment between physical state and virtual replica continuously. The interface is implemented as a React single-page application and is organized around a row of real-time KPI cards and a coordinated set of operational charts.

4.3.1. Operational Charts and KPI Cards

Real-time KPI cards expose the four headline indicators of the system—instantaneous PV power, hydrogen production rate, steam demand, and avoided CO_2—at every refresh interval. Six coordinated charts present the operational state in greater depth: the Renewable Energy chart overlays measured and predicted PV (and, optionally, wind) generation; the Hydrogen Production chart shows the electrolyzer output computed from the live and forecast power; the H_2/CH_4 Blending chart reports the recommended volumetric hydrogen fraction; the Fuel Cost chart compares the operating cost of the blended configuration against a natural-gas-only baseline; the CO_2 Emissions chart traces the avoided emissions over time; and the Boiler Efficiency chart reports the thermal performance of the dual-fuel burner against the steam-demand profile. Side-by-side rendering of measured and predicted quantities enables continuous validation of the Digital Twin: sustained divergence between the two is itself a diagnostic signal, flagging either sensor anomalies or model drift requiring retraining.

4.3.2. Blending Optimization

At every time step, the platform computes the volumetric hydrogen fraction that satisfies the steam demand at minimum methane consumption, subject to the available renewable power and the technical limits of the dual-fuel boiler. The optimization is performed deterministically, using the equations of Section 2.3, and produces, for each operating point, the H_2 and CH_4 flow rates, the equivalent natural-gas displacement, and the avoided CO_2 emissions. For the reference configuration of the pilot installation, driven by measured on-site PV generation, the computed reduction in CO_2 emissions reaches 34% at 15 min interval level during high-solar, low-demand periods—with daily aggregates of approximately 5–15% under the measured factory-demand profile—compared to a natural-gas-only baseline, while maintaining the steam-generation thermal output within the operational tolerance of the boiler [8,11].

Formally, at each 15 min step the platform minimizes q_CH_4 subject to (i) thermal demand, $(q\_H_2 \cdot LHV\_H_2,vol + q\_CH_4 \cdot LHV\_CH_4,vol) \cdot \eta\_boiler = \dot{Q}\_steam$; (ii) hydrogen availability, $0 \leq q\_H_2 \leq q\_H_2$, availability from Equation (1); (iii) the certified firing envelope of the dual-fuel burner (20–100% H_2); and (iv) non-negativity of the flows. Because the objective is monotonically decreasing in q_H_2 and the constraints are linear in the two flow variables, the optimum admits a closed-form solution: all available hydrogen is allocated up to the burner envelope, and the balancing methane flow follows as $q\_CH_4^* = (\dot{Q}\_steam/\eta\_boiler - q\_H_2^* \cdot LHV\_H_2,vol)/LHV\_CH_4,vol$. No iterative solver is required; the solution is evaluated deterministically at each refresh interval.

4.3.3. Scenario and Cases Module

A dedicated cases module allows operators and analysts to define and compare alternative scenarios. Each case specifies a steam-demand profile, an electrolyzer efficiency, fuel- and electricity-cost assumptions and emission factors. The platform then evaluates the case against the live forecast and against a natural-gas-only baseline, reporting the difference in fuel consumption, operating cost and CO_2 emissions. This functionality is intended to support informed operational and planning decisions by SME operators who lack the in-house engineering capacity to perform such analyses manually [9], and complements the real-time operational view with a forward-looking what-if analysis.

4.3.4. Integration and Deployment

The DT Dashboard is deployed as a containerized web application and is integrated with the data-acquisition layer through asynchronous API calls. Inference is performed server-side on the exported LightGBM models, with predictions cached at 15 min intervals to bound latency. The entire stack—data ingestion, forecasting, optimization and visualization—runs without operator intervention and is accessed through a standard web browser, satisfying the requirement that the system be deployable in food-industry SMEs without dedicated information-technology infrastructure.

4.3.5. Parametric Simulation Scenarios

To illustrate the operating envelope of the platform across a range of conditions that go beyond the experimental test week, five parametric scenarios were pre-configured in the cases module of the dashboard with parameter ranges informed by operational measurements from the food industry: a sunny summer day, a cloudy winter day, a high-steam-demand campaign, a weekend low-steam-demand period and a scaled-up installation with larger photovoltaic capacity. Each scenario specifies a 24 h profile of photovoltaic output and steam demand at 15 min resolution, with the photovoltaic envelopes derived from the operational range observed at the steam-demand of the food industry. These scenarios are not field-measured outcomes; rather, they exercise the energy-flow model end-to-end to characterize the platform's sensitivity to renewable availability and demand structure. The results are shown in Table 10.

