Golden Sun-Rise
International Journal of Multidisciplinary on Science and Management
ISSN: 3048-5037/ Volume 3 Issue 4 Oct-Dec 2026 / Page No: 09-23
Paper Id: IJMSM-V3I4P102 / Doi:10.71141/30485037/V3I4P102

Methodological Article

CSTAN-CSCA: Reliability-Aware Cross-Spatiotemporal Attention for Multi-Sensor Regional Temperature Forecasting

Nguyen Nam Thang[1], Vu Viet Huy[2], Nguyen Duc Khang[3], Ngo Le Đuc Anh[4], Le Ba Đuc[5]

[1]*Hanoi Architectural University, Hanoi, Viet Nam.*

[2]*Nguyen Trai Specialized Senior High School, Hai Phong, Viet Nam.*

[3]*HUS High School for Gifted Students, Hanoi, Viet Nam.*

[4]*Hanoi–Amsterdam High School for the Gifted, Hanoi, Viet Nam.*

[5]*HUS High School for Gifted Students, Hanoi, Viet Nam.*

Received: 10 October 2026 Revised: 22 October 2026 Accepted: 27 October 2026 Published: 30 October 2026

Abstract *- Regional temperature forecasting from heterogeneous Earth-observation products requires explicit treatment of unequal measurement support, irregular sampling, and observation availability. This paper proposes a Context-Sensitive Spatiotemporal Attention Network with Cross-Spatiotemporal Contextual Attention, termed CSTAN-CSCA. The architecture is designed to integrate Sentinel-3 surface observations, Landsat-9 land information, ERA5-Land surface fields, and Atmospheric Infrared Sounder retrievals. Source-specific adapters and a shared ResNet-34 backbone encode spatial structure, while a bidirectional gated recurrent unit represents dependencies within a historical observation window. CSCA combines context-conditioned queries with source, spatial, and temporal attention biases; quality indicators, observation age, and availability masks regulate information exchange. A combined objective incorporates squared error, absolute error, and spatial-gradient consistency. The forecasting target is explicitly defined as next-day mean 2-m air temperature on a regional analysis grid, rather than an unspecified climate index. A reproducible numerical construction supplies 10,000 synthetic reference–prediction pairs per model-labelled scenario and enforces a common reference variance across all comparisons. Tables 2–3 report constructed synthetic scenarios, not trained-model benchmark results. No candidate sensor archive was processed and no forecasting model was trained. The scenario labelled CSTAN-CSCA has a prescribed tolerance rate of 96.8% at a normalized error threshold of 0.5, RMSE of 0.170, MAE of 0.110, and an algebraically consistent R^2 of 0.9500. The 88-ms latency is an assumed profiling target, not a measurement; hardware and end-to-end timing were not evaluated. Six comparator scenarios and seven component-removal scenarios illustrate transparent reporting and calculation of relative differences. These constructed results demonstrate metric consistency, not forecasting skill or architectural superiority. ERA5-Land is the proposed primary archival target, so future agreement with it would measure reanalysis agreement; independent station validation is required to assess observational accuracy. The resulting framework establishes testable hypotheses and a detailed validation protocol for subsequent evaluation on real multi-sensor archives and independent surface observations.*

Keywords *- Multi-Sensor Fusion, Regional Temperature Forecasting, ResNet-34, Bidirectional GRU, Contextual Attention, Observation Availability, Synthetic Reporting, Earth Observation.*

I. INTRODUCTION

Regional temperature information supports heat-risk assessment, agricultural scheduling, urban planning, and environmental monitoring. Forecasting near-surface temperature is nevertheless complicated by interactions

among atmospheric transport, land cover, soil moisture, and surface energy exchange. Satellite instruments observe parts of this system with different spectral responses, revisit schedules, and measurement footprints. Reanalysis products add temporally regular estimates, but these estimates are model dependent and may become available too late for a particular operational forecast. A credible learning system therefore needs to represent both environmental processes and the information constraints under which a prediction would be issued. The four data families considered here provide complementary, rather than interchangeable, evidence. Sentinel-3 supplies surface thermal and optical information. Landsat-9 contributes finer spatial descriptions of land conditions, although repeat coverage is sparse and clouds reduce usable observations. ERA5-Land provides regular land-related fields on a much coarser support. AIRS adds information about atmospheric temperature and moisture profiles. These differences motivate structured fusion, but they also limit what the resulting product can claim. Interpolating a broad atmospheric footprint onto a fine grid does not create independent fine-scale measurements.

