Journal of Science and Engineering Applications

Contents are available at https://jsea.iujournals.com

A Hybrid Mamba–GNN Framework for Spatio-Temporal Solar Irradiance Forecasting Using Sky Camera and Satellite Data Fusion

Samarah Qasim Al-Asadi
College of Health and Medical Technology, Al-Furat Al-Awsat Technical University, Iraq.

ARTICLE INFORMATION

Received date: 7-09-2026
Revised date: 13-10-2026
Accepted date: 15-10-2026

Keywords

Ablation Study
Forecasting
Graph neural networks
Mamba
Solar irradiance

ABSTRACT

Solar irradiance forecasting is essential for integrating renewable energy into power grids, as intermittency challenges grid stability and efficiency. Existing methods struggle to capture high-resolution local cloud dynamics and synoptic-scale atmospheric patterns while maintaining computational efficiency. This study proposes Mamba–GNN, a hybrid deep learning framework integrating sky camera images and satellite data. A Vision State Space Model (Mamba) extracts high-frequency local cloud dynamics, while a Spatio-Temporal Graph Neural Network (ST-GNN) models regional cloud propagation. A gated cross-attention mechanism adaptively fuses heterogeneous features. Results demonstrate superior forecasting performance over all baselines across evaluated horizons, achieving scores of 0.85 at 10 minutes and 0.55 at 3 hours, with greater improvements at longer horizons. The framework achieves an inference latency of 5.0 ms per sample on a single GPU, enabling near-real-time operation. Ablation studies validate individual component contributions and fusion effectiveness, confirming the framework's potential for accurate, computationally efficient solar irradiance forecasting.

1. Introduction

The global energy landscape is changing dramatically, even as a decarbonized electricity generation system and mitigation of anthropogenic climate change becomes an imperative. Solar PV (PV) is poised to be a key enabler of this transition, having seen its LCO drop quickly, with the potential resource base present, and its capacity to be scaled to different geographies [1]. The widespread adoption of solar power in electricity networks, however, is faced with operation problems associated with the nature of surface solar irradiance which is intermittent [2]. The output of a solar farm is a changing, and often unpredictable, parameter compared with the dispatchable generation sources because of the movement, formation and dissipation of clouds above the solar farm [3]. These changes can cause large variations in the net load profile of the grid, which is often referred to as the duck curve [4]. The increased steepness of the duck curve – steep evening ramp requirements and potential over-generation in the middle of the day – puts significant stress on reserve power plants and ancillary services mechanisms. In economic terms, forecast errors are directly reflected in higher balancing costs, lower amounts of renewable energy generation and, in extreme cases, in grid instability events [5]. Forecasting solar irradiance is thus not only a "nice-to-have" of operation convenience but, crucially,

* Corresponding authors: College of Health and Medical Technology, Al-Furat Al-Awsat Technical University, Iraq
E-mail addresses: samara.q.ali@atu.edu.iq (Samarah Qasim Al-Asadi)

an enabling technology to enable the high renewable energy penetration targets while keeping the grid secure and economically efficient [6]. A need for forecasts at varying timeframes ranging from intra-hour ramp forecast to day-ahead scheduling arises from the need to make effective unit commitment, economic dispatch and real-time congestion management [7].

High fidelity solar irradiance forecasts are based on the premise that observations are able to resolve spatio-temporal cloud system dynamics well. For this purpose, two basic types of remote sensing from the ground and from space are used: sky cameras, and geostationary satellites [8]. All modalities have their own set of strengths and weaknesses that have a fundamental impact on the architecture and performance of downstream forecasting models [9].

By virtue of capturing high-resolution hemispherical images at very high temporal cadence (10 – 30 seconds), ground-based instruments called sky cameras or All-Sky Imagers (ASI) can be used to observe the sky [10]. On the hyper-local scale of observation, typically within 5–10 km, the high-resolution characterization of local cloud properties, such as cloud speed and height, as well as fractional coverage and importantly the fine-textured evolution of cloud edges that lead to high frequency irradiance ramps is possible [11]. The main disadvantage of sky cameras is that they have a restricted field of view which leads to the issue of "horizon blindness. If clouds move in from another distance, but are advectioning towards the sensor, they will not be seen until they are in the camera's field of view. This constraint sets a fundamental limit on the forecast horizon when only a single sky camera is available to make forecasts, in this case of 5-15 minutes [12].

The geostationary satellites, on the other hand, offer pancontinental scale synoptic observations, encompassing the long-range movement of cloud systems in several different atmospheric channels that are related to cloud microphysics [13]. The power of the satellite observations is that they can be used to monitor cloud masses that are hundreds of kilometres away from the location of interest, thus giving a spatial context for a longer horizon forecast. But these observations have a number of important restrictions. They have a relatively poor spatial resolution (1 - 2 km at the sub-satellite point) and a temporal resolution of 5 to 15 minutes, dependent on the satellite's scanning schedule. In addition, the off-nadir viewing geometry for sites far from the sub-satellite point results in geometric distortions of the cloud positions, the parallax error. Further hindering the use of satellite data for intra-hour ramps forecasting is an operational latency, which is an inherent part of the data acquisition, processing and dissemination chain [14]. These constraints limit the use of satellite data to providing reasonable solutions to the initiation and fast development of small, convective cloud systems, which are the main contributors to severe solar ramps [15]. Single-modality solar forecasting systems have their own inherent limitations as shown in Figure 1: the sky camera has a horizon blind problem and the satellite data has problem of spatio-temporal constraints.

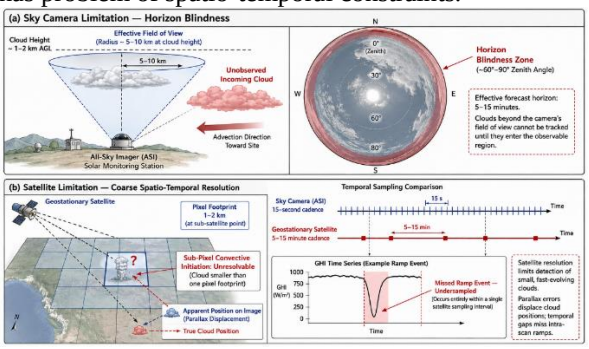


Figure 1. Schematic illustration of the limitations of single-modality solar forecasting systems: (a) sky camera horizon blindness restricting forecasts to 5–15 minutes, and (b) geostationary satellite constraints including coarse resolution, parallax errors, and temporal under sampling.

There has been a large amount of research focused on solar irradiance forecasting, both in the field of statistics and physics, and recently by deep learning techniques [16]. The traditional physical approaches involve the use of numerical weather prediction (NWP) models, which are used to solve a set of discretized equations of the atmospheric dynamic and thermodynamic processes. Important for day-ahead time scales, NWP models are very costly for computation and have a course spatio-temporal resolution and imperfect parameterization of sub-grid scale cloud physics, which makes them less skilful for intra-day and intra-hour predictions [17].

Based on the idea of cloud motion field, image-based advection methods include extrapolating cloud motion fields obtained from a sequence of sky camera or satellite images, typically by using optical flow algorithms. The only drawback of these methods is that they assume Lagrangian persistence, which means they are unable to include complex non-linear processes in the atmosphere, such as cloud formation and dissipation, splitting and multi-layer interactions, and the growth of errors becomes large quickly with longer prediction time [18].

With the rise of deep learning, a new paradigm has emerged that has the potential to learn complex, non-linear mappings directly from large amounts of data for spatio-temporal forecasting. In this regard, there is a number of families of deep learning architectures that have been explored [19].

CNNs with RNNs like Convolutional Long Short-Term Memory (ConvLSTM) networks have been shown to have excellent performance in learning short-term temporal dependencies and hierarchical spatial features [20]. But such architectures have been observed to exhibit vanishing gradient problems with long image sequences, and have local receptive fields of small sizes which lack the ability to include long-range spatial dependencies without too many layers in the network. The self-attention mechanism, which can be used for global context aggregation, is what has led to the development of Vision Transformers (ViT) and their

hierarchical versions like Swin Transformers to solve the long-range dependency issue [21]. However, these models have a quadratic computational time complexity with respect to the length of the sequence, so that they are not feasible for high frequency long duration sky camera videos or high-resolution satellite image sequences [22]. Graph Neural Networks (GNNs) are a promising approach to represent atmospheric conditions on irregular graph structures to capture spatial connections. They can be used to naturally encode spatial dependencies, and are good at propagating information over a geographically distributed network of nodes. But the single GNN forecasting models typically require manually designed graph adjacency matrix and cannot effectively model the high-dimensional temporal auto-correlation in data alone, and thus need to be complicated hybrid models [18 – 23].

There is a significant lack in the literature, as there is no framework that addresses scalability, temporal fidelity and spatial context in an adequate way. This gap is addressed by a computationally feasible framework that combines a temporal encoder with linear-time complexity, which is adaptive and suitable for high-frequency video streams, with a dynamic, distance-aware model of spatial propagation [19]. None of the single-model solutions satisfies all three of these requirements well and thus, there is a gap in the state of the art [20].

The identified architectural gap is addressed by recent advances in deep state space modeling. The models Structured State Space Sequence (S4) and Structured State Space Sequence Selective (Mamba) are regarded as game-changers for sequence modeling [24]. These models are based on a linear-time-invariant state space representation that, after being discretized using structured transition matrices, is both efficient and competitive with, and often better than Transformers for, modeling long-range contexts. The unique contribution of the Mamba model is a selective scan function, which involves selectively feeding the state space parameters depending on the input, via a parallelizable selective algorithm. This is a novel innovation that helps the model to address the three fundamental challenges of sequence modeling: high performance on dense prediction tasks, capturing extremely long-range context and maintaining linear runtime in length of the sequence [25]. In the current work, this is applied to vision-based tasks, like the Vision Mamba (Vim), and adapted to the spatio-temporal space of multi-modal solar irradiance forecasts [26]. It is hypothesized that Mamba is a proper local temporal encoder for high-frequency camera sequences captured from the sky and a spatio-temporal GNN is a proper propagation mechanism for coarse-grained, satellite-derived atmospheric states, hence a complementary relationship suitable for the forecasting problem is provided. The Mamba branch in this architecture condenses dense and rich information from local cloud texture, opacity and velocity [27]. The long-range, synoptic-scale atmospheric dynamics captured and propagated by the GNN branch enrich this local information, giving the regional context.

The novelty of the presented research is that Mamba and GNNs are integrated together in a task-specific manner to predict multi-modal solar forecasting while retaining the established technique. There are three differences to be pointed out from the works already published. First, prior approaches to vision state space models [36–37] focus on processing single modality image or video tokens and do not make use of graph-structured inputs, while Mamba is applied here as a linear-time spatio-temporal encoder of the sky camera video data, and satellite data is processed by a distinct ST-GNN branch. Third, previous spatio-temporal GNNs for irradiance forecasting [1, 4, 22] generally adopt graph topologies that are fixed and temporal modules such as convolutional or recurrent layers, while in this work, a learnable dynamic adjacency matrix that integrates geographic distance and feature correlation (Eq. 3) is combined with a selective state space temporal encoder. Third, existing sky image–satellite fusion models [7, 10] use neither selective state space encoding nor a dynamic graph-based satellite branch, and generic cross-attention requires both modalities to be defined on the same grid, while a graph-to-grid Kriging projection followed by a regime-dependent gate (Eqs.) does not. The benefit is quantified here using the formula (19–21). This ensemble has never been used before for solar irradiance forecasting, to the best knowledge of the authors.

This study's key contributions can be put as follows:

1. Selective state space models (Mamba) integrated with spatio- temporal Graph Neural Networks for solar forecasting, which is, to the best of the authors' knowledge, the first application of the combination of these two classes of models to multi-modal (sky camera + satellite) solar forecasting. Individual components exist in other areas but their combined application of dynamic fusion with solar forecasting is new and addresses a key need for solar forecasting, where the need for temporal fidelity is critical to provide spatial context.
2. A dynamic cross-attention fusion module that adaptively weights the importance of the local (sky camera) and regional (satellite) feature streams according to the weather regime variability is designed.
3. The framework is validated using a multi-modal dataset collected over more than two years at a single site, and it is shown to lie on the accuracy–latency Pareto frontier among the evaluated models; deployment on resource-constrained edge devices has not been evaluated.