Table 10. Summary of the resulting aggregated indicators.

Scenario	H_2 Produced (Nm^3/day)	H_2 Consumed (Nm^3/day)	CH_4 Displaced (Nm^3/day)	CO_2 Avoided (kg/day)	CO_2 Reduction (%)	Time-Averaged H_2 Fraction (%)
Sunny summer day	27.6	27.6	8.3	16.2	25.0	52.6
Cloudy winter day	3.6	3.6	1.1	2.1	3.3	10.1
High steam-demand campaign	26.4	26.4	7.9	15.6	9.6	26.1
Weekend low-demand period	27.6	25.9	7.8	15.2	62.7	84.9
Scaled-up installation (large PV)	202.5	85.9	25.8	50.6	77.7	92.1

The time-averaged volumetric hydrogen fraction ranged from 10% under solar-poor conditions to 92% when electrolyzer output exceeded contemporaneous demand, with a representative summer-day value of 53%. The corresponding CO_2 reduction scaled from 3% to 78%. In the scaled-up scenario, 117 of the 202 Nm^3/day produced exceed the same-interval fuel demand, and must be buffered or curtailed, quantifying the storage capacity required for full utilization. All scenario values are computed with the corrected fuel-flow equation (including η_boiler) and the corrected Equation (1). The figures confirm that the contribution of the blending strategy is dominated by the daily envelope of renewable availability and by the ratio of installed PV capacity to steam-demand load—two parameters that the platform exposes explicitly through the cases module, enabling SME operators to quantify the operational and environmental impact of prospective capacity upgrades without external engineering analysis.

5. Discussion

The results presented in this study demonstrate that the integration of machine-learning-based forecasting with an energy-flow constitutes a technically viable and operationally practical pathway for green-hydrogen adoption in the food industry. Three findings stand out as having particular significance for the decarbonization of the industry.

5.1. Accurate Prediction of Green Energy Availability as a Prior Condition Required for Decision-Making

The forecasting layer directly addresses one of the most challenging parts of day-ahead photovoltaic forecasting, and the accompanying Production-Split regime analysis identifies regime predictability as the dominant source of residual day-ahead error. This level of predictive accuracy is not only a technical achievement, but also a requirement for the decision-support function of the platform. At the 15 min horizon, an R^2 of 0.868 provides sufficient accuracy for real-time electrolyzer dispatch and blending control. At the 24 h horizon, an R^2 of 0.701 enables the user-operator to design a steam production schedule and plan one day in advance—a function that conventional SCADA systems, which operate only on historical measurements, cannot support. The absence of reliable day-ahead forecasts, as illustrated by the weak same-time-yesterday persistence baseline (R^2 = 0.413), would make this planning framework far less dependable. Therefore, a feature of focus of this platform is accurate green energy-availability forecasting.

5.2. Alternative Pathways for Green H_2 and RES Penetration

The parametric scenario analysis presented in this work reveals that under current pilot-scale conditions, CO_2 reductions are achievable depending on seasonal renewable availability and steam demand, without any modification to the existing infrastructure. The cases module of the platform is specifically designed to make this platform useful for SME operators who lack the engineering expertise to plan manually. An operator evaluating whether to expand installed PV capacity or add hydrogen storage to buffer weekend surplus can do so within the platform by specifying the relevant parameters and observing the resulting shift in the CO_2 reduction profile. This function proves that the platform is an operational tool and a strategic planning instrument for the progressive penetration of green hydrogen and renewable energy within the food-industry energy system. As a result, the framework is compatible with a range of deployment scales and investment horizons—from the near-term actions available to a typical food-industry SME today to the medium-term transitions enabled by EU green-hydrogen funding programs.