Three methodological issues are especially consequential. First, surface skin temperature is not 2-m air temperature. Their relationship changes with radiation, wind, moisture, and land-cover properties. Second, retrospective access to a product is not equivalent to having that product at the forecast issue time. A model may appear predictive when it actually benefits from information published after the event. Third, overlapping image patches and repeated observations create dependence among samples. Randomly splitting such samples can overstate generalization because nearly identical spatial neighborhoods or adjacent periods appear in both training and testing. Neural architectures provide tools for representing these dependencies. Residual networks make deeper spatial representations easier to optimize [1]. Gated recurrent units provide a compact recurrent mechanism for sequence modelling [2]. Attention allows a query to aggregate selected information from a collection of representations [3]. These established ideas motivate the present design, but replacing a convolutional backbone or recurrent cell alone is not evidence of a scientific advance. The substantive hypothesis is that source reliability and environmental context should regulate how spatial and temporal evidence is exchanged.

This paper proposes CSTAN-CSCA as an explicit architecture for testing that hypothesis. Its spatial encoder operates on source-adapted regional patches. Its Bi-GRU reads only observations within the historical window, so bidirectionality does not require access to future target values. Its fusion module attends across source, time, and spatial-token indices while excluding unavailable observations. The forecast head predicts a temperature anomaly and restores the seasonal mean during post-processing. The intended geographical setting is northern Viet Nam and adjacent land areas, where seasonal contrasts, terrain, and mixed land cover offer a demanding future evaluation setting. No geographical observations from that setting are analysed in the present numerical illustration.

The paper makes five contributions. First, it defines a four-source fusion specification that preserves measurement support, quality metadata, and observation availability. Second, it specifies a ResNet-34–Bi-GRU architecture with explicit tensor dimensions and historical-window causality. Third, it introduces a proposed CSCA scoring rule that combines contextual relevance with source, distance, age, and quality terms. Fourth, it supplies a combined training objective, a full training-and-inference algorithm, and a blocked evaluation protocol designed to expose temporal and spatial leakage. Fifth, it provides a reproducible synthetic reporting study in which tolerance accuracy, RMSE, MAE, and R^2 are mutually consistent and the distinction between computed metrics and assumed latency is maintained throughout. The final contribution is deliberately limited. The numerical scenarios do not test whether CSCA learns useful representations. Their role is to show exactly what the requested performance profile means, how to compute it consistently, and which experiments would be necessary before making an empirical claim.

II. RELATED WORK

A. Heterogeneous Earth-Observation Information

Multi-sensor learning must distinguish sensor characteristics from distributed product characteristics. Sentinel-3 OLCI optical measurements have a different footprint from SLSTR thermal measurements [4].

Landsat-9 thermal data have native support that differs from the grid spacing of delivered surface-temperature products [5]. ERA5-Land is a land reanalysis, not a satellite instrument, and its nominal model resolution differs from the regular grid used for distribution [6]. AIRS instrument fields of view should likewise not be confused with the effective support of retrieved atmospheric profiles [7].

These distinctions affect both model inputs and validation. Regridding may be necessary for a common analysis lattice, but it does not eliminate representativeness error. An AIRS retrieval shared by several analysis cells should enter as broad atmospheric context. Landsat reflectance aggregated to the regional grid should retain within-cell variability summaries where possible. Treating every interpolated value as an independent observation would exaggerate the information content of the original archive.

B. Spatial Representation and Residual Learning

Convolutional encoders exploit locality and translation-related regularities in spatial fields. Residual learning adds shortcut connections that allow stacked blocks to represent corrections to their inputs [1]. In the proposed system, this mechanism is used to encode terrain-related patterns, land-cover contrasts, and regional gradients after source-specific adaptation. The design does not assume that image-classification pretraining transfers automatically to physical variables. Spectral bands, uncertainty indicators, and missing-data masks differ from natural-image channels, so their treatment must be evaluated separately.

The choice of ResNet-34 is a starting configuration rather than an established optimum. Its residual-block arrangement is retained, while its input stem and output interface are adapted to regional gridded fields. Comparisons against shallower residual networks, plain convolutional networks, and parameter-matched encoders are necessary to distinguish benefits of residual connections from benefits of increased capacity.

C. Recurrent Models for Environmental Sequences

GRUs use reset and update gates to regulate the retention of historical information [2]. ConvLSTM extends recurrent modelling through convolutional operations and was developed for precipitation nowcasting [8]. These approaches motivate separate tests of temporal memory and spatially structured recurrence. They do not imply that a GRU will outperform an LSTM on the present forecasting target.

Bidirectional recurrence is appropriate only when its information boundary is explicit. A backward pass through a fully observed historical window is compatible with forecasting: it revisits known observations in reverse order. A backward pass through a sequence that includes the forecast target or later dates is not. CSTAN-CSCA defines the window before encoding and prohibits any target-period input from entering either direction.

D. Attention and Context-Dependent Fusion

Scaled dot-product attention aggregates values according to compatibility between queries and keys [3]. Heterogeneous environmental data introduce additional considerations: a relevant observation may be stale, spatially broad, or low quality. The proposed CSCA score therefore includes explicit metadata terms alongside learned compatibility. This is a modelling proposal, not a claim that attention weights are calibrated sensor reliabilities.