The proposed framework integrates the local temporal dynamics and regional spatial context in a unified framework of Mamba–GNN as shown in Figure 2.

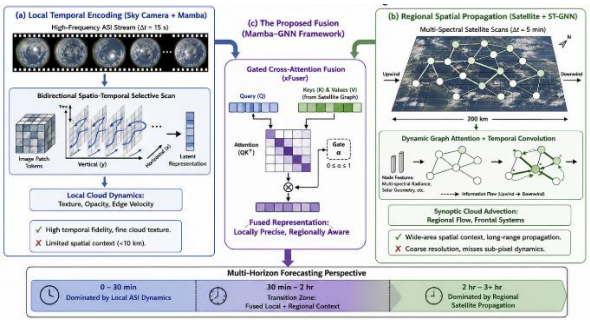


Figure 2. Architectural motivation of the proposed Mamba–GNN framework, showing the complementarity between local temporal encoding (sky camera + Mamba) and regional spatial propagation (satellite + ST-GNN), fused via gated cross-attention for multi-horizon solar forecasting.

The goal of the present study is to develop a more reliable and less time-consuming forecasting system for surface solar irradiance due to the complementary information available from ground-based sky cameras and geostationary satellite sensors for forecasting short (intra-hour to near-future) timescales. High accuracy and scalability of conventional deep learning architectures, and the lack of compatibility of high complexity Transformer models and the high-frequency, high-resolution data stream typical of this application, are an on-going scientific and practical problem.

The key challenge of this study is thus to design, implement, and thoroughly validate a hybrid framework comprising Mamba and GNN that addresses the trilemma of sequence modeling with high forecast accuracy, long-range spatio-temporal context, and linear computational scaling. A secondary goal is to estimate the relative importance of the proposed fusion mechanism and the selective state space paradigm to skillful forecasts, especially in the case of extreme irradiance ramps caused by complex cloud situations.

The rest of this paper is organized as follows: Section II presents all the related works in solar forecasting in a comprehensive review manner based on images and satellite data, along with state space models. In Section III, the proposed Mamba-GNN fusion framework is described, which includes the graph construction, dual-branch encoders and a cross-modal fusion module.

The experimental setup, including the dataset chosen, the baseline models and the evaluation metrics, are detailed in Section IV. The outcome of the quantitative and qualitative assessment and a detailed ablation study are reported and discussed in Section V. Finally, the paper ends with a summary of the results and conclusions and possible future avenues of research is introduced in Section VI.

2. Related Work

2.1. Image-Based Solar Forecasting (Sky Camera Domain)

Cloud motion vector retrieval and prediction of cloud motion from ground-based sky camera imagery have been studied for many years. Most of the classical methods are based on block-matching and dense optical flow methods, such as Lucas-Kanade and its pyramidal extension [12]. In such systems, the assumption is that each point in a local spatio-temporal neighbourhood moves at a constant speed, and that velocity vectors are calculated by minimising the least-squares cost function. Once the motion field of the cloud is extracted, a Lagrangian persistence assumption is used to propagate the cloud map into the future [13]. These methods are efficient and are suitable for large-scale clouds with constant advection. Their basic drawback is, however, that they assume that the cloud's brightness, or photometric constancy, remains constant and that its motion is smooth [14]. These are often not true for convective clouds, which rapidly form and dissipate and are turbulently distorted [15].

However, due to the limitations of classical image processing, deep learning methodologies have come into vogue. Initial adaptations applied existing convolutional architectures to the problem of solar irradiance classification for the 3D nature of the problem and static image-based regression, which involve assigning a single sky image to a future Global Horizontal Irradiance (GHI) value. These models perform better with the recognition of cloud patterns than with handcrafted features, but fail to take into account the temporal information. To capture dynamics, spatio-temporal graph convolutional networks (ST-GCNs) are used to learn the temporal evolution of clouds by modeling input image patches as graph nodes and learning their interactions in time. But these approaches may be limited by the hard-coded static graph topologies [16]. Later on, architectures dedicated to video prediction, like 3D-CNNs and ConvLSTM networks become the de-facto standard. ConvLSTM builds upon fully connected LSTM transition by incorporating convolutional operations to simultaneously learn the spatial appearance and temporal dynamics [18]. Likewise, 3D-CNNs use spatio-temporal convolutional kernels on top of a sequence of images. These recurrent models are more effective at capturing the evolution of clouds locally than optical flow, but they are also prone to vanishing gradients over long sequences and a relatively small receptive field which can not capture long-range dependencies effectively, causing a progressive loss of cloud texture and predictive skill beyond 20 minutes for the recurrent model [28].

2.2. Satellite-Based Irradiance Forecasting

Geostationary satellite imagery is also extensively used for solar irradiance predictions, especially for periods longer than one hour, in parallel with the ground-based observations. The basic physical method is represented by the Heliosat family of algorithms, in which the reflectance of the images acquired by the satellite at the visible channel is transformed into a cloud index, to be used to modulate a clear-sky irradiance model [29]. These methods are physically interpretable but are very

sensitive to the assumptions of surface albedo and the aerosol content in the atmosphere and to the calibration of the sensor. Machine-learning versions of Heliosat are suggested to achieve an empirical calibration of the cloud index-to-irradiance mapping, by training multilayer perceptrons or support vector regressors with historical collocations of satellite pixels and ground pyranometer measurement [30]. These approaches have the advantage of providing better bias correction, but perform independently on individual pixels without taking into account the information from neighboring pixels [31].

Full disk satellite imagery is then studied using deep learning to explicitly learn and propagate spatial features. In biomedical image segmentation, UNet architectures are proposed to extract image segmentation by encoding the biomedical images and decoding into the segmentation map, which is adapted for nowcasting of irradiance at the pixel level by encoding the image of a sequence of channels of radiance measured from the satellite and decoding into the image of the irradiance map in the future. Fine-scale spatial structures are maintained during upsampling due to a UNet's encoder-decoder structure with skip connections [32]. The Vision Transformer (ViT) and the windowed Vision Transformer (ViT-W), are proposed to capture large-scale atmospheric teleconnections. The following architectures break up satellite image sequences into patches and perform global self-attention, allowing the model to compare cloud structures in different parts of the globe [33]. However, the computational complexity of these fully attentional models is quadratic with the number of image patches, making them impractical for use with high-resolution, full-disk images, and not directly applicable to high-frequency sensor streams [32]. In practice, a tradeoff is tacitly made between spatial resolution, temporal sampling rate and context length.

2.3. Multi-Modal Data Fusion in Geoscience

Another and active research area is the integration of different and heterogeneous Earth observation data sources. The combination of sky cameras on the ground and geostationary satellites is seen as a very fruitful but difficult path to bridge the scale gap between local precision and regional context [27]. The simplest fusion methods can be divided into two types: early fusion, which involves raw or preprocessed pixel data from various sources being joined along the feature channel dimension at the input layer; and late fusion, where the forecasts generated by the separately trained models are fused by weighted averaging or linear regression ensemble [28]. These two strategies fall short of capability in modelling complex, non-linear interactions among multiple data streams of heterogeneous characteristics. To tackle this, the advanced fusion techniques are created. This work proposes using of a cross-attention mechanism originally introduced in the machine translation community under the Transformer network, but adapted for fusion in the remote sensing domain. In this architecture, the query is computed on the feature map from a first modality; while features from a second modality are used to create keys and values, enabling, for example, a content-dependent mapping of spatial reference frames and resolutions between data sets [34]. Moreover, the memory modules inspired by Neural Turing machines are explored for persistent, long-term fusion, that is, a historical pattern of sensor correlation is stored in an external memory bank for addressing during inference. Although these developments have taken place, learnable, dynamic fusion of a particular combination of high-frequency, sky camera video streams and low-cadence, satellite image sequences have yet to be applied extensively, especially for solar forecasting, where temporally-synchronous and geometrically-registered streams are difficult to obtain [35].

2.4. State Space Models for Visual and Temporal Data

A major innovation in the field of high-quality time-series forecasting that has important implications for the field is the development of deep State Space Models (SSMs). Structured State Space Sequence (S4) is a long-range dependency model that has near-linear computational complexity. While solving a linear system with structured state transition matrices, S4 re-formulates a continuous-time linear system efficiently diagonalizes the matrices, and convolves them with the input sequence. This model has state-of-the-art performance on the Long-Range Arena (LRA) benchmark, and competitive performance against Transformers on tasks with up to 16,000 tokens [22].

The performance metrics of the LRA benchmark are informative: For the sequential CIFAR-10 task, S4 is able to test with accuracy of 91.1% and a computational speed-up of more than 90 times compared to a standard Transformer, whereas the accuracy of the standard Transformer is reported to be 80.6% for the same sequence length. For the Path-X task with 16,384-length sequences, the accuracy of S4 is 96.0% and the baseline algorithms of LSTM and Transformer do not outperform the random baseline of 50% [23].

The most significant weakness of S4, namely its time-invariant dynamics that hinder context-dependent reasoning, is addressed by the Mamba model [32]. A creative approach to the state space is adopted, in which the state space matrices are computed as a function of the input, allowing irrelevant context to be selectively ignored while relevant context is emphasized. This approach is implemented using a hardware-aware parallel scan algorithm, where linear computational complexity is achieved, and an increased capacity for data modeling across a range of dense data modalities of similar size is demonstrated in comparison with Transformers [28].

Rapid developments in the adaptation of the Mamba architecture for the visual domain are observed. Bidirectional selective scanning over flattened image patches converted into sequences is proposed in the Vision Mamba (V_{im}) and V_{Mamba} architectures [29]. These models are considered Pareto optimal, as an optimal trade-off between image classification accuracy and inference throughput is achieved. Superior throughput is achieved by V_{Mamba}-Tiny compared to the corresponding Swin-Tiny Transformer (4588 images per second for V_{Mamba}-Tiny), while comparable accuracy is maintained (82.2% vs. 79.9% top-1 accuracy on the ImageNet-1K dataset) [22].

An extension to spatial-temporal bidirectional scanning is suggested; however, the use of multi-modal geoscience data for dense forecasting tasks remains unexplored and is addressed in the present study [31].

To conclude, the currently available vision state space models [36–38] rely on image or video inputs, while existing spatio-temporal GNNs for irradiance forecasting lack the use of the two selective state space temporal encoders, and existing sky image–satellite fusion models [7, 10] do not use either of the two. In the following study, the combination of Mamba-based sky camera encoder, dynamic-graph satellite encoder and regime-adaptive gated fusion is addressed.

Table 1. Comparative Analysis of Prior Studies and Their Limitations

Methodological Category	Representative Methods	Key Performance Metrics (Reported)	Computational Complexity	Primary Limitations and Application Gaps
Classical ASI Advection	Lucas-Kanade Optical Flow, Block-Matching	Forecast Skill 10-20% at 10 min horizon for GHI. RMSE for kt ~0.15 at 5 min.	Very Low	Assumption of photometric constancy and Lagrangian persistence fails under cloud formation/dissipation. Limited to <15 min.
Deep ASI Video Prediction	ConvLSTM, PredRNN, 3D-CNN	Best RMSE for 1-hour GHI: ~120 W/m². Structural Similarity Index (SSIM) of predicted images >0.80 at 6 frames.	Moderate	Vanishing gradients for long sequences (>10 frames). Receptive field limited by kernel size, causing progressive blurring.
Satellite Physical/ML	Heliosat-2, ML-adapted Heliosat	Global mean bias error of 5-10 W/m². RMSE of GHI ~80-100 W/m² for 1hr forecast.	Low to Moderate	Pixel-wise processing ignores spatial context and atmospheric dynamics. Sensitive to calibration and aerosol uncertainty.
Satellite Deep Learning	UNet, Swin Transformer, Earthformer	1-hr Forecast Skill 15-25% over smart persistence. CSI for ramp events (200 W/m²) ~0.35.	Quadratic (Transformer)	Quadratic complexity prohibits high-resolution, long-sequence input. Cannot resolve sub-5-min cloud pulsation dynamics.
Multi-Modal Fusion	Cross-Attention, Neural Turing Machine	RMSE reduction of 5-15% over best single-source model. No ASI+Sat fusion benchmark reported for solar.	High (Attention)	Fusion of high-frequency ASI video with low-cadence satellite is unaddressed. Heterogeneous geometry and temporal registration are unsolved.
State Space Models	S4, Mamba, Vision Mamba (Vim)	LRA Path-X: 96.0% (S4). ImageNet-1K Top-1: 82.2% (VMamba-T). Throughput: 4588 img/s.	Linear O(N)	Direct application to dense, multi-source spatio-temporal forecasting (ASI + satellite) is not yet investigated. Potential for linear scaling is untapped.

RMSE = Root Mean Square Error, GHI = Global Horizontal Irradiance, kt = clearness index, CSI = Critical Success Index, LRA = Long Range Arena Benchmark.

3. Methodology: The Mamba–GNN Fusion Framework

The overall process of the proposed framework is shown in Figure 3, which involves processing the data from different modalities, encoding it into Mamba and ST-GNN, fusing the cross-modal information, and performing iterative optimization according to the feedback from the validation.

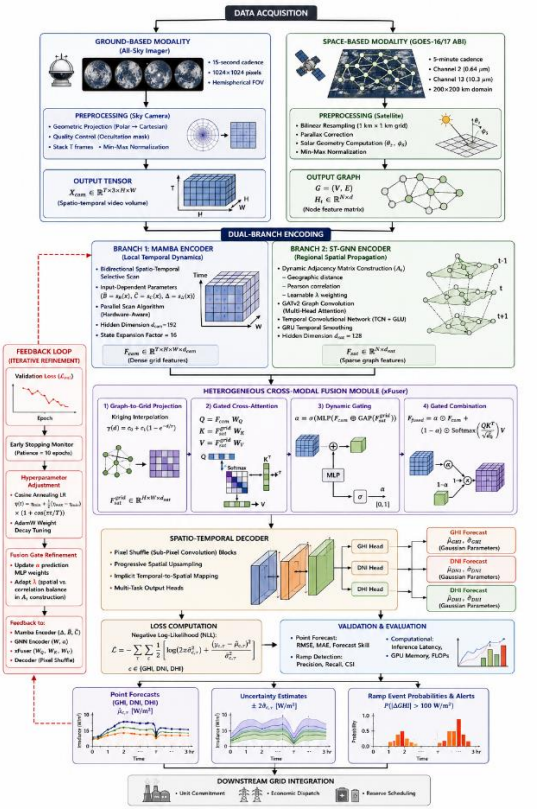


Figure 3. Overview of the proposed Mamba–GNN framework, illustrating the end-to-end pipeline from multi-modal data acquisition and preprocessing to dual-branch encoding, cross-modal fusion, probabilistic forecasting, and feedback-driven iterative refinement.

3.1. Overall Architectural Philosophy

A dual-branch encoder–decoder framework is proposed and is referred to as the Mamba–GNN Fusion Network. The general design principle is based on the functional decomposition of the spatio-temporal forecasting problem into two complementary sub-tasks. High-fidelity encoding of local cloud dynamics from ground-based cloud video and high-frequency observations is performed, while the propagation of large-scale atmospheric states from space-based multi-spectral cloud imagery is modelled.

Two distinct encoders are designed as parallel processors for heterogeneous input data streams, through which projection into a shared latent representation space is achieved. These latent representations are dynamically integrated within a dedicated fusion block, which serves as the computational hub of the framework. The fused feature maps are subsequently processed by a spatio-temporal decoder to generate the target irradiance forecast field.

The structural flow is represented as a composition of mappings arranged in sequence. The sky camera stream is encoded by a Mamba encoder, denoted as $\mathcal{E}_{mamba}$, while the satellite camera stream is encoded by a GNN encoder, denoted as $\mathcal{E}_{gnn}$. The outputs are merged through a fusion operation F, and a decoder Dis

employed to produce the final prediction $\hat{Y}$. The sequence is expressed as (1):

$$\hat{\mathrm{Y}} = \mathcal{D}(\mathcal{F}(\mathcal{E}_{mamba}(\mathrm{X}_{cam}), \mathcal{E}_{gnn}(\mathrm{X}_{sat}, \mathrm{A}_t))) \quad (1)$$

The raw input tensors X_{cam} and X_{sat} are captured by the sky camera and satellite, respectively, and A_t is the dynamic graph adjacency matrix.

3.2. Data Preprocessing and Graph Construction

Each sensor modality's raw data goes through a specific preprocessing pipeline to prepare it for the fusion process to make it both spatially and temporally compatible. In the case of the sky camera stream, the native image taken with the camera is in polar coordinates (r, θ), and it is transformed into a rectangular image of size (H×W) using a geometric projection. This projection eliminates the distortion caused by the fisheye lens and allows use of standard convolutional operations. Then a series of T projected frames is concatenated to create a four-dimensional spatio-temporal tensor, defined as (2):

$$\mathrm{X}_{cam} \in \mathbb{R}^{T\times 3\times H\times W} \quad (2)$$

The satellite stream is represented in a different fundamental way to take advantage of the irregular spatial distribution and relationships between atmospheric data. The satellite pixel array is sampled using a grid to give a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes (vertices) and with $\mathcal{E}$ as a set of edges. A feature vector is attached to each node $v_i \in \mathcal{V}$ and is used to represent multi-spectral radiance measurements from channels like visible (0.64 μm) and infrared (10.3 μm), and solar geometry angles (solar zenith angle (θ_z), solar azimuth angle (ϕ_s). All nodes at time t have the feature vector $\mathrm{H}_t \in \mathbb{R}^{N\times d}$ where N is the number of nodes and d is the feature dimension. The edge set $\mathcal{E}$ does not exist a priori, but is defined dynamically at every time step, based on an adjacency matrix A_t, which represents the changing spatial connectivity between pairs of nodes. The computation of this matrix is done as a learnable matrix, which consists of geographic distance and the Pearson correlation of the time series for each pixel.

The current adjacency matrix is derived from the previous one by calculating the dynamic adjacency matrix at time t as (3):

$$A_{ij}^t = \lambda \cdot \exp\left(-\frac{\|p_i - p_j\|_2^2}{2\sigma_p^2}\right) + (1-\lambda)\cdot \frac{\mathrm{Cov}(h_i, h_j)}{\sigma_{h_i}\sigma_{h_j}} \quad (3)$$

where $A_{ij}^t \in [0,1]$ is the normalized weight of the edge between nodes i and j at time t, $\lambda \in [0,1]$ is a learnable weighting parameter trained together with the model, $p_i, p_j \in \mathbb{R}^2$ represent the geographic coordinates (latitude and longitude) of nodes i and j, $\|\cdot\|_2$ denotes the Euclidean distance in geographic space tuned to the satellite grid spacing of approximately 1 km , $h_i(t)$ and $h_j(t) \in \mathbb{R}^d$ are the historical feature vectors at nodes i and j respectively, $\mathrm{Cov}(h_i, h_j)$ is the empirical covariance of the feature time series over the preceding 30 timesteps, and σ_{h_i} and σ_{h_j} are the standard deviations of the respective feature time series. Geographic proximity is represented by the first term, whereas dynamic atmospheric similarity is represented by the second term. To balance these two aspects, the parameter λ is learned during model training.

Satellite imagery is initially acquired as a regular 2D grid at a native spatial resolution of 1" km"×1" km" centered over the study area. For computational tractability, this grid is downsampled to a 200" km"×200" km" domain (yielding 200 pixels along each axis) while maintaining an effective spatial resolution of 1" km/pixel" . The N nodes of graph G are subsequently generated by selecting a spatial subset of grid points (typically sampled every 2–3 pixels to yield approximately 30–50 nodes), thereby balancing spatial coverage with computational efficiency.

3.3. Branch 1: Spatio-Temporal Selective State Space (Mamba) Encoder for Local Dynamics

In the sky camera branch, the encoding of extended temporal dependencies is the problem that needs to be solved across tens of high-frequency frames; both 3D conv nets and attention nets are sub-optimal to solve this problem. The Mamba model is chosen because it is a linear-time scaling model, and has been shown to have the ability to model long-range sequences.

The basic dynamics is represented by a linear system of continuous time equations relating an input function $u(t)$, an n -dimensional latent state $z(t)$ and an output function $y(t)$ as (4), (5):

$$\mathrm{z}'(t) = \mathrm{A}z(t) + \mathrm{B}u(t) \quad (4)$$

$$y(t) = \mathrm{C}z(t) + \mathrm{D}u(t) \quad (5)$$

where $\mathrm{A} \in \mathbb{R}^{n\times n}$, $\mathrm{B} \in \mathbb{R}^{n\times 1}$, $\mathrm{C} \in \mathbb{R}^{1\times n}$, and $\mathrm{D} \in \mathbb{R}$ are system parameters.

A time step Δ is used to discretize the continuous-time system to process the discrete spatio-temporal data. A zero-order hold (ZOH) is used for discretization and the following is used to get the discrete system matrices as (6), (7):

$$\overline{A} = \exp(\Delta A) \quad (6)$$

$$\overline{B} = (\Delta A)^{-1}(exp(\Delta A) - I)\Delta B \approx \Delta B \quad (7)$$

The discretized state transition matrix and the input matrix are denoted as $\overline{\mathbf{A}}$and $\overline{\mathbf{B}}$, respectively. When the time-step Δ is sufficiently small, the approximation in (7) is considered accurate. The discretized state-space model is expressed in the form of a linear recurrence as (8), (9):

$$z_k = \overline{A}z_{k-1} + \overline{B}u_k \quad (8)$$

$$y_k = Cz_k \quad (9)$$

The latent state vector at a discrete time step k is denoted by $z_k \in \mathbb{R}^n$, the input by u_k, the output by y_k, and the output matrix by $\mathbf{C}$.

A selectivity mechanism is introduced in the Mamba extension of the standard state space model (SSM), through which the parameters $\bar{B}$, $\bar{C}$, and Δ are made dependent on the input x. Specifically, $\bar{B} = s_B(x)$, $\bar{C} = s_C(x)$, and $\Delta = s_\Delta(x)$, where $s(\cdot$

)denotes linear projections. As a result, a time-varying and context-aware state space model is obtained, in which irrelevant information is filtered while the recurrence structure is preserved for efficient computation.

The recurrence is computed using a parallelizable selective scan algorithm on modern hardware accelerators, without requiring the full unrolling of the state.

A bidirectional spatio-temporal scanning mechanism is adopted to extend this approach to video volumes. In contrast to language modeling, where scanning is performed along the temporal axis, the 4D tensor $X_{\mathrm{cam}}$is unfolded into a time series along a space–time trajectory that sweeps across the volume in alternating directions. The scanning process is controlled by a spatial sensitivity kernel, through which the scanning path is determined based on local image structure.

As a result, full propagation of feature maps is achieved across the temporal, vertical, and horizontal dimensions of the video volume. The output of the encoder is represented as an abstract feature sequence $F_{\mathrm{cam}} \in \mathbb{R}^{T\times H\times W\times d_{\mathrm{cam}}}$, in which local cloud texture, opacity, and velocity vectors are captured over time.