5.3. Hydrogen Blending as a Viable and Compatible First Stage of Industrial Decarbonization

The most significant conclusion of this work is that the near-term integration of hydrogen blends into existing gas infrastructure is not only technically feasible, but is demonstrated at pilot scale in terms of system integration and real-time operation, with emission reductions quantified through the mass–energy balance of the installed equipment. All five parametric scenarios, as well as the reference blending result of 34% CO_2 reduction, were achieved using a commercially available PEM electrolyzer, a standard dual-fuel gas boiler, and on-site photovoltaic generation—none of which required modification of the existing gas-distribution network. This compatibility is fundamental to the initial goal deployment case: food-industry SMEs do not need to wait for the development of a hydrogen distribution infrastructure, nor do they need to replace their existing thermal equipment, to begin reducing their carbon footprint. This is consistent with the position advanced in the recent literature, which is that hydrogen blending serves as a critical transitional technology bridging the gap between today's fossil-fuel-dependent industrial thermal systems and the fully decarbonized energy systems of the future [8,11]. The Hy4GreenSteam Project framework operationalizes this transition by providing the digital intelligence—accurate forecasting, real-time optimization, and scenario planning—that transforms a technically feasible concept into a practically manageable operational strategy.

The hydrogen blending results are consistent with the theoretical expectation that CO_2 emissions decrease proportionally to the hydrogen fraction in the fuel mixture when the fraction is defined on an energy basis, and that a volumetric blend of 30–40% H_2 corresponds to a significantly smaller energy displacement, due to the lower volumetric LHV of hydrogen. The demonstrated reduction under the reference operating conditions of the pilot unit is significant from an industrial perspective, as it falls within the range that would allow food-industry operators to meet near-term EU decarbonization targets without capital-intensive infrastructure replacement. The fuel-blending model employed here assumes complete combustion and a fixed boiler thermal efficiency; future work should account for the variation in flame speed and adiabatic flame temperature as the hydrogen fraction increases, which may require burner-level adjustments beyond those already incorporated in the pilot installation.

Several limitations should be acknowledged. The predictive models were trained on data from a single Mediterranean site; transferability to installations at higher latitudes or with significantly different cloud-cover statistics has not been evaluated. The economic analysis embedded in the platform uses static fuel- and electricity-price assumptions, whereas actual industrial operators face time-varying tariffs and carbon costs that may substantially alter the optimal blending strategy. The platform allows operators to input site-specific fuel prices, electricity tariffs, and emission factors, to estimate scenario-dependent fuel cost savings. A detailed economic assessment, including CAPEX/OPEX and payback analysis, is outside the scope of this study, and is identified as future work. The reported CO_2 reductions quantify direct combustion emissions only—the stoichiometric CO_2 of the displaced natural gas—and do not constitute a lifecycle assessment; upstream emissions associated with equipment manufacturing, water treatment and auxiliary electricity consumption are outside the scope of this study. Finally, the current implementation does not account for electrolyzer degradation over time, which reduces production efficiency and may shift the optimal operating point as the system ages. Incorporating degradation models informed by real operational data is identified as a priority for future development.

Despite these limitations, the framework represents a TRL 6 demonstration—technology demonstrated in a relevant environment—that can be deployed in food-industry SMEs without dedicated IT infrastructure, addressing the technology-access gap identified in the literature for small and medium enterprises seeking to comply with EU energy and climate regulations. This TRL 6 assessment is supported by the demonstrated operation of the platform at the Lavrion pilot installation, integrating a commercial 15.39 kWp PV array, a PEM electrolyzer, and a 400 kWth dual-fuel boiler operating on H_2/CH_4 blends, consistent with the European Commission's TRL 6 definition of technology demonstrated in a relevant environment; extended-duration operation and independent performance verification remain necessary steps toward higher readiness levels. As a result, the combination of accurate green energy-vailability prediction, structured alternative pathway analysis, and demonstrated compatibility with existing gas infrastructure makes the Hy4GreenSteam platform a useful model for digitally enabled industrial decarbonization across the food sector.

6. Conclusions

This study presented an integrated digital decision-support framework for green hydrogen-based steam production in the food industry, developed within the Hy4GreenSteam project. The framework combines multi-horizon photovoltaic power forecasting, a deterministic hydrogen-blending energy-flow model, and a real-time Digital Twin Dashboard into a single, browser-deployable platform targeting the food industry.

The principal technical contribution is the integrated forecasting and decision-support layer, evaluated leakage-free on a two-year dataset from the 15.39 kWp Lavrion reference pilot installation: the day-ahead model achieves an R^2 of 0.701 using only information available at forecast issue time—a 70% relative improvement over same-time-yesterday persistence—while the standard LightGBM regressors achieve R^2 values of 0.868 (15 min) and 0.832 (1 h), confirming the suitability of gradient-boosting methods for operational solar forecasting under Mediterranean conditions. The accompanying Production-Split regime analysis shows that the production regime is predictable with 90.7% accuracy from issue-time information, and quantifies an oracle headroom of R^2 0.794, identifying gating robustness as the key lever for further day-ahead gains. This result confirms the fact that hydrogen blending into existing gas boiler infrastructure, enabled by on-site renewable generation and real-time decision support, represents a technically viable and near-term deployable transitional decarbonization strategy for the food sector.