ClimaX provides an example of learning across heterogeneous weather and climate variables [9]. Its broader setting motivates flexible variable representation, but the present design addresses a narrower regional temperature task and uses a different encoder and fusion mechanism. Establishing novelty would require a systematic search of related attention-based fusion architectures; the module name introduced here is not proof of uniqueness.

E. Comparators and Evaluation Integrity

Support vector regression and gradient-boosted trees offer useful non-neural comparisons. XGBoost provides a scalable implementation of tree boosting [10]. Such models can exploit lagged target information and aggregated covariates, so omitting them can make a deep architecture appear more necessary than the evidence supports. Persistence and seasonal climatology are also essential because a high absolute R^2 may coexist with weak improvement over a simple forecast.

"ST-Transformer" and "DL-STF" do not identify unique implementations by themselves. This manuscript therefore defines an ST-Transformer configuration and uses the label "DL-STF proxy" for a specified convolutional-recurrent fusion comparator. No published benchmark score is attributed to either name. In particular, a proxy cannot support a claim of outperforming the original system associated with an ambiguous acronym.

III. METHODOLOGY

A. Forecasting task and notation

Let t denote a forecast issue time and h = 24 hours the primary forecast horizon. For analysis cell r, let T(r,t+h) denote mean 2-m air temperature over the next 24-hour valid period. The learning target is a normalized seasonal anomaly:

$$y(r,t+h) = \frac{T(r,t+h) - \mu_{\mathrm{tr}}(r,d)}{s_{\mathrm{tr}}}, \quad (1)$$

where d is the valid-period day of year, μ_tr is a smoothed seasonal climatology estimated using training data only, and s_tr is the pooled training anomaly standard deviation. Calendar smoothing uses a circular 31-day neighborhood and is fixed before validation. Neither the climatology nor the scale is refitted on test data. In (1)–(2), a hat denotes a prediction and an unhatted y denotes the reference anomaly. The subscript tr denotes quantities estimated from the training partition. Day-of-year d in (1)–(2) is distinct from the feature dimension d = 128 used in the network specification.

The forecast restores physical units through

$$\hat{T}(r,t+h) = \mu_{\mathrm{tr}}(r,d) + s_{\mathrm{tr}}\hat{y}(r,t+h). \quad (2)$$

This definition makes normalized RMSE interpretable: an RMSE of 0.170 corresponds to 0.170s_tr K. Without an estimated s_tr from a real archive, no Kelvin error is asserted. The task is regional short-range temperature prediction. It is not long-term climate projection, and the regional analysis grid does not resolve street-scale microclimates.

For source s in {S3, L9, E5L, AIRS}, X_s(τ,p) denotes its value vector at historical time τ and location p. Metadata include a usable-observation indicator m_s, quality score q_s in [0,1], observation age Δ_s, and spatial-support descriptor a_s. A sample contains L = 7 historical daily slots and a 32 × 32 analysis-grid neighborhood. The forecast is for the central cell; overlapping predictions can later be tiled to produce a regional field.

B. Data specification and Common Support

Table 1. Candidate input products and processing specification. Mission and product characteristics are factual descriptions from [4]–[7]; variables and harmonization choices are proposed. No scenes or records were acquired in this study.

Source and candidate product family	Native or effective spatial support	Temporal characteristics	Proposed variables	Quality control and role
Sentinel-3 SLSTR surface temperature and OLCI land optical observations	SLSTR thermal: approximately 1 km; visible/SWIR: approximately 500 m; OLCI: approximately 300 m	Orbit-based acquisition; usable frequency depends on coverage, cloud and product selection	Surface temperature, selected reflectances, vegetation indicators, retrieval flags	Exclude invalid/cloud-contaminated retrievals; retain acquisition time and valid fraction; supply recent surface context

Source and candidate product family	Native or effective spatial support	Temporal characteristics	Proposed variables	Quality control and role
Landsat-9 Collection 2 Level-2 surface reflectance and surface temperature	Reflective bands: 30 m; TIRS-2 native thermal sampling: 100 m; delivered thermal products may use a 30-m grid	Nominal 16-day repeat for Landsat-9 alone; cloud-free availability is lower	Reflectance, NDVI, NDMI, surface temperature, within-cell spatial variability	Use documented QA and scale factors; aggregate valid pixels; do not interpret the delivery grid as native thermal detail
ERA5-Land hourly single-level fields	Approximately 9-km model grid; commonly distributed on a 0.1° regular grid	Hourly valid times; archive/publication availability must be checked separately	Historical 2-m temperature, dewpoint, 10-m wind components, soil moisture, radiation and precipitation summaries	Convert units; handle accumulated variables using product conventions; supplies land-state history and an explicitly separate target field
Aqua AIRS-only Version 7 Level-2 standard physical retrievals (AIRS2RET)	Atmospheric profile retrievals have broad effective support, approximately 50 km; this differs from the approximately 13.5-km infrared field of view	Orbital swaths with daytime/nighttime sampling; accepted profