Machine Learning with Applications 24 (2026) 100886

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Bridging data gaps in smart greenhouses: Outdoor-to-indoor mapping for synthetic climate forecasting

Juan Bonastre-Egea [a], Andrés Bueno-Crespo [a], Virginia C. Sánchez [b], José M. Cecilia [b], Juan Morales-García [c,*]

[a] Escuela Politécnica Superior, Universidad Católica de Murcia, Murcia, Spain
[b] Department of Computer Engineering (DISCA), Universitat Politècnica de València, Valencia, Spain
[c] Department of Software and Computing Systems, Universidad de Alicante, Alicante, Spain

ARTICLE INFO

Keywords:
Smart agriculture
Smart greenhouses
Climate control systems
Data-driven modeling
Multi-model deep learning

ABSTRACT

In order to make reliable forecasts of greenhouse climate variables, it is often necessary to have a long history of indoor sensor data, but newly constructed facilities often lack such records. In contrast, multi-year outdoor weather series are usually available. This paper introduces a two-stage deep learning pipeline to address this data scarcity. First, outdoor-to-indoor mapping models are trained to translate outdoor measurements of temperature, humidity, and radiation into synthetic indoor series. Secondly, these synthetic indoor series are used to train prediction models, which are then compared with their counterparts trained with real indoor data. Experiments conducted on six greenhouses across four countries with six deep learning architectures demonstrate that synthetic indoor climate series, generated from weather records, can effectively substitute for missing sensor histories. This approach enables the rapid deployment of forecasting systems in data-limited greenhouses and provides a practical AIoT strategy to mitigate information gaps in precision agriculture.

1. Introduction

The agricultural sector faces unprecedented challenges driven by two global trends: the increasing frequency of extreme weather events (Schmitt et al., 2022) and the rapid growth of the world population, projected to reach 10 billion by 2050 (Sadigov et al., 2022). Meeting future food demand requires a transformation towards more efficient and resilient production systems. In this context, the integration of Artificial Intelligence (AI) and the Internet of Things (IoT) into agricultural practices has given rise to the concept of *Smart Greenhouses*, where environmental conditions can be monitored and controlled in real time to improve yields and resource efficiency (Ardiansah et al., 2020; Nakhaei et al., 2023). This transition aligns with broader global technological trends; as recent systematic analyses highlight, the exponential growth of IoT data across domains such as climate and agriculture has rendered manual data processing impractical, necessitating advanced machine learning frameworks to generate meaningful, actionable insights (Chahal et al., 2024).

Forecasting indoor climate variables such as temperature, humidity, and radiation is key to greenhouse management, as these factors directly influence crop growth and determine the decisions of climate control systems and decision-support systems (DSS) (Maraveas, 2022; Zhai et al., 2020). A wide range of forecasting approaches has been investigated, spanning statistical models (Sharma et al., 2022; Sun et al., 2019), machine learning methods (Kaneda & Mineno, 2016; Tsai et al., 2020), and deep learning architectures (Jin et al., 2021; Liu et al., 2023). While many of these solutions successfully forecast individual climate variables (Codeluppi et al., 2020; Oh et al., 2023), and more recently multivariate forecasting models for greenhouse environments have emerged (Morales-García et al., 2024), they all rely fundamentally on the presence of historical greenhouse data for training. Some research has explored reinforcement learning-based approaches for greenhouse climate control, but even these methods require substantial exploration and adaptation periods to learn internal thermodynamics, making them non–plug-and-play and costly in both time and resources (Mallick et al., 2025; Morcego et al., 2023).

Fortunately, the internal dynamics of a greenhouse are strongly influenced by the local exterior weather regime. This observation motivates the main research questions of this work: (i) can we train a global climate-control ML model by matching outdoor weather variables (as input) with indoor greenhouse behavior (as output)? and (ii) can we use this model to bootstrap forecasting in newly instrumented greenhouses lacking interior histories?

* Corresponding author.
E-mail addresses: jbonastre@ucam.edu (J. Bonastre-Egea), abueno@ucam.edu (A. Bueno-Crespo), vcassan@disca.upv.es (V.C. Sánchez), jmcecilia@disca.upv.es (J.M. Cecilia), juan.morales@ua.es (J. Morales-García).

https://doi.org/10.1016/j.mlwa.2026.100886
Received 29 January 2026; Received in revised form 5 March 2026; Accepted 13 March 2026
Available online 23 March 2026

Unlike conventional greenhouse forecasting approaches, which assume the availability of historical indoor observations for model training, we reformulate indoor climate prediction as a bootstrapped cross-domain learning problem. Specifically, we address the scenario of newly instrumented greenhouses where no prior indoor sensor history exists. Instead of relying on target-domain data, we propose to leverage long-term outdoor meteorological records to synthesize structurally coherent indoor climate series through a learned outdoor-to-indoor transformation. This formulation enables what we define as *zero-history indoor forecasting*, where predictive models can be deployed without requiring historical indoor measurements from the target greenhouse. To the best of our knowledge, this problem has not been previously formalized in the greenhouse forecasting literature, which predominantly assumes data-rich indoor environments.

To operationalize this hypothesis, we propose a *two-stage deep learning pipeline*. In the first stage, *outdoor-to-indoor mapping models* learn to translate external measurements (temperature, humidity, radiation) into synthetic indoor series that implicitly encode site-specific buffering and control responses. In the second stage, these synthetic series serve as training data for forecasting models, which are then rigorously benchmarked against identical forecasters trained on real indoor data. Experiments over six greenhouses spanning four countries and six neural architectures (MLP, LSTM, CNNLSTM, LSTM with Attention, CNNLSTM with Attention, Transformer) validate that synthetic interior series can effectively stand in for missing sensor histories.

It is important to emphasize that this research serves as a proof-of-concept for a *zero-history* deployment strategy. Unlike existing hybrid control strategies that optimize performance in established greenhouses, the primary objective here is to enable the immediate generation of predictive models for newly instrumented facilities that lack historical sensor data. By framing this as a data-driven bridge between global meteorological records and local indoor dynamics, we aim to demonstrate the feasibility of "cold-starting" AIoT systems, a necessary precursor to physical implementation in the agricultural sector.

The major contributions of this paper are the following:

- We introduce a novel reformulation of greenhouse climate forecasting as a *zero-history indoor forecasting* problem, explicitly addressing the practical constraint of newly deployed or sensor-limited greenhouses where historical indoor data are unavailable.
- We propose a two-stage exogenous bootstrapping framework that decomposes the learning process into: (i) a structural cross-domain mapping from outdoor to indoor climate variables, capturing thermodynamic buffering and control-system effects; and (ii) a temporal forecasting model trained on synthetically generated indoor sequences.
- We conceptually distinguish our approach from conventional synthetic data augmentation, domain adaptation, and surrogate modeling paradigms by demonstrating that the proposed method does not require access to target-domain indoor observations for model training.
- We provide extensive multi-site empirical validation across six heterogeneous greenhouses in four countries, showing that forecasting models trained on synthetically generated indoor series can approach, and in some cases surpass, the performance of models trained on real indoor histories.
- We characterize the conditions under which the structural mapping enables reliable forecasting, thereby identifying the limits and robustness properties of exogenous bootstrapped indoor climate prediction.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details the use-case setting, datasets, problem formulation, and models; Section 4 presents evaluation and results; Section 5 discusses the findings; Section 6 concludes and outlines future directions.

2. Related work

Several studies have focused on short-term forecasting of greenhouse climatic conditions through models that predict a single target variable, namely the indoor air temperature. Ruiz et al. (2022) evaluated different time-series forecasting libraries (Prophet, Greykite, and TPOT) using only historical indoor temperature as input. After data sanitization, prediction errors ranged from 1.5 °C to 3 °C, with Greykite showing the best performance. Morales-García et al. (2023) compared AR, ARIMA, K-Nearest Neighbors, and Random Forest regressors for indoor temperature prediction in an operational smart greenhouse, with Random Forest achieving MAE < 1 °C and $R^2 \approx 0.97$ at 12 h and 24 h horizons. These models primarily focused on isolated, data-rich scenarios without closed-loop control, serving as the foundation for later, more integrated approaches.

As greenhouse systems are strongly influenced by interacting climatic variables, later research shifted towards multi-input and multivariate approaches to capture the coupled effects of temperature, humidity, and radiation. García-Vázquez et al. (2023) applied linear regression and SVR to predict indoor air temperature, incorporating indoor and outdoor humidity and temperature as explanatory variables, and reported $R^2 > 0.95$. Liu et al. (2025) compared multiple linear regression, random forest, SVR, LSTM, and GRU across 21 prediction horizons (15 min–24 h) for solar greenhouses, with GRU consistently outperforming alternatives ($R^2 \approx 0.99$ for 24 h forecasts, reducing RMSE by 12%–27% compared to LSTM). Yu et al. (2025b) developed a dual Transformer-BiLSTM architecture that integrates relative positional encoding and sparse attention for winter prediction in northern China solar greenhouses. Their model achieved an average MSE of 0.0599 $(°C)^2$ and $R^2 = 0.9989$, remaining robust across horizons from 6 to 48 h and under variable feature combinations, and highlighting vapor pressure deficit (VPD) and radiation as critical factors. Expanding to multiple outputs, Eraliev and Lee (2023) tested DNN, LSTM, and 1D-CNN for simultaneous prediction of temperature, humidity, and CO_2 concentration in hydroponic greenhouses, with LSTM yielding $R^2 \approx 0.99$ for temperature and 0.97 for humidity.

Domain-specific frameworks have also been introduced. Liu et al. (2022) proposed GCP-LSTM for short-term prediction of six climatic factors (temperature, humidity, illumination, CO_2 concentration, soil temperature, soil humidity), achieving robustness against noisy sensor inputs. Jung et al. (2022) integrated evapotranspiration estimates with LSTM-based modeling to improve humidity prediction and irrigation scheduling in tomato greenhouses. Jeon et al. (2024) benchmarked MLR, SVM, ANN, and XGBoost for melon greenhouse datasets, with XGBoost attaining R^2 up to 0.993. Riskiawan et al. (2023) deployed IoT sensors combined with LSTM predictions to drive actuator control (ventilation, misting, irrigation), enabling adaptive responses. Similarly, Soheli et al. (2022) coupled IoT sensing with ANFIS-based fuzzy inference for automated microclimate regulation and secure data handling.

More complex deep architectures have emerged to improve multi-step robustness. Yang et al. (2023) introduced FAM-LSTM with attention for 12–48 h forecasting, enhancing accuracy in both temperature and humidity. Ma et al. (2024) proposed a fuzzy-adaptive normalized encoder–decoder network that improved resilience under data volatility. Morales-García et al. (2024) presented a multi-model ensemble allocating different architectures to distinct variables, achieving balanced prediction performance across datasets. Wang et al. (2025) extended the scope to irrigation management in cherry tomato greenhouses, showing that XGBoost integrating light, temperature, and humidity data improved water-use efficiency and crop quality.

Beyond forecasting, a growing body of work targets control-oriented strategies such as Model Predictive Control (MPC), Reinforcement Learning (RL), and hybrid architectures. Morcego et al. (2023) compared MPC and RL in greenhouse climate control, showing that RL algorithms such as DDPG handle continuous action spaces with robust learning, whereas MPC ensures stronger constraint satisfaction. Mallick

et al. (2025) integrated MPC as a function approximator for RL, reducing conservatism and improving growth and resource efficiency in lettuce greenhouses. Mansour et al. (2025) advanced this with a hierarchical system: an upper-level economic MPC optimizing trajectories under dynamic energy prices, and a lower-level deep RL agent ensuring robust tracking even under actuator failures. Hamidane et al. (2024) applied constrained MPC with subspace identification (N4SID) to regulate temperature (15 °C–30 °C) and humidity (50%–70% RH) in *Schefflera arboricola*, ensuring control under actuator voltage limits while reducing computational complexity. Naagarajan and Streif (2025) coupled MPC with interpretable AI through an Adaptive Retrieval-Augmented Generation (ARAG) framework, which explained MPC outputs in natural language. Their approach improved explanation quality (BERTScore +12.1%) without degrading control performance, bridging optimization with practical decision-making.

A complementary direction is the integration of physical models with data-driven approaches, known as Physics-Informed Machine Learning (PIML). Nghiem et al. (2023) reviewed PIML methods for dynamical systems, highlighting how embedding conservation laws into neural architectures or loss functions enhances robustness and interpretability. Ullah et al. (2022) proposed a hybrid AI-IoT system integrating ANN-based modules for prediction and optimization, achieving ∼62% energy reduction and ∼68% cost savings compared to baselines, while preserving yield. Fink et al. (2025) developed Neural Predictive Control (NPC) based on feedforward neural networks trained on 81 days of Mediterranean greenhouse data, achieving < 3 K temperature tracking error and keeping humidity within bounds 81% of the time, demonstrating adaptability to seasonal and structural changes. PIML thus bridges data-driven and physics-based reasoning, offering interpretability yet still relying on extensive measurement datasets for calibration.

Finally, several articles introduced reviews on this topic. Vanegas-Ayala et al. (2022) analyzed 93 studies on fuzzy inference for greenhouse humidity control, concluding that Mamdani-type fuzzy controllers combined with clustering and optimization achieve high robustness. Maraveas (2023) surveyed AI in smart greenhouses, spanning robotics, bio-inspired optimization, UAV-based pest control, and irrigation, while stressing barriers such as cost and energy demand. Goldenits et al. (2024) reviewed reinforcement learning and digital twins in agriculture, categorizing applications in irrigation, pest detection, crop management, and greenhouse optimization, and highlighting the need for explainable and data-efficient RL solutions. Yu et al. (2025a) provided a comprehensive review of temperature prediction and control, showing deep learning's strength in forecasting and fuzzy logic, MPC, and RL in control, with digital twins and plant-centric AI as promising future pathways.

While the aforementioned studies demonstrate the power of deep learning for greenhouse climate control, a critical gap remains in the practical deployment phase. Existing research predominantly assumes the availability of high-quality, long-term historical datasets from within the target facility to train these models. However, this "data-rich" assumption fails in real-world agricultural expansion, where new greenhouses often lack any sensor history.

Unlike traditional virtual sensing approaches that focus on spatial interpolation within a single instrumented site, or GAN-based augmentation that requires a seed of real data to function, our proposed framework introduces a fundamental shift: decoupling the learning of greenhouse physics from the specific target site. By framing the problem as an "outdoor-to-indoor mapping" problem, we move beyond the descriptive modeling found in current literature towards a zero-history deployment strategy. This approach allows us to exploit the abundance of global meteorological data to "cold-start" forecasting systems in new facilities, a critical transition from empirical experimentation to scalable, industrial AIoT application.

It is important to note that the evolution from univariate regressors to multivariate deep architectures, and further into reinforcement learning, predictive control, and physics-informed hybrids, reflects a clear convergence towards intelligent and sustainable greenhouse management. Yet, challenges remain in interpretability, robustness across climates, computational efficiency, and integration with real-time decision support; gaps that motivate the approach presented in this work. Specifically, most studies assume the existence of dense and long-term indoor sensor datasets, which limits their deployment in data-scarce greenhouses. Building upon these advances, our forthcoming work aims to address one of the main limitations identified in current greenhouse forecasting and control research; the dependence on long-term indoor datasets for model training. We will develop a two-stage methodology in which outdoor-to-indoor mapping models translate external meteorological data into synthetic indoor climate series, enabling the creation of realistic training datasets even for newly instrumented or data-scarce greenhouses. These synthetic sequences will then feed multivariate deep learning models for temperature, humidity, and radiation forecasting. By systematically evaluating multiple architectures across geographically diverse sites, the project seeks to establish outdoor-to-indoor mapping as a generalizable and scalable strategy for rapid AI deployment in smart greenhouse environments. In doing so, it will bridge the current gap between purely data-driven prediction and the practical constraints of real-world agricultural systems. The following section details the design and implementation of this proposed outdoor-to-indoor mapping framework.

3. Materials & methods

This section outlines the use-case setting in which the proposed solution has been tested, the description of the target prediction problem along with all deep learning models used and the integrated learning methodology adopted to solve it.

3.1. Use-case setting

This article uses paired indoor and outdoor time series of temperature, relative humidity, and global radiation collected from six commercial greenhouses located in Alhama de Murcia (Murcia, Spain), Falces (Navarra, Spain), Sonora (Mexico), Miranda (Cartagena, Spain), Melipilla (Chile), and Antalia (Turkey). This heterogeneous set of facilities offers a diverse experimental framework encompassing a range of climatic conditions, control system complexities, and geographical locations, which is essential for validating the generalizability and robustness of the proposed mapping and forecasting models.

The outdoor datasets were obtained from the paid Weatherbit API service (Weatherbit, 2025), providing continuous meteorological observations with a temporal resolution of fifteen minutes and a coverage of approximately five years per site. The indoor datasets were acquired through Nutricontrol greenhouse sensors (Nutricontrol, 2025) as part of the research project SERGIoT (Project, 2025). The availability of indoor data varies between sites, ranging from one to five years, depending on the commissioning date and maintenance of the sensing infrastructure.

The indoor sensors were integrated with Nutricontrol climate controllers from the V4 series, namely the *Mithra Clima Pro*, *Mithra Clima*, and *Mastia V* models (Nutricontrol, 2022). These controllers operate as embedded systems designed for automated environmental regulation in greenhouses, performing closed-loop control of actuators such as ventilation windows, heating circuits, humidification systems, and shading or thermal screens based on continuous sensor feedback. The *Mithra Clima Pro* unit, deployed at the most instrumented sites (Alhama de Murcia, Sonora, and Falces), supports up to nine independent compartments and incorporates advanced multi-variable proportional–integral–derivative (PID) control algorithms. This device enables high-resolution regulation of air temperature, relative humidity, and CO_2 concentration by coordinating heating, ventilation, and screen systems. Additionally, it includes modular hardware expansion,

programmable time-based control, and remote communication capabilities for supervisory management. The *Mithra Clima* controller, installed in medium-complexity greenhouses such as Melipilla and Miranda, provides integrated control of temperature, humidity, and radiation-dependent screen positioning. It features hourly scheduling, dynamic control loops, and humidity-compensated modulation to maintain stable climatic conditions under variable exterior conditions. Finally, the *Mastia V* controller, used in conventional installations such as the Antalia greenhouse, delivers essential climate management functions, primarily temperature and ventilation control. It operates through predefined actuation stages and safety thresholds, offering a reliable yet cost-efficient solution for smaller or less instrumented facilities.

3.2. Dataset characterization

Table 1 shows the main characteristics and quality indicators of the indoor and outdoor datasets collected from the six greenhouses previously explained. The datasets on outdoor activities show a consistent pattern of missing data across all locations, typically around 20% of total observations (e.g., 19.96 per cent), with the longest consecutive intervals spanning up to 96 time intervals. These periodic gaps correspond to structured missing days that repeat approximately every five days, as shown by the short window line plots in Figs. A.4–A.6. This recurring pattern suggests systematic interruptions in external data acquisition rather than random losses, which required specific imputation and synchronization strategies during preprocessing.

In contrast, indoor datasets show greater heterogeneity in both completeness and statistical behavior. Sites such as Alhama de Murcia and Miranda exhibit relatively high data integrity, with missing rates below 6%, while Melipilla and Antalia show substantial data loss (up to 100% for humidity in Melipilla) and irregular coverage across variables. The maximum consecutive missing sequences in indoor data range from several hundred samples in well-instrumented sites to several thousand in those with limited maintenance, reflecting differences in hardware reliability and operational continuity. Regarding statistical properties, mean temperature values range from 14.2 °C (Melipilla exterior) to 23.5 °C (Sonora interior), highlighting the climatic diversity among locations. Humidity distributions are also markedly site-dependent, with interior averages between 57% and 72%, typically higher than outdoor levels due to evapotranspiration and ventilation effects. Radiation data show the largest dispersion (standard deviations exceeding 250 W/m^2 in some cases), capturing strong diurnal cycles and occasional sensor saturation events.

Comparative sample of the indoor and outdoor time series (see Figs. A.4–A.6) show a constant attenuation of amplitude and time lag in the indoor climate, indicating the buffering capacity of the greenhouse structure and its control systems. These effects are particularly pronounced in the case of temperature and radiation, where indoor signals tend to smooth out outdoor peaks. Similarly, the box plots and histograms in Figs. A.10–A.15 show clear discrepancies in the distribution between indoor and outdoor variables. For example, at the Sonora greenhouse, the indoor and outdoor temperature distributions are almost inversely shaped, highlighting the strong thermodynamic decoupling between the two domains. It is important to note that the datasets encompass a broad range of climatic conditions, data quality profiles, and control-system configurations. This diversity provides a robust testbed for evaluating the generalizability of the proposed outdoor-to-indoor mapping models and their capacity to handle noisy, heterogeneous, and partially missing real-world data.

3.3. Problem formulation

The proposed approach addresses two interconnected learning problems: (i) the *outdoor-to-indoor mapping* problem, and (ii) the *indoor climate forecasting* problem. Both are formulated in terms of learning functions that approximate the underlying relationships among meteorological and greenhouse variables observed over time. This two-stage formulation enables the development of forecasting systems even in sensor-limited greenhouses, where indoor data are not yet available, by leveraging outdoor observations to synthesize reliable indoor training series.

3.3.1. Outdoor-to-indoor mapping

The first challenge to address is the outdoor-to-indoor mapping problem, which consists of translating outdoor meteorological measurements into their indoor counterparts. This task does not involve forecasting future conditions, but rather learning a synchronous transformation that captures how the greenhouse structure and its control systems buffer, filter, and modify external influences. By establishing this mapping, outdoor records can be used to generate consistent synthetic indoor series, even in sensor-limited greenhouses where direct indoor observations are unavailable.

Let $\mathcal{X}_{1:t} = \{x_1, x_2, \dots, x_t\}$ denote the historical outdoor measurements within a time window of length t, where each $x_i \in \mathbb{R}^d$ represents a d-dimensional vector of outdoor variables such as temperature, relative humidity, and global radiation:

$$x_i = [T_i^{\text{out}}, H_i^{\text{out}}, R_i^{\text{out}}].$$

Similarly, let $\mathcal{Y}_{1:t} = \{y_1, y_2, \dots, y_t\}$ represent the corresponding indoor observations,

$$y_i = [T_i^{\text{in}}, H_i^{\text{in}}, R_i^{\text{in}}].$$

The goal of the mapping stage is to learn a transformation function

$$\mathcal{M} : \mathcal{X}_{1:t} \to \mathcal{Y}_{1:t},$$

such that $\mathcal{M}$ minimizes the discrepancy between the predicted indoor variables $\hat{\mathcal{Y}}_{1:t} = \mathcal{M}(\mathcal{X}_{1:t})$ and the real indoor observations $\mathcal{Y}_{1:t}$. This task is not a forecasting problem but rather a synchronous translation between outdoor and indoor conditions that captures the thermodynamic buffering and control responses of each greenhouse.

3.3.2. Indoor climate forecasting

Once the mapping model $\mathcal{M}$ has been trained, it can be used to generate synthetic indoor series $\tilde{\mathcal{Y}}_{1:t}$ from long-term outdoor records. The second stage involves forecasting future indoor conditions from either real or synthetic indoor data. Formally, given historical indoor measurements $\mathcal{Y}_{1:t}$ (real) or $\tilde{\mathcal{Y}}_{1:t}$ (synthetic), the forecasting model $\mathcal{F}$ learns to predict the next T-step-ahead sequence of indoor variables:

$$\mathcal{F}(\mathcal{Y}_{1:t}) \to \mathcal{Y}_{t:t+T}, \quad \text{or} \quad \mathcal{F}(\tilde{\mathcal{Y}}_{1:t}) \to \mathcal{Y}_{t:t+T}.$$

Here, $\mathcal{Y}_{t:t+T}$ denotes the true future indoor values over the forecasting horizon T. The objective is to minimize the error between the predicted and actual values across all variables and timestamps, typically using metrics such as MAE, RMSE, and the coefficient of determination (R^2).

3.4. Machine learning and deep learning models

To address the two-stage formulation described in Section 3.3, several neural architectures, $\mathcal{P}$, are employed to model both the outdoor-to-indoor mapping and the indoor forecasting tasks. Specifically, six representative deep learning models are considered: MultiLayer Perceptron (MLP), Long Short-Term Memory (LSTM), Convolutional Long Short-Term Memory (CNNLSTM), Long Short-Term Memory with Attention (LSTM-Attention), Convolutional Long Short-Term Memory with Attention (CNNLSTM-Attention), and Transformer. In the first stage, these models act as *mappers*, receiving synchronized outdoor variables (temperature, humidity, and radiation) as inputs and producing the corresponding indoor variables at matching timestamps. In the second stage, the same models are used as *forecasters*, learning to predict future indoor conditions from either real or synthetic indoor series.

Table 1
Dataset characteristics and quality metrics for indoor and outdoor greenhouse data across different locations.