The Digital Twin Dashboard operationalizes these capabilities through continuous side-by-side rendering of measured and predicted system state, real-time KPI monitoring, blending optimization, and scenario analysis, providing operators with decision-relevant insights. The overall framework constitutes a practical solution that directly addresses the technology-access gap faced by the EU's food-industries in meeting the decarbonization requirements of the "Fit for 55" legislative package. Future work will focus on multi-site validation at higher latitudes, incorporation of electrolyzer-degradation models, integration of the co-located wind-generation array, and extension of the economic module to account for time-varying electricity tariffs and EU carbon pricing.

Author Contributions: Conceptualization, C.P.; methodology, A.P. (Andreas Poyias); software, A.P. (Andreas Poyias); validation, D.P. and A.P. (Antonis Peppas); formal analysis, C.P.; investigation, D.G.; resources, D.G.; data curation, C.P.; writing—original draft preparation, D.G.; writing—review and editing, A.P. (Antonis Peppas) and C.P.; visualization, D.G.; supervision, P.M.; project administration, A.P. (Antonis Peppas); funding acquisition, A.P. (Antonis Peppas). All authors have read and agreed to the published version of the manuscript.

Funding: The work presented was part of the research performed under the Greek research program Hy4GreenSteam, funded by Clusters of Research Excellence-CREs under ESPA. Action code: 16289/TA 5180519. The action is implemented within the framework of the National Recovery and Resilience Plan "Greece 2.0", funded by the European Union—Next GenerationEU.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The photovoltaic power and meteorological datasets used in this study were collected at the Lavrion pilot site and contain proprietary operational information; they are available from the corresponding author upon request. The revised data-ingestion and feature-engineering pipeline, all training and evaluation scripts, and the trained models for every configuration reported are openly available in a public repository at https://github.com/mgd-energy/Hy4GreenSteam (accessed on 9 August 2026).

Acknowledgments: The authors thank Yiotis S.A. for the valuable contribution to the project. During the preparation of this manuscript, the authors used Anthropic Claude (claude-opus-4-7) for drafting, restructuring and language review of the Sections 3 and 4. The authors have reviewed and edited all model-generated output and take full responsibility for the content of this publication. Regarding figures: Figures 4 and 5 were generated directly by the released analysis code without generative AI; Figure 1 was redrawn with the assistance of Anthropic Claude and has been reviewed and approved by the authors.

Conflicts of Interest: Authors Andreas Poyias, Panayiotis Mourtopallas and Diamanto Platanou are employed by MGD Energy, Maroussi, Greece Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

References

1. Meitz, S.; Reiter, J.; Fluch, J.; Tugores, C.R. Decarbonization of the Food Industry—The Solution for System Design and Operation. *Sustainability* **2023**, *15*, 1426. [CrossRef]
2. World Energy Balances Database Documentation. 2025. Available online: https://www.iea.org/data-and-statistics/data-product/world-energy-balances (accessed on 9 August 2026).
3. European Commission. *Energy Efficiency Directive*; European Commission: Brussels, Belgium, 2023.
4. Office of the European Union I (Legislative Acts). Directives Directive (EU) 2023/1791 of the European Parliament and of the Council of 13 September 2023 on Energy Efficiency and Amending Regulation (EU) 2023/955 (recast) (Text with EEA Relevance). Available online: https://eur-lex.europa.eu/eli/dir/2023/1791/oj/eng (accessed on 9 August 2026).
5. Köhl, M.; Linser, S.; Prins, K.; Talarczyk, A. The EU climate package 'Fit for 55'—a double-edged sword for Europeans and their forests and timber industry. *For. Policy Econ.* **2021**, *132*, 102596. [CrossRef]
6. Sebbagh, T.; Şahin, M.E.; Beldjaatit, C. *Green Hydrogen Revolution for a Sustainable Energy Future*; Springer Science and Business Media Deutschland, GmbH: Berlin/Heidelberg, Germany, 2024. [CrossRef]
7. 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]
8. Franco, A.; Rocca, M. Industrial Decarbonization through Blended Combustion of Natural Gas and Hydrogen. *Hydrogen* **2024**, *5*, 519–539. [CrossRef]

9. Sovacool, B.K.; Bazilian, M.; Griffiths, S.; Kim, J.; Foley, A.; Rooney, D. Decarbonizing the food and beverages industry: A critical and systematic review of developments, sociotechnical systems and policy options. *Renew. Sustain. Energy* **2021**, *143*, 110856. [CrossRef]
10. Famiglietti, A.; Abbas, R. Steam Generation for Industry Using Linear Fresnel Solar Collectors and PV-Driven High-Temperature Heat Pumps: Techno-Economic Analysis. *Solar* **2025**, *5*, 27. [CrossRef]
11. Wang, T.; Zhang, H.; Zhang, Y.; Wang, H.; Lyu, J.; Yue, G. Efficiency and emissions of gas-fired industrial boiler fueled with hydrogen-enriched nature gas: A case study of 108 t/h steam boiler. *Int. J. Hydrogen Energy* **2022**, *47*, 28188–28203. [CrossRef]
12. Yavari, A.; Harrison, C.J.; Gorji, S.A.; Shafiei, M. Hydrogen 4.0: A Cyber–Physical System for Renewable Hydrogen Energy Plants. *Sensors* **2024**, *24*, 3239. [CrossRef] [PubMed]
13. Abualigah, L.; Abu Zitar, R.; Almotairi, K.H.; Hussein, A.M.; Elaziz, M.A.; Nikoo, M.R.; Gandomi, A.H. Wind, Solar, and Photovoltaic Renewable Energy Systems with and without Energy Storage Optimization: A Survey of Advanced Machine Learning and Deep Learning Techniques. *Energies* **2022**, *15*, 578. [CrossRef]
14. IRENA. *Innovation Landscape for Smart Electrification: Power to Hydrogen—Digital Backbone for Green Hydrogen Production*; International Renewable Energy Agency: Abu Dhabi, United Arab Emirates, 2023.
15. Ba, L.; Tangour, F.; El Abbassi, I.; Absi, R. Analysis of Digital Twin Applications in Energy Efficiency: A Systematic Review. *Sustainability* **2025**, *17*, 3560. [CrossRef]
16. Kim, M.; Ghobadi, F.; Charmchi, A.S.T.; Lee, M.; Lee, J. Digital Twins for Clean Energy Systems: A State-of-the-Art Review of Applications, Integrated Technologies, and Key Challenges. *Sustainability* **2026**, *18*, 43. [CrossRef]
17. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*, 5th ed.; Wiley: Hoboken, NJ, USA, 2015.
18. Ali, M.; Prasad, R.; Xiang, Y.; Khan, M.; Farooque, A.A.; Zong, T.; Yaseen, Z.M. Variational mode decomposition based random forest model for solar radiation forecasting: New emerging machine learning technology. *Energy Rep.* **2021**, *7*, 6700–6717. [CrossRef]
19. Erbay, C. Machine learning models for solar forecasting and impact on green hydrogen production costs. *Int. J. Hydrogen Energy* **2025**, *132*, 225–238. [CrossRef]
20. Persson, C.; Bacher, P.; Shiga, T.; Madsen, H. Multi-site solar power forecasting using gradient boosted regression trees. *Sol. Energy* **2017**, *150*, 423–436. [CrossRef]
21. Upadhyay, D.; Manero, J.; Zaman, M.; Sampalli, S. Gradient Boosting Feature Selection with Machine Learning Classifiers for Intrusion Detection on Power Grids. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 1104–1116. [CrossRef]
22. Hokmabad, H.N.; Husev, O.; Belikov, J. Day-Ahead Solar Power Forecasting Using LightGBM and Self-Attention Based Encoder-Decoder Networks. *IEEE Trans. Sustain. Energy* **2025**, *16*, 866–879. [CrossRef]
23. Zhang, Z.; Hong, W.C.; Dong, Y. Multi-hyperplane twin support vector regression guided with fuzzy clustering. *Inf. Sci.* **2024**, *666*, 120435. [CrossRef]
24. Intergovernmental Panel on Climate Change (IPCC). *IPCC Guidelines for National Greenhouse Gas Inventories*; IPCC: Geneva, Switzerland, 2006; Volume 2.
25. European Environment Agency. *EMEP/EEA Air Pollutant Emission Inventory Guidebook*; EEA: Copenhagen, Denmark, 2023.
26. 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. Available online: https://github.com/Microsoft/LightGBM (accessed on 9 August 2026).
27. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 13–17 August 2016; pp. 785–794. [CrossRef]