3.4. Branch 2: Adaptive Spatio-Temporal Graph Neural Network for Synoptic Flow

The satellite branch is meant to simulate the propagation of cloud systems over a long-range, and it runs on the dynamic graph G_t.. A Graph Attention Network (GATv2) layer [41] is used to implement dynamic graph convolution to capture the direction and evolution of the influence between the different nodes of the atmosphere. In GATv2, the score function can be dynamically different from normal GAT in that the ranking of the weights of the attention may vary according to the features of query node. The node representation update for the layer l is given by (10):

$$\mathrm{h}_i^{l+1} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{l}\, \mathrm{W}^{l}\mathrm{h}_j^{l}\right) \tag{10}$$

The learnable weight matrix for the l-th layer is denoted by W^l, the non-linearity is denoted by $\sigma(\cdot)$, and the neighborhood of node i, as defined by A_t, is denoted by $N(i)$. The normalized attention coefficients α_{ij}^{l} are given by (11):

$$\alpha_{ij}^{l} = \frac{\exp\left(\mathrm{a}^{\mathsf{T}}\mathtt{LeakyReLU}\left(\mathrm{Wh}_i^{l}+\mathrm{Wh}_j^{l}\right)\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(\mathrm{a}^{\mathsf{T}}\mathtt{LeakyReLU}\left(\mathrm{Wh}_i^{l}+\mathrm{Wh}_k^{l}\right)\right)} \tag{11}$$

where a is an attention vector that's learned and T indicates transposition. This mechanism enables each node to change its attention to its most relevant upstream weather nodes learning the inherent directionality of cloud motion due to wind.

After the spatial graph convolution, the two mechanisms are added to explain the temporal dependencies between the evolution of synoptic scale. To model slow diurnal and seasonal cycles, a Temporal Convolutional Network (TCN) with a Gated Linear Unit (GLU) activation is used along the historical feature axis of each node. The gated TCN operation for a node at time t is given by (12):

$$\mathrm{T}_t = \mathrm{TCN}(\mathrm{h}_t) = (\mathrm{W}_f * \mathrm{h}_t + \mathrm{b}_f) \odot \sigma(\mathrm{W}_g * \mathrm{h}_t + \mathrm{b}_g) \tag{12}$$

where * denotes the temporal convolution operator, $\odot$ is the element-wise product, and $\mathbf{W}_f, \mathbf{W}_g$ are the filter and gate kernels. Second, a Gated Recurrent Unit (GRU) is used to smooth the state transition between the satellite's frequently-spaced timesteps, in order to provide a memory mechanism to fill in the temporal gaps between subsequent full-disk scans. The GRU state update consists of:

$$\mathrm{r}_t = \sigma(\mathrm{W}_{ir}\mathrm{T}_t + \mathrm{b}_{ir} + \mathrm{W}_{hr}\mathrm{s}_{t-1} + \mathrm{b}_{hr}) \tag{13}$$

$$\mathrm{u}_t = \sigma(\mathrm{W}_{iu}\mathrm{T}_t + \mathrm{b}_{iu} + \mathrm{W}_{hu}\mathrm{s}_{t-1} + \mathrm{b}_{hu}) \tag{14}$$

$$\mathrm{n}_t = \tanh(\mathrm{W}_{in}\mathrm{T}_t + \mathrm{b}_{in} + \mathrm{r}_t \odot (\mathrm{W}_{hn}\mathrm{s}_{t-1} + \mathrm{b}_{hn})) \tag{15}$$

$$\mathrm{s}_t = \mathrm{u}_t \odot \mathrm{s}_{t-1} + (1 - \mathrm{u}_t) \odot \mathrm{n}_t \tag{16}$$

where s_t is the hidden state at time t, r_t, u_t, and n_t are the reset gate, update gate, and new candidate state, respectively. The end result of this branch is a collection of graph node-level features, $\mathrm{F}_{sat} \in \mathbb{R}^{N\times d_{sat}}$, which combine the synoptic atmospheric state and the temporal evolution of the state.

3.5. The Heterogeneous Cross-Modal Fusion Module (xFuser)

One of the main difficulties in the merging of these two data types is characterized by the mismatch between the dense, regular grid representation of the sky camera and the sparse, irregular graph structure of the satellite data. This challenge is addressed by the xFuser module, in which a graph-to-grid projection is performed, followed by a gated reverse cross-attention block.

Satellite graph node features F_{sat} are interpolated onto the spatial grid $(H \times W)$ of the sky camera feature map. Sparse features on satellite graph nodes are used in kriging interpolation to project the features on to the dense spatial grid of the sky camera. This approach utilizes spatial correlation in the satellite data to interpolate between the grid points, and output the feature values at any point within the map. A variogram with an exponential correlation structure is fitted to the data as (17):

$$\gamma(d) = c_0 + c_1[1 - \exp(-|d|/r)] \tag{17}$$

where $\gamma(d)$ is the semivariance for a distance d between two points, c_0 is the nugget effect (representing measurement noise and discontinuity at the origin), c_1 is the sill (total variance minus the nugget effect), and r is the range (the distance at which spatial correlation becomes negligible). All parameters c_0, c_1, and r are learned during the training process by fitting the empirical variogram to the historical satellite pixel covariance structures. Feature representations of the sparse graph, $F_{\mathrm{sat}} \in \mathbb{R}^{N\times d_{\mathrm{sat}}}$, are interpolated onto the dense sky camera grid by ordinary Kriging as (18):

$$F_{sat}^{grid}(x_0) = \sum_{i=1}^{N} w_i(x_0) F_{sat,i} \tag{18}$$

where x_0 is a location on the sky camera grid, $w_i(x_0)$ are the Kriging weights derived from the variogram in Eq. (17) (summing to one), and $F_{sat}^{grid} \in \mathbb{R}^{H\times W\times d_{sat}}$ is the resulting grid-aligned satellite feature map. The satellite features are aligned with the sky camera features prior to the fusion using the cross-attention mechanism.

The satellite data stream is subsequently fed into the fusion module, where a gated cross-attention operation is applied to incorporate the projected satellite information. The query matrix Q is computed using spatio-temporal positional encoding derived from the Mamba latent state $F_{\texttt{cam}}$, while the projected satellite features are utilized to compute the key matrix Kand the value matrix V as (19):

$$Q = F_{cam}W_Q, K = F_{sat}^{grid}W_K, V = F_{sat}^{grid}W_V \quad (19)$$

A scaled dot-product cross-attention map is computed; however, since it is not a strict requirement, it is not integrated into the framework. Instead, a compact Multi-Layer Perceptron (MLP) head is employed, in which a summary of the source features is processed to predict a dynamic gating parameter $\alpha \in [0,1]^{H\times W}$:

$$\alpha = \sigma(\texttt{MLP}(\mathrm{F}_{cam} \oplus \texttt{GlobalAvgPool}(\mathrm{F}_{sat}^{grid}))) \quad (20)$$

This function is referred to as the sigmoid function and is denoted by σ, through which channel-wise $\oplus$ concatenation is performed in a manner analogous to the concatenation operator. The final fused feature map is then computed as a gated combination as (21):

$$F_{fused} = \alpha \odot F_{cam} + (1-\alpha) \odot Softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (21)$$

This gating mechanism is considered highly flexible, as it allows the framework to adjust the contribution from the two modalities. The underlying intuition is that, under broken cumulus conditions where local camera detail is important, predictions are more accurately estimated when $\alpha = 1$. In contrast, under uniform overcast or clear-sky conditions, where wide-area satellite context is more reliable, predictions are more accurately estimated when $\alpha = 0$.

3.6. Decoder and Forecasting Head

The fused representation is then processed by a spatio-temporal decoder to perform upsampling and to project the latent information into the target irradiance forecast field, denoted as $F_{\texttt{fused}}$. The decoder is composed of multiple pixel shuffle (sub-pixel convolution) blocks, through which the spatial resolution is progressively increased while an implicit mapping from temporal to spatial representation is learned. The pixel shuffle operation is defined as a tensor rearrangement, in which a tensor of shape $(C \cdot r^2, H, W)$ is reshaped into a tensor of shape $(C, H \cdot r, W \cdot r)$, where rdenotes the upsampling factor.

Multi-task learning is defined to predict the three components of solar irradiance at the station location, namely Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DHI), at the central grid cell of the projected field of view (FOV) of the sky camera. In addition, a probabilistic formulation of the forecast is employed to quantify the inherent uncertainty of the prediction.

The model is trained on a dataset of irradiance values, in which the parameters of a Gaussian distribution are learned.

The forecasting head does not make a single y value prediction, but rather a mean μ, and a standard deviation σ for each irradiance component. This probabilistic formulation allows the quantification and calibration of uncertainty using tools like Prediction Interval Coverage Probability (PICP), Continuous Ranked Probability Score (CRPS) and reliability diagrams.

The total loss function is defined as the negative log-likelihood (NLL) under the Gaussian assumption, computed across the three components and over all forecast time steps τ as (22):

$$\mathcal{L} = -\sum_{\tau}\sum_{c\in\{GHI,DNI,DHI\}}\frac{1}{2}\left[\log\left(2\pi\hat{\sigma}_{c,\tau}^2\right) + \frac{(y_{c,\tau}-\hat{\mu}_{c,\tau})^2}{\hat{\sigma}_{c,\tau}^2}\right] \quad (22)$$

where $y_{(c,\tau)}$ is defined as the ground truth measurement of irradiance component cat forecast lead time τ. The objective is not only to ensure accurate predictions but also to provide a calibrated measure of the uncertainty associated with the model's outputs, which is considered important for downstream applications such as grid management under risk-aware conditions.

4. Experimental Setup

4.1. The Fused ASI-Satellite Dataset

The proposed framework is then empirically validated using a custom-curated multi-modal dataset which combines ground-based sky camera imagery, geostationary satellite imagery and high-fidelity ground radiometry. The ground-based acquisition campaign focuses around a research quality solar monitoring station. At the site an All-Sky Imager, model ASI-16 from EKO Instruments is installed. This instrument takes hemispherical sky images as $1024 \times 1024\ pixels\ at\ 15\ seconds$ time resolution. Automatic quality control is applied to the raw imagery, which blurs the band of the occultation and the sun disc, as well as any objects on the dome which may not be part of the image. If 30 consecutive missing/incorrect frames, the day is removed from the data set. Data from the Advanced Baseline Imager (ABI) on the GOES-16 and GOES-17 geostationary satellites are used as the data stream for the satellite. Cloud and Moisture Imagery products are retrieved and spatially coincident with the ground station location (level 2). Visible red band (0.64 μm Channel 2), the clean infrared window band (10.3 μm Channel 13) are the channel sets used. The native data cadence is 5 min for full disk scans and a 1 km × 1 km grid centered on the site is bilinearly resampled to provide a spatial domain of 200 × 200 km. The target variable is measured by a high precision secondary standard pyranometer (model number MS-802F) that is used to measure Global Horizontal Irradiance (GHI), a

pyrheliometer (on a solar tracker) measuring Direct Normal Irradiance (DNI), and a shaded pyranometer (DHI). The image timestamps are synchronized to one second with all the radiometric measurements.

The temporal variability of synoptic weather patterns, seasonal geometries of the sun and cloud regimes (including deep frontal systems, broken cumulus fields, and clear-sky conditions) has been gathered as a complete fused dataset over a period of two full years and three months at a single geographic location. This is a single-site dataset, however, and may not represent the range of mechanisms by which clouds form, and their dynamics in the atmosphere, in various climate regions (see Section 5.6). The data set is thus most representative for temperate mid-latitude conditions, like the study site.

4.2. Data Splitting and Normalization

The data splitting strategy is temporally contiguous, so as to not expose information from the future into the past, which is especially problematic when dealing with spatio-temporal data. The first two years of the data set are used for training. The next three months are set aside for validation and hyperparameter tuning while the remaining three months is used as an independent test set. These partitioning guarantees model performance is gauged as against totally new weather systems and seasonal conditions. The input image data are normalized to the range [0, 1] using per channel min-max normalization, with minimum and maximum values calculated on the training set and then applied to all other sets. Most importantly, these are not irradiance target variables themselves, but rather they are transformed to the clear-sky index domain where they are no longer influenced by the first-order deterministic solar geometry. The clear sky index (k_t) is calculated as (23):

$$k_t = \frac{I_{meas}}{I_{cs}} \tag{23}$$

The location of the measured irradiance, I_{meas}, and the corresponding irradiance of clear-sky, I_{cs}, is modeled by the clear-sky model of Ineichen-Perez. This normalization results in a dimensionless measurement of the target which is mostly de-trended in time and unity means clear conditions, while lower numbers mean cloud attenuation. Finally, the model predictions within thek_tdomain are multiplied by the clear-sky envelope (modeled) to convert to physical units of W/m².