Location	Type	Temperature					Humidity					Radiation				
		Missing (%)	Max Cons.	Valid Points	Mean	Std	Missing (%)	Max Cons.	Valid Points	Mean	Std	Missing (%)	Max Cons.	Valid Points	Mean	Std
Alhama de Murcia	Indoor	1.18	1646	207 940	18.74	7.55	1.45	1647	207 378	66.77	22.87	1.29	1646	207 715	189.86	270.52
	Outdoor	19.96	96	139 776	17.84	7.39	19.96	96	139 776	63.90	18.96	19.96	96	139 776	188.90	269.30
Falces	Indoor	5.92	2976	120 155	15.22	8.14	18.18	2981	104 496	71.16	20.31	41.44	2985	74 783	267.60	277.35
	Outdoor	19.96	96	139 776	9.80	7.59	19.96	96	139 776	69.69	16.94	19.96	96	139 776	167.68	246.00
Sonora	Indoor	7.07	2976	86 113	23.49	7.46	14.83	2976	78 921	68.89	19.21	42.82	2981	52 984	397.06	336.11
	Outdoor	19.96	96	139 776	21.41	8.85	19.96	96	139 776	42.44	22.61	19.96	96	139 776	232.12	315.41
Miranda	Indoor	4.32	482	27 580	20.32	5.67	8.03	482	26 511	66.79	19.59	39.29	482	17 500	278.73	269.51
	outdoor	19.96	96	139 776	19.40	6.24	19.96	96	139 776	67.07	16.48	19.96	96	139 776	191.08	270.87
Melipilla	Indoor	4.58	48	13 422	16.38	4.70	100.00	0	0	–	–	32.20	58	9537	257.76	287.15
	Outdoor	19.96	96	139 776	14.20	6.34	19.96	96	139 776	71.80	20.98	19.96	96	139 776	219.19	314.29
Antalia	Indoor	16.39	2976	51 573	22.23	9.63	35.71	16 720	39 660	57.51	26.92	47.98	2977	32 092	329.24	305.89
	Outdoor	19.96	96	139 776	20.81	7.07	19.96	96	139 776	55.42	23.51	19.96	96	139 776	223.16	296.81

Note: Max Cons. = Maximum consecutive missing values; – = No data available (NaN values replaced for clarity)

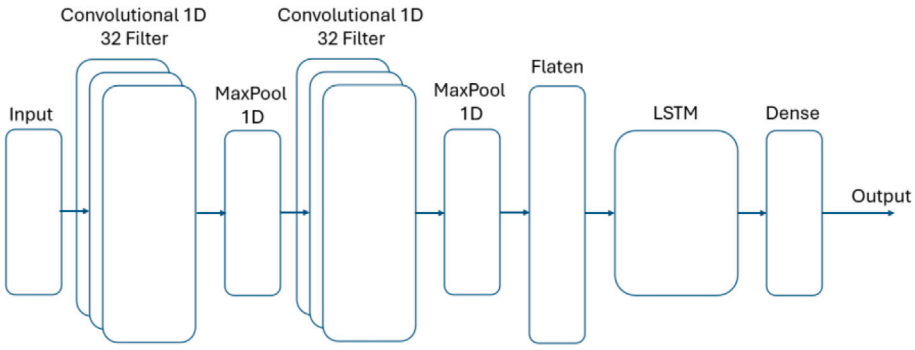

Fig. 1. Inner architecture of the CNNLSTM model used in the study.

This unified family of predictors has been selected for its ability to capture complex nonlinear dependencies, temporal dynamics, and site-specific thermodynamic effects inherent to greenhouse environments. Each architecture contributes a different learning bias, ranging from feedforward representation learning (MLP) to sequence modeling (LSTM) and long-range dependency extraction (Transformer), allowing a comprehensive evaluation of model suitability for both mapping and forecasting tasks.

1. **MLP**: This model represents the simplest form of a feedforward neural network, in which information flows unidirectionally from input to output through one or more hidden layers. Although originally developed for classification, it is equally applicable to nonlinear regression tasks. Each neuron performs a weighted sum of its inputs followed by a nonlinear activation, enabling the network to approximate complex input–output mappings. In this work, the MLP ingests flattened windows of environmental features (e.g., temperature, humidity, radiation) of length L with F variables (input size $F \times L$). The network uses three hidden fully connected layers with sizes $64 \to 128 \to 64$, each followed by `LeakyReLU` (negative slope 0.01) and dropout ($p = 0.1$). A final fully connected layer maps to $T \times O$ outputs, followed by a `Sigmoid`; predictions are reshaped to T future steps and O target indoor variables. The sigmoid bounds outputs to $[0, 1]$ (Rumelhart et al., 1986; Zhang, 2003).
2. **LSTM**: This model is a specialized recurrent neural network architecture designed to capture both short- and long-term temporal dependencies in sequential data. Each LSTM cell contains a set of gating mechanisms (input, forget, and output gates) that regulate the flow of information through the cell state. This design allows the model to preserve or discard information adaptively, avoiding vanishing or exploding gradient problems commonly found in standard RNNs. In the context of greenhouse modeling, LSTMs are particularly effective for representing delayed responses between outdoor and indoor variables and for forecasting smooth climatic transitions across time. In this work, we use a stacked `LSTM` with $K = 3$ layers (hidden size 256) that processes sequences of length L with F features; inter-layer dropout is 0.2 (active since $K > 1$). Hidden and cell states are initialized to zero; the last time-step representation is normalized with `LayerNorm` and passed through a dense head $256 \to 512 \to T \cdot O$ with `GELU` and dropout 0.1, producing T future steps for O target variables (reshaped to (T, O)). No activation is applied at the output (unconstrained regression) (Greff et al., 2016; Hochreiter & Schmidhuber, 1997).
3. **CNNLSTM**: This model is a hybrid neural architecture that combines convolutional and recurrent learning mechanisms to exploit both spatial and temporal dependencies within multivariate time series. As illustrated in Fig. 1, convolutional layers first perform local feature extraction from the input sequences, automatically identifying short-term trends and correlations among adjacent time steps or environmental variables. The extracted feature maps are then passed to an LSTM network, which integrates these representations over time and models their long-term dynamics. In this configuration, the CNN acts as an adaptive feature encoder, while the LSTM aggregates and contextualizes the extracted information to generate temporally consistent predictions. Within the proposed framework, CNNLSTMs are applied to both the outdoor-to-indoor mapping and indoor forecasting stages, leveraging their dual capacity to capture transient greenhouse fluctuations and sustained climatic patterns. Concretely, sequences of length L with F variables are permuted to channels-first and processed by two 1D convolutional blocks: $\texttt{Conv1d}(F, 32, \text{kernel} = 3, \text{pad} = 1) \to \texttt{LeakyReLU}(0.01) \to$

MaxPool1d(2), then Conv1d(32, 64, kernel = 3, pad = 1) → LeakyReLU(0.01) → MaxPool1d(2), yielding 64-channel feature maps at a downsampled temporal resolution. These features are fed to a stacked LSTM with num_layers = 2 and hidden_size = 128 (batch_first), and the last time-step representation is passed through a linear layer to produce $T \times O$ outputs, reshaped to T future steps and O target variables. No activation is applied at the output (unconstrained regression) (Borovykh et al., 2017; Sainath et al., 2015; Shi et al., 2015).

4. **LSTM-Attention:** While standard LSTMs effectively capture long-term dependencies in sequential data, they assign equal importance to all past timesteps, which can limit their responsiveness to sudden environmental changes. To address this limitation, this architecture integrates an attention mechanism that dynamically weights historical observations according to their relevance to the current prediction. The attention layer computes a set of context-dependent coefficients that emphasize the most influential temporal patterns, allowing the model to focus selectively on critical segments of the sequence. In the context of greenhouse modeling, this mechanism enhances the model's ability to detect short-term perturbations or abrupt transitions driven by external weather or internal control actions, improving interpretability and predictive performance in both mapping and forecasting tasks. Concretely, sequences of length L with F variables are encoded by a bidirectional LSTM with hidden_size = 256, num_layers = 3, batch_first, and dropout = 0.2 (active since num_layers > 1), producing 512-dimensional timestep embeddings. A Bahdanau (additive) attention module computes weights over the encoded sequence using the last hidden state, yielding a 512-dimensional context vector. The context and last hidden state are concatenated (size 1024), normalized with LayerNorm, and decoded by an MLP $1024 \to 1024 \to 512 \to T \cdot O$ with GELU activations and dropout (0.2, 0.1). A residual output projection is added before reshaping predictions to T future steps and O target variables; no activation is applied at the output (unconstrained regression) (Bahdanau et al., 2014; Qin et al., 2017).
5. **CNNLSTM-Attention:** This architecture extends the previous hybrid model by introducing an attention layer on top of the CNNLSTM backbone. In this configuration, the CNN performs localized feature extraction across time, the LSTM captures sequential dependencies, and the attention mechanism adaptively highlights the most relevant temporal representations before generating the final output. This multi-stage learning process combines hierarchical feature encoding with dynamic temporal focusing, enabling the network to effectively model complex nonstationary patterns. Within the proposed pipeline, CNNLSTM-Attention models have shown strong performance in capturing both gradual and abrupt indoor climate variations, particularly in greenhouses subject to fluctuating external weather conditions. Concretely, sequences of length L with F variables are permuted to channels-first and passed through three 1D convolutional blocks with residual connections: for channels [32, 64, 128] and kernel sizes [7, 5, 3] (padding = kernel/2), each block applies Conv1d → BatchNorm1d → GELU → Dropout(0.2) → Conv1d → BatchNorm1d → GELU, followed by MaxPool1d(2, padding = 1) on the first two blocks (identity on the last). The resulting features feed a bidirectional LSTM with hidden_size = 128, num_layers = 2, batch_first, and dropout = 0.2. A multi-head self-attention layer with 8 heads operates on the LSTM outputs; residual connections and LayerNorm are applied pre/post attention. Global average pooling over time yields a 256-dimensional context vector that is decoded via an MLP $256 \to 512 \to 256$ with GELU and Dropout (0.2, 0.1) and a skip connection. A final linear projection maps to $T \times O$ outputs, reshaped to T future steps and O targets, with no output activation (unconstrained regression) (Lai et al., 2018; Woo et al., 2018).
6. **Transformer:** This model represents a fully attention-based architecture that dispenses with recurrence altogether. Instead, it employs multi-head self-attention to capture dependencies across all positions in the input sequence simultaneously, enabling efficient learning of both short- and long-range temporal relationships. Positional encodings are used to preserve the order of the time steps, while multi-head attention layers facilitate the extraction of diverse contextual representations. Transformers have recently demonstrated state-of-the-art results in a variety of time-series forecasting tasks due to their parallelization efficiency and ability to model global dependencies. In this study, the Transformer architecture is applied to both the mapping and forecasting problems, where it effectively learns global correspondences between outdoor and indoor climatic variables and provides robust generalization across multiple greenhouse locations. Concretely, sequences of length L with F variables are projected to a $d_{\text{model}} = 256$ space via a linear layer, then augmented with sinusoidal positional encodings (dropout 0.1). The encoder comprises 4 layers of pre-normalized transformer blocks with 8 attention heads, feedforward dimension 1024, GELU activations, and dropout 0.1 (batch_first, norm_first). The final representation from the last time step is decoded by an MLP $256 \to 1024 \to 512 \to T \cdot O$ with GELU and dropout 0.1, and outputs are reshaped to T future steps and O target variables. No activation is applied at the output (unconstrained regression) (Vaswani et al., 2017; Wu et al., 2021).

It is important to note that each of these architectures operates under a different structural configuration, enabling them to capture complementary patterns and dependencies in the input sequences. We instantiate one model per architecture in $\mathscr{P}$ and train all models under a unified protocol: up to 1000 epochs with early stopping (*patience*=50, *min_delta*=0.0001), AdamW optimizer, Mean Squared Error (MSE) loss, learning rate 0.001 with *ReduceLROnPlateau* (factor 0.5, patience 10), and batch size 64. Model-specific architectural and regularization hyperparameters are summarized in Table 2; detailed architecture descriptions are provided above.

4. Evaluation & results

4.1. Execution environment

To run all the tests presented in this manuscript, a server (named "Mercurio") has been utilized, with the following hardware characteristics: An Intel® Xeon® Gold 6226R CPU, with 16 cores at 2.90 GHz, 196 GigaBytes DDR4 2933 MHz of RAM memory and 22 MB of cache memory. Two Nvidia® Quadro® RTX 5000 GPUs, with 16 GB GDDR6 384 Tensor cores, 3072 CUDA cores and NVLink® PCI Express x16 3.0. A Solid State Drive with 15 TeraBytes. Regarding the software characteristics, it should be mentioned that "Mercurio" runs on an Ubuntu 20.04 LTS operating system, which has Python version 3.8 with TensorFlow and Keras version 2.12 installed. In addition, for running all the code, the Jupyter development environment was used for the tests, more specifically, the Jupyter Notebooks.

4.2. Evaluation metrics

To assess the predictive performance of the models described in Section 3.4, five complementary evaluation metrics are employed; i.e., Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Coefficient of Variation of the RMSE (CVRMSE), and the Coefficient of Determination (R^2). These metrics collectively provide a comprehensive evaluation of model accuracy, robustness, and proportionality between predicted and observed values.

Table 2
Summary of hyperparameters for each deep learning model. All models use default configurations from the implementation, with consistent training settings across experiments. CNN filters and kernel sizes are listed in order from first to last layer.

Hyperparameter	Description	MLP	LSTM	CNNLSTM	LSTM+Att.	CNNL-STM+Att.	Transformer
Hidden size/Units	Number of neurons in hidden layers	64, 128, 64	256, 256, 256	128, 128	256, 256, 256	128, 128	256, 256, 256, 256
CNN filters	Number of feature detectors per layer	–	–	32, 64	–	32, 64, 128	–
CNN kernel size	Size of convolutional filters	–	–	3, 3	–	7, 5, 3	–
LSTM layers	Number of stacked LSTM layers	–	3	2	3	2	–
Transformer layers	Number of encoder layers	–	–	–	–	–	4
Feedforward dim.	Transformer feedforward dimension	–	–	–	–	–	1024, 1024, 1024, 1024
Attention heads	Number of multi-head attention heads	–	–	–	–	8	8, 8, 8, 8
Bidirectional	Bidirectional LSTM processing	–	No	No	Yes	Yes	–
Dropout rate	Dropout regularization rate	0.1, 0.1, 0.1	0.1, 0.1, 0.1	0.0	0.2, 0.1	0.2	0.1
Activation function	Primary activation function	LeakyReLU, Sigmoid	GELU	LeakyReLU	GELU	GELU	GELU
Batch size	Number of samples per training batch	64					
Epochs	Maximum number of training epochs	1000 (+ *EarlyStopping (patience=50, min_delta=0.0001)*)					
Optimizer	Weight update optimization algorithm	AdamW					
Loss function	Objective function minimized during training	Mean Squared Error (MSE)					
Learning rate	Step size for weight updates	0.001 (+ *ReduceLROnPlateau, factor=0.5, patience=10*)					

1. **MAE:** The Mean Absolute Error measures the average magnitude of the prediction errors, regardless of their direction. It provides a direct and interpretable indication of the average deviation between the predicted and actual values:

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

2. **RMSE:** The Root Mean Squared Error quantifies the standard deviation of the residuals, penalizing larger errors more heavily than MAE. It is defined as:

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

3. **MAPE:** The Mean Absolute Percentage Error expresses prediction accuracy as a percentage, providing a scale-independent measure that facilitates comparison across variables with different units.

$$\mathrm{MAPE}(y, \hat{y}) = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

   Although MAPE is intuitive, it can be unstable when y_i approaches zero; therefore, it is primarily used here for comparative rather than absolute interpretation.

4. **CVRMSE:** The Coefficient of Variation of the RMSE normalizes the RMSE by the mean of the observed values, yielding a dimensionless indicator of relative model error:

$$\mathrm{CVRMSE}(y, \hat{y}) = \frac{\mathrm{RMSE}(y, \hat{y})}{\bar{y}} \times 100\%.$$

   This metric is particularly useful for comparing models across datasets with different scales or magnitudes.

5. **Coefficient of Determination** (R^2): This metric evaluates the proportion of variance in the observed data that is explained by the model predictions. An R^2 value close to 1 indicates strong predictive agreement:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Each metric operates on two vectors: (1) $y = [y_1, y_2, \ldots, y_N]$, representing the observed ground-truth values, and (2) $\hat{y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N]$, corresponding to the model predictions. Together, these indicators ensure a robust and multidimensional evaluation of forecasting accuracy across all experimental conditions.

4.3. Evaluation of mapping models

As introduced in Section 3.3, the outdoor-to-indoor mapping problem aims to learn a transformation function $\mathscr{M} : \mathscr{X}_{1:t} \to \mathscr{Y}_{1:t}$ that accurately translates outdoor climatic conditions into their corresponding indoor responses. It is basically a regression problem in which the model must capture the nonlinear dependencies between outdoor temperature, humidity, and radiation, and their interior counterparts, effectively learning the thermodynamic buffering and control responses of each greenhouse.

Table 3 shows the performance of all architectures in estimating the mapping function $\mathscr{M}$ across the three variables. The results confirm that all models achieve low MAE and RMSE values and relatively high R^2, showing a strong ability to reconstruct indoor series from outdoor inputs. Across metrics, no single architecture consistently dominates all variables, suggesting that the learned mapping is not model-specific but rather a function of the data characteristics and site dynamics. The MLP shows consistently strong performance, ranking among the top three in MAE for all variables and achieving the best (tied) accuracy for radiation. Attention-based networks (LSTM-Attention and CNNLSTM-Attention) slightly outperform other architectures in terms of average R^2, although the differences are not very significant. As shown in Figs.

Table 3
Performance of the mapping models $\mathscr{M}$ across all variables (temperature, humidity, and radiation). Each column represents one of the six neural architectures evaluated. Values correspond to the mean ± standard deviation averaged across all datasets. Lower MAE, RMSE, and CVRMSE values indicate higher accuracy, while higher R^2 values denote stronger correlation between predicted and observed indoor variables.

Metric	MLP	LSTM	LSTM+Att.	CNNLSTM	CNNLSTM+Att.	Transformer
MAE (Temperature)	0.053 ± 0.028	0.058 ± 0.040	0.059 ± 0.043	0.062 ± 0.036	0.068 ± 0.051	**0.052 ± 0.028**
MAE (Humidity)	0.091 ± 0.031	0.095 ± 0.030	**0.084 ± 0.025**	0.100 ± 0.030	0.098 ± 0.034	0.092 ± 0.032
MAE (Radiation)	**0.057 ± 0.030**	0.059 ± 0.031	**0.057 ± 0.030**	0.064 ± 0.032	0.066 ± 0.039	0.060 ± 0.026
RMSE (Temperature)	**0.067 ± 0.037**	0.074 ± 0.055	0.071 ± 0.047	0.078 ± 0.050	0.081 ± 0.054	**0.067 ± 0.038**
RMSE (Humidity)	0.116 ± 0.040	0.122 ± 0.036	**0.110 ± 0.032**	0.125 ± 0.034	0.123 ± 0.039	0.119 ± 0.041
RMSE (Radiation)	**0.100 ± 0.050**	0.103 ± 0.050	0.102 ± 0.052	0.107 ± 0.049	0.113 ± 0.065	0.104 ± 0.042
MAPE (Temperature)	14.4 ± 8.2	16.9 ± 10.2	15.1 ± 9.5	14.3 ± 8.8	**13.8 ± 8.3**	15.5 ± 10.3
MAPE (Humidity)	15.4 ± 7.0	15.2 ± 5.8	**13.7 ± 5.8**	18.6 ± 8.8	17.5 ± 8.0	14.7 ± 6.3
MAPE (Radiation)	255.0 ± 338.5	140.8 ± 94.7	159.7 ± 153.9	147.3 ± 102.5	150.9 ± 67.2	**133.6 ± 69.3**
CVRMSE (Temperature)	18.199 ± 10.964	18.617 ± 11.070	19.451 ± 12.588	**16.323 ± 10.198**	16.983 ± 11.230	18.037 ± 10.924
CVRMSE (Humidity)	18.699 ± 10.606	19.545 ± 10.171	**17.702 ± 9.650**	21.835 ± 10.545	21.443 ± 10.428	19.259 ± 11.509
CVRMSE (Radiation)	77.707 ± 47.334	80.485 ± 49.245	79.184 ± 47.784	**59.933 ± 33.549**	63.856 ± 42.799	81.352 ± 47.350
R^2 (Temperature)	0.637 ± 0.238	0.573 ± 0.402	0.623 ± 0.268	0.589 ± 0.368	**0.646 ± 0.241**	0.616 ± 0.273
R^2 (Humidity)	0.608 ± 0.272	0.593 ± 0.230	**0.668 ± 0.194**	0.589 ± 0.203	0.586 ± 0.235	0.622 ± 0.239
R^2 (Radiation)	0.697 ± 0.282	0.681 ± 0.298	0.684 ± 0.295	**0.771 ± 0.229**	0.708 ± 0.358	0.685 ± 0.271

Table 4
Performance of the mapping models $\mathscr{M}$ for each dataset and variable. Each column corresponds to one greenhouse location. Metrics are averaged across models and variables, quantifying the local correspondence between outdoor and indoor dynamics. High R^2 and low MAE/RMSE values indicate effective learning of the outdoor-to-indoor mapping at that site.

Metric	Alhama	Antalia	Falces	Melipilla	Miranda	Sonora
MAE (Temperature)	0.036 ± 0.003	0.040 ± 0.003	0.053 ± 0.008	0.130 ± 0.021	0.063 ± 0.007	**0.029 ± 0.003**
MAE (Humidity)	0.080 ± 0.008	**0.063 ± 0.005**	0.122 ± 0.012	N/A	0.128 ± 0.008	0.073 ± 0.004
MAE (Radiation)	0.036 ± 0.005	0.043 ± 0.002	0.086 ± 0.011	0.109 ± 0.008	**0.030 ± 0.005**	0.058 ± 0.005
RMSE (Temperature)	0.046 ± 0.002	0.051 ± 0.003	0.067 ± 0.010	0.164 ± 0.019	0.074 ± 0.008	**0.037 ± 0.004**
RMSE (Humidity)	0.102 ± 0.009	**0.084 ± 0.004**	0.158 ± 0.012	N/A	0.159 ± 0.009	0.093 ± 0.005
RMSE (Radiation)	0.066 ± 0.007	0.081 ± 0.004	0.146 ± 0.011	0.179 ± 0.015	**0.045 ± 0.007**	0.112 ± 0.003
MAPE (Temperature)	9.5 ± 0.7	23.6 ± 1.1	21.1 ± 7.8	22.9 ± 3.0	9.0 ± 1.3	**4.0 ± 0.4**
MAPE (Humidity)	12.1 ± 2.5	**9.6 ± 0.7**	23.3 ± 4.3	N/A	23.1 ± 1.3	11.1 ± 0.8
MAPE (Radiation)	251.6 ± 292.2	161.0 ± 60.6	215.3 ± 92.7	214.3 ± 146.6	73.9 ± 18.0	**71.2 ± 19.4**
CVRMSE (Temperature)	11.108 ± 0.400	26.984 ± 1.647	24.350 ± 9.161	29.788 ± 3.456	**10.099 ± 1.093**	5.280 ± 0.511
CVRMSE (Humidity)	15.092 ± 3.070	**10.355 ± 0.521**	25.328 ± 3.925	N/A	35.063 ± 1.897	12.897 ± 0.637
CVRMSE (Radiation)	77.321 ± 18.070	63.455 ± 2.963	119.739 ± 37.970	122.118 ± 10.151	**18.581 ± 3.096**	41.303 ± 1.165
R^2 (Temperature)	0.755 ± 0.012	**0.810 ± 0.019**	0.473 ± 0.140	0.093 ± 0.206	0.789 ± 0.038	0.766 ± 0.043
R^2 (Humidity)	0.743 ± 0.050	**0.867 ± 0.013**	0.411 ± 0.088	N/A	0.345 ± 0.068	0.688 ± 0.035
R^2 (Radiation)	0.791 ± 0.058	0.831 ± 0.015	0.449 ± 0.177	0.282 ± 0.132	**0.977 ± 0.008**	0.897 ± 0.006

B.1–B.3, model variance is limited for temperature but increases for humidity and radiation, reflecting greater dataset sensitivity and data completeness issues in those variables.

At the architecture level, the Transformer trained on synthetic indoor data shows consistent gains for temperature and humidity: temperature MAE decreases from 2.48 to 2.40 (≈ −3%) and humidity MAE from 14.84 to 12.84 (≈ −13%), while R^2 increases from 0.622 to 0.633 for temperature and from 0.204 to 0.309 for humidity. These improvements across all metrics indicate that the Transformer particularly benefits from synthetic training, achieving superior accuracy and explanatory power relative to its real-trained counterpart.

Table 4 shows the performance of the mapping function across locations. The results reveal substantial heterogeneity among sites, mainly driven by local climatic contrasts and data quality. High-performing locations such as Alhama de Murcia, Antalia, Miranda, and Sonora exhibit tight coupling between outdoor and indoor dynamics, achieving temperature R^2 values between 0.79 and 0.81. In contrast, Melipilla and, to a lesser extent, Falces show weaker alignment between outdoor and indoor series, with lower R^2 (0.09–0.47) and higher error dispersion. In Melipilla, the temperature R^2 drops to 0.093, nearly 0.7 points below the best-performing site (Antalia, $R^2 \approx 0.81$), highlighting the difficulty of learning $\mathscr{M}$ when the relationship between indoor and outdoor variables is attenuated by control systems, poor data coverage, or missing variables (e.g., humidity). Moreover, the higher variance observed at these sites shows that model stability is worse when the training data are incomplete or poorly correlated between modalities.

Figs. B.4–B.6 show predictions of the indoor variables $\hat{\mathscr{Y}}_{1:t} = \mathscr{M}(\mathscr{X}_{1:t})$ across datasets and models. Some patterns are observed: (i) in Antalia, Melipilla, and Sonora, the indoor series are temporally shifted relative to outdoor signals, yet all models successfully capture this phase delay, aligning predictions with the correct indoor timing; (ii) despite differences in amplitude or waveform between outdoor and indoor patterns (e.g., in humidity for Antalia or Falces), the learned mappings accurately reproduce the overall dynamics; (iii) humidity mapping remains the most challenging variable due to the high rate of missing data and stronger nonlinear control interactions, especially in Falces and Miranda, where data gaps and weak correlations dominate, and models diverge significantly, leading to degraded mapping accuracy. These patterns confirm that $\mathscr{M}$ generalizes well when exterior–interior coupling is strong, but its fidelity diminishes under high data noise or low signal coherence.

4.4. Evaluation of forecasting models: Real vs. Synthetic training

Following the learning of the mapping function $\mathscr{M}$, the forecasting problem assesses whether models trained on synthetic indoor data can achieve predictive performance comparable to those trained on real indoor measurements. As formulated in Section 3.3, this forecasting problem aims to learn a temporal function $\mathscr{F}$ that predicts future indoor conditions $\mathscr{Y}_{t:t+T}$ based on past indoor observations $\mathscr{Y}_{1:t}$. Using synthetic data, the training data $\mathscr{Y}^{\text{synth}}_{1:t} = \mathscr{M}(\mathscr{X}_{1:t})$ are generated using the previously trained mappers. However, it is important to note

Table 5
Comparison of model performance across architectures. Each column corresponds to a model, and each row shows a performance metric averaged across variables (temperature, humidity, and radiation). Lower MAE, RMSE, and CVRMSE indicate better accuracy, while higher R^2 represents stronger correspondence between predicted and observed indoor variables.

Metric	CNNLSTM (R)	CNNLSTM (S)	CNNLSTM+Att (R)	CNNLSTM+Att (S)	LSTM (R)	LSTM (S)	LSTM+Att (R)	LSTM+Att (S)	MLP (R)	MLP (S)	Transformer (R)	Transformer (S)
MAE (Temp.)	2.47±0.38	2.79±0.77	2.57±0.36	2.78±1.22	2.58±0.40	3.59±1.72	**2.36±0.37**	3.01±1.36	2.44±0.29	3.31±1.17	2.48±0.29	2.40±0.62
MAE (Hum.)	13.47±3.27	14.24±3.82	13.39±3.74	12.83±3.47	13.52±3.77	14.96±4.31	12.79±3.77	13.90±3.68	**12.67±3.80**	14.83±2.51	14.84±5.23	12.84±3.43
MAE (Rad.)	55.22±28.65	77.23±57.24	48.79±32.27	59.98±48.44	54.29±31.28	121.57±122.92	**47.22±31.02**	82.17±80.53	51.13±37.44	69.75±60.91	50.03±29.15	68.99±69.11
RMSE (Temp.)	3.00±0.43	3.35±0.87	3.07±0.42	3.31±1.37	3.12±0.46	4.25±1.96	**2.85±0.43**	3.60±1.58	2.97±0.36	3.91±1.30	3.00±0.34	2.90±0.74
RMSE (Hum.)	16.86±3.81	17.51±4.07	16.70±4.16	16.03±4.32	16.99±4.20	18.43±4.87	**16.03±4.05**	17.05±4.39	15.78±4.21	18.00±3.12	18.63±6.34	16.04±3.92
RMSE (Rad.)	96.11±46.20	121.60±84.32	86.87±50.03	97.07±69.46	95.12±52.96	171.76±154.96	**84.58±52.06**	120.35±109.68	89.28±57.89	115.19±90.20	89.31±48.53	111.33±102.24
MAPE (Temp.)	**15.8±8.4%**	17.1±6.3%	20.0±18.1%	17.1±7.9%	21.2±17.9%	19.7±7.8%	18.5±13.4%	17.4±6.9%	17.5±8.2%	21.8±9.3%	23.7±24.6%	15.9±6.9%
MAPE (Hum.)	22.1±7.3%	24.9±10.2%	21.5±6.6%	20.5±6.9%	22.4±7.7%	23.2±7.3%	21.0±7.2%	21.6±6.9%	21.2±7.1%	24.0±6.1%	23.2±8.8%	**20.5±6.8%**
MAPE (Rad.)	N/A	N/A	N/A	**233.7±221.6%**	N/A	N/A	N/A	252.8±409.9%	N/A	N/A	N/A	N/A
CVRMSE (Temp.)	**17.95±6.51**	20.99±8.59	18.32±6.07	20.63±9.64	19.82±7.38	26.26±11.46	18.01±6.62	22.14±10.27	18.84±6.47	23.47±7.92	18.79±6.43	18.13±7.32
CVRMSE (Hum.)	30.35±15.65	31.42±13.83	30.07±14.93	30.69±19.25	31.89±16.99	34.20±17.02	29.82±15.08	32.37±18.54	**29.42±15.31**	33.52±17.28	36.11±25.13	30.24±16.93
CVRMSE (Rad.)	63.65±25.08	80.86±35.19	**57.52±25.70**	64.81±30.57	65.33±28.59	104.93±53.65	58.23±28.02	71.98±35.63	61.32±28.47	75.57±39.70	62.35±29.22	70.26±34.53
R^2 (Temp.)	0.621±0.182	0.550±0.232	0.614±0.194	0.576±0.262	0.605±0.208	0.394±0.409	**0.648±0.209**	0.520±0.299	0.647±0.162	0.497±0.271	0.622±0.208	0.633±0.202
R^2 (Hum.)	0.289±0.225	0.249±0.185	0.317±0.217	0.319±0.230	0.241±0.196	0.182±0.251	0.318±0.179	0.280±0.204	**0.331±0.213**	0.275±0.191	0.204±0.254	0.309±0.220
R^2 (Rad.)	0.773±0.147	0.662±0.226	**0.802±0.145**	0.767±0.176	0.762±0.165	0.377±0.555	0.799±0.152	0.690±0.292	0.786±0.149	0.705±0.230	0.778±0.156	0.726±0.229

Table 6
Performance of mapping models across datasets. Each column represents a greenhouse location (with Real and Synthetic data regimes). Metrics are expressed as mean ± standard deviation across models. Lower MAE, RMSE, and CVRMSE values indicate higher accuracy, while higher R^2 reflects stronger correspondence between predicted and observed indoor variables.

Metric	Alhama (R)	Alhama (S)	Antalia (R)	Antalia (S)	Falces (R)	Falces (S)	Melipilla (R)	Melipilla (S)	Miranda (R)	Miranda (S)	Sonora (R)	Sonora (S)	Avg (R)	Avg (S)
MAE (Temp.)	2.34±0.11	3.30±1.28	2.64±0.22	**1.84±0.09**	3.01±0.14	2.88±0.10	2.47±0.14	4.44±1.14	2.59±0.19	3.54±0.24	2.06±0.13	1.89±0.18	2.49±0.34	2.98±1.18
MAE (Hum.)	13.09±0.50	15.93±3.07	15.94±2.53	15.98±1.31	**7.92±0.20**	9.81±1.38	N/A	N/A	18.66±0.77	16.81±0.52	12.16±0.39	11.14±0.53	13.44±3.67	13.93±3.37
MAE (Rad.)	48.47±4.07	85.70±51.92	28.30±5.48	28.55±3.23	55.18±3.46	64.30±5.26	115.49±8.57	220.19±58.88	**25.12±1.82**	34.65±6.27	39.30±4.39	45.36±8.55	51.20±29.67	79.95±74.21
RMSE (Temp.)	2.84±0.12	3.92±1.40	3.12±0.23	2.21±0.12	3.59±0.16	3.43±0.11	3.07±0.14	5.21±1.27	3.17±0.24	4.25±0.29	2.48±0.14	**2.30±0.22**	3.01±0.39	3.55±1.34
RMSE (Hum.)	16.19±0.65	19.09±3.47	20.94±3.14	20.66±1.42	**10.85±0.27**	12.53±1.35	N/A	N/A	22.09±0.81	19.86±0.66	14.91±0.54	13.76±0.60	16.83±4.21	17.18±3.89
RMSE (Rad.)	87.43±5.63	127.05±60.17	47.75±7.89	**45.96±4.51**	99.05±5.31	108.22±8.70	192.61±9.63	317.81±66.74	46.85±2.27	55.92±9.81	73.46±7.21	81.00±12.00	90.34±47.85	122.88±100.34
MAPE (Temp.)	12.3±0.6%	17.4±8.3%	15.4±1.0%	11.5±0.6%	51.2±15.3%	26.5±1.7%	13.1±0.6%	21.2±4.0%	18.9±2.2%	22.8±1.1%	10.1±0.9%	**9.5±0.7%**	19.3±15.1%	18.2±7.3%
MAPE (Hum.)	20.9±1.3%	26.0±6.8%	32.1±2.6%	30.4±1.7%	**10.9±0.4%**	13.7±2.1%	N/A	N/A	26.2±0.8%	24.6±1.3%	20.3±1.1%	17.5±0.7%	21.9±6.9%	22.4±7.0%
MAPE (Rad.)	233.5±35.3%	466.6±637.3%	72.0±4.3%	**68.0±2.4%**	318.2±92.5%	567.3±682.4%	227.2±92.5%	N/A	95.7±7.5%	1133.8±2103.7%	102.0±10.4%	190.5±178.9%	171.2±107.5%	1264.6±3505.9%
CVRMSE (Temp.)	14.75±0.62	19.72±7.11	18.49±1.36	13.04±0.70	30.54±1.03	29.19±1.25	16.92±0.83	28.76±7.05	22.31±1.77	29.82±2.13	11.92±0.73	**11.08±1.09**	18.57±6.16	21.94±8.99
CVRMSE (Hum.)	25.68±0.90	29.90±5.49	59.82±8.86	58.89±3.93	**12.87±0.31**	14.64±1.47	N/A	N/A	38.55±1.53	35.32±1.11	23.28±0.82	21.61±0.90	31.15±16.05	32.08±15.74
CVRMSE (Rad.)	54.28±3.21	78.22±37.93	33.88±5.62	**32.61±3.19**	110.73±5.66	124.33±12.43	67.86±2.97	107.49±22.24	65.38±3.15	77.86±13.43	41.32±3.56	45.82±6.42	61.32±25.66	78.07±38.36
R^2 (Temp.)	0.551±0.029	0.381±0.196	0.790±0.024	**0.887±0.012**	0.260±0.047	0.305±0.030	0.629±0.022	0.256±0.225	0.712±0.035	0.536±0.046	0.782±0.018	0.805±0.029	0.625±0.182	0.528±0.276
R^2 (Hum.)	0.401±0.041	0.277±0.172	0.023±0.115	0.093±0.043	0.196±0.034	0.178±0.048	N/A	N/A	0.134±0.020	0.191±0.023	0.549±0.033	**0.607±0.024**	0.286±0.204	0.269±0.201
R^2 (Rad.)	0.817±0.018	0.621±0.344	0.938±0.020	**0.948±0.010**	0.554±0.043	0.483±0.078	0.639±0.023	0.250±0.323	0.826±0.015	0.757±0.076	0.903±0.015	0.882±0.030	0.784±0.143	0.654±0.316

that the assessment is always carried out using actual data from the greenhouses under study.

Six forecasting architectures were employed under both regimes: MLP, LSTM, CNNLSTM, LSTM with Attention, CNNLSTM with Attention, and Transformer. Table 5 shows an aggregated view of their performance across temperature, humidity, and radiation variables. Overall, models trained on synthetic indoor series exhibit higher error and slightly lower explanatory power than their real-trained counterparts, but remain competitive in absolute terms. On average across architectures, the use of synthetic training data increases MAE by approximately +0.495 °C for temperature (+16%), +0.489 for humidity (+3%), and +28.746 W/m^2 for radiation (+56%). Correspondingly, the average R^2 decreases by −0.10 (−16%) for temperature, −0.017 (−5%) for humidity, and −0.13 (−16%) for radiation, as observed in the "Average" rows of Table 6. These results demonstrate that while real indoor data still generate the most accurate forecasting models, synthetic series generated through $\mathscr{M}$ can serve as a viable substitute for initialization and model training in sensor-limited settings.

A key observation from both tables is that variance increases significantly with synthetic training. Standard deviations nearly double for temperature and radiation (e.g., MAE standard deviations of 1.177 vs. 0.339 for temperature and 74.205 vs. 29.671 for radiation), with similar inflation observed in R^2. This dispersion indicates that synthetic series, although informative, introduce additional noise and site-dependent variability. As shown in Figs. C.1–C.3 show that the variability is especially pronounced for the LSTM and attention-based models, while hybrid architectures such as CNNLSTM and the Transformer demonstrate more stable performance under synthetic training. This suggests that models with spatially aware components or attention mechanisms can better accommodate the residual distortions of synthetic indoor dynamics.

Despite the overall performance gap, several datasets show cases where synthetic-trained models outperform those trained on real data. In particular, temperature forecasting improves under synthetic training in Antalia (MAE reduction of −0.802 °C, −30%; R^2 increase of +0.097, +12%) and in Falces (MAE reduction of −0.128 °C, ∼4%; R^2 increase of +0.045, ∼17%). Sonora also exhibits enhanced performance, with temperature MAE improving by −0.167 °C (∼8%) and humidity by −1.021 (∼8%), accompanied by modest gains in R^2 (+0.023 and +0.058, respectively). These cases correspond to sites with high-quality outdoor–indoor coupling and limited control interference, where $\mathscr{M}$ reproduces the physical transfer function with high fidelity. By contrast, performance degradation is more notable in Melipilla, where data sparsity and weak correlation between modalities hinder both mapping and subsequent forecasting.

MAPE values are less reliable for radiation due to division by low denominators in early morning or evening hours, making MAE, RMSE, CVRMSE, and R^2 the preferred metrics for assessing consistency across regimes. Visual inspection of the qualitative predictions in Figs. C.4, C.5 and C.21 confirms the main findings. Synthetic-trained models reproduce the main dynamics and seasonal patterns of indoor temperature and humidity but tend to show higher residuals in transient regimes and under abrupt environmental changes. These results collectively indicate that synthetic data generated through learned mappings provide a robust foundation for rapid model deployment in new greenhouses, effectively bridging the data gap when historical interior records are unavailable.

4.5. Statistical significance and robustness analysis

To rigorously evaluate the predictive performance and validate the robustness of the synthetic data substitution, we conducted a non-parametric Wilcoxon signed-rank test. This test compared the absolute forecasting errors of the models trained on synthetically generated data against those trained on actual historical sensor data. The full statistical results across all 102 experimental combinations (six sites,

six architectures, and three variables) are detailed in Appendix D, Table D.1.

The statistical analysis yields several critical insights regarding the viability of exogenous bootstrapping for zero-history indoor forecasting:

Overall viability of synthetic substitution. In 51% of all evaluated scenarios (52 out of 102), the models trained purely on synthetically mapped data either performed statistically identically ($p \geq 0.05$, 24 cases) or significantly outperformed ($p < 0.05$, 28 cases) the models trained on actual historical sensor data. This provides empirical evidence supporting the core hypothesis of the proposed framework: synthetic indoor climate series can act as a highly effective and robust substitute for real sensor histories in newly instrumented environments.

Performance by climate variable. The ablation study reveals that the synthetic mapping is exceptionally robust for thermodynamic buffering, specifically temperature. For temperature forecasting, synthetic-trained models equaled or surpassed real-trained baselines in 58% of the cases. Humidity prediction showed a perfectly balanced outcome (50% equivalent or better). However, radiation proved to be the most challenging variable to substitute synthetically. Real-trained models yielded statistically lower errors in 20 out of 36 radiation tests, likely because localized cloud cover dynamics and physical shading structures are difficult to infer purely from external meteorological signals without direct indoor history.

Architectural resilience. The statistical evaluation demonstrates a clear dependency on the underlying deep learning architecture. Advanced attention-based architectures, specifically the Transformer, exhibited extraordinary resilience to synthetic data substitution. The Transformer trained on synthetic data equaled or outperformed its real-trained counterpart in 70% of its deployments (12 out of 17 cases). This suggests that the self-attention mechanism is highly capable of filtering out minor synthesis artifacts and focusing on the underlying structural climate patterns. Conversely, simpler recurrent (LSTM) or feed-forward (MLP) architectures suffered higher rates of performance degradation when exposed to synthetic inputs.

Geographic and structural generalizability. Finally, the Wilcoxon test quantifies the geographic generalizability of the framework. In strongly coupled environments, such as the Antalia site, the synthetic approach successfully replaced real data in over 60% of the evaluated models. Conversely, sites with potentially weaker outdoor–indoor coupling or more aggressive manual climate interventions, such as the Miranda greenhouse, showed a higher reliance on real historical data (where real models performed better in 10 out of 18 cases). This highlights the operational boundaries of exogenous bootstrapping, indicating that synthetic mapping is most reliable in environments governed primarily by passive thermodynamic responses rather than unpredictable manual interventions.

5. Discussion

The results obtained in the two-stage process, comprising outdoor-to-indoor mappers and indoor predictors, demonstrate that greenhouse dynamics can be effectively inferred from external weather conditions and that synthetic indoor series can serve as a practical substitute for missing historical sensor data.

The mapping experiments confirmed that all deep learning architectures were able to learn the outdoor–indoor transformations with high accuracy, achieving low MAE and RMSE values and strong determination coefficients across most sites. This supports the hypothesis that a significant portion of greenhouse thermal and radiative behavior can be explained by local weather patterns filtered through the structure's physical and control responses.

Among the mapping architectures, recurrent and attention-based models (notably LSTM-Attention and CNNLSTM-Attention) achieved the most stable performance across diverse climatic conditions, capturing both short-term fluctuations and longer-term dependencies between outdoor and indoor variables. The Transformer-based mapper also displayed strong generalization, particularly when longer outdoor datasets were available, confirming the suitability of attention mechanisms for representing greenhouse thermodynamics. However, mapping accuracy varied with site-specific characteristics: locations with strong outdoor–indoor discrepancies (e.g., high buffering or distinct ventilation regimes) showed greater residual error, indicating that local control strategies and greenhouse design can limit the generalizability of purely data-driven mappings.

Furthermore, it is important to contextualize the lower R^2 and higher variance observed specifically in humidity predictions compared to temperature. Humidity is heavily influenced by complex biological processes, such as plant evapotranspiration and condensation, making it more challenging to map using only exogenous meteorological data. This difficulty is exacerbated by the presence of missing data points and sensor gaps in the raw datasets. Finally, this variance must be evaluated within the context of our 3-day (72-hour) forecasting horizon, which naturally accumulates significantly more uncertainty than the 1-to-24-hour horizons typically evaluated in existing literature.

The prediction stage provided further evidence of the practical viability of the proposed approach. Models trained with synthetic indoor data consistently achieved performance similar to those trained with real indoor data, with error profiles that are, in some cases, practically indistinguishable. Moreover, across several datasets, predictors trained with synthetic data outperformed their counterparts trained with real data, suggesting that synthetic data can introduce beneficial regularization effects or mitigate overfitting to noisy historical datasets. This finding is particularly significant, as it demonstrates that greenhouse prediction systems can be implemented even when no previous indoor data exists, using only long-term meteorological records and a trained mapper.

From an architectural perspective, the LSTM and CNNLSTM families maintained strong predictive capacity across both training regimes, while the attention-enhanced variants generally delivered the best overall balance between accuracy and robustness. Notably, the Transformer forecaster trained on synthetic indoor data achieved better results than its counterpart trained on real indoor data. These outcomes highlight that attention-based temporal modeling, when combined with synthetic data generation, constitutes a powerful strategy for climate forecasting in data-limited greenhouses.

Regarding the choice of baselines, our experimental design specifically isolates the impact of synthetic data by comparing it directly against the "gold standard" of real-sensor data across a spectrum of architectures, from simple MLPs to complex Transformers. We deliberately exclude domain adaptation techniques and purely physics-informed models from this comparison, as they inherently violate the "zero-history" constraint of our problem formulation. Domain adaptation typically requires a sample of target-domain data to align feature spaces, which is unavailable in a newly instrumented facility. Similarly, purely physical models require precise structural metadata (e.g., thermal transmittance, precise greenhouse volume) that is rarely available at scale. Therefore, the simple MLP serves as our baseline for algorithmic complexity, while the real-data trained models serve as the upper-bound baseline for data quality, providing a clean ablation of the synthetic mapping's effectiveness.

Interpreting these empirical results through a domain-specific physical lens, our synthetic mapping can be viewed as a data-driven surrogate for classical greenhouse energy balance dynamics. The high reliability observed in temperature mapping confirms that attention-based networks implicitly learn thermodynamic buffering coefficients, such as thermal inertia and time-lagged heat dissipation, directly from exogenous meteorological signals. Conversely, the degraded performance in radiation mapping reflects the physical reality that solar transmission is highly non-linear and easily disrupted by unmeasured

local phenomena, such as condensation-altered refraction or mechanical shading screens. Therefore, this approach establishes a clear physical boundary for AIoT deployment: data-driven mapping is highly reliable for variables governed by passive thermal mass (temperature, humidity), but inherently limited for those dictated by immediate physical obstructions.

To address concerns regarding bias amplification and error propagation, we explicitly analyzed the relationship between mapping residuals (Stage 1) and downstream forecasting errors (Stage 2). Our analysis reveals that there is no substantial linear correlation between the magnitude of the mapping error and the final predictive performance. This suggests that the forecasting models exhibit a degree of "error absorption" rather than amplification. We attribute this stability to two factors. First, the mapping stage acts as a physics-informed filter that smooths high-frequency sensor noise, providing a more consistent signal for the forecaster to learn. Second, the use of attention-based architectures like the Transformer provides an implicit form of uncertainty management, as the model learns to prioritize reliable temporal features over those that may have been poorly mapped. While we acknowledge that a formal uncertainty quantification (e.g., Bayesian neural networks or conformal prediction) would provide a deeper calibration analysis, the current empirical stability across diverse sites suggests that the synthetic mapping stage provides a robust enough representation of greenhouse dynamics to support reliable downstream forecasting without catastrophic error propagation.

A practical consideration when deploying deep learning regression models in physical environments is the potential for generating physically unrealistic predictions, such as negative solar radiation or relative humidity exceeding 100%. Because standard neural networks perform unconstrained regression, they do not inherently respect these domain boundaries. In this study, we deliberately evaluated and reported the "raw" model outputs to transparently assess the unconstrained learning capacity, bias, and error profiles of each architecture. However, for real-world AIoT deployment, a deterministic post-processing layer is essential. Implementing simple boundary constraints, such as clipping relative humidity to a maximum of 100% and applying a zero-floor to radiation predictions, would easily prevent these physical violations and further reduce operational error without degrading the underlying predictive power of the models.

Finally, it is important to note that the experimental findings validate the feasibility of the proposed pipeline as a bridge between external weather information and indoor greenhouse forecasting. The approach enables new greenhouses, often lacking interior sensor histories, to immediately benefit from predictive models, facilitating rapid deployment of AI-driven control systems without long calibration phases.

6. Conclusions and future work

This study proposed and validated a two-stage deep learning framework to address the scarcity of historical indoor data in greenhouse environments. The first stage learns an outdoor-to-indoor mapping that generates synthetic microclimate variables (temperature, humidity, radiation) from meteorological records, while the second stage uses these synthetic series to train forecasting models for short-term prediction of indoor conditions. Results obtained across six geographically diverse greenhouses and six deep learning architectures demonstrate that (i) outdoor–indoor dynamics can be accurately modeled through neural mappings, and (ii) forecasters trained on synthetic indoor data achieve performance comparable to those trained on real measurements. Attention-based models consistently exhibited superior generalization, while simpler architectures such as MLPs offered competitive, computationally efficient baselines. Crucially, this study provides actionable guidelines for deployment: attention-based architectures, specifically Transformers, should be prioritized as they demonstrate a unique capacity to filter synthetic artifacts that cause simpler recurrent models to degrade. Furthermore, while the framework is highly robust for thermodynamically buffered variables like temperature, its application to radiation is physically bounded by unmeasured local shading dynamics. Notably, the Transformer forecaster trained on synthetic indoor data outperformed its counterpart trained on real measurements (particularly for temperature and humidity and at longer horizons) highlighting the practical advantage of attention-based models under synthetic training. These findings confirm that synthetic indoor data can serve as a viable surrogate for real historical data, enabling cost-effective deployment of forecasting systems in newly instrumented or data-sparse facilities.

The implications extend beyond greenhouse management. By leveraging abundant outdoor data to emulate indoor conditions, this framework provides a scalable foundation for AIoT-based climate control and decision-support systems. It also offers a practical route towards the rapid deployment of intelligent control in precision agriculture, accelerating its digital transformation. Ultimately, while this work establishes the viability of synthetic data substitution, uncertainty quantification and the integration of diverse control scenarios remain critical next steps to ensure operational robustness. Future research will therefore extend the comparative analysis to hybrid physical—data-driven baselines, enabling a more granular assessment of the trade-offs between synthetic-trained models and conventional site-specific approaches in high-intervention agricultural systems. In parallel, efforts will focus on improving model generalization and integration. Developing a general-purpose mapper capable of transferring across greenhouses without site-specific retraining would enable on-demand generation of synthetic indoor data from external weather records. Incorporating actuator and control variables (e.g., ventilation, heating, irrigation) is also expected to enhance physical realism and forecasting accuracy. Additionally, since the current experiments were conducted offline on high-performance computing servers, an important step towards practical deployment will be benchmarking inference times and computational footprints on resource-constrained IoT hardware. Integrating the optimized framework within digital twin architectures and enabling edge-based inference would ultimately support real-time, adaptive climate prediction in operational greenhouses.

CRediT authorship contribution statement

Juan Bonastre-Egea: Methodology, Software, Investigation, Writing – original draft, Visualization. **Andrés Bueno-Crespo:** Methodology, Validation, Investigation, Writing – review & editing, Funding acquisition. **Virginia C. Sánchez:** Software, Writing – original draft. **José M. Cecilia:** Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Juan Morales-García:** Conceptualization, Methodology, Software, Validation, Investigation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Consent to publish

All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the SERGIoT project (CPP2023-010458), funded by MICIU/AEI/10.13039/501100011033 (Spain) and the European Union through FEDER; the SATRAI project (CIGE/2024/199), funded by the Conselleria de Educación, Cultura, Universidades y Empleo (Generalitat Valenciana); and the AM-DS project (INREED/2024/1 and INREED/2024/7), funded by the Conselleria de Innovación, Industria, Comercio y Turismo (Generalitat Valenciana) through the GVANEXT programme and the European Union through NextGenerationEU/PRTR.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Data description

See Figs. A.1–A.15.

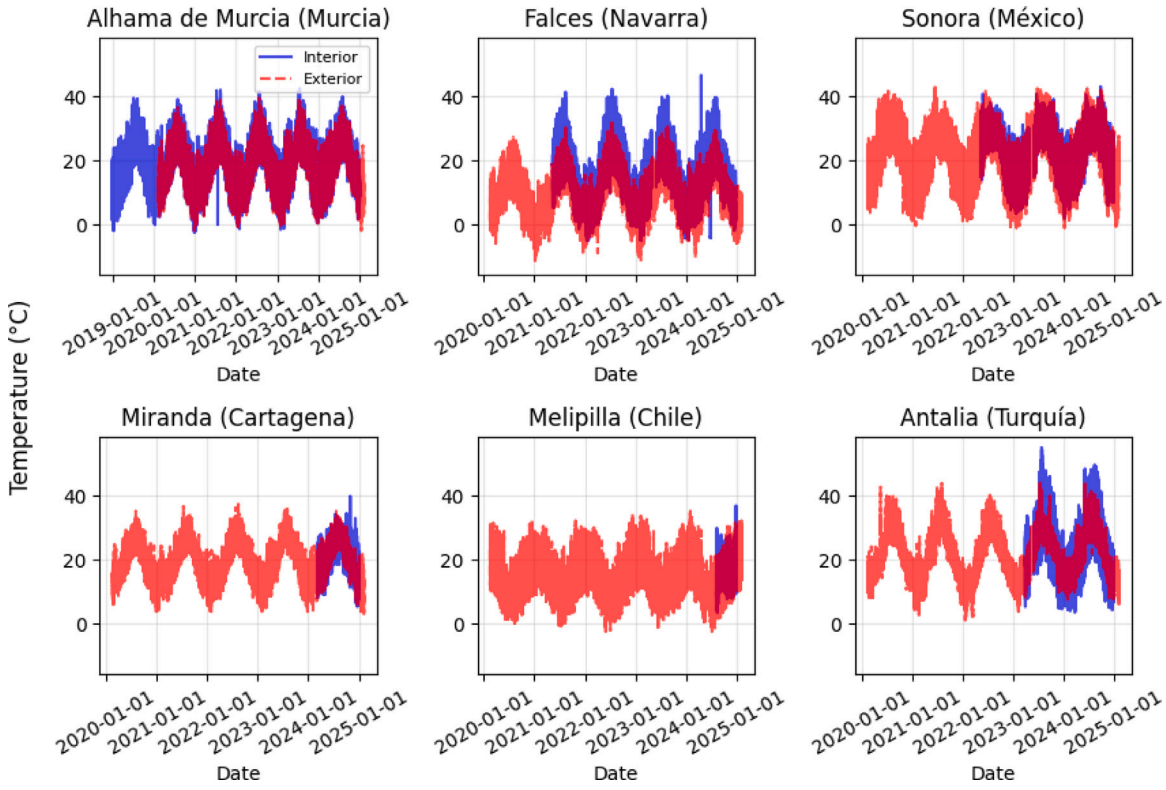

Fig. A.1. Temporal evolution of indoor and outdoor temperature across all greenhouse locations. The interior series exhibit a smoother profile and reduced amplitude compared to the outdoor signals, reflecting the thermal buffering effect of the greenhouse structure and control systems.

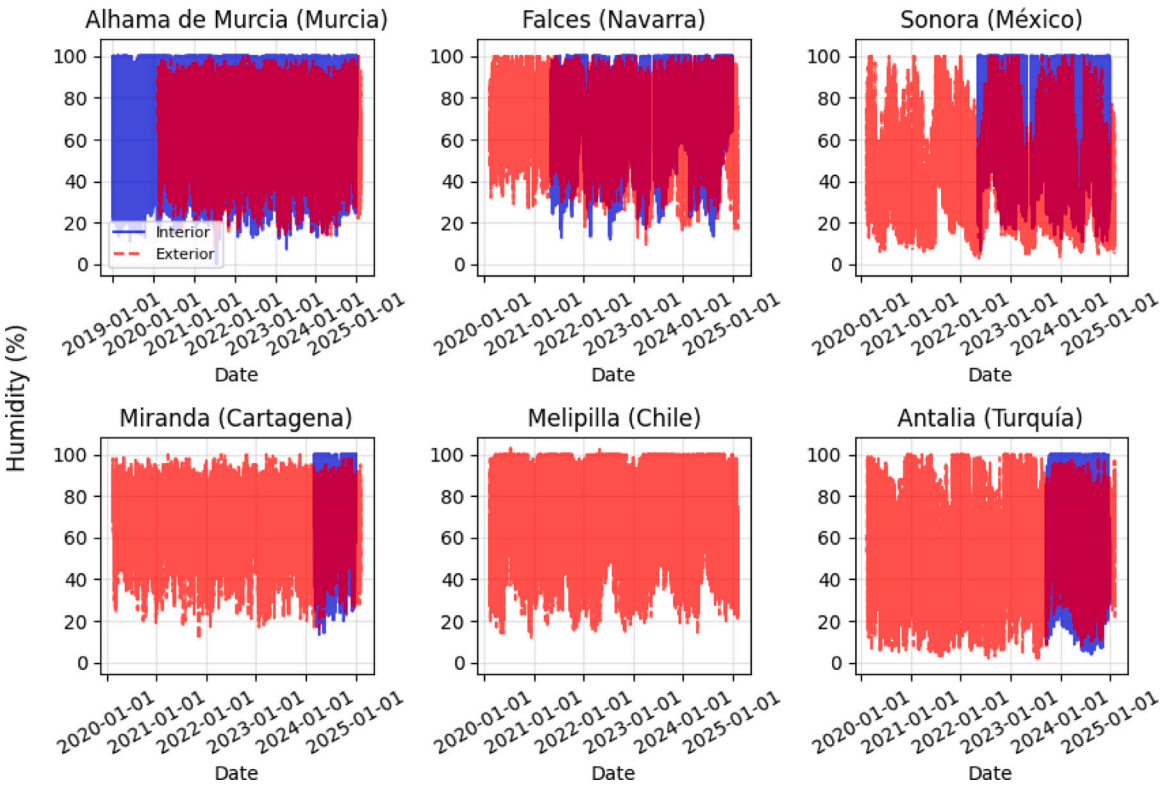

Fig. A.2. Time series of indoor and outdoor relative humidity for all sites. Indoor humidity values remain consistently higher and less variable due to evapotranspiration and controlled ventilation, while outdoor conditions show greater short-term fluctuations.

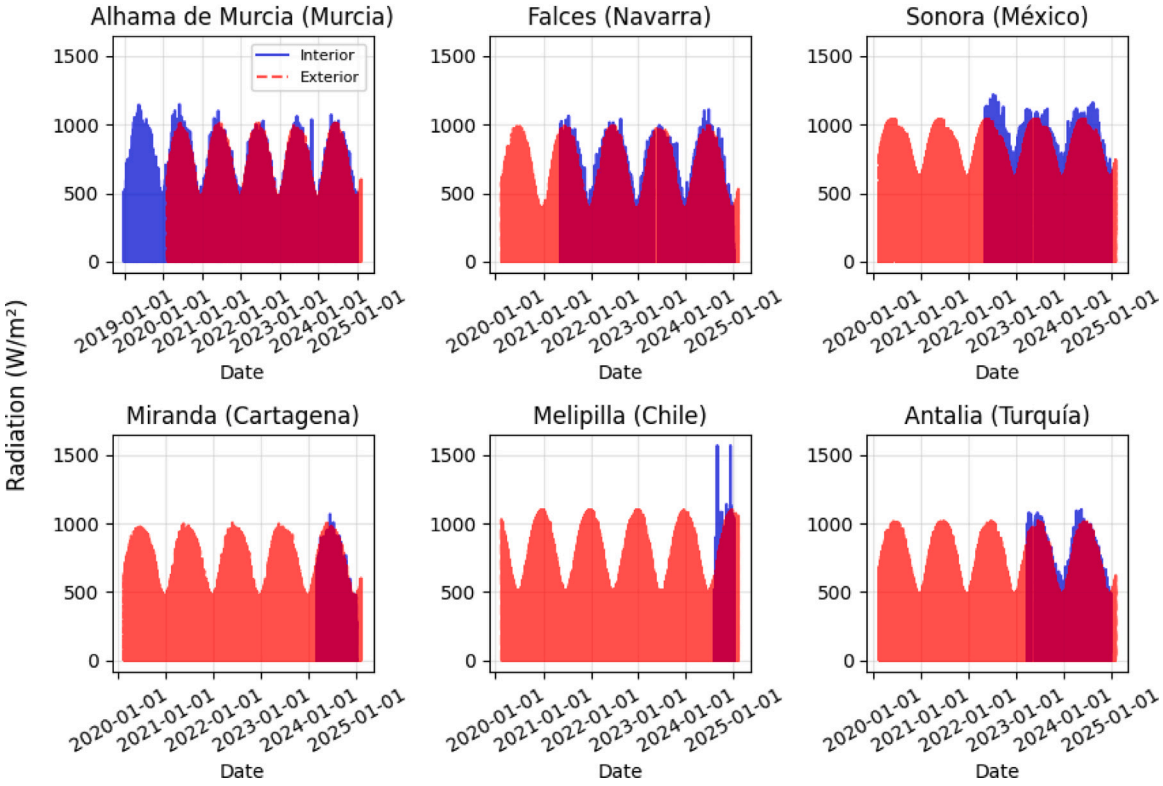

Fig. A.3. Comparison between indoor and outdoor global radiation measurements. The indoor radiation signals display attenuated peaks and reduced variance, indicating the shading and transmissivity effects of greenhouse covers and screen management systems.

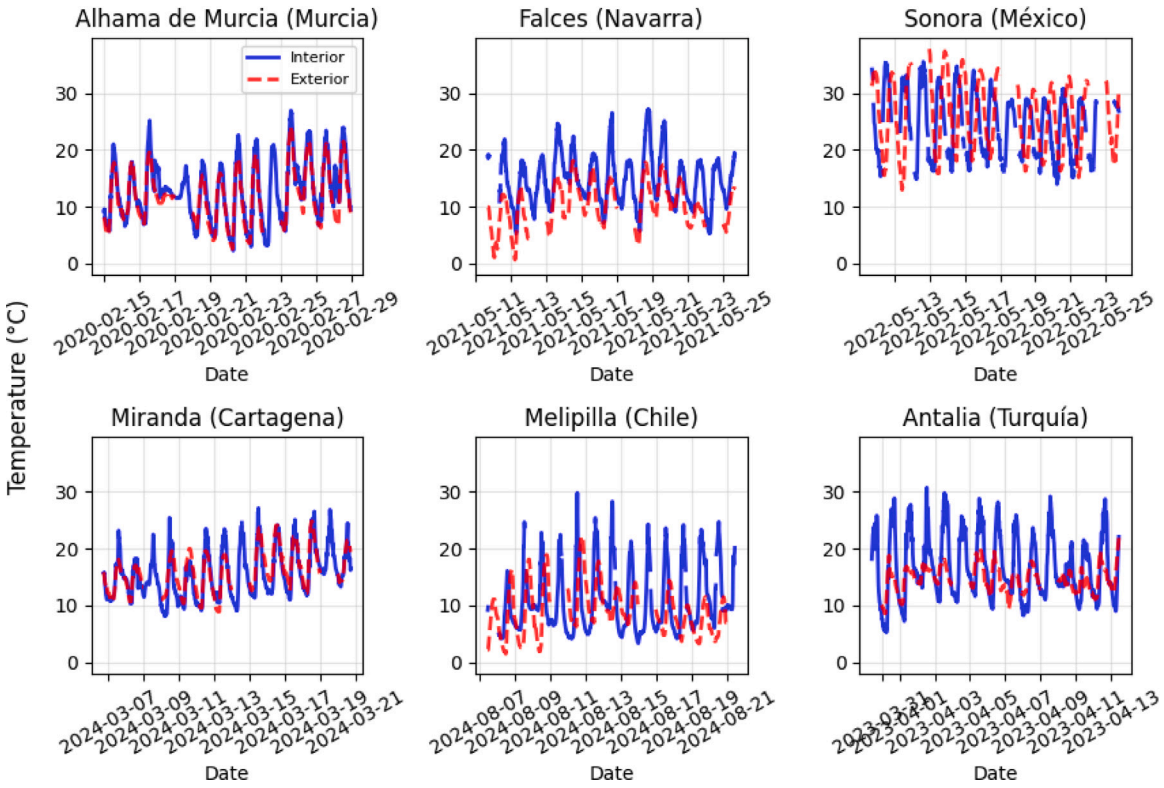

Fig. A.4. Short-window time series of indoor and outdoor temperature during the first 20 days. Indoor temperatures show smoother fluctuations and delayed responses relative to outdoor dynamics, illustrating the thermal inertia and regulation mechanisms within greenhouse environments.

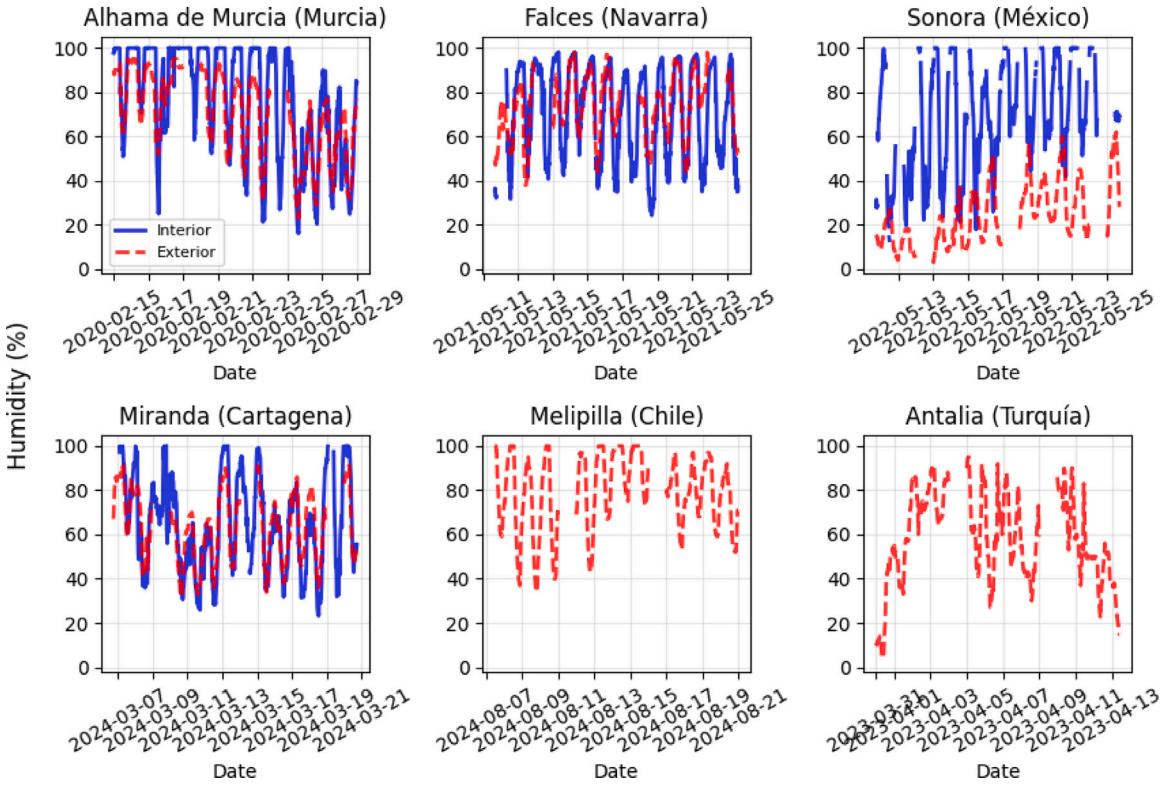

Fig. A.5. Time series of indoor and outdoor relative humidity over a 20-day period. The indoor humidity exhibits higher baseline levels and attenuated variability compared to the outdoor conditions, reflecting the combined effects of evapotranspiration and humidity control systems.

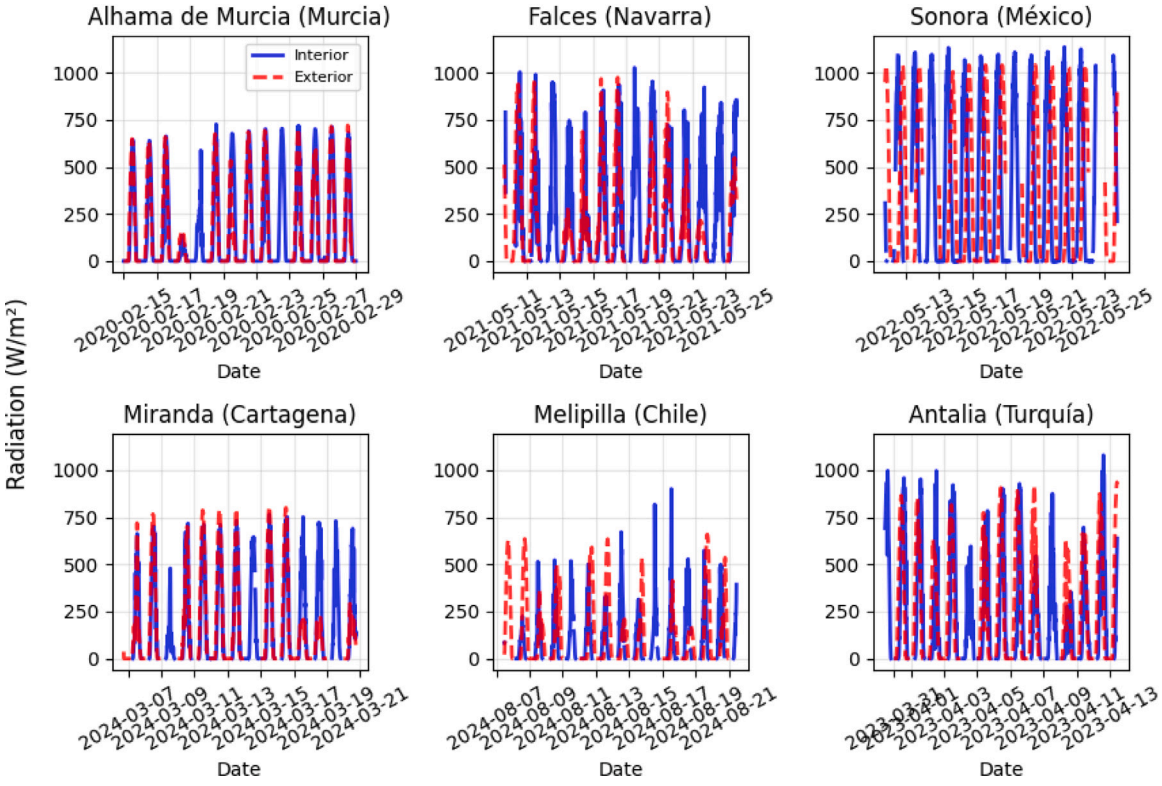

Fig. A.6. Short-term comparison of indoor and outdoor global radiation across the first 20 days. The attenuation of radiation peaks within the greenhouse reveals the influence of structural shading, cover transmissivity, and screen management strategies.

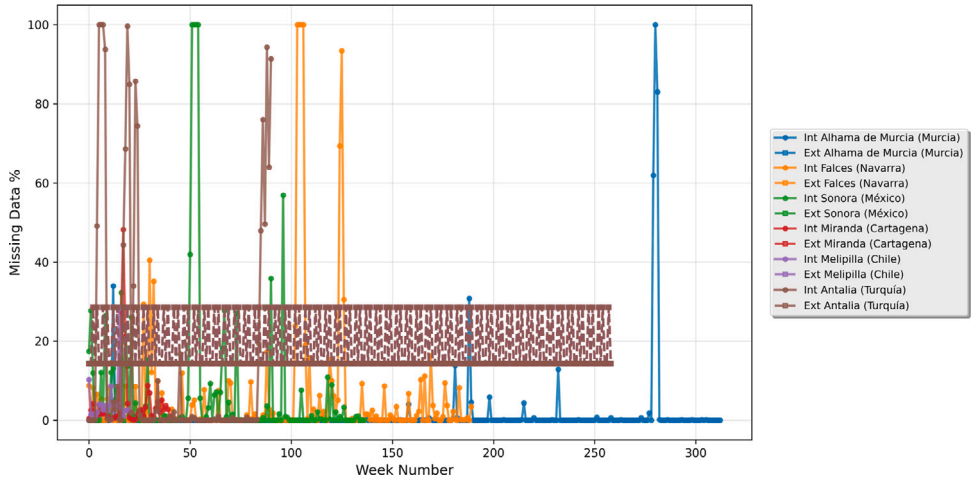

Fig. A.7. Temporal distribution of missing data in indoor and outdoor temperature measurements across all sites. The pattern reveals periodic interruptions, particularly in the outdoor series, likely associated with scheduled downtime or data acquisition failures.

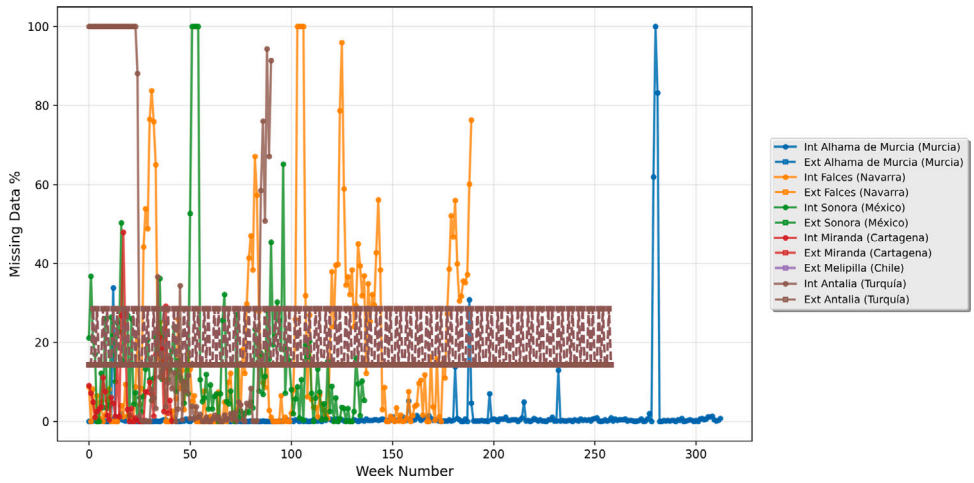

Fig. A.8. Missing data pattern for relative humidity in indoor and outdoor datasets. The interior sensors show variable completeness depending on site maintenance, while outdoor series display more systematic gaps due to network-level interruptions.

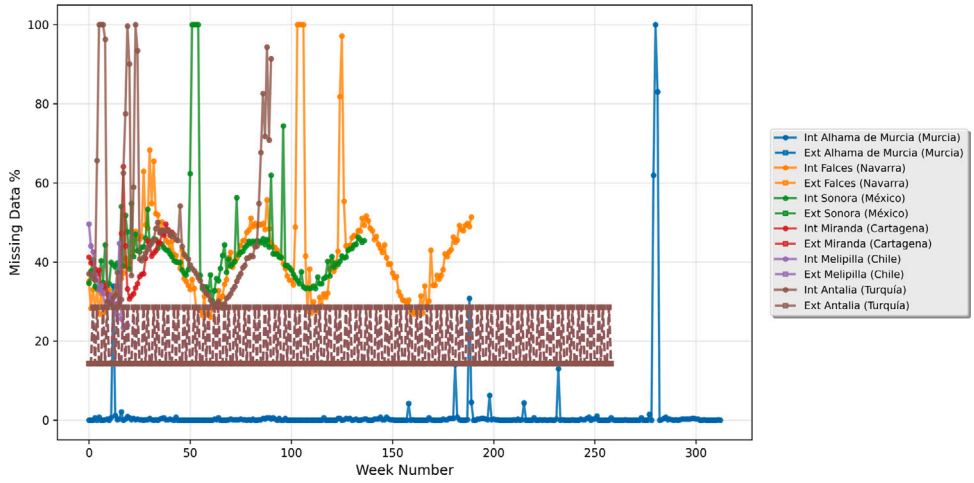

Fig. A.9. Missing data distribution for global radiation across indoor and outdoor measurements. The interior radiation records exhibit greater discontinuities, reflecting sensor limitations and the influence of control screens on signal acquisition.

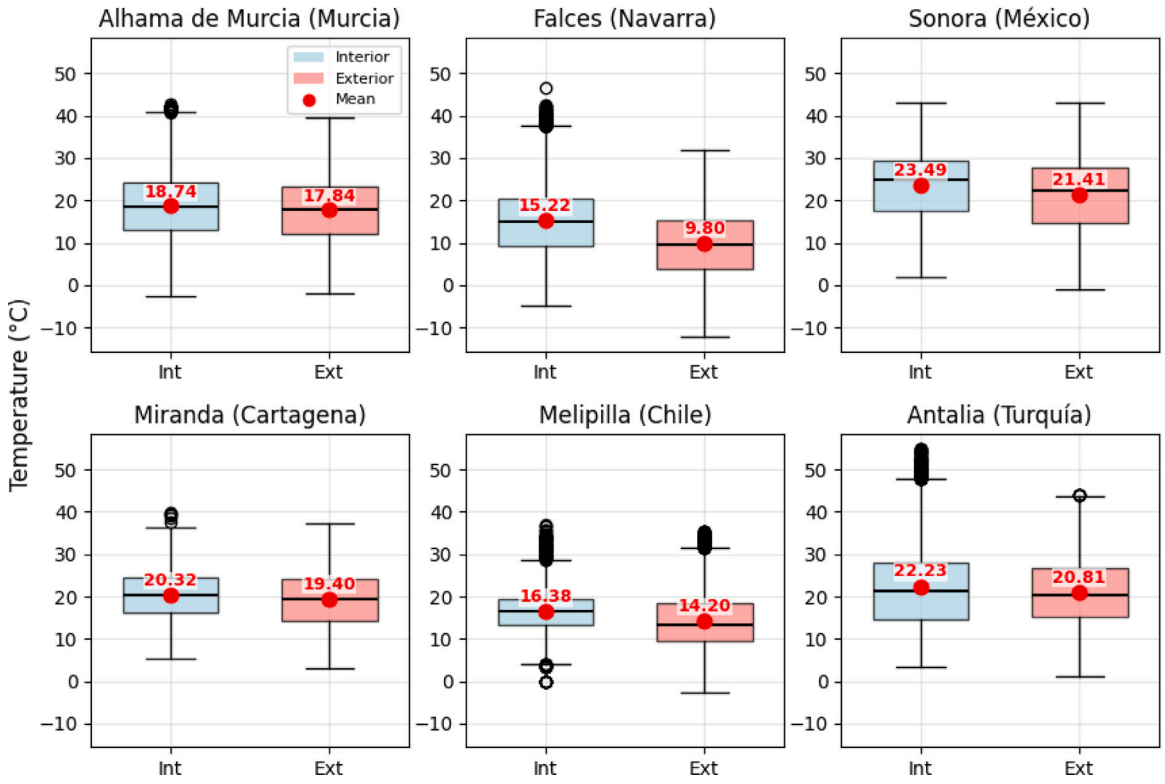

Fig. A.10. Boxplots comparing the distribution of indoor and outdoor temperature across all greenhouse sites. Indoor temperature values exhibit lower variability and narrower interquartile ranges than outdoor data, reflecting the thermal buffering effect of greenhouse structures and climate control systems.

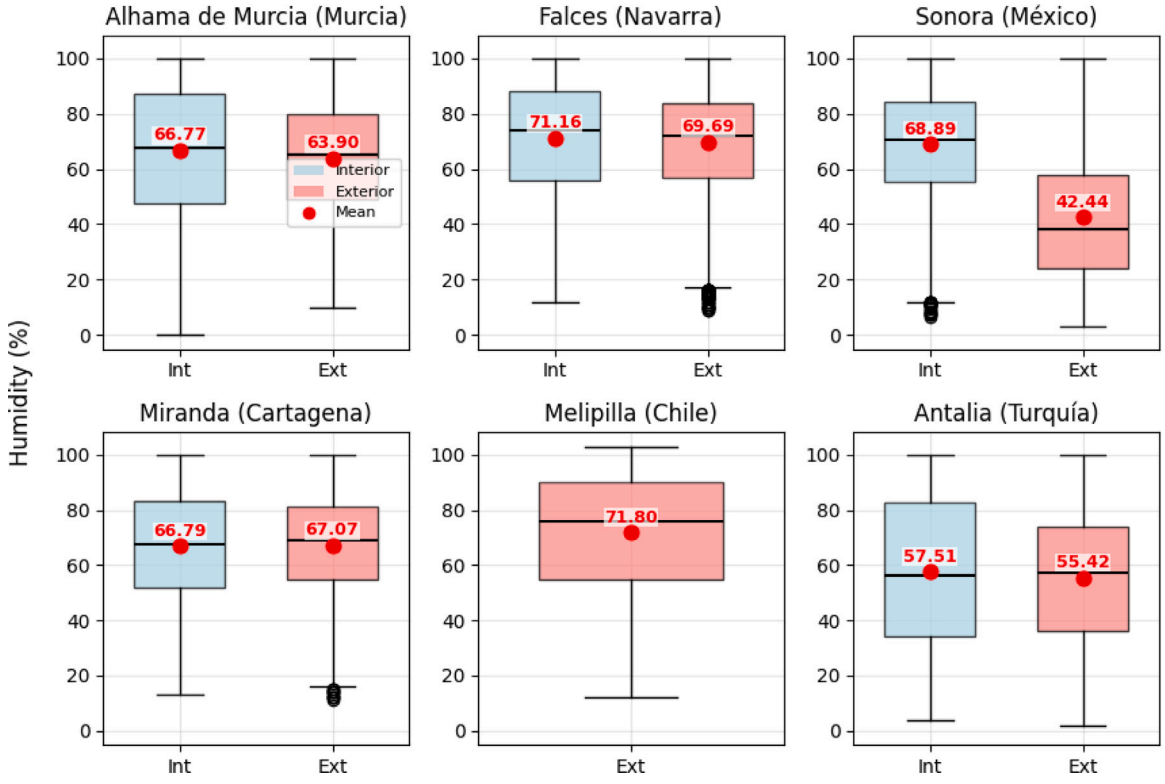

Fig. A.11. Boxplots of indoor and outdoor relative humidity for the analyzed locations. Indoor humidity distributions are generally higher and more stable, indicating controlled vapor dynamics and limited exposure to external fluctuations.

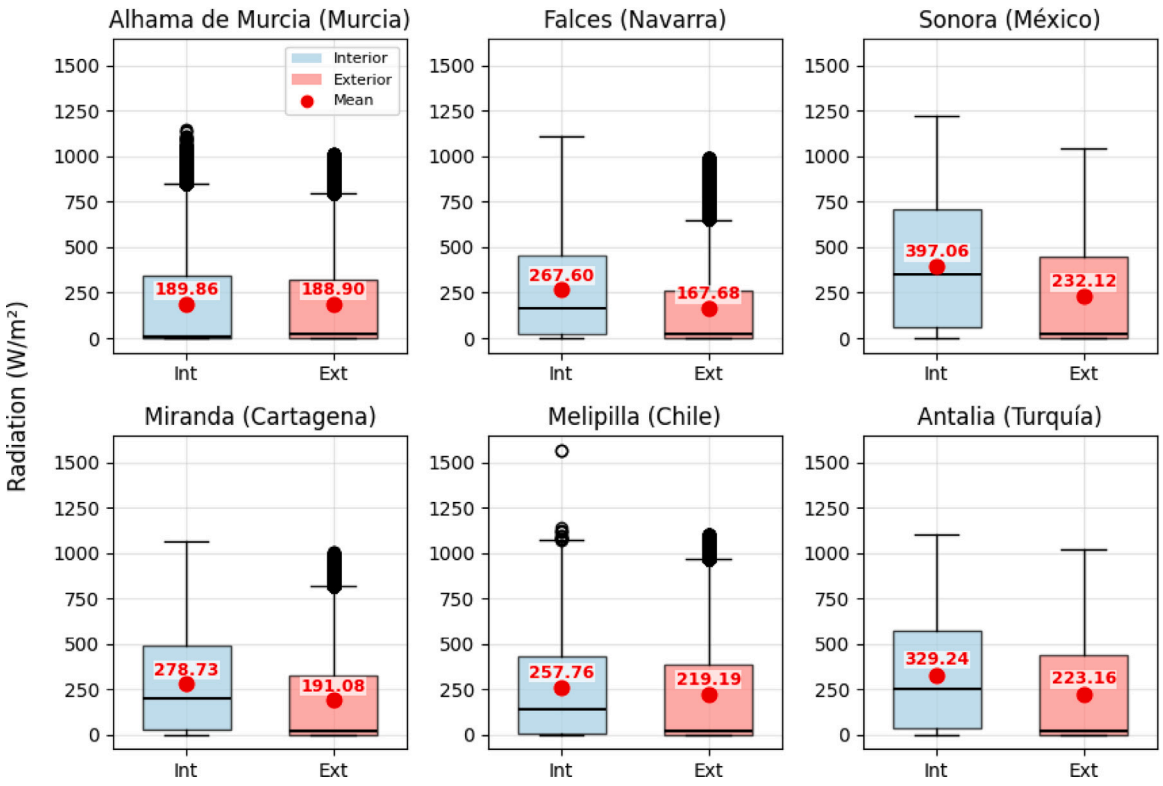

Fig. A.12. Boxplots comparing indoor and outdoor global radiation. The indoor measurements show significantly lower medians and compressed ranges due to shading effects, cover transmissivity, and dynamic screen management within the greenhouse environment.

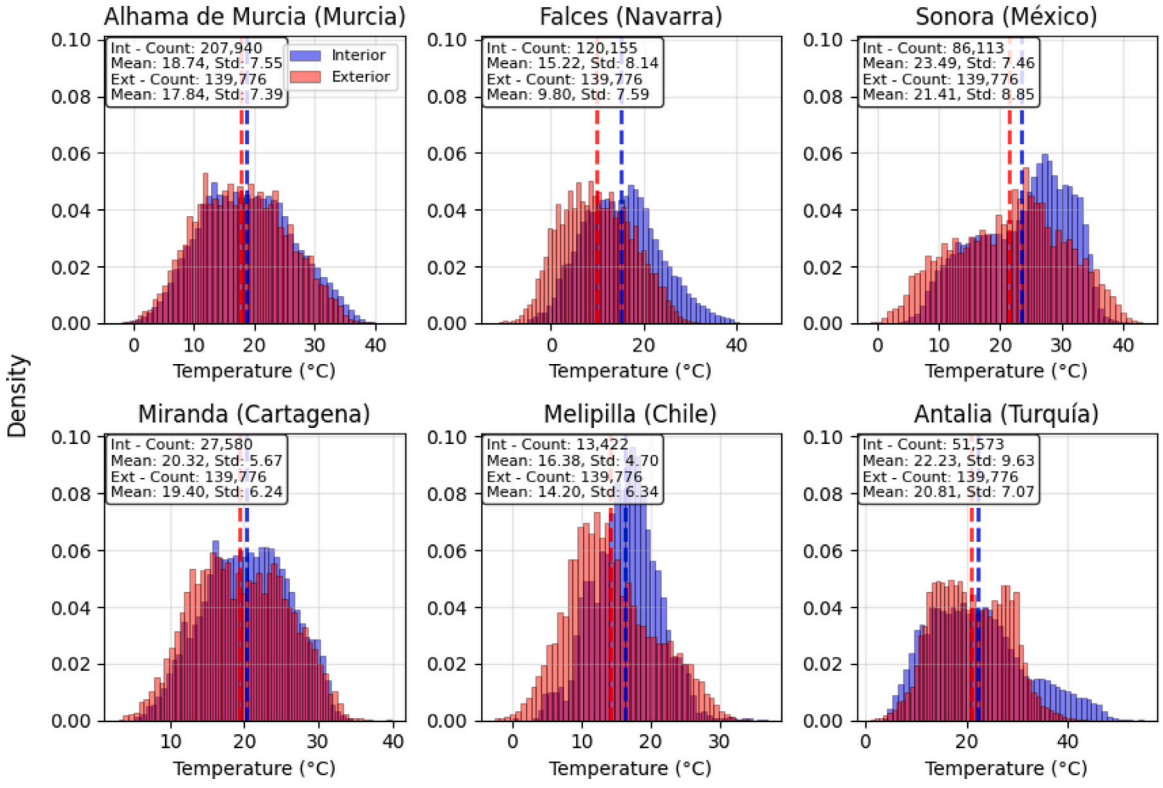

Fig. A.13. Histograms comparing the distribution of indoor and outdoor temperature across all greenhouse sites. Indoor temperature values concentrate around narrower ranges, while outdoor temperatures exhibit broader and more bimodal distributions, highlighting the moderating influence of greenhouse structures and climate control.

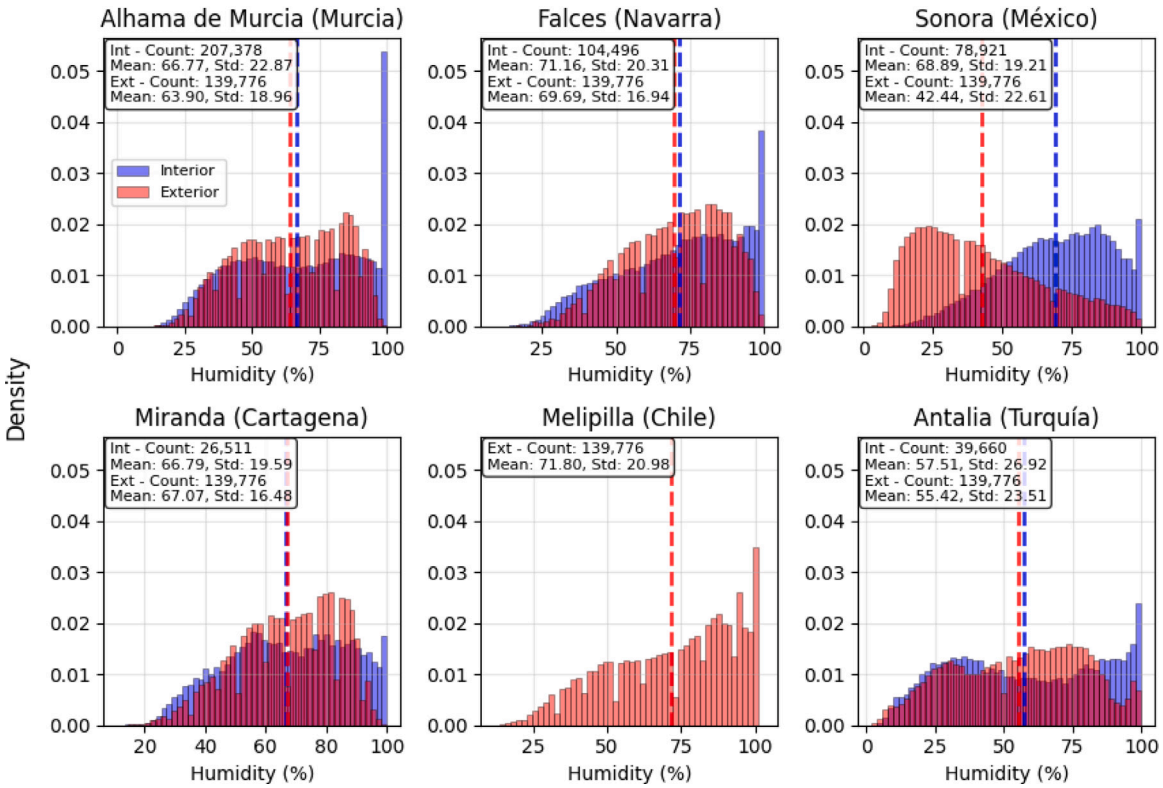

Fig. A.14. Histograms of indoor and outdoor relative humidity. Indoor humidity distributions are generally skewed towards higher values and reduced variance, reflecting the stabilizing effect of controlled ventilation and transpiration processes within the greenhouse environment.

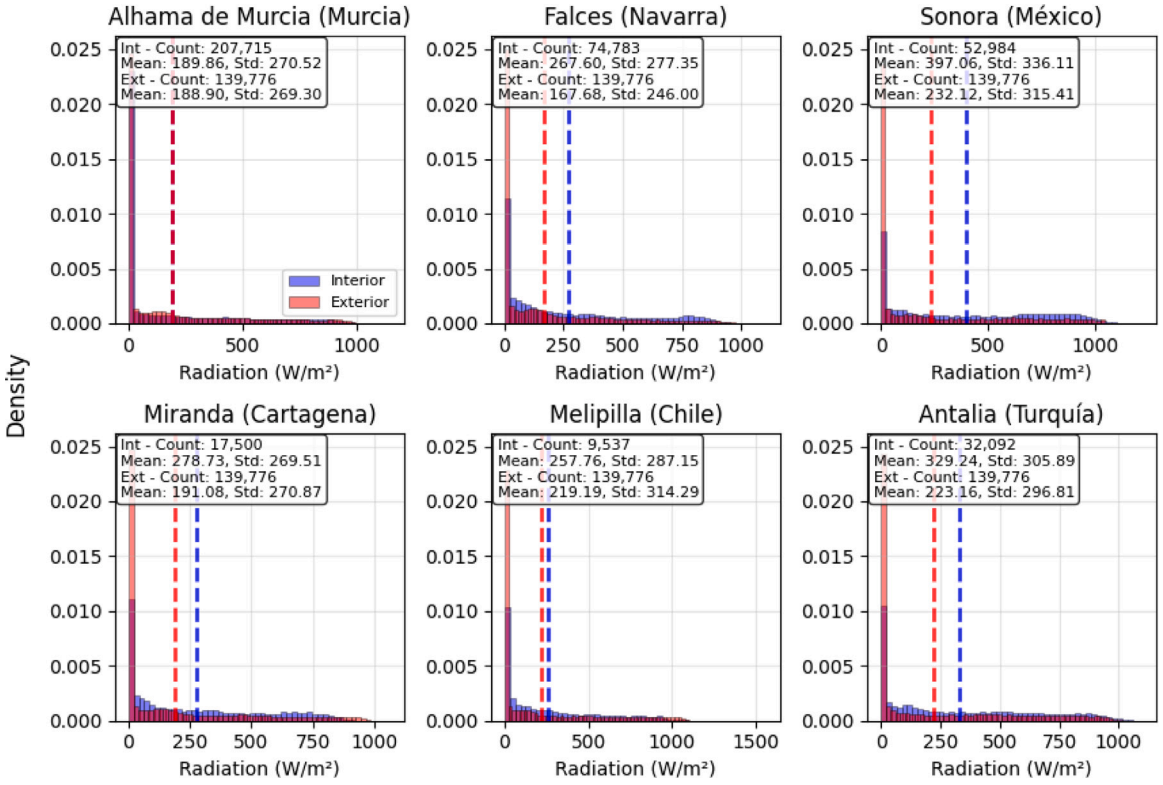

Fig. A.15. Histograms of indoor and outdoor global radiation measurements. Outdoor radiation exhibits high variability and pronounced right-skewed distributions due to diurnal peaks, while indoor measurements show truncated ranges caused by shading, screen control, and cover transmissivity losses.

Appendix B. Evaluation figures of mapping models

See Figs. B.1–B.6.

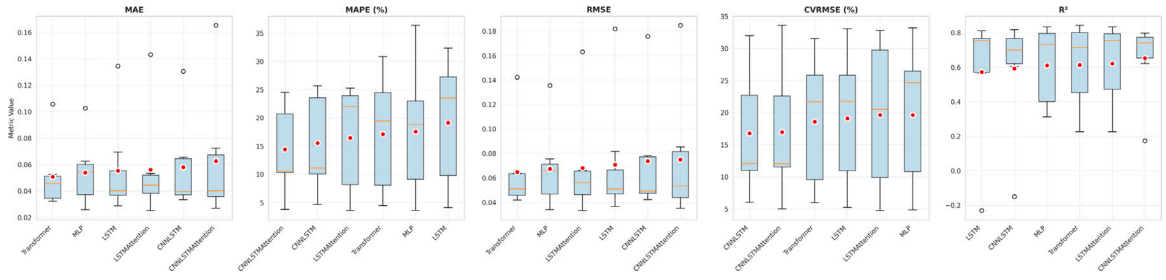

Fig. B.1. Distribution of mapping performance across models for the temperature variable. Each boxplot represents the dispersion of evaluation metrics (MAE, RMSE, R^2) across all datasets, highlighting cross-site consistency. Results show that all architectures achieve comparable accuracy, with limited variance and no single dominant model.

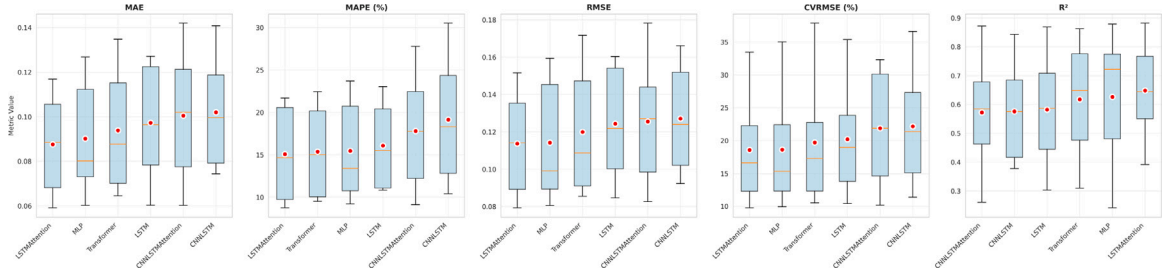

Fig. B.2. Distribution of mapping performance across models for relative humidity. Larger interquartile ranges indicate higher variability among datasets, reflecting the influence of missing data and nonlinear control processes. Attention-based architectures exhibit slightly higher R^2 and lower error dispersion compared to baseline models.

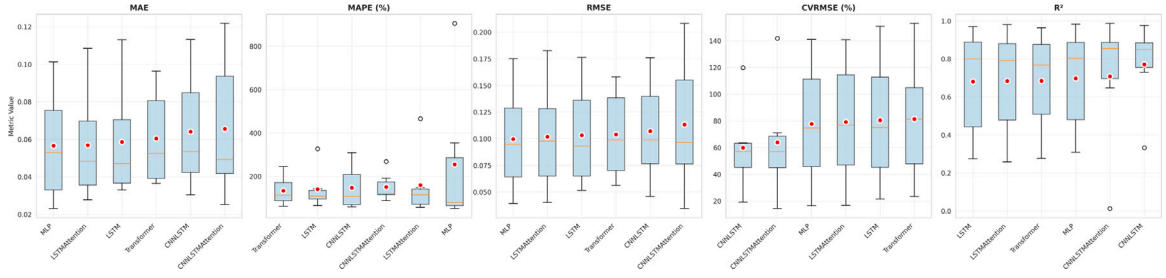

Fig. B.3. Distribution of mapping performance across models for global radiation. Radiation presents the highest overall variability, partly due to sensor saturation and shading effects within greenhouses. Despite this, all models maintain robust alignment between outdoor and indoor radiation patterns.

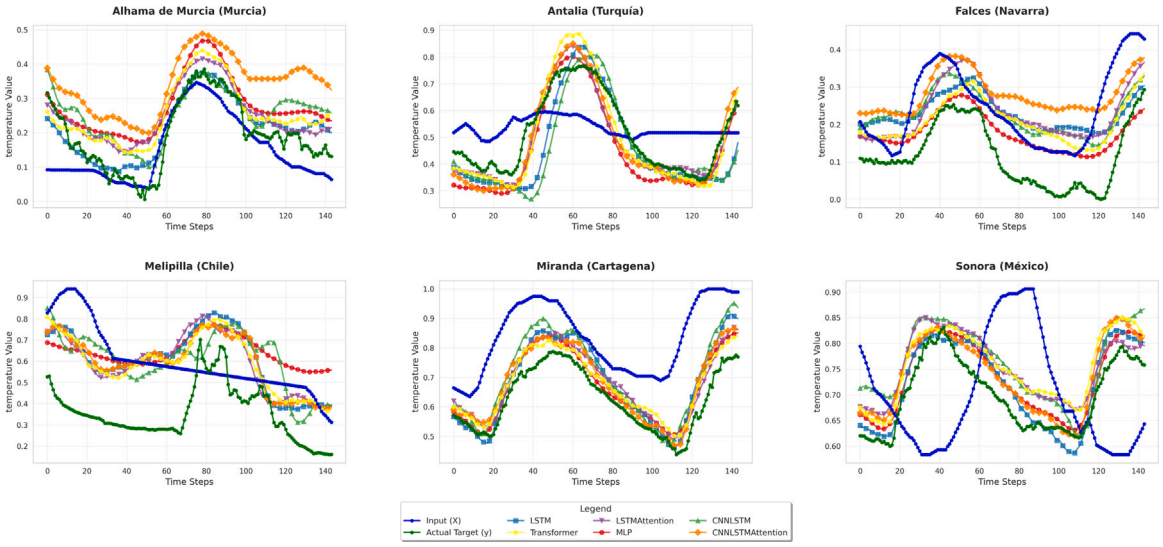

Fig. B.4. Predicted versus observed indoor temperature for all datasets. The models successfully reproduce the temporal structure and phase lag between outdoor inputs and indoor responses, demonstrating the learned thermal translation in $\mathscr{M}$. Discrepancies are primarily associated with data gaps or control-driven deviations in sites such as Melipilla.

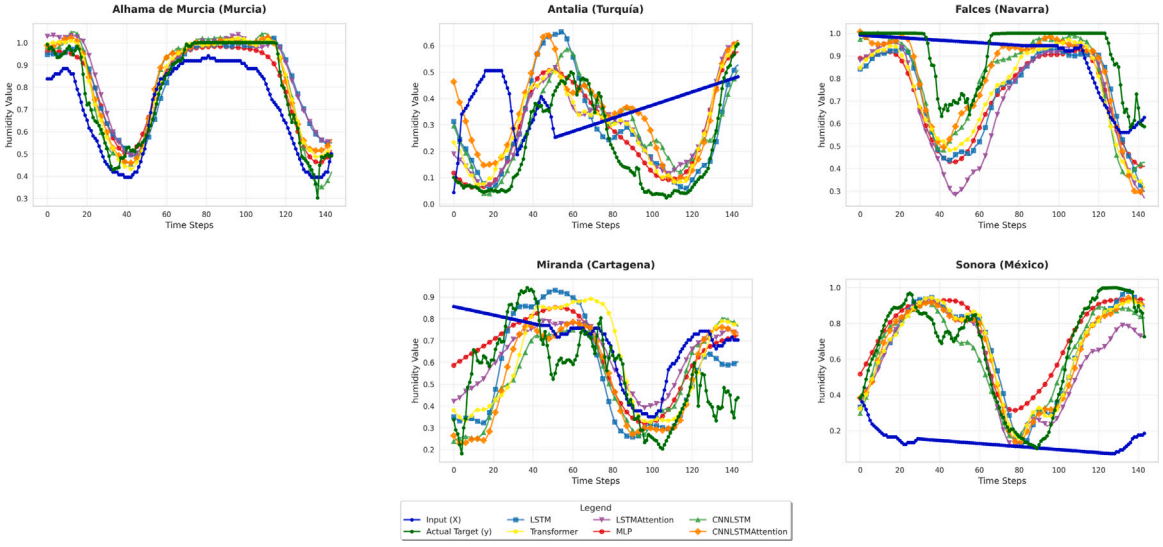

Fig. B.5. Predicted versus observed indoor relative humidity across all locations. The learned mappings capture the main humidity dynamics despite data sparsity and nonlinear ventilation effects. Performance degradation is most evident in datasets with higher missingness (e.g., Alhama).

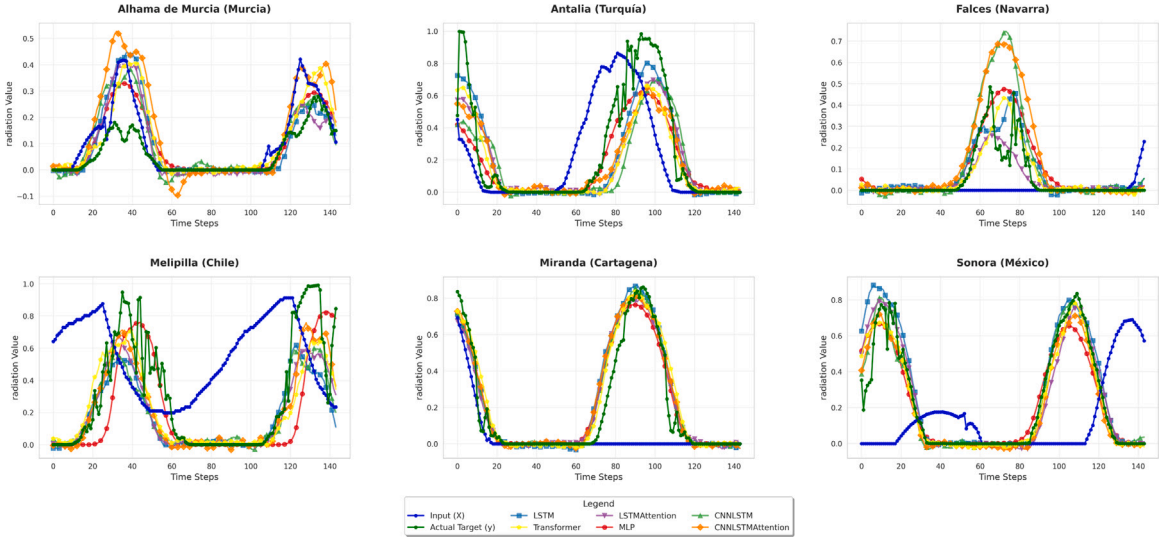

Fig. B.6. Predicted versus observed indoor global radiation for all greenhouse sites. The mapping models accurately reproduce the attenuated radiation peaks caused by greenhouse covers and dynamic shading, confirming their ability to learn structural transmissivity effects.

Appendix C. Evaluation of forecasting models

See Figs. C.1–C.21.

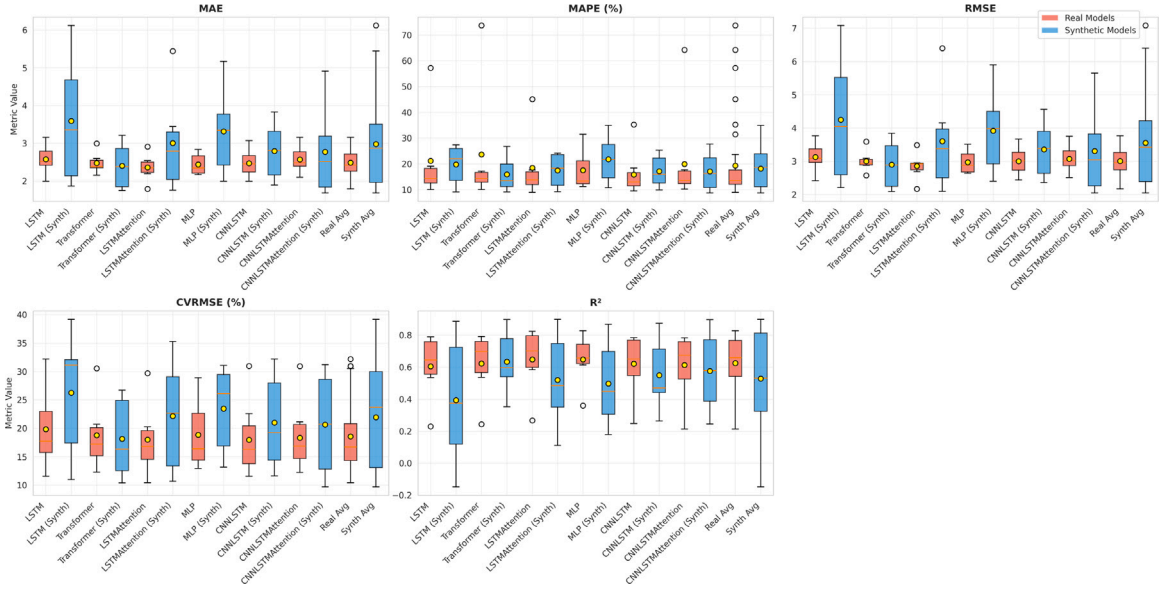

Fig. C.1. Comparison of forecasting performance across architectures for temperature under real and synthetic training regimes. The boxplots summarize the distribution of error and accuracy metrics (MAE, RMSE, R^2) across all datasets. Models trained on real data exhibit slightly lower variance and higher median accuracy, while synthetic-trained models maintain competitive performance with moderate dispersion, confirming the transferability of learned thermal dynamics.

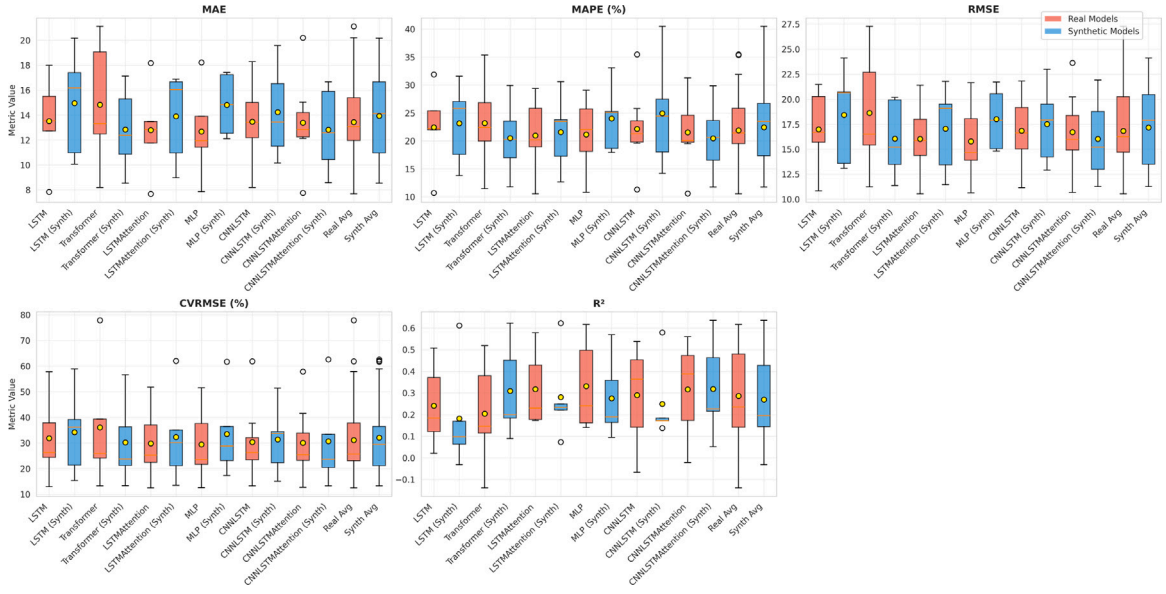

Fig. C.2. Forecasting performance boxplots for relative humidity under real and synthetic data training. Humidity predictions show greater variability across both regimes due to missing data and nonlinear greenhouse control effects. Attention-based architectures (LSTM-Attention and CNNLSTM-Attention) achieve slightly higher R^2 and lower median error, particularly when trained on synthetic indoor series with well-preserved humidity dynamics.

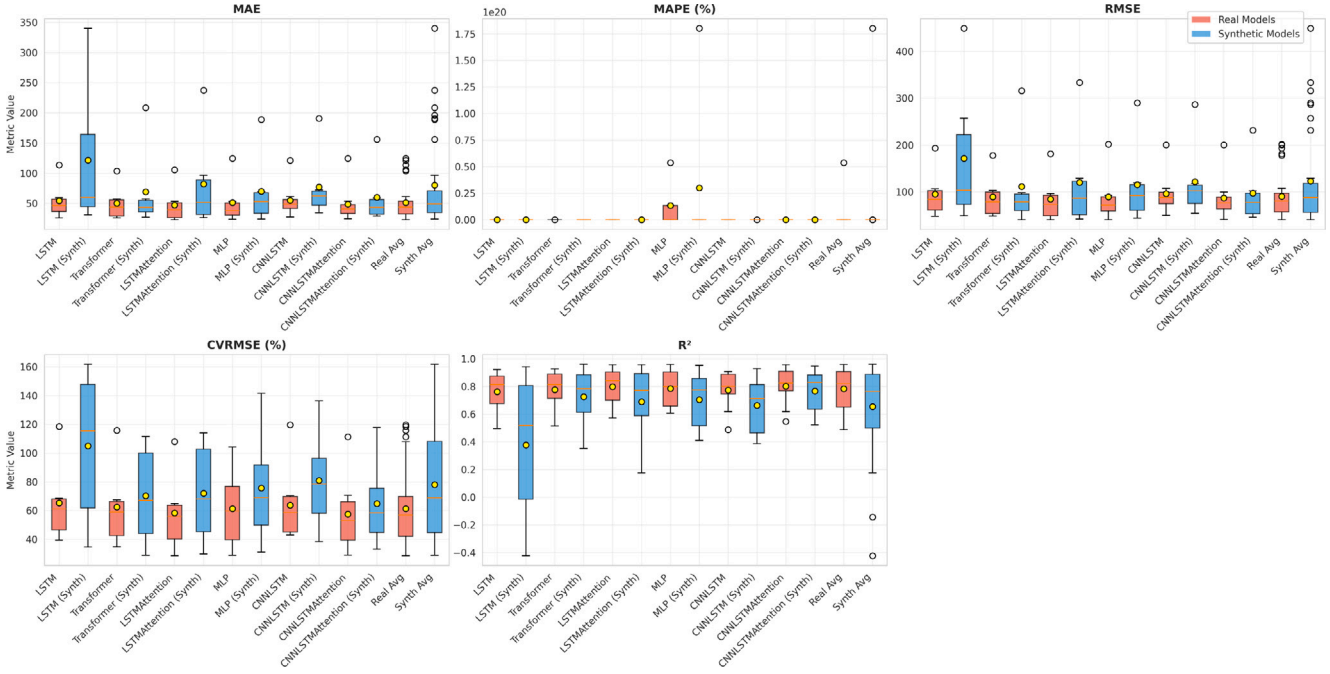

Fig. C.3. Performance comparison for global radiation forecasts using real and synthetic training data. Radiation exhibits the largest dispersion among variables, reflecting the strong influence of screen management, transmissivity, and shading in each site. Synthetic-trained models capture the dominant diurnal trends but show increased variability, consistent with the higher standard deviations observed in Table 5.

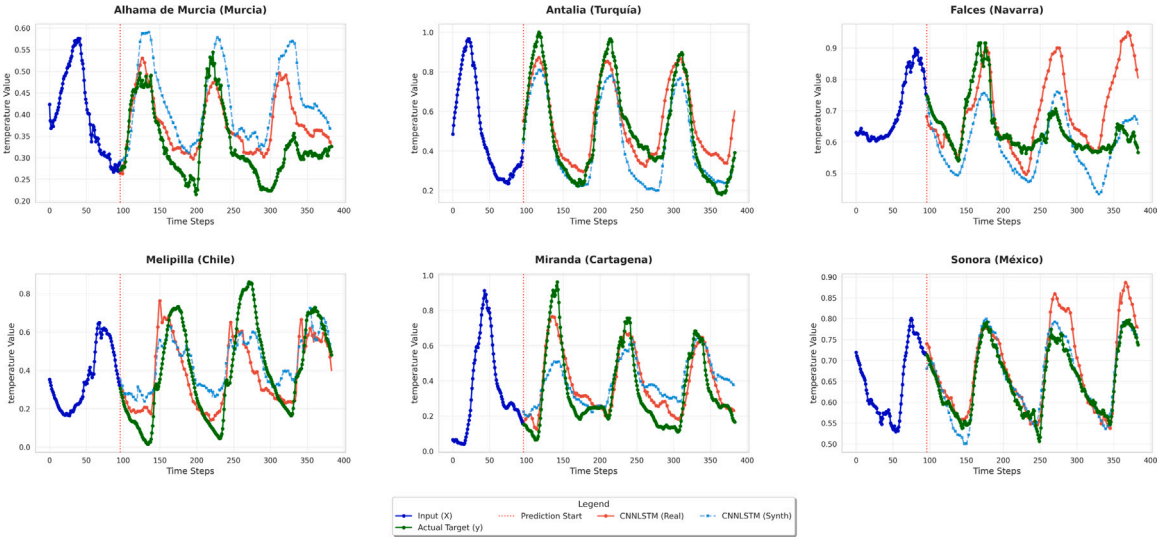

Fig. C.4. Forecasting comparison for indoor temperature using the PyTorch CNNLSTM architecture trained on real and synthetic data. The synthetic-trained model reproduces the diurnal oscillations and temporal phase of the indoor temperature series with only a slight increase in residual amplitude. Both regimes align closely with the observed indoor patterns, confirming that thermal dynamics are well captured by the mapping-forecasting pipeline.

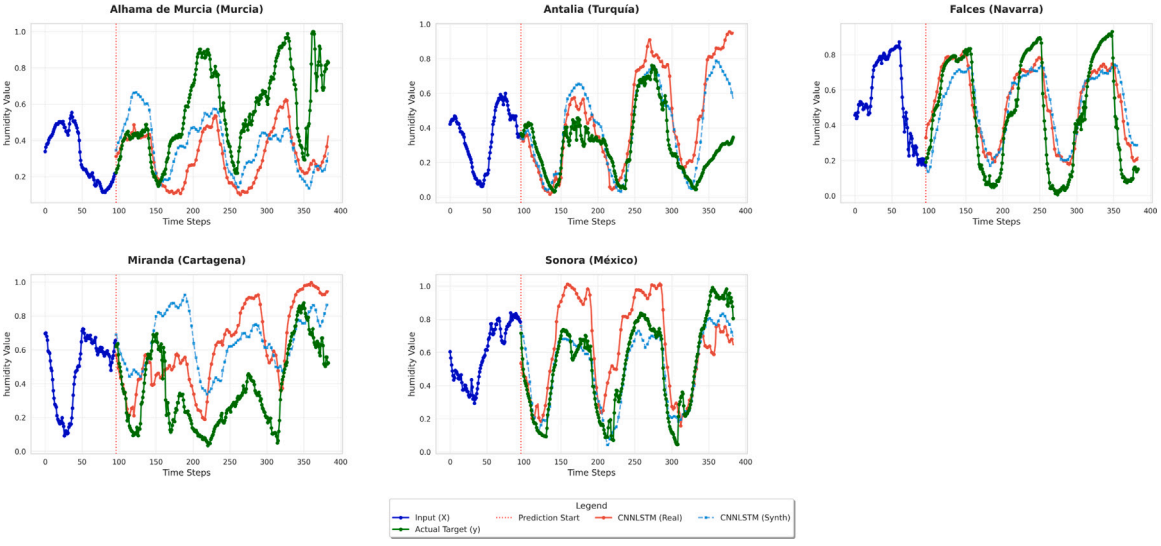

Fig. C.5. Predicted versus observed indoor humidity using the PyTorch CNNLSTM under real and synthetic training regimes. While both models capture the dominant humidity cycles, the synthetic-trained version exhibits slightly delayed responses and higher variance in high-humidity periods. These deviations reflect the greater nonlinearity and control-dependence of humidity dynamics compared with temperature.

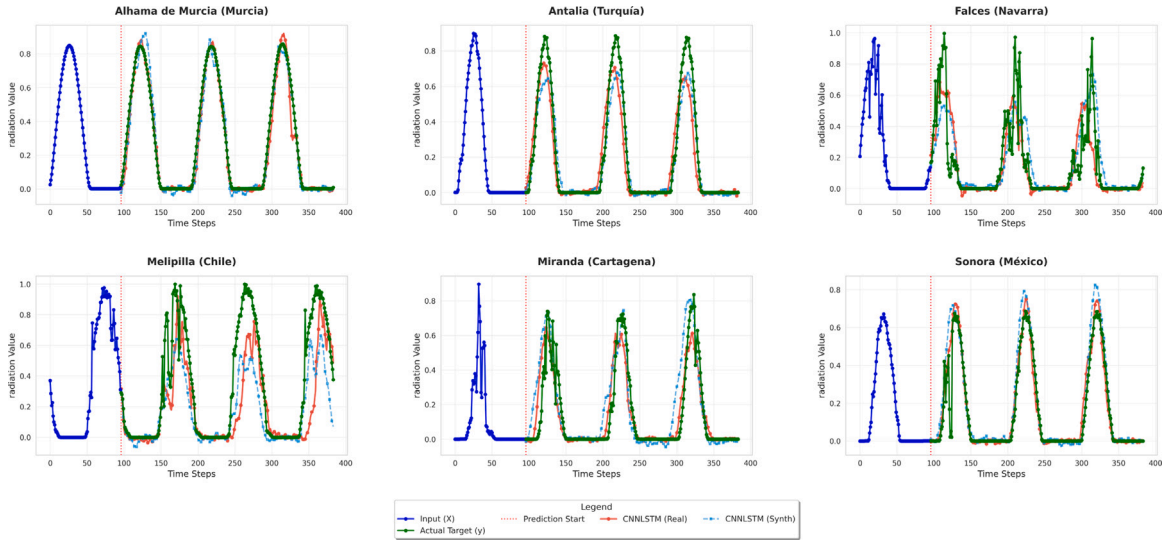

Fig. C.6. Forecasting of indoor radiation using the PyTorch CNNLSTM trained with real and synthetic indoor data. The model trained on synthetic data successfully reproduces the main irradiance peaks but shows larger residuals during low-irradiance periods, when shading and transmissivity effects dominate. Overall, both training regimes capture the temporal envelope of radiation, demonstrating the CNNLSTM's robustness to synthetic data variability.

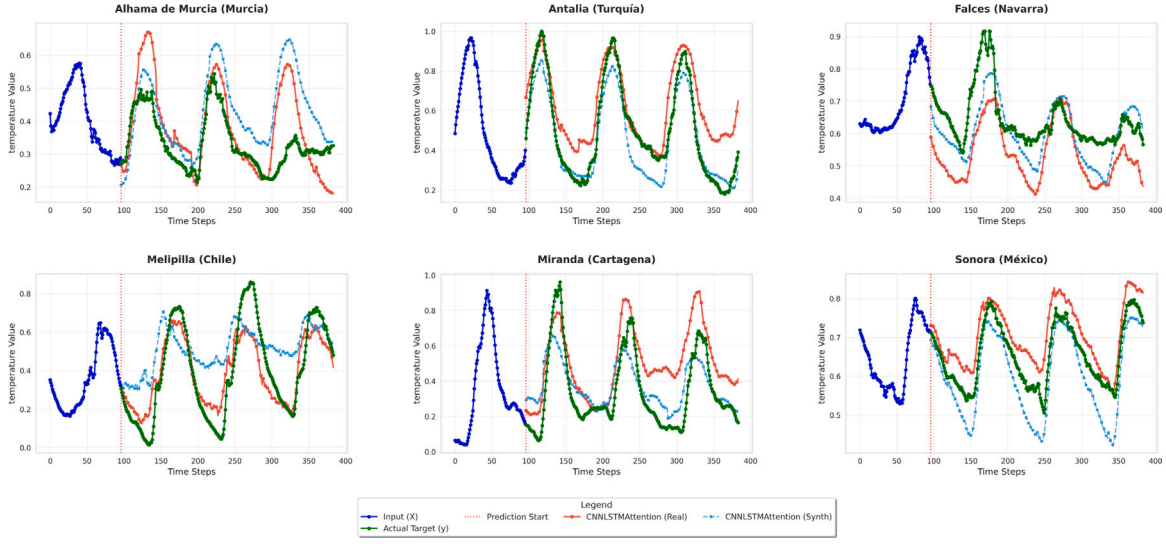

Fig. C.7. Forecasting of indoor temperature using the PyTorch CNNLSTM with Attention mechanism trained on real and synthetic indoor data. Both models closely reproduce the temporal dynamics and amplitude of the observed indoor temperature series. The attention layer improves phase alignment and reduces short-term deviations, allowing the synthetic-trained model to achieve accuracy levels comparable to those of the real-trained counterpart.

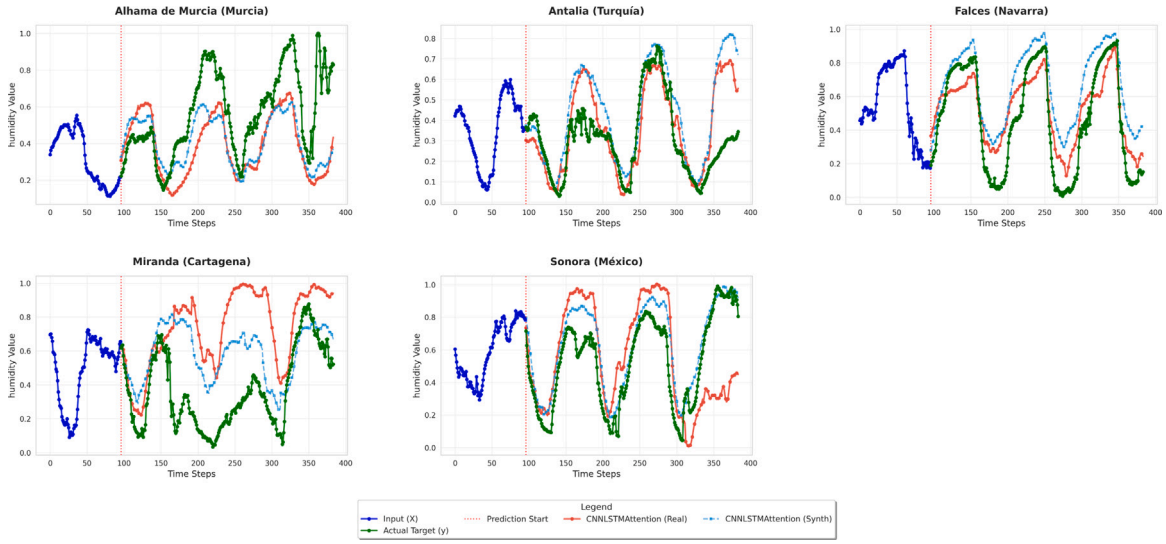

Fig. C.8. Predicted and observed indoor humidity using the CNNLSTM with Attention architecture under real and synthetic training regimes. The attention-enhanced model captures the delayed and nonlinear humidity transitions characteristic of greenhouse control systems. Although the synthetic-trained version shows slightly higher residuals during abrupt fluctuations, both maintain coherent long-term humidity trends and daily cycles.

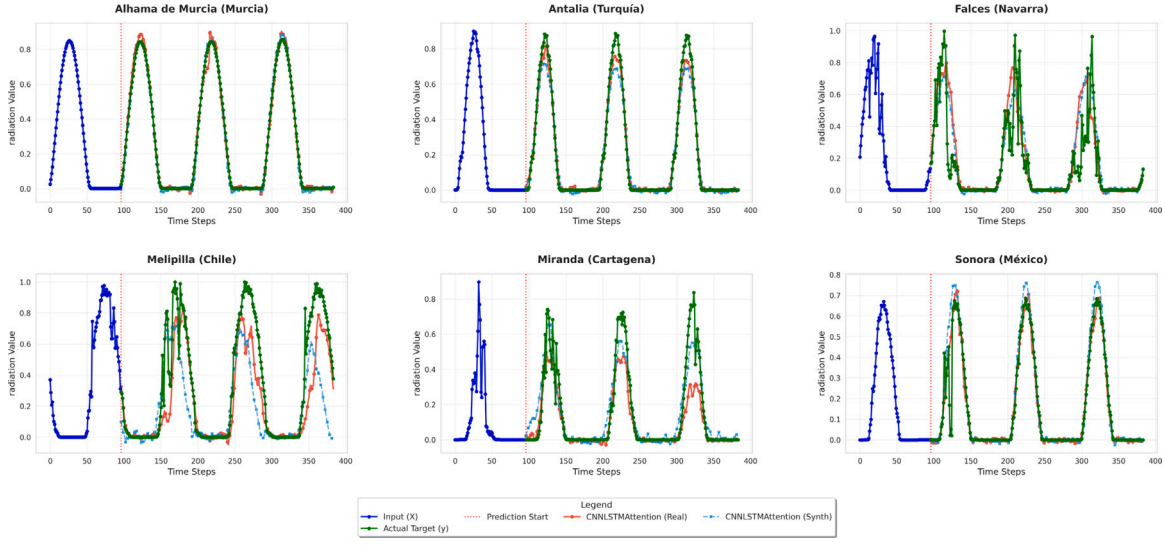

Fig. C.9. Comparison of indoor radiation forecasts produced by the CNNLSTM with Attention model trained on real and synthetic datasets. The attention mechanism enhances the reconstruction of irradiance peaks and transitions, particularly under synthetic training. Minor discrepancies appear in low-radiation intervals due to cumulative propagation errors, but overall agreement remains strong across regimes, confirming the robustness of attention-based architectures.

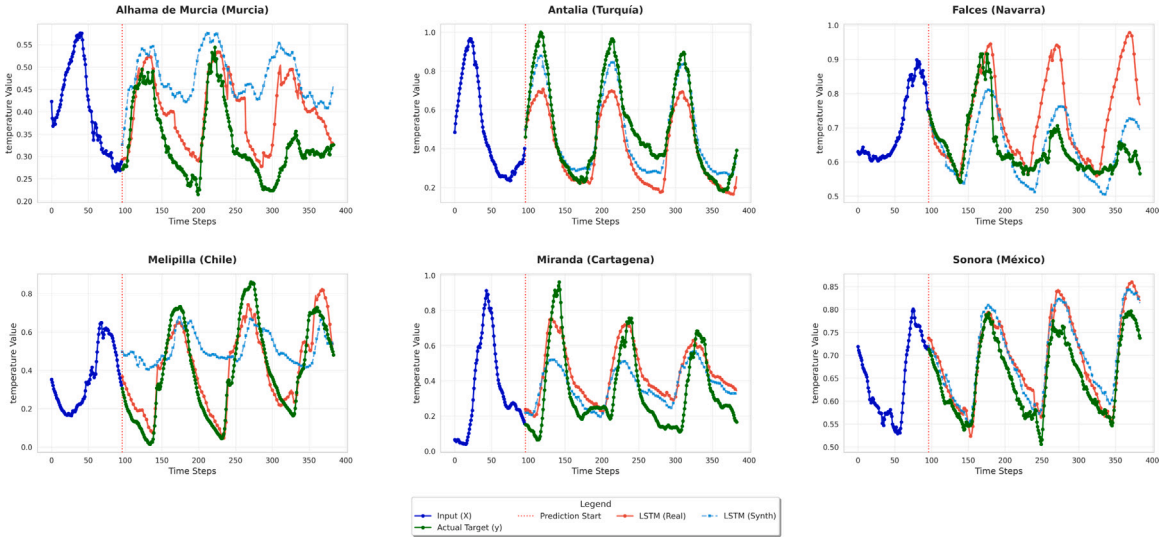

Fig. C.10. Forecasted versus observed indoor temperature using the PyTorch LSTM architecture trained on real and synthetic data. The LSTM accurately captures diurnal thermal fluctuations and maintains good phase alignment across both training regimes. Synthetic-trained models show a mild amplitude attenuation but preserve the temporal shape of the series, confirming the model's ability to generalize temporal dependencies even when trained on generated data.