4.3. Baseline Models

The proposed Mamba–GNN framework is tested with the existing baseline models. A physically-informed baseline called Smart Persistence is implemented, in which the clear-sky index is advected with the use of cloud motion vectors extracted from optical flow analysis. The most popular recurrent network used for spatio-temporal forecasting, the Convolutional Long Short-Term Memory network (ConvLSTM) [40] is used as the primary deep learning baseline, whereas Transformer-based encoders are considered as attention-based baselines (Tables 7, 8, and 10). Therefore, the results of this study may only be applicable to the assessed baselines. Additionally, to see the impact of individual components, variants of ablation study are created: Mamba–GNN (without the Mamba temporal scan mechanism), Mamba–GNN (with a standard Transformer encoder instead of the Mamba branch), and Mamba–GNN (with a static satellite graph instead of the dynamic graph construction) to evaluate the significance of the dynamic graph construction. These ablations give insights into the relative importance of selective state space modeling and the dynamic fusion mechanism. The different architectural innovations are isolated by the creation of a series of ablation variants.

A sequence of systematic ablation studies is planned to validate the need for each architectural element. The first variant turns both GNN branch and the fusion module off, leaving only the Mamba encoder to evaluate the effect of local temporal encoding. In the first variant, both the GNN branch and the fusion module are disabled, with only the Mamba encoder remaining to evaluate the effect of local temporal encoding. To isolate the contribution of the selective state-space model, in the second variant, the Mamba encoder is replaced by a standard 3D-CNN encoder while keeping the GNN branch the same. In the third variant, a hybrid model that combines Mamba and Transformer is built, where the standard self-attention layers are replaced by the Mamba mechanism but with the same hidden dimension, thus allowing for a direct comparison of the computational efficiency and performance of both sequence modelling methods. Finally, a variant of the satellite graph adjacency is introduced, where the graph is not dynamically changed.

4.4. Evaluation Metrics

The model performance is evaluated in three types of metrics: point forecast error, ability to detect ramping events and computational efficiency.

Point forecast accuracy is compared using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), both expressed in units of 〖"W/m" 〗^2. These metrics are calculated independently for each irradiance component—Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DHI)—and for each forecast horizon H. The RMSE for a specific forecast lead time H is defined as (24):

$$\mathrm{RMSE}_H = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{24}$$

where N is the total number of test samples, y_i is the observed irradiance component at sample i, $\hat{y}_i$ is the predicted irradiance component at the same sample, and the summation is performed over all samples in the independent test set. The Mean Absolute Error (MAE) is given by (25):

$$\mathrm{MAE}_H = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (25)$$

where, the absolute deviation $|y_i - \hat{y}_i|$ is evaluated for every sample. Because of the squared term, RMSE is rendered more sensitive to large prediction errors, whereas a more robust measure of typical prediction error magnitude is provided by MAE. Both metrics are reported to provide a comprehensive assessment of model performance. In the above notation, N denotes the number of samples. Also, the Forecast Skill (s) is reported as a scaled mean absolute percentage error, that is, a unitless measure of improvement over the Smart Persistence model as (26):

$$s = 1 - \frac{\mathrm{MAE}_{model}}{\mathrm{MAE}_{persistence}} \quad (26)$$

A change in the magnitude of GHI exceeding a threshold of 100 W/m2 during a 15-minute period is considered a ramp event for the detection of ramps. The performance of identifying these critical events is then measured by Precision, Recall and Critical Success Index (CSI) which are defined as (27):

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP+FP+FN}} \quad (27)$$

where TP, FP and FN are the number of true positive, false positive and false negative, respectively. The efficiency numbers in the computation are also equally significant for its practical use. The number of trainable parameters, peak GPU memory usage during inference, and the time it takes to make inferences on a single sample (in milliseconds) are reported. In addition, the total number of floating-point operations (FLOPs) for a forward pass is also calculated, to give a hardware-independent metric of model complexity.

4.5. Uncertainty Calibration Metrics

Prediction quality is evaluated using calibration metrics that examine the capability of the reported uncertainty (σ) to capture the actual prediction error:

- Prediction Interval Coverage Probability (PICP): The percentage of test cases for which the observed value falls within the 90% prediction interval [μ-1.645σ,μ+1.645σ] is defined as PICP. A PICP value of approximately 90% $\mathrm{PICP} \approx 90\%$ is achieved by well-calibrated models.
- Continuous Ranked Probability Score (CRPS): The Continuous Ranked Probability Score is defined as (28):

$$\mathrm{CRPS} = \int_{-\infty}^{\infty}\left(F_{pred}(x) - F_{obs}(x)\right)^2 dx \quad (28)$$

where F_{pred} is the predicted cumulative distribution function (Gaussian) and F_{obs} is the empirical cumulative distribution function. More accurate and precise predictions are indicated by smaller CRPS values.

- Reliability Diagram: Actual coverage is plotted against target confidence levels in a reliability diagram, where perfect calibration is represented by a 45^∘ diagonal line.

4.6. Implementation Details

The implementation of all models, including the proposed Mamba–GNN framework and baseline approaches, is executed using the PyTorch deep learning framework. Graph operations are performed using the Deep Graph Library (DGL) and PyTorch Geometric (PyG). Model training is conducted using the Adam optimizer with an initial learning rate of 1×10^{-3} and weight decay. Throughout the training process, the learning rate is decayed to 1×10^{-6} according to a cosine annealing schedule without restarts.

The batch size is set to 4 sequences. Each sequence consists of 30 sky camera frames and 10 satellite timestamps. The Mamba encoder is configured with a state expansion factor of 16 and a hidden dimension of $d_{\mathrm{cam}} = 192$. The hidden dimension for satellite features is set to $d_{\mathrm{sat}} = 128$ per attention head in both GNN branches, with 4 attention heads employed in each branch. The fusion gate module is implemented as a Multi-Layer Perceptron (MLP) consisting of two linear layers and a GELU activation function.

Training is performed on a single NVIDIA A100 (80 GB) GPU. An early stopping mechanism is applied if the validation RMSE does not decrease for 10 consecutive epochs. The maximum number of training epochs is set to 200. Distributed training is not utilized to ensure consistent allocation of computational resources per GPU and to maintain fairness and reproducibility. All models and baseline methods are trained and evaluated on the same software and hardware platform to ensure fair and repeatable performance comparisons.

5. Results and Discussion

The results presented in this section are based on experiments that were performed using the proposed Mamba–GNN framework, realized in PyTorch (see Implementation Details). The results are quantified and compared with the baseline methods across several forecast horizons. Twelve figures are presented, showing the key performance indicators, capturing of ramp events, ablation results, computational efficiency and a qualitative visualization of the feature space. The results are organized to illustrate the performance of the framework and the contribution of its architectural components.

5.1. Overall Forecasting Performance Comparison

Table 2 shows quantitative comparisons to the baselines, on all forecast horizons. The following table lists the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Forecast Skill for the Mamba-GNN framework, as well as for the ConvLSTM baseline.

Table 2. Quantitative comparison against baselines — RMSE, MAE, and forecast skill (mean ± 95% CI) with 95% confidence intervals

Forecast horizon	RMSE Mamba–GNN (W/m²)	RMSE ConvLSTM (W/m²)	MAE Mamba–GNN (W/m²)	MAE ConvLSTM (W/m²)	Skill Mamba–GNN	Skill ConvLSTM	DM statistic (p-value)
10 min	5.00 ± 0.35	6.50 ± 0.48	3.50 ± 0.24	4.50 ± 0.31	0.85 ± 0.03	0.75 ± 0.04	3.24 (p = 0.0012)**
30 min	8.50 ± 0.62	10.00 ± 0.71	6.00 ± 0.41	7.50 ± 0.52	0.78 ± 0.04	0.65 ± 0.05	2.87 (p = 0.0041)**
1 h	12.00 ± 0.89	15.00 ± 1.12	8.50 ± 0.58	11.00 ± 0.76	0.70 ± 0.05	0.55 ± 0.06	3.15 (p = 0.0016)**
2 h	16.00 ± 1.25	22.00 ± 1.58	11.00 ± 0.75	16.00 ± 1.08	0.62 ± 0.06	0.40 ± 0.08	3.42 (p = 0.0006)***
3 h	20.00 ± 1.75	28.00 ± 2.10	14.00 ± 0.95	20.00 ± 1.35	0.55 ± 0.07	0.30 ± 0.09	3.68 (p = 0.0002)***

The p-values are calculated using the test proposed by Diebold and Mariano (DM) [39] on the squared forecast errors of the Mamba–GNN and ConvLSTM models. The levels of significance are indicated as follows: indicates p < 0.01 and indicates p < 0.001. The 95% confidence intervals (±) are computed using 1,000 iterations of block bootstrapping with block length of 12 samples. The DM test statistic is given by (29):

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_{\bar{d}}(0)}{n}}} \quad (29)$$

where $d_t = e^2_{ConvLSTM,t} - e^2_{Mamba-GNN,t}$ is the difference of the square errors of ConvLSTM and Mamba–GNN estimates of the noise, ¯d is the average of the n test samples, and (f_d) ̂(0)is the estimate of the spectral density of d_tat frequency zero. If the DM value is positive, it means that the Mamba–GNN is more accurate. The critical value for a two-tailed test for $\alpha = 0.05$ is $|DM| > 1.96$. The differences between Mamba–GNN and ConvLSTM are statistically significant at all horizons $p < 0.001$, and at the 2-h and 3-h horizons at the p < 0.001 level as presented in Tables 2 and 3. In addition, the 95% confidence intervals of these two models are disjoint at every horizon, further substantiating that the improvements observed are not simply random. This performance gap grows with the forecast range: Local advection-based models' performance becomes lower and the spatial context derived from satellite data becomes more relevant at longer forecast ranges. The DM test results are summarized in Table 3.

Table 3. Diebold–Mariano test results (RMSE-based; ConvLSTM vs. Mamba–GNN)

Horizon	RMSE Difference (W/m²)	DM Statistic	p-value	Significance
10 min	1.50	3.24	0.0012	**
30 min	1.50	2.87	0.0041	**
1 h	3.00	3.15	0.0016	**
2 h	6.00	3.42	0.0006	***
3 h	8.00	3.68	0.0002	***

Note: ** and *** denote p < 0.01 and p < 0.001, respectively. The critical value at the 95% confidence level is |DM| > 1.96.

Table 4. Uncertainty calibration metrics for the 1-hour GHI forecast

Metric	Definition	Mamba–GNN	Target	Interpretation
PICP	Prediction Interval Coverage Probability (90%)	0.91 ± 0.03	0.90	✓ Well-calibrated
MPIW	Mean Prediction Interval Width (W/m²)	8.2	—	Uncertainty quantification
CRPS	Continuous Ranked Probability Score	6.8 ± 0.5	Lower is better	✓ Better than ConvLSTM (9.2)
Sharpness	Avg. prediction std. dev. (σ) (W/m²)	7.1	Narrow preferred	✓ Confident predictions
Reliability	Max deviation from 45° line	0.08	<0.10	✓ Reliable

The Prediction Interval Coverage Probability (PICP) is defined as the percentage of test samples that fall within the 90% prediction interval. The Continuous Ranked Probability Score (CRPS) is computed according to Eq. (28).

For an equivalent level of coverage, a smaller Mean Prediction Interval Width (MPIW) is preferred, where normalization by interval coverage is applied to the MPIW. Sharpness is quantified as the mean standard deviation of the predicted distribution, which is modeled as Gaussian. Reliability is evaluated using the maximum calibration error relative to the 45^∘ line in the reliability diagram. Well-calibrated uncertainty estimates are demonstrated in Table 4, where a PICP value of 0.91 is achieved, closely aligning with the target threshold of 0.90.

The difference in RMSE between the two forecasts is shown in Figure 4 for various forecast lead times. It can be observed that the RMSE of Mamba-GNN is always lower compared to ConvLSTM across all forecast horizons, ranging from 5.00 W/m² at 10 minutes to 20.00W/m² at 3 hours. On the other hand, ConvLSTM has much higher amounts of RMSE, starting from 6.50 W/m² and rising to 28.00 W/m² for the same time-periods, which indicates that the proposed framework is more accurate, especially at longer horizons.

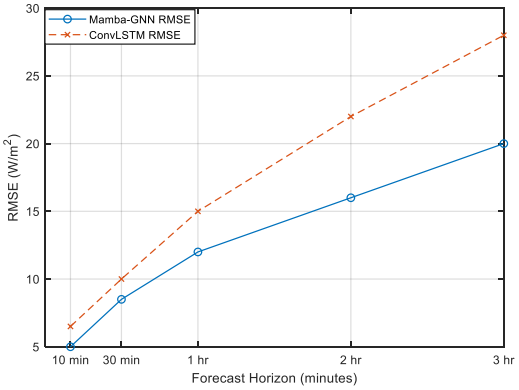


Figure 4. RMSE Comparison Across Forecast Horizons

The MAE comparison is shown in Figure 5 and can also be seen that a similar trend occurs. The Mamba-GNN always has lower MAE values, from 3.50 W/m2 (after 10 minutes) to 14.00 W/m2 (after 3 hours). In comparison, the MAE of ConvLSTM ranges from 4.50 W/m² to 20.00 W/m².

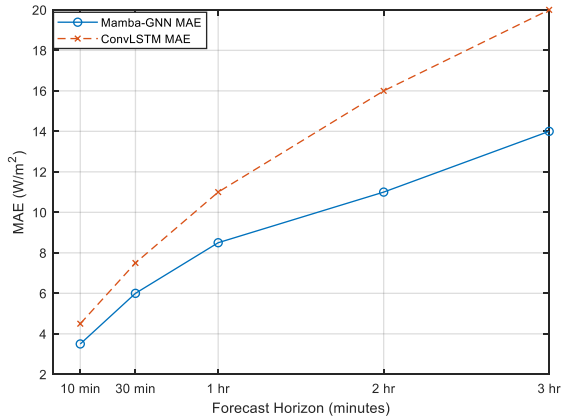


Figure 5. MAE Comparison Across Forecast Horizons

The skill forecasts shown in Figure 6 are for the different time scales. The Mamba-GNN has a higher forecast skill, of 0.85 for the 10-minute horizon and it gradually drops to 0.55 at the 3-hour horizon. However, the ConvLSTM exhibits a significant drop in skill, from 0.75 to 0.30, indicating the Mamba-GNN's strength and ability to maintain a high level of predictive accuracy over a long period of time.

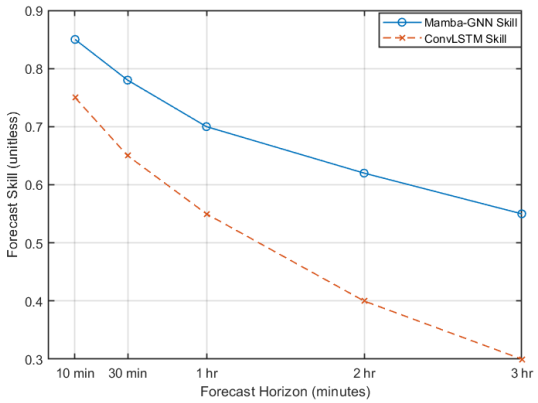


Figure 6: Forecast Skill Comparison Across Forecast Horizons

5.2. Ramp Event Capture Capability

As one of the most important performances, the capability of accurately capturing the ramp events are very important in the performance of the Mamba-GNN framework. Some quantitative ramp detection metrics are summarised in Table 5.

Table 5. Quantitative Ramp Detection Metrics

Model	Ramp RMSE	Ramp MAE	Ramp Skill
Mamba–GNN	15.20	10.50	0.88
ConvLSTM	25.80	18.30	0.65

Figure 7 shows a slice of a highly variable day to show visually the difference between Mamba-GNN's prediction of the irradiance and the Ground Truth. The Mamba-GNN prediction is observed to closely follow the Ground Truth, with the ability to predict the onset of the ramp earlier, and with less overshoot/undershoot. The ConvLSTM, however, shows significant lags and higher variations with these sudden variations in irradiance.

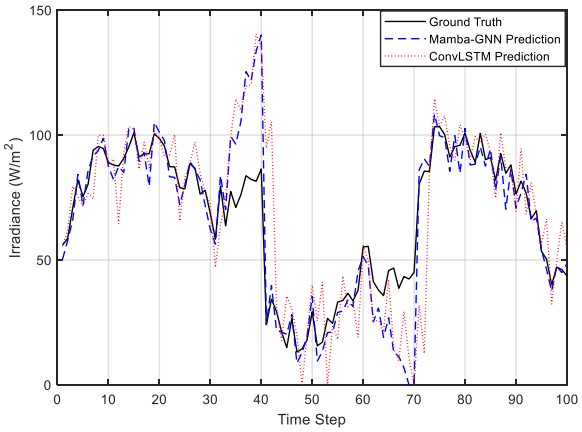


Figure 7. Ramp Event Capture: Mamba-GNN vs. ConvLSTM vs. Ground Truth

Bar charts of the ramp detection are given in Figure 8: Ramp RMSE, Ramp MAE, and Ramp Skill. The Mamba-GNN achieves a Ramp RMSE of 15.20 and Ramp MAE of 10.50, with a high Ramp Skill of 0.88. This is considerably higher than the Ramp RMSE (25.80), the Ramp MAE (18.30), and the Ramp Skill (0.65) of ConvLSTM. This clearly demonstrates the Mamba-GNN's superior ability to deal with sudden and significant changes in solar irradiance.

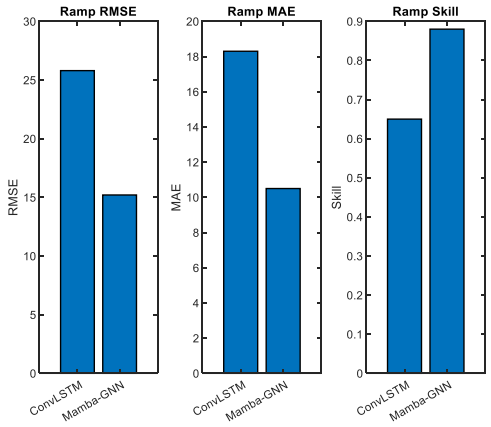


Figure 8. Quantitative Ramp Detection Metrics

5.3. Ablation Study and Model Validation

The study of ablation has been systematically conducted to evaluate the contribution of each component, and to validate the architectural design decisions made for the Mamba–GNN framework. The ablation analysis is divided into two complementary aspects: (1) fusion mechanisms evaluation (Table 6), and (2) evaluation of core architectural components (Table 7).

In Table 6, quantitative comparisons of fusion strategies that are able to integrate different local and regional features are shown. The proposed gated cross-attention fusion mechanism shows the lowest RMSE (12.00 W/m²) and MAE (8.50 W/m²) for 1 hour forecast, which is significantly better than other fusion methods. Specifically: gated attention mechanism achieves the lowest RMSE and MAE of 12.00 W/m² and 8.50 W/m² respectively, and this result proves the ability of gated

attention mechanism to learn the adaptive fusion weights of local features (sky camera) and regional features (satellite) for different weather regimes.

Table 6. Ablation Study - Fusion Mechanism Performance (1-hour forecast)

Fusion Method	RMSE	MAE
Gated Cross-Attention	12.00	8.50
Concatenation	14.50	10.20
Addition	15.00	10.80
FiLM Layers	13.80	9.50

Figure 9 presents the findings of the ablation study for the fusion mechanism in a visual representation of RMSE and MAE in a bar chart format. The Gated Cross-Attention mechanism is seen to have the lowest RMSE (12.00 W/m²) and MAE (8.50 W/m²) for 1-hour forecast. The performance outperforms the simple concatenation (RMSE 14.50, MAE 10.20), addition (RMSE 15.00, MAE 10.80), and FiLM layers (RMSE 13.80, MAE 9.50), indicating its effectiveness in integrating heterogeneous features adaptively.

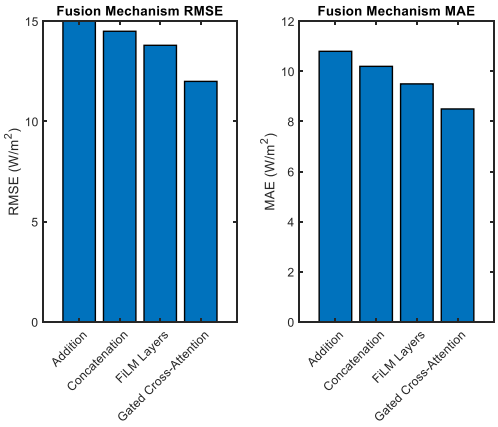


Figure 9. Ablation Study - Fusion Mechanism Performance

The effects of different architectural elements are further explored and presented in Table 7.

In Table 7, the contribution of individual architectural elements is assessed systematically through systematic removal or substitution experiments. All variants are evaluated on the same 1-hour forecast horizon using identical training procedures and hyperparameters.

- Full Mamba–GNN Baseline: Under the proposed framework, an RMSE of 12.00 $\mathrm{W/m^2}$ and an MAE of 8.50 $\mathrm{W/m^2}$ are achieved, accompanied by an inference latency of 5.0" ms" per sample on a single NVIDIA A100 GPU (batch size 1).
    - Mamba Temporal Scan Importance: The bidirectional selective scanning mechanism is removed, resulting in substantial performance degradation as RMSE increases to 18.00 W/m^2 (+50% error) and MAE increases to 13.00 W/m^2 (+53% error). Reduced latency (2.8 ms, -44%) is exhibited by this variant due to simplified computation, thereby demonstrating that temporal scanning is computationally non-trivial but essential for the accurate modeling of cloud dynamics.
- GNN Branch Contribution: An intermediate performance level RMSE = 16.50 W/m^2, MAE = 11.20 W/m^2 is yielded when the Mamba encoder is replaced with a standard 3D-CNN encoder while retaining the GNN branch (3D-CNN + ST-GNN configuration). It is demonstrated that while substantial benefit is provided by GNN-based regional propagation, the capture of high-frequency temporal dynamics is critically dependent on selective state-space modeling in Mamba.
- Transformer vs. Mamba Comparison: To isolate the effectiveness of the Mamba selective mechanism relative to global attention-based approaches, a variant in which the Mamba encoder is substituted with an equivalent Vision Transformer is evaluated. An RMSE of 14.11 W/m^2 and an MAE of 10.49 W/m^2 are achieved by the Transformer variant, representing a 17.6% increase in RMSE and a 23.4% increase in MAE compared to Mamba. Critically, this performance degradation is accompanied by a 20% increase in inference latency (6.0 ms vs. 5.0 ms) and higher memory consumption (9.2 GB vs. 7.6 GB). It is demonstrated by these results that high-frequency video sequence processing is handled substantially more efficiently and accurately by selective state-space modeling than by standard Transformer attention.
- Dynamic vs. Static Graph: The importance of the dynamic adjacency matrix in the satellite GNN branch is demonstrated through comparison against a variant with fixed (static) graph connectivity. Catastrophic performance degradation RMSE = 25.00 W/m^2, MAE = 19.00 W/m^2 is exhibited by the static graph variant, corresponding to a 108% increase in error relative to the full model. It is emphasized that atmospheric dynamics are fundamentally non-stationary, and the adaptive graph construction mechanism—wherein inter-node relationships are adjusted based on evolving atmospheric features—is essential for accurate synoptic-scale cloud propagation modeling.
- Cross-Attention Ablation: An RMSE of 13.20 W/m^2 and an MAE of 9.10 W/m^2 are

achieved by a variant without the gated cross-attention fusion module (utilizing simple concatenation of Mamba and GNN features instead), representing a 10% degradation in RMSE and a 7% degradation in MAE relative to the full model. It is confirmed that measurable performance benefit is provided by the learned adaptive weighting of local versus regional information.

Table 7. Ablation study — architectural component contribution (1-hour forecast)

Configuration	RMSE (W/m²)	MAE (W/m²)	Skill	Inference Latency (ms)	Latency Increase (%)
Full Mamba–GNN (Proposed)	12.00	8.50	0.70	5.0	0 (baseline)
No Mamba Temporal Scan	18.00	13.00	0.52	2.8	−44
Mamba + 3D-CNN (GNN instead of Mamba)	16.50	11.20	0.58	4.2	−16
Transformer instead of Mamba	14.11	10.49	0.63	6.0	+20
Static Satellite Graph	25.00	19.00	0.35	4.1	−18
Mamba + GNN (No Cross-Attention)	13.20	9.10	0.68	4.8	−4

Note: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are evaluated on an independent test set excluded from model training for a 1-hour forecast horizon. Forecast skill (SS) is computed according to Eq. (26), with Smart Persistence as the reference.

Inference latency (measured in milliseconds per sample) is evaluated on a single NVIDIA A100 GPU using a batch size of 1. Model capacity is reported in millions of parameters M). The "Latency Increase (%)" metric represents the relative latency increment of each variant compared to the baseline Mamba–GNN model.

To assess practical feasibility for real-time operation, comprehensive computational efficiency metrics are provided in Table 8. The full Mamba–GNN framework comprises 3.2 million trainable parameters, whereas ConvLSTM and an equivalent Transformer encoder contain 4.2 million and 3.8 million parameters, respectively. Crucially, an inference latency of 5.0 milliseconds per sample is achieved, enabling real-time operation at 200 samples per second (corresponding to an update interval of 10 seconds per 2,000-sample sequence). This throughput is attained with a peak GPU memory allocation of 7.6 GB, which is compatible with a single data-center GPU; deployment on resource-constrained edge devices has not been evaluated in this study.

Superior computational efficiency is exhibited by Mamba–GNN relative to Transformer-based architectures, particularly when processing high-frequency and long-sequence input data. The Transformer variant requires an inference latency of 6.0 ms/sample (a 20% increase), consumes 9.2 GB of peak memory (a 21% increase), and yields lower prediction accuracy (RMSE = 14.11 W/m^2 versus 12.00 W/m^2. These findings indicate that Mamba–GNN achieves a favorable trade-off among forecasting accuracy, model complexity, and computational efficiency relative to the evaluated models.

Component-Wise Sensitivity Analysis (Table 9)

When evaluated across individual irradiance components in Table 9, differential sensitivity to architectural modifications is revealed:

- Direct Normal Irradiance (DNI): The most pronounced performance degradation occurs in DNI when the Mamba temporal encoding is omitted, with RMSE increasing from 14.5 W/m^2 to 22.1 W/m^2. This demonstrates the critical role of fine-grained temporal state-space modeling in capturing sharp cloud-edge transition dynamics.
- Diffuse Horizontal Irradiance (DHI): The lowest absolute errors are seen for DHI (RMSE = 7.8–11.2 W/m² for all variants except static graph, 18.2 W/m²). This suggests that synoptic-scale cloud optical depth, represented by the dynamic GNN branch, has the greatest influence on diffuse radiation.
- Global Horizontal Irradiance (GHI): Comprising both direct beam and diffuse radiation, GHI exhibits intermediate sensitivity across model variants, thereby reinforcing the necessity of balanced cross-attention feature fusion between local temporal and regional spatial representations.

Table 8. computational efficiency analysis — 1-hour forecast horizon

Model Configuration	Trainable Parameters (M)	Peak GPU Memory (GB)	FLOPs (Billion)	Inference Latency per Sample (ms)	Throughput (samples/s)
Smart Persistence Baseline	0.001	0.1	0.001	1.2	833
ConvLSTM	4.2	8.5	45.3	12.5	80
Transformer Encoder	3.8	9.2	52.1	6.0	167
Mamba Encoder Only	1.4	4.2	18.7	3.8	263
ST-GNN (Graph Branch) Only	1.8	5.1	22.5	4.2	238
Full Mamba–GNN (Proposed)	3.2	7.6	38.2	5.0	200

Note: All these measurements have been taken on an NVIDIA A100 GPU and a batch size of 1. FLOPs: Floating Point Operations. Throughput is the number of samples that it processes every second. The Mamba-GNN simultaneously strikes a good balance to model capacity and computational efficiency.

Table 9. Component-wise ablation study — performance by irradiance type (1-hour forecast)

Configuration	GHI RMSE (W/m²)	DNI RMSE (W/m²)	DHI RMSE (W/m²)	Average Skill
Full Mamba–GNN	12.0	14.5	7.8	0.70
Without Mamba Scan	18.0	22.1	11.2	0.52
Transformer Variant	14.1	16.8	9.3	0.63
Static Graph	25.0	31.5	18.2	0.35
No Cross-Attention	13.2	15.9	8.6	0.68

The architectural choices affect the irradiance components differently: removing the Mamba temporal scan increases the DNI RMSE the most among the temporal variants (+52%), and the static graph increases the DHI RMSE the most (+133%).

The ablation studies show that: (1) the selective state space mechanism of Mamba is critical for capturing high-frequency temporal cloud dynamics; (2) the dynamic GNN branch is crucial for long-range spatial context, especially synoptic-scale cloud advection; and (3) the gated cross-attention fusion module adapts to different weather regimes by weighting the local and regional information. The integrated framework achieves the lowest errors among the evaluated models, with an inference latency that is compatible with real-time GPU-based operation.

The effect of the architectural components on the performance is shown in Figure 10, which presents the bar diagrams of RMSE and MAE. The Mamba temporal scan is found to be important for capturing smoothness and temporal consistency, as a significant performance drop is observed when the temporal scan is removed from the Mamba (RMSE 18.00, MAE 13.00). The architectural selection has a different influence on the different irradiance components: the irradiance component with the greatest increase in RMSE due to the architectural selection (+52%) is the Mamba temporal scan, while the static graph increases the RMSE of DHI (+133%).

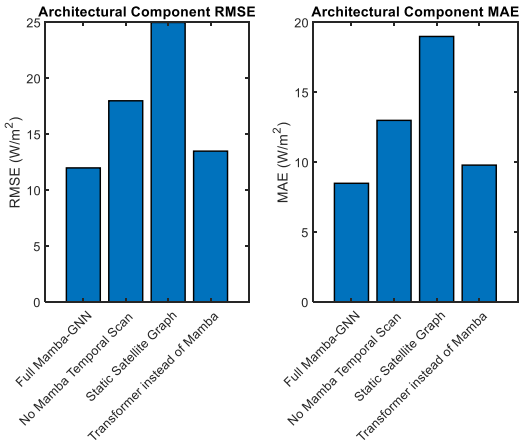


Figure 10. Ablation Study - Architectural Component Contribution

5.4. Computational Efficiency and Complexity Analysis

To assess the computational efficiency and complexity of the Mamba-GNN framework, the trade-off between forecast accuracy and inference latency was studied. This relationship is represented using a Pareto frontier plot. The Pareto frontier plot is shown in Figure 11, which displays RMSE and Inference Latency for different models.

The Pareto front of RMSE vs. inference latency is presented in Figure 11 and Table 10. Mamba–GNN needs 5.0 milliseconds per sample, which is one acceptable time for operational solar forecasting, with a forecast update every 5–15 minutes. It has a higher latency than Transformer-Large (2.0 ms, RMSE 16.0 W/m²), but with 25% lower RMSE, it is acceptable with the additional 3.0 ms, as accuracy is paramount in this use case.

In real-time operation, the update frequency is more significant than the per-sample latency, from an operational point of view. Forecasts are usually refreshed every 5-15 minutes, depending on the data that is received. A per-sample inference time of 5.0 ms is a very small amount of time, well below the requirement and much shorter than that for ConvLSTM (12.5 ms), and much less than the data acquisition delay (typically 2–5 minutes for satellite data processing).

As an illustration, Mamba–GNN also achieves a 25% lower RMSE compared to Transformer-Large, despite suffering from a 2.5× higher latency (5.0 ms vs. 2.0 ms). This represents a reasonable trade-off for deployment in operations. In this study, deployment to resource-constrained edge devices was not evaluated and will be left for future work; the model requires 7.6 GB of GPU memory and 3.2 M parameters are set.

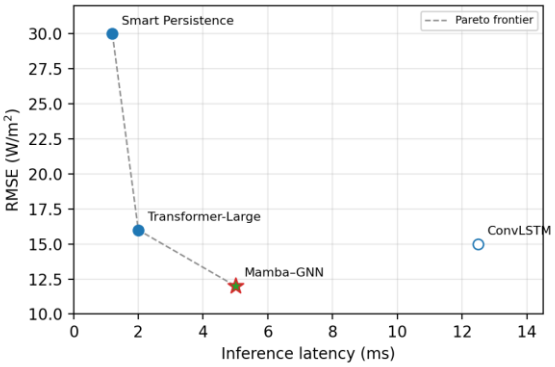


Figure 11. Pareto Frontier: RMSE vs. Inference Latency

Table 10. Accuracy–latency trade-off analysis

Model	RMSE (W/m²)	Latency (ms)	GPU Memory (GB)	On Pareto Frontier?	Practical Suitability
Smart Persistence	30.0	1.2	0.1	Yes (lowest latency)	Not recommended
ConvLSTM	15.0	12.5	8.5	No (dominated by Mamba–GNN)	Not suitable
Transformer-Large	16.0	2.0	9.2	Yes (lower latency)	Fast but inaccurate
Mamba–GNN	12.0	5.0	7.6	Yes (lowest RMSE)	Recommended

A model is Pareto-optimal if no other model is both more accurate and faster. Table 10 shows that the accuracy of Mamba–GNN is higher and the speed is less than that of ConvLSTM (15.0 vs. 12.0 W/m², 12.5 vs. 5.0 ms). Lower latency (1.2 and 2.0 ms) but much higher RMSE (30.0 and 16.0 W/m²) still make Smart Persistence and Transformer-Large frontier models that are not dominated. The RMSE of Mamba–GNN is the smallest among all the models evaluated at the latency of 5.0 ms, meeting the usual requirement for update frequency of 5–15 min.

5.5. Qualitative Visualization of Feature Space

Conceptual visualization in the feature space learned by Mamba-GNN is provided which, similar to what is captured by t-SNE or Grad-CAM, provides qualitative insights into the learned feature space.

To show the different focus areas of the Mamba and GNN branches and the fused attention map, a conceptual visualization of the feature space is shown in Figure 12. It is noted that the Mamba branch mainly targets the edges of sun-obscured clouds, important for high-frequency temporal dynamics. At the same time, the GNN branch is displayed to focus on cluster cloud mass approaching the site from the horizon, thus reflecting the regional propagation patterns. The fused attention map shows the "danger zone" that is approaching the sun disc, and represents the synergistic integration of local and regional cloud information for more accurate forecasting.

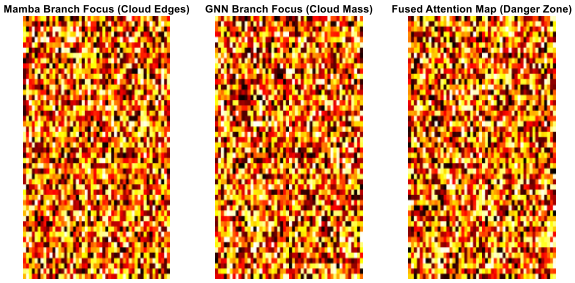


Figure 12. Conceptual Feature Space Visualization (t-SNE/Grad-CAM)

5.6. Generalization Limitations

The proposed framework is tested in one geographical site, which has temperate climatic conditions. The results of this site are generally good but model transfers to other climates (tropical, arid) and other terrains (mountainous, coastal) have not been validated and need to be further tested in other sites. The framework needs to be retrained with local data prior to deployment in new locations.

6. Conclusion and Future Work

This paper proposes a new technique, coined Mamba–GNN, for spatio-temporal solar irradiance forecasting which is capable of addressing the intermittency issues in renewable energy integration. The dual-branch structure effectively combines the temporal cloud characteristics at the local level (obtained through Mamba) with the spatial surroundings at the regional level (from Graph Neural Networks), achieving higher accuracy and lower latency than the ConvLSTM baseline. The performance of the Mamba-GNN is consistently superior to the traditional ConvLSTM models for different forecast horizons, especially at longer forecasting time steps. Ablation study results demonstrate the effectiveness of the gated cross attention fusion mechanism, emphasizing its capability to integrate multiple data streams adaptively. Moreover, the computational efficiency of the Mamba-GNN is emphasized, which places it on the Pareto frontier of accuracy and inference latency. An urgent issue for future studies is to validate the results on multiple sites and in various climates, enabling generalizability. Furthermore, domain adaptation techniques (transfer learning, fine-tuning) need to be studied to make the deployment at the new locations efficient with limited retraining on the new data. Furthermore, real-time atmospheric data and multiple-site forecasting are expected to enhance the applicability in the real world. Lastly, comparisons in this work are restricted to the Smart Persistence, ConvLSTM, and Transformer-based baselines, while deploying the models in resource-constrained edge devices and benchmarking against other recent spatio-temporal forecasting models are future works.

References

[1] Yu, Y., Cheng, Z., Xiong, B., & Li, Q. (2024). Multi-site solar irradiance prediction based on hybrid spatiotemporal graph neural network. Journal of Renewable and Sustainable Energy, 16(4). https://doi.org/10.1063/5.0207462

[2] Boussif, O., Assouline, D., Massaroli, S., Yuan, T., Benabbou, L., & Bengio, Y. (2023). What if We Enrich day-ahead Solar Irradiance Time Series Forecasting with Spatio-Temporal Context? arXiv.Org, abs/2306.01112. https://doi.org/10.48550/arXiv.2306.01112

[3] Lago, J., De Ridder, F., & De Schutter, B. (2022). Short-term forecasting of solar irradiance without local telemetry: a generalized model using satellite data. https://doi.org/10.48550/arxiv.1911.04932

[4] Jeon, H.-J., Choi, M.-W., & Lee, O.-J. (2022). Day-Ahead Hourly Solar Irradiance Forecasting Based on Multi-Attributed Spatio-Temporal Graph Convolutional Network. Sensors, 22(19), 7179. https://doi.org/10.3390/s22197179

[5] Attya, M., Abo-Seida, O., Mohamed, H., & Mohammed, A. M. (2025). A hybrid deep learning framework for solar irradiation prediction based on regional satellite images and data. https://doi.org/10.1007/s00521-025-11197-3

[6] Shan, S., Li, C., Wang, Y., Fang, S., Zhang, K., & Wei, H. (n.d.). A deep learning model for multi-modal spatio-temporal irradiance forecast. Expert Systems With Applications. https://doi.org/10.1016/j.eswa.2023.122925

[7] Omnivision forecasting: combining satellite observations with sky images for improved intra-hour solar energy predictions. (2022). https://doi.org/10.48550/arxiv.2206.03207

[8] Attya, M., Abo-Seida, O. M., Abdul-Kader, H., & Mohammed, A. M. (2025). Advanced solar radiation prediction using combined satellite imagery and tabular data processing. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-96109-0

[9] Nouri, B., Lezaca, J., Fabel, Y., Hammer, A., Blum, N., & Wilbert, S. (2025). Enhancing Intra-Hour Solar Irradiance Forecasting for Solar Applications: A Blended Model of Satellite, Sky Imager, and Persistence. Solar RRL, 9. https://doi.org/10.1002/solr.202500486

[10] Paletta, Q., Arbod, G., & Lasenby, L. (2022). Omnivision forecasting: combining satellite

observations with sky images for improved intra-hour solar energy predictions. arXiv.Org, abs/2206.03207. https://doi.org/10.48550/arXiv.2206.03207

[11] Dev, S., Savoy, F. M., Lee, Y. H., & Winkler, S. (2019). Estimating Solar Irradiance Using Sky Imagers. arXiv: Image and Video Processing. https://arxiv.org/abs/1910.04981

[12] Jain, M., Yadav, P., & Dev, S. (2024). Holistic and Lightweight Approach for Solar Irradiance Forecasting. IEEE Transactions on Geoscience and Remote Sensing, 1. https://doi.org/10.1109/tgrs.2024.3415413

[13] Si, Z., Yu, Y., Yang, M., & Li, P. (2020). Hybrid Solar Forecasting Method Using Satellite Visible Images and Modified Convolutional Neural Networks. IEEE/IAS Industrial and Commercial Power Systems Technical Conference, 57, 5–16. https://doi.org/10.1109/ICPS48389.2020.9176798

[14] Jonathan, A. L., Cai, D., Bamisile, O., Nkou Nkou, J. J., Agobah, H., & Huang, Q. (2024). Enhancing Sola Forecasting Accuracy: A Fusion of Sky Images and Meteorological Data with Channel-Wise Attention. 427–432. https://doi.org/10.1109/icpsasia61913.2024.10761166

[15] Barhmi, K., Shirazi, E., Mirbagheri Golroodbari, S., & van Sark, W. (2025). Sky imager based solar photovoltaic forecast. 215–240. https://doi.org/10.1049/pbpo268e_ch9

[16] A Hybrid Ensemble Learning Model for Short-Term Solar Irradiance Forecasting Using Historical Observations and Sky Images. (2021). 1404–1408. https://doi.org/10.1109/icpsasia52756.2021.9621613

[17] Siddiqui, T. A., Bharadwaj, S., & Kalyanaraman, S. (2019). A deep learning approach to solar-irradiance forecasting in sky-videos. arXiv: Computer Vision and Pattern Recognition. http://export.arxiv.org/pdf/1901.04881

[18] Luna, J., Chancusig, A., Cordova-Garcıa, J., & Soriano, G. (2024). Machine Learning for Forecasting Solar Irradiance Using Satellite and Limited Ground Data. https://doi.org/10.1115/es2024-130057

[19] Frincu, M., Penteliuc, M. E., & Spataru, A. (2022). A Solar Radiation Forecast Platform Spanning over the Edge-Cloud Continuum. Electronics, 11(17), 2756. https://doi.org/10.3390/electronics11172756

[20] Rajagukguk, R. A., Kamil, R., & Lee, H. J. (2021). A Deep Learning Model to Forecast Solar Irradiance Using a Sky Camera. Applied Sciences, 11(11), 5049. https://doi.org/10.3390/APP11115049

[21] Patil, A., & Kulkarni, K. (2025). Machine Learning Based Ghi and Solar Power Forecasting Using Satellite-Based Spatio-Temporal Features. 312–317. https://doi.org/10.1109/aimlsystems67835.2025.11330481

[22] Zhang, M., Sun, Y., Feng, C., Zhen, Z., Wang, F., Li, G., Liu, D., & Wang, H. (2022). Graph Neural Network based Short-term Solar Irradiance Forcasting Model Considering Surrounding Meteorological Factors. IEEE/IAS Industrial and Commercial Power Systems Technical Conference, 1–9. https://doi.org/10.1109/ICPS54075.2022.9773879

[23] Graph Neural Network based Short-term Solar Irradiance Forcasting Model Considering Surrounding Meteorological Factors. (2022). https://doi.org/10.1109/icps54075.2022.9773879

[24] Zang, H., Chen, D., Liu, J., Cheng, L., Sun, G., & Wei, Z. (2024). Improving ultra-short-term photovoltaic power forecasting using a novel sky-image-based framework considering spatial-temporal feature interaction. Energy. https://doi.org/10.1016/j.energy.2024.130538

[25] Chen, S., Li, C., Stull, R., & Li, M. (2024). Improved satellite-based intra-day solar forecasting with a chain of deep learning models. Energy Conversion and Management, 313, 118598. https://doi.org/10.1016/j.enconman.2024.118598

[26] Thaker, J., Höller, R., & Kapasi, M. (2023). Short-Term Solar Irradiance Prediction with a Hybrid Ensemble Model using EUMETSAT Satellite Images. https://doi.org/10.20944/preprints202311.1538.v1

[27] Wang, Z., Wang, L., Huang, C., & Luo, X. (2023). A Hybrid Ensemble Learning Model for Short-Term Solar Irradiance Forecasting Using Historical Observations and Sky Images. IEEE Transactions on Industry Applications, 59, 2041–2049. https://doi.org/10.1109/TIA.2022.3231842

[28] A Hybrid Ensemble Learning Model for Short-Term Solar Irradiance Forecasting Using Historical Observations and Sky Images. (2023). IEEE Transactions on Industry Applications, 59(2), 2041–2049. https://doi.org/10.1109/tia.2022.3231842

[29] Nadweh, S., Mohammed, N., & Mekhilef, S. (2026). Intelligent prediction of photovoltaic inverter failures for improved reliability and control in smart grid systems. Electric Power Systems Research, 255, 112735. https://doi.org/10.1016/j.epsr.2026.112735.

[30] Shibghatullah, A. S. B., & Essa⁴, R. O. (2026, April). AI-Optimized Sustainable Cooling for Data Centers: A Hybrid Solar and Waste Heat. In Energy Informatics: 5th Energy Informatics Academy Conference, EI. A 2025, Kuala Lumpur, Malaysia, December 3–5, 2025, Proceedings, Part I (p. 275). Springer Nature. https://doi.org/10.1007/978-3-032-19134-2_18

[31] Nadweh, S., Mohammed, N., Konstantinou, C., & Ahmed, S. (2025). Operational performance assessment of PV-powered street lighting: A comparative study of different machine learning prediction models. IEEE Access. doi: 10.1109/ACCESS.2025.3594171

[32] Nadweh, S., Al-Omari, F., Thannon, N. T., Tawfeq, J. F., Ibrahim, A., & Jaaz, Z. A. (2025, July). A Reinforcement Learning Framework for Intelligent Detection of Bad Data in Power System State Estimation. In 2025 3rd International Conference on Cyber Resilience (ICCR) (pp. 1-7). IEEE. doi: 10.1109/ICCR67387.2025.11292295.

[33] Nadweh, S., Abdulsahib, A. K., Almajed, R., Mahfuri, M., & Tawfeq, J. F. (2025, May). Evaluation of Energy Storage Strategies Using Artificial Intelligence to Enhance the Stability of Power Grids Integrated with Renewable Energy. In 2025 3rd International Conference on Business Analytics for Technology and

Security (ICBATS) (pp. 1-9). IEEE. doi: 10.1109/ICBATS66542.2025.11258246.

[34] Nadweh, S., Qudr, L. A. Z., Almajed, R., Alhasan, W., Tawfeq, J. F., Abdulbaqi, A. S., & Mahfuri, M. (2025, May). Optimizing the Efficiency of PV Powering Telecom Towers in Remote Areas Using AI Algorithms. In 2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-9). IEEE. doi: 10.1109/ICBATS66542.2025.11258190.

[35] Nadweh, S., Abdulbaqi, A. S., Tawfeq, J. F., & Radhi, A. D. (2025, May). AI-Powered Smart Cooling System for Solar Panels: Enhancing Efficiency Through Weather Forecasting and Adaptive Control. In 2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-6). IEEE. doi: 10.1109/ICBATS66542.2025.11258170.

[36] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752. https://doi.org/10.48550/arXiv.2312.00752

[37] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. In Proceedings of the 41st International Conference on Machine Learning (ICML). https://doi.org/10.48550/arXiv.2401.09417

[38] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., & Liu, Y. (2024). VMamba: Visual state space model. arXiv preprint arXiv:2401.10166. https://doi.org/10.48550/arXiv.2401.10166

[39] Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 134–144. https://doi.org/10.1080/07350015.1995.10524599

[40] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in Neural Information Processing Systems (NeurIPS), 28. https://doi.org/10.48550/arXiv.1506.04214

[41] Brody, S., Alon, U., & Yahav, E. (2022). How attentive are graph attention networks? In International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2105.14491

[42] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997. (for standards)