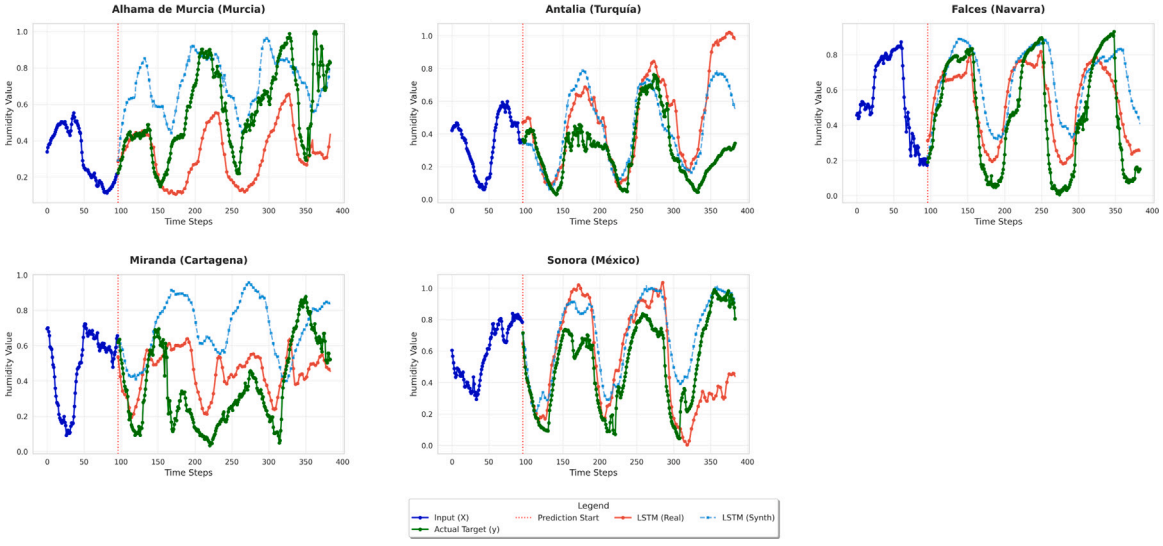

Fig. C.11. Forecasting of indoor humidity using the PyTorch LSTM trained with real and synthetic data. While the real-trained model reproduces both slow and abrupt humidity variations more faithfully, the synthetic-trained version tends to smooth rapid transitions and overestimate recovery phases. This behavior reflects the increased sensitivity of humidity to noise and unmeasured control variables.

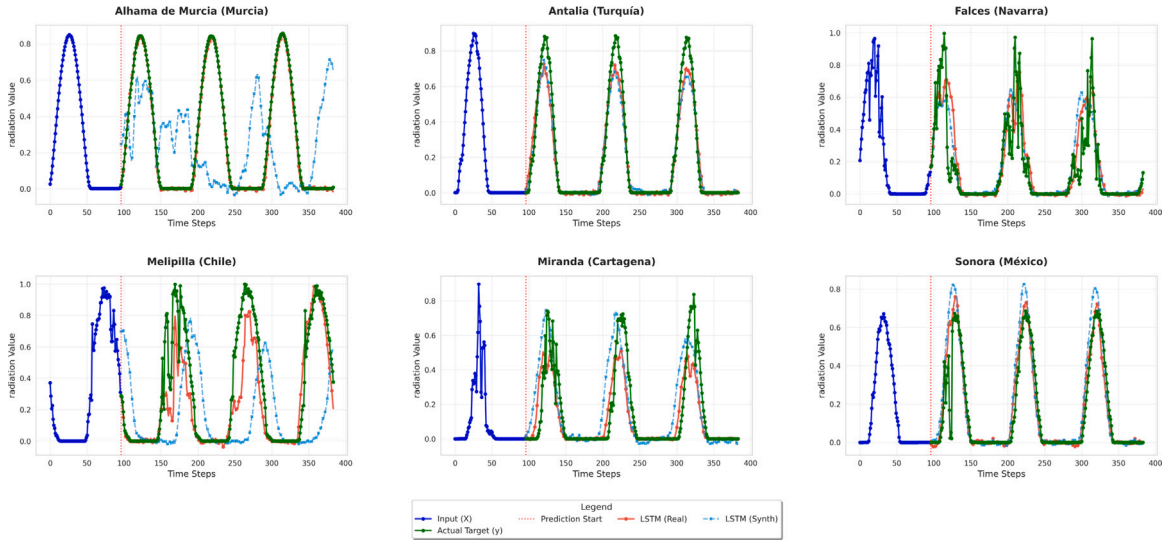

Fig. C.12. Predictions of indoor radiation obtained with the PyTorch LSTM model trained under real and synthetic data regimes. Both models reproduce the main irradiance cycles but the synthetic-trained version exhibits larger residuals near peak radiation levels, indicating partial loss of fine-grained variability. Nevertheless, the overall dynamics and phase of the radiation signal remain consistent, underscoring the robustness of recurrent architectures to synthetic inputs.

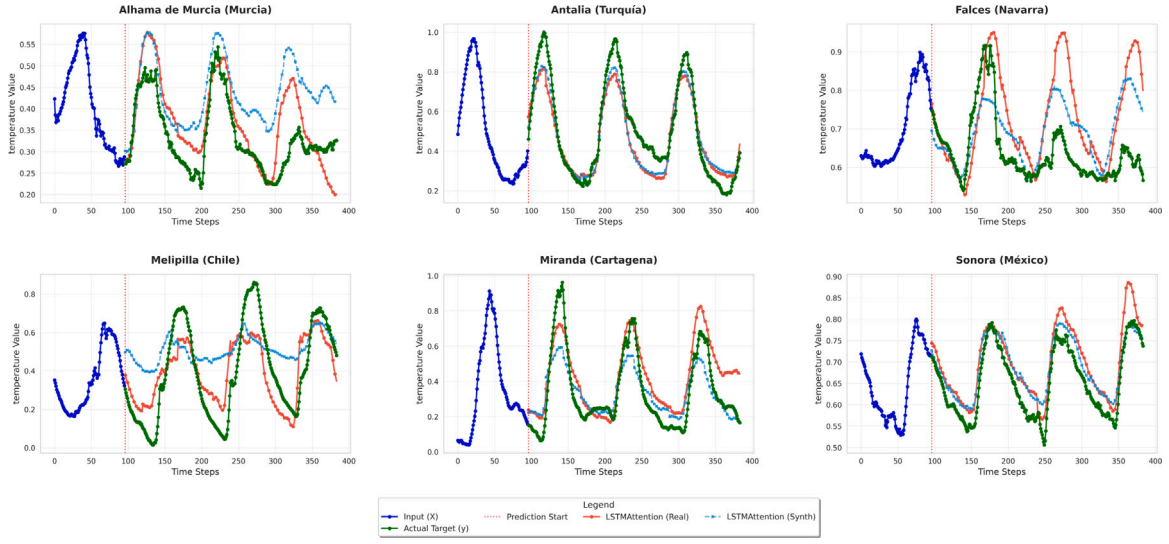

Fig. C.13. Forecasted versus observed indoor temperature using the PyTorch LSTM with Attention architecture trained on real and synthetic data. The attention mechanism enhances short-term sensitivity and improves alignment during rapid temperature fluctuations. Synthetic-trained models maintain strong temporal coherence with only a modest loss of amplitude precision, demonstrating the stability of the attention-enhanced recurrent design.

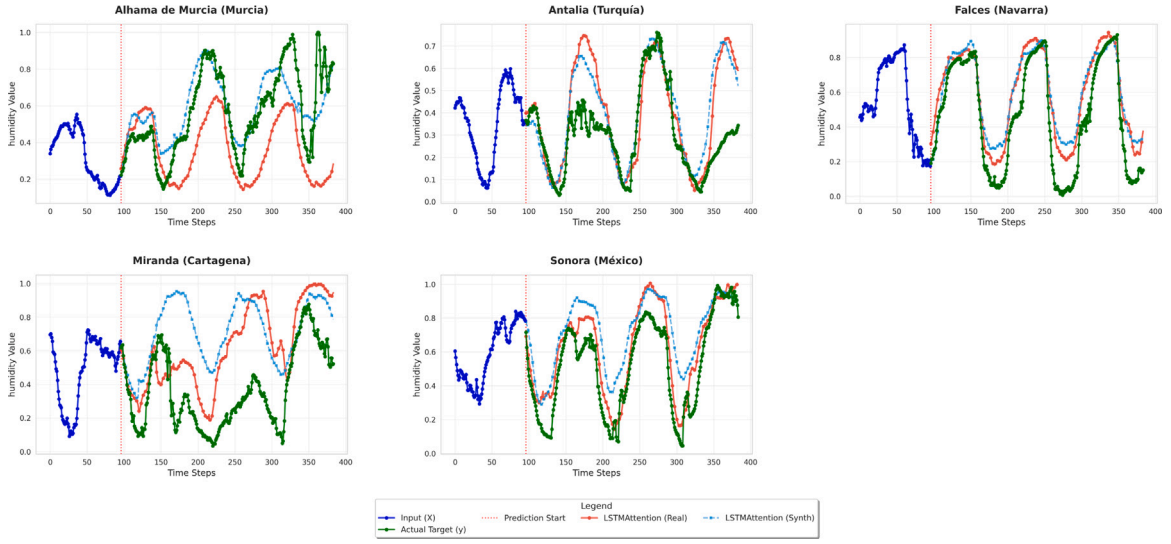

Fig. C.14. Indoor humidity predictions obtained with the PyTorch LSTM with Attention model under real and synthetic data training. Attention weighting improves tracking of the delayed humidity response typical of greenhouse systems, especially around control-induced transitions. The synthetic-trained version slightly over-smooths humidity peaks, suggesting minor attenuation of variability but preserved overall pattern fidelity.

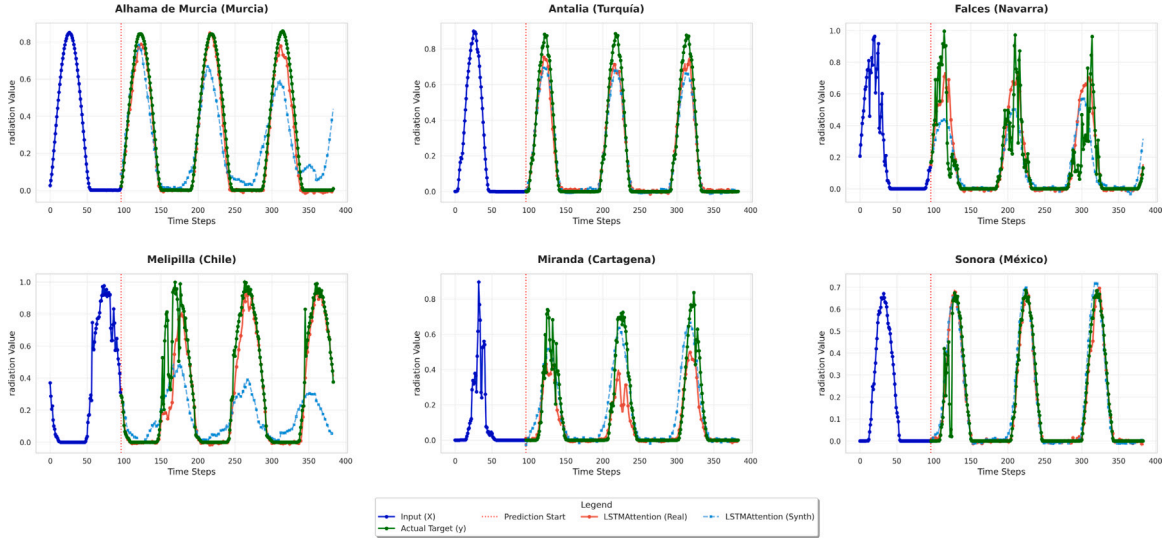

Fig. C.15. Predicted and observed indoor radiation using the PyTorch LSTM with Attention mechanism trained on real and synthetic data. The model effectively reproduces diurnal radiation patterns and captures phase shifts between real and synthetic regimes. Slight underestimation in low-irradiance periods is observed for synthetic training, but the overall energy trend and daily envelope remain consistent, confirming good generalization.

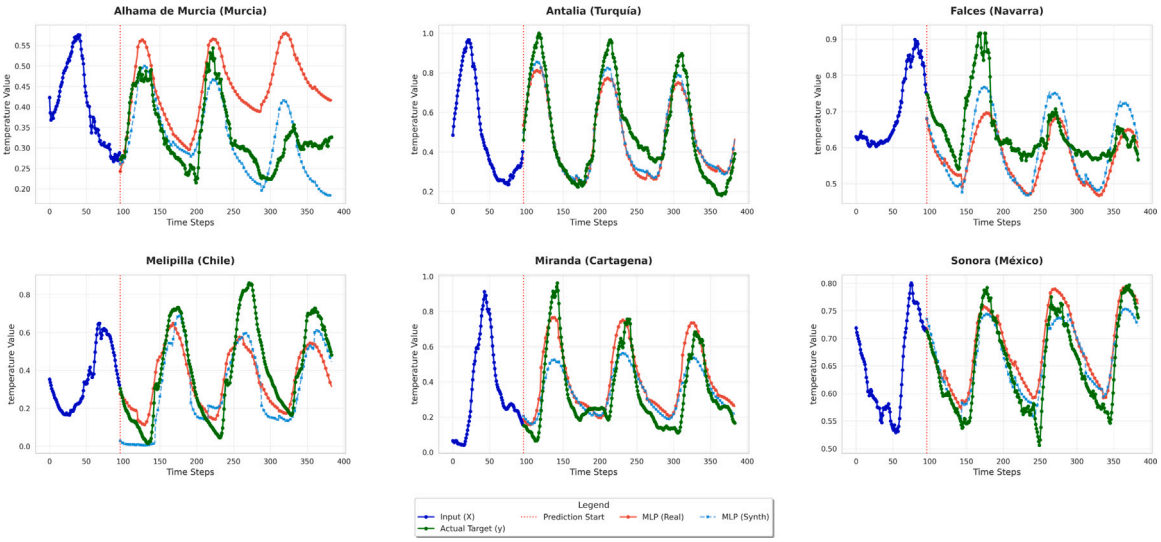

Fig. C.16. Forecasted and observed indoor temperature using the PyTorch MLP model trained on real and synthetic data. The MLP, despite lacking temporal recurrence, effectively captures the overall temperature evolution and daily oscillations. Synthetic-trained models show a slight bias and reduced sharpness in transitions, yet retain high correlation with ground-truth signals, confirming that static nonlinear mappings can approximate thermal dynamics when temporal coupling is limited.

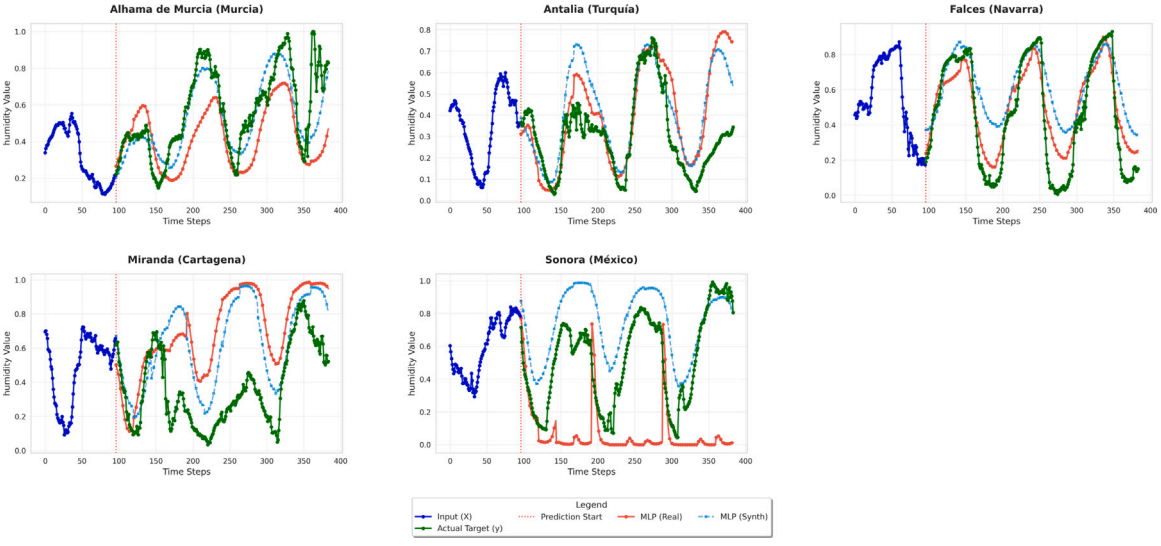

Fig. C.17. Indoor humidity predictions obtained with the PyTorch MLP trained on real and synthetic data. The model reproduces average humidity trends but struggles with abrupt control-induced fluctuations. Synthetic training increases smoothing effects and amplitude damping, suggesting that static feed-forward architectures are more sensitive to noise and data completeness in this variable.

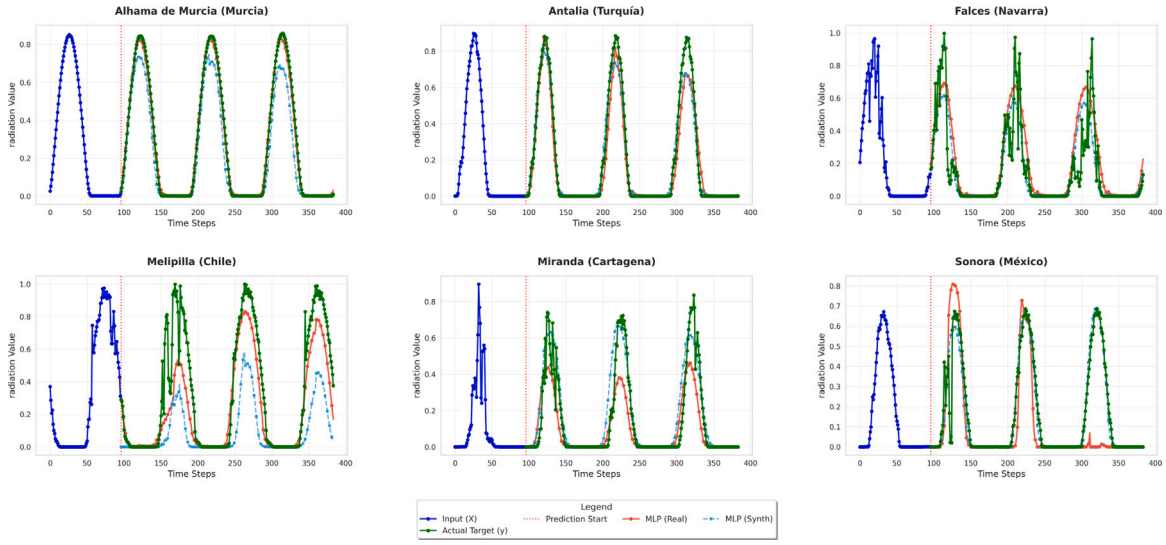

Fig. C.18. Predicted versus observed indoor radiation using the PyTorch MLP trained with real and synthetic datasets. Radiation patterns are generally well reproduced, including diurnal cycles and attenuation during cloudy periods. However, synthetic-trained models exhibit higher residual errors during irradiance peaks, reflecting their limited ability to capture fast transients without explicit temporal context.

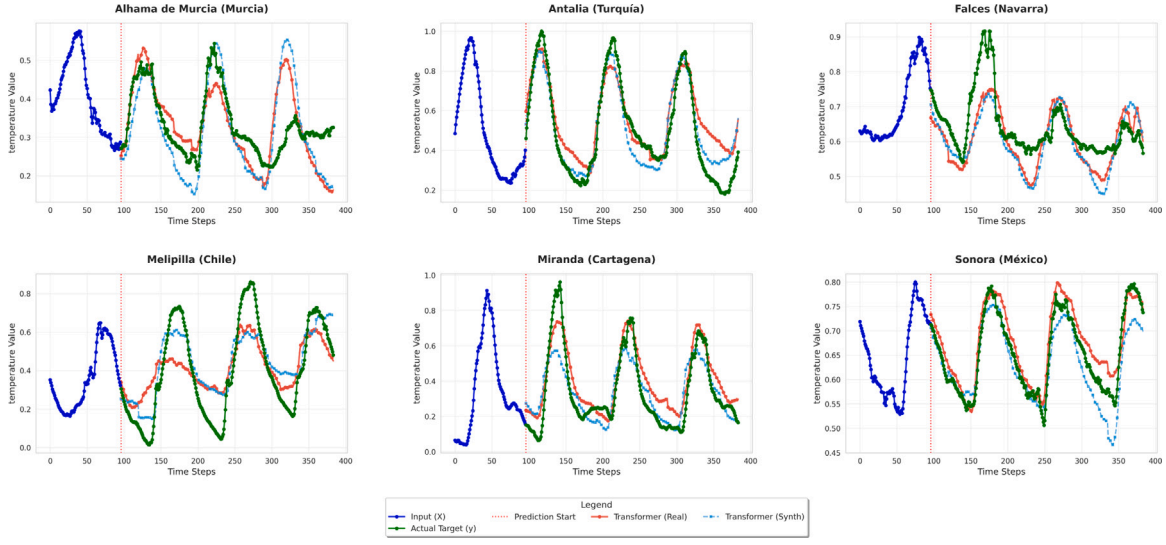

Fig. C.19. Forecasted and observed indoor temperature using the PyTorch Transformer architecture trained on real and synthetic data. The Transformer effectively captures long-range temporal dependencies and maintains accurate phase alignment across multi-day horizons. Synthetic-trained models slightly underpredict peak amplitudes, yet preserve the overall structure of temperature cycles, demonstrating robust generalization despite reduced physical signal coherence.

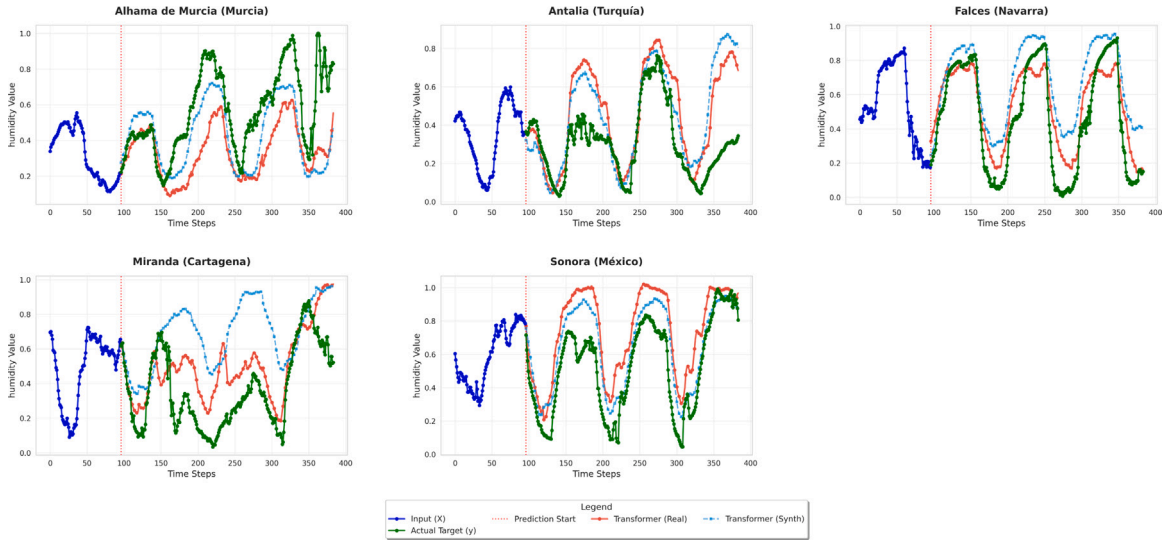

Fig. C.20. Indoor humidity predictions produced by the PyTorch Transformer trained on real and synthetic data. The self-attention mechanism enables the model to reproduce gradual humidity variations and delayed responses to ventilation or irrigation events. Synthetic-trained versions show increased smoothing and phase lag, suggesting that longer contextual windows improve tracking stability but attenuate short-term fluctuations.

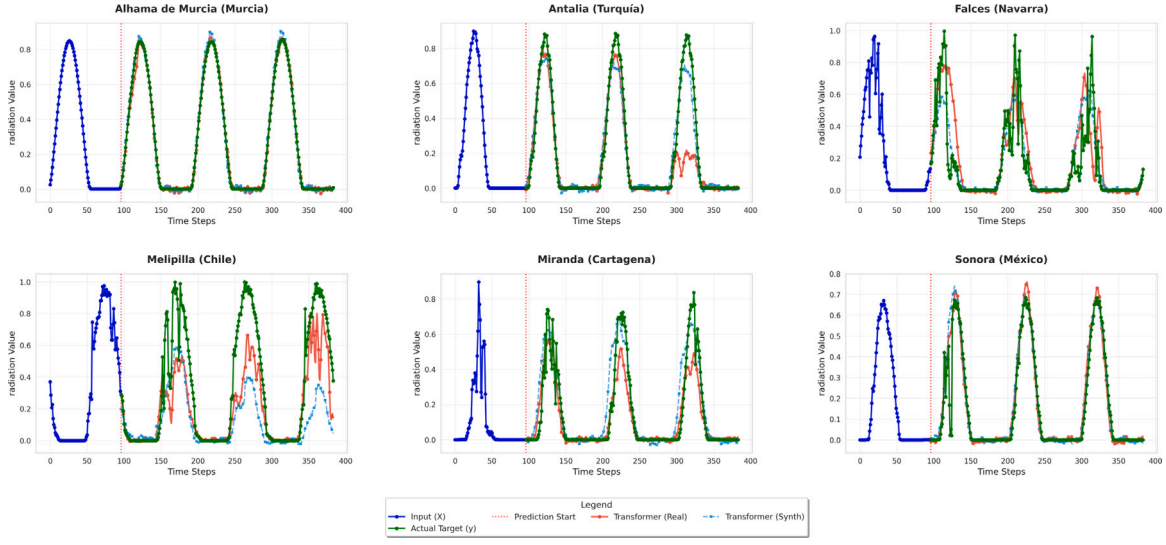

Fig. C.21. Predicted and observed indoor radiation using the PyTorch Transformer trained with real and synthetic datasets. The model accurately reconstructs diurnal irradiance cycles and their attenuation due to shading or weather conditions. Under synthetic training, peak irradiance tends to be underestimated and residual variance increases, but the Transformer still reproduces consistent temporal patterns, confirming its resilience to data synthesis artifacts.

Appendix D. Wilcoxon results

See Table D.1.

Table D.1
Wilcoxon Signed-Rank Test Results comparing Real vs Synthetic errors across all datasets, models, and features.

Dataset	Model	Feature	Mean_Error_Real	Mean_Error_Synth	Wilcoxon_Stat	p_value	Significant_Diff ($p < 0.05$)
Alhama de Murcia (Murcia)	MLP	temperature	0.077104	0.095627	16 268.000000	0.001330	True
Alhama de Murcia (Murcia)	LSTM	temperature	0.115311	0.167398	10 636.000000	0.000000	True
Alhama de Murcia (Murcia)	CNNLSTM	temperature	0.096211	0.077773	17 217.000000	0.011131	True
Alhama de Murcia (Murcia)	LSTMAttention	temperature	0.079961	0.149532	2204.000000	0.000000	True
Alhama de Murcia (Murcia)	CNNLSTMAttention	temperature	0.124800	0.075379	6446.000000	0.000000	True
Alhama de Murcia (Murcia)	Transformer	temperature	0.112663	0.073766	13 196.000000	0.000000	True
Antalia (Turquía)	MLP	temperature	0.081361	0.067175	13 794.000000	0.000001	True
Antalia (Turquía)	LSTM	temperature	0.100124	0.069294	14 081.000000	0.000002	True
Antalia (Turquía)	CNNLSTM	temperature	0.129124	0.112164	18 362.000000	0.083785	False
Antalia (Turquía)	LSTMAttention	temperature	0.083056	0.072801	18 142.000000	0.059476	False
Antalia (Turquía)	CNNLSTMAttention	temperature	0.109180	0.066832	13 048.000000	0.000000	True
Antalia (Turquía)	Transformer	temperature	0.078350	0.053800	13 102.000000	0.000000	True
Falces (Navarra)	MLP	temperature	0.154312	0.099347	2381.000000	0.000000	True
Falces (Navarra)	LSTM	temperature	0.075993	0.156806	1615.000000	0.000000	True
Falces (Navarra)	CNNLSTM	temperature	0.129334	0.087046	6632.000000	0.000000	True
Falces (Navarra)	LSTMAttention	temperature	0.082302	0.152649	2469.000000	0.000000	True
Falces (Navarra)	CNNLSTMAttention	temperature	0.075354	0.091900	11 081.000000	0.000000	True
Falces (Navarra)	Transformer	temperature	0.100589	0.082276	11 567.000000	0.000000	True
Melipilla (Chile)	MLP	temperature	0.144191	0.257425	2133.000000	0.000000	True
Melipilla (Chile)	LSTM	temperature	0.089172	0.273606	4451.000000	0.000000	True
Melipilla (Chile)	CNNLSTM	temperature	0.167889	0.139275	13 724.000000	0.000001	True
Melipilla (Chile)	LSTMAttention	temperature	0.103645	0.127495	15 802.000000	0.000402	True
Melipilla (Chile)	CNNLSTMAttention	temperature	0.197691	0.207508	15 835.000000	0.000439	True
Melipilla (Chile)	Transformer	temperature	0.153118	0.116124	12 243.000000	0.000000	True
Miranda (Cartagena)	MLP	temperature	0.062826	0.087530	11 181.000000	0.000000	True
Miranda (Cartagena)	LSTM	temperature	0.131268	0.076136	9205.000000	0.000000	True
Miranda (Cartagena)	CNNLSTM	temperature	0.067010	0.096080	13 507.000000	0.000000	True
Miranda (Cartagena)	LSTMAttention	temperature	0.061372	0.062501	20 736.000000	0.959406	False
Miranda (Cartagena)	CNNLSTMAttention	temperature	0.080233	0.076092	19 803.000000	0.477420	False
Miranda (Cartagena)	Transformer	temperature	0.078951	0.071535	19 187.000000	0.251827	False
Sonora (México)	MLP	temperature	0.056885	0.059354	19 208.000000	0.258022	False
Sonora (México)	LSTM	temperature	0.052259	0.047874	17 282.000000	0.012681	True
Sonora (México)	CNNLSTM	temperature	0.063805	0.056259	12 011.000000	0.000000	True
Sonora (México)	LSTMAttention	temperature	0.055169	0.061857	13 634.000000	0.000000	True
Sonora (México)	CNNLSTMAttention	temperature	0.039072	0.047071	15 232.000000	0.000081	True
Sonora (México)	Transformer	temperature	0.030413	0.050774	11 026.000000	0.000000	True
Alhama de Murcia (Murcia)	MLP	humidity	0.123106	0.177756	11 524.000000	0.000000	True
Alhama de Murcia (Murcia)	LSTM	humidity	0.163026	0.239101	13 320.000000	0.000000	True
Alhama de Murcia (Murcia)	CNNLSTM	humidity	0.121136	0.199527	9751.000000	0.000000	True
Alhama de Murcia (Murcia)	LSTMAttention	humidity	0.121695	0.194641	10 271.000000	0.000000	True
Alhama de Murcia (Murcia)	CNNLSTMAttention	humidity	0.212792	0.154309	10 247.000000	0.000000	True
Alhama de Murcia (Murcia)	Transformer	humidity	0.215429	0.107862	6855.000000	0.000000	True
Antalia (Turquía)	MLP	humidity	0.123361	0.165075	5595.000000	0.000000	True
Antalia (Turquía)	LSTM	humidity	0.151886	0.218959	9201.000000	0.000000	True
Antalia (Turquía)	CNNLSTM	humidity	0.086677	0.249229	7788.000000	0.000000	True
Antalia (Turquía)	LSTMAttention	humidity	0.113080	0.210131	9875.000000	0.000000	True
Antalia (Turquía)	CNNLSTMAttention	humidity	0.117039	0.195621	10 628.000000	0.000000	True
Antalia (Turquía)	Transformer	humidity	0.198610	0.211614	18 806.000000	0.156992	False
Falces (Navarra)	MLP	humidity	0.191795	0.158475	11 003.000000	0.000000	True
Falces (Navarra)	LSTM	humidity	0.146001	0.164970	15 951.000000	0.000596	True
Falces (Navarra)	CNNLSTM	humidity	0.182193	0.219349	9748.000000	0.000000	True
Falces (Navarra)	LSTMAttention	humidity	0.192183	0.147540	13 186.000000	0.000000	True
Falces (Navarra)	CNNLSTMAttention	humidity	0.159747	0.155030	20 634.000000	0.902103	False
Falces (Navarra)	Transformer	humidity	0.172302	0.174666	20 415.000000	0.781150	False
Miranda (Cartagena)	MLP	humidity	0.173413	0.196348	17 261.000000	0.012160	True
Miranda (Cartagena)	LSTM	humidity	0.139200	0.212699	10 382.000000	0.000000	True

(*continued on next page*)

Table D.1 (*continued*).

Dataset	Model	Feature	Mean_Error_Real	Mean_Error_Synth	Wilcoxon_Stat	p_value	Significant_Diff ($p < 0.05$)
Miranda (Cartagena)	CNNLSTM	humidity	0.178930	0.196404	19 763.000000	0.460068	False
Miranda (Cartagena)	LSTMAttention	humidity	0.153023	0.160782	18 815.000000	0.158865	False
Miranda (Cartagena)	CNNLSTMAttention	humidity	0.152090	0.164088	18 764.000000	0.148471	False
Miranda (Cartagena)	Transformer	humidity	0.163778	0.209982	7359.000000	0.000000	True
Sonora (México)	MLP	humidity	0.101833	0.131871	8899.000000	0.000000	True
Sonora (México)	LSTM	humidity	0.076447	0.063945	15 229.000000	0.000080	True
Sonora (México)	CNNLSTM	humidity	0.078701	0.082643	18 732.000000	0.142220	False
Sonora (México)	LSTMAttention	humidity	0.099880	0.060874	11 065.000000	0.000000	True
Sonora (México)	CNNLSTMAttention	humidity	0.095747	0.093421	17 491.000000	0.019034	True
Sonora (México)	Transformer	humidity	0.177372	0.066850	4478.000000	0.000000	True
Alhama de Murcia (Murcia)	MLP	radiation	0.075653	0.048970	17 877.000000	0.038266	True
Alhama de Murcia (Murcia)	LSTM	radiation	0.060039	0.232763	3964.000000	0.000000	True
Alhama de Murcia (Murcia)	CNNLSTM	radiation	0.066879	0.092243	10 923.000000	0.000000	True
Alhama de Murcia (Murcia)	LSTMAttention	radiation	0.079008	0.127235	11 341.000000	0.000000	True
Alhama de Murcia (Murcia)	CNNLSTMAttention	radiation	0.068645	0.083941	8956.000000	0.000000	True
Alhama de Murcia (Murcia)	Transformer	radiation	0.072060	0.055544	10 350.000000	0.000000	True
Antalia (Turquía)	MLP	radiation	0.040241	0.049986	16 910.000000	0.005859	True
Antalia (Turquía)	LSTM	radiation	0.083360	0.058290	18 191.000000	0.064311	False
Antalia (Turquía)	CNNLSTM	radiation	0.084830	0.064573	19 570.000000	0.381481	False
Antalia (Turquía)	LSTMAttention	radiation	0.055777	0.049919	19 957.000000	0.547446	False
Antalia (Turquía)	CNNLSTMAttention	radiation	0.061615	0.054565	20 333.000000	0.737030	False
Antalia (Turquía)	Transformer	radiation	0.035748	0.043936	15 538.000000	0.000195	True
Falces (Navarra)	MLP	radiation	0.087294	0.057421	6459.000000	0.000000	True
Falces (Navarra)	LSTM	radiation	0.082652	0.089041	18 318.000000	0.078367	False
Falces (Navarra)	CNNLSTM	radiation	0.093248	0.095638	16 401.000000	0.001837	True
Falces (Navarra)	LSTMAttention	radiation	0.065633	0.092596	10 874.000000	0.000000	True
Falces (Navarra)	CNNLSTMAttention	radiation	0.085385	0.082360	18 915.000000	0.180829	False
Falces (Navarra)	Transformer	radiation	0.076225	0.077653	20 004.000000	0.569786	False
Melipilla (Chile)	MLP	radiation	0.096039	0.110827	19 279.000000	0.279748	False
Melipilla (Chile)	LSTM	radiation	0.137307	0.264735	9729.000000	0.000000	True
Melipilla (Chile)	CNNLSTM	radiation	0.134419	0.134313	16 044.000000	0.000758	True
Melipilla (Chile)	LSTMAttention	radiation	0.146316	0.173471	13 116.000000	0.000000	True
Melipilla (Chile)	CNNLSTMAttention	radiation	0.129664	0.107501	16 589.000000	0.002859	True
Melipilla (Chile)	Transformer	radiation	0.154506	0.118204	20 170.000000	0.651977	False
Miranda (Cartagena)	MLP	radiation	0.060931	0.073470	14 025.000000	0.000002	True
Miranda (Cartagena)	LSTM	radiation	0.075434	0.103363	13 253.000000	0.000000	True
Miranda (Cartagena)	CNNLSTM	radiation	0.048607	0.085699	8569.000000	0.000000	True
Miranda (Cartagena)	LSTMAttention	radiation	0.044043	0.093286	7104.000000	0.000000	True
Miranda (Cartagena)	CNNLSTMAttention	radiation	0.075162	0.061953	19 793.000000	0.473049	False
Miranda (Cartagena)	Transformer	radiation	0.055860	0.077610	8622.000000	0.000000	True
Sonora (México)	MLP	radiation	0.027679	0.062743	6596.000000	0.000000	True
Sonora (México)	LSTM	radiation	0.051200	0.052058	18 096.000000	0.055215	False
Sonora (México)	CNNLSTM	radiation	0.054312	0.062874	15 859.000000	0.000468	True
Sonora (México)	LSTMAttention	radiation	0.028872	0.034277	15 701.000000	0.000306	True
Sonora (México)	CNNLSTMAttention	radiation	0.037138	0.052185	15 520.000000	0.000185	True
Sonora (México)	Transformer	radiation	0.035257	0.054101	10 888.000000	0.000000	True

Data availability

Data will be made available on request.

References

Ardiansah, I., Bafdal, N., Suryadi, E., & Bono, A. (2020). Greenhouse monitoring and automation using arduino: a review on precision farming and internet of things (IoT). *International Journal on Advanced Science, Engineering and Information Technology*, *10*(2), 703–709.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017). Conditional time series forecasting with convolutional neural networks. arXiv preprint arXiv:1703.04691.

Chahal, A., Addula, S. R., Jain, A., Gulia, P., Gill, N. S., et al. (2024). Systematic analysis based on conflux of machine learning and internet of things using bibliometric analysis.. *Journal of Intelligent Systems & Internet of Things*, *13*(1).

Codeluppi, G., Cilfone, A., Davoli, L., & Ferrari, G. (2020). Ai at the edge: a smart gateway for greenhouse air temperature forecasting. In *2020 IEEE international workshop on metrology for agriculture and forestry* (pp. 348–353). IEEE.

Eraliev, O., & Lee, C.-H. (2023). Performance analysis of time series deep learning models for climate prediction in indoor hydroponic greenhouses at different time intervals. *Plants*, *12*(12), 2316.

Fink, M., Daniels, A., García-Mañas, F., Rodríguez, F., Leibold, M., & Wollherr, D. (2025). Learning-based model identification for greenhouse climate control. *At-Automatisierungstechnik*, *73*(6), 451–465. http://dx.doi.org/10.1515/auto-2024-0163, Publisher Copyright: © 2025 the author(s), published by De Gruyter..

García-Vázquez, F., Ponce-González, J. R., Guerrero-Osuna, H. A., Carrasco-Navarro, R., Luque-Vega, L. F., Mata-Romero, M. E., Martínez-Blanco, M. d. R., Castañeda-Miranda, C. L., & Díaz-Flórez, G. (2023). Prediction of internal temperature in greenhouses using the supervised learning techniques: Linear and support vector regressions. *Applied Sciences*, *13*(14), http://dx.doi.org/10.3390/app13148531.

Goldenits, G., Mallinger, K., Raubitzek, S., & Neubauer, T. (2024). Current applications and potential future directions of reinforcement learning-based digital twins in agriculture. *Smart Agricultural Technology*, *8*, Article 100512. http://dx.doi.org/10.1016/j.atech.2024.100512.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232.

Hamidane, H., EL Faiz, S., Rkik, I., El Khayat, M., Guerbaoui, M., Ed-Dahhak, A., & Lachhab, A. (2024). Constrained temperature and relative humidity predictive control: Agricultural greenhouse case of study. *Information Processing in Agriculture*, *11*(3), 409–420. http://dx.doi.org/10.1016/j.inpa.2023.04.003.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jeon, Y.-J., Kim, J. Y., Hwang, K.-S., Cho, W.-J., Kim, H.-J., & Jung, D.-H. (2024). Machine learning-powered forecasting of climate conditions in smart greenhouse containing netted melons. *Agronomy*, *14*(5), http://dx.doi.org/10.3390/agronomy14051070.

Jin, X.-B., Zheng, W.-Z., Kong, J.-L., Wang, X.-Y., Zuo, M., Zhang, Q.-C., & Lin, S. (2021). Deep-learning temporal predictor via bidirectional self-attentive encoder–decoder framework for IOT-based environmental sensing in intelligent greenhouse. *Agriculture*, *11*(8), 802.

Jung, D.-H., Lee, T. S., Kim, K., & Park, S. H. (2022). A deep learning model to predict evapotranspiration and relative humidity for moisture control in tomato greenhouses. *Agronomy*, *12*(9), 2169.

Kaneda, Y., & Mineno, H. (2016). Sliding window-based support vector regression for predicting micrometeorological data. *Expert Systems with Applications*, *59*, 217–225.

Lai, G., Chang, W.-C., Yang, Y., & Liu, H. (2018). Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 95–104).

Liu, W., Han, T., Wang, C., Zhang, F., & Xu, Z. (2025). Predicting indoor temperature of solar green house by machine learning algorithms: A comparative analysis and a practical approach. *Smart Agricultural Technology*, *12*, Article 101096. http://dx.doi.org/10.1016/j.atech.2025.101096.

Liu, Y., Li, D., Wan, S., Wang, F., Dou, W., Xu, X., Li, S., Ma, R., & Qi, L. (2022). A long short-term memory-based model for greenhouse climate prediction. *International Journal of Intelligent Systems*, *37*(1), 135–151.

Liu, J., Shao, M., et al. (2023). The forecast of power consumption and freshwater generation in a solar-assisted seawater greenhouse system using a multi-layer perceptron neural network. *Expert Systems with Applications*, *213*, Article 119289.

Ma, H.-J., Jin, X.-B., Li, Z.-M., & Bai, Y.-T. (2024). Fuzzy adaptive-normalized deep encoder-decoder network: Medium and long-term predictor of temperature and humidity in smart greenhouses. *Computers and Electronics in Agriculture*, *226*, Article 109480. http://dx.doi.org/10.1016/j.compag.2024.109480.

Mallick, S., Airaldi, F., Dabiri, A., Sun, C., & De Schutter, B. (2025). Reinforcement learning-based model predictive control for greenhouse climate control. *Smart Agricultural Technology*, *10*, Article 100751.

Mansour, M., Sathyanarayanan, K. K., Sauerteig, P., & Streif, S. (2025). Adaptive robust greenhouse climate control: Combining deep reinforcement learning and economic optimization. *Smart Agricultural Technology*, *12*, Article 101327. http://dx.doi.org/10.1016/j.atech.2025.101327.

Maraveas, C. (2022). Incorporating artificial intelligence technology in smart greenhouses: Current state of the art. *Applied Sciences*, *13*(1), 14.

Maraveas, C. (2023). Incorporating artificial intelligence technology in smart greenhouses: Current state of the art. *Applied Sciences*, *13*(1).

Morales-García, J., Bueno-Crespo, A., Martínez-España, R., & Cecilia, J. M. (2023). Data-driven evaluation of machine learning models for climate control in operational smart greenhouses. *Journal of Ambient Intelligence and Smart Environments*, (Preprint), 1–15.

Morales-García, J., Terroso-Sáenz, F., & Cecilia, J. M. (2024). A multi-model deep learning approach to address prediction imbalances in smart greenhouses. *Computers and Electronics in Agriculture*, *216*, Article 108537.

Morcego, B., Yin, W., Boersma, S., Van Henten, E., Puig, V., & Sun, C. (2023). Reinforcement learning versus model predictive control on greenhouse climate control. *Computers and Electronics in Agriculture*, *215*, Article 108372.

Naagarajan, R. A., & Streif, S. (2025). Enhancing greenhouse management with interpretable AI: A natural language interface for advanced and optimization-based control. *Smart Agricultural Technology*, *11*, Article 101041. http://dx.doi.org/10.1016/j.atech.2025.101041.

Nakhaei, M., Ahmadi, A., Gheibi, M., Chahkandi, B., Hajiaghaei-Keshteli, M., & Behzadian, K. (2023). A smart sustainable decision support system for water management of power plants in Water Stress Regions. *Expert Systems with Applications*, Article 120752.

Nghiem, T. X., Drgoňa, J., Jones, C., Nagy, Z., Schwan, R., Dey, B., Chakrabarty, A., Di Cairano, S., Paulson, J. A., Carron, A., Zeilinger, M. N., Cortez, W. S., & Vrabie, D. L. (2023). Physics-informed machine learning for modeling and control of dynamical systems. arXiv:2306.13867.

Nutricontrol (2022). HP_CLIMA_V4_Tri_V01.02. Technical manual. URL: https://nutricontrol.com/wp-content/uploads/2022/10/HP_CLIMA_V4_Tri_V01.02_s.pdf, (Accessed 13 October 2025).

Nutricontrol (2025). Nutricontrol – greenhouse climate and fertigation control solutions. URL https://nutricontrol.com/es/home/, (Accessed 7 October 2025).

Oh, K. C., Kim, S. J., Park, S. Y., Cho, L., Lee, C. G., & Kim, D. H. (2023). Development of greenhouse internal temperature prediction model based on data characteristics using machine learning. Available At SSRN 4329492.

Project, S. (2025). SergIoT: Digital twin strategy for efficient greenhouse management. URL https://sergiot.com/, (Accessed 7 October 2025).

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. arXiv preprint arXiv:1704.02971.

Riskiawan, H., Gupta, N., Setyohadi, D., Anwar, S., Agus Kurniasari, A., Hariono, B., Firmansyah, M. H., Yutadi, Y., Mansur, A., & Basori, A. H. (2023). Artificial intelligence enabled smart monitoring and controlling of IoT-green house. *Arabian Journal for Science and Engineering*, *49*, http://dx.doi.org/10.1007/s13369-023-07887-6.

Ruiz, S., Morales-García, J., Calafate, C. T., Cano, J.-C., Manzoni, P., & Cecilia, J. M. (2022). Evaluation of time-series libraries for temperature prediction in smart greenhouses. In *2022 18th international conference on intelligent environments* (pp. 1–7). IEEE.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Sadigov, R., et al. (2022). Rapid growth of the world population and its socioeconomic results. *The Scientific World Journal*, *2022*.

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing* (pp. 4580–4584). Ieee.

Schmitt, J., Offermann, F., Söder, M., Frühauf, C., & Finger, R. (2022). Extreme weather events cause significant crop yield losses at the farm level in german agriculture. *Food Policy*, *112*, Article 102359.

Sharma, S., Saxena, A. K., & Bansal, M. (2022). Forecasting of GHG (greenhouse gas) emission using (ARIMA) data driven intelligent time series predicting approach. In *2022 7th international conference on communication and electronics systems* (pp. 315–322). IEEE.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, *28*.

Soheli, S. J., Jahan, N., Hossain, M., Adhikary, A., Khan, D. M. A., & Wahiduzzaman, M. (2022). Smart greenhouse monitoring system using internet of things and artificial intelligence. *Wireless Personal Communications*, *124*, http://dx.doi.org/10.1007/s11277-022-09528-x.

Sun, Q., Liao, B., & Tao, Q. (2019). Ecological agriculture development and spatial and temporal characteristics of carbon emissions of land use.. *Applied Ecology & Environmental Research*, *17*(5).

Tsai, Y.-Z., Hsu, K.-S., Wu, H.-Y., Lin, S.-I., Yu, H.-L., Huang, K.-T., Hu, M.-C., & Hsu, S.-Y. (2020). Application of random forest and ICON models combined with weather forecasts to predict soil temperature and water content in a greenhouse. *Water*, *12*(4), 1176.

Ullah, I., Fayaz, M., Aman, M., & Kim, D. (2022). Toward autonomous farming—A novel scheme based on learning to prediction and optimization for smart greenhouse environment control. *IEEE Internet of Things Journal*, *9*(24), 25300–25323. http://dx.doi.org/10.1109/JIOT.2022.3196053.

Vanegas-Ayala, S.-C., Barón-Velandia, J., & Leal-Lara, D.-D. (2022). A systematic review of greenhouse humidity prediction and control models using fuzzy inference systems. *Advances in Human-Computer Interaction*, *2022*(1), Article 8483003.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, Z., Liu, Z., Yuan, M., Yin, W., Zhang, C., Zhang, Z., & Hu, X. (2025). A machine learning-based irrigation prediction model for cherry tomatoes in greenhouses: Leveraging optimal growth data for precision irrigation. *Computers and Electronics in Agriculture*, *237*, Article 110558. http://dx.doi.org/10.1016/j.compag.2025.110558.

Weatherbit (2025). Weatherbit API. URL https://www.weatherbit.io/, (Accessed 7 October 2025).

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, *34*, 22419–22430.

Yang, Y., Gao, P., Sun, Z., Wang, H., Lu, M., Liu, Y., & Hu, J. (2023). Multistep ahead prediction of temperature and humidity in solar greenhouse based on FAM-LSTM model. *Computers and Electronics in Agriculture*, *213*, Article 108261. http://dx.doi.org/10.1016/j.compag.2023.108261.

Yu, J., Sun, C., Zhao, J., Ma, L., Zheng, W., Xie, Q., & Wei, X. (2025a). Prediction and control of greenhouse temperature: Methods, applications, and future directions. *Computers and Electronics in Agriculture*, *237*, Article 110603.

Yu, J., Zhao, J., Sun, C., Zhang, R., Zheng, W., Xu, L., & Wei, X. (2025b). A dual deep learning approach for winter temperature prediction in solar greenhouses in northern China. *Computers and Electronics in Agriculture*, *229*, Article 109807. http://dx.doi.org/10.1016/j.compag.2024.109807.

Zhai, Z., Martínez, J. F., Beltran, V., & Martínez, N. L. (2020). Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, *170*, Article 105256.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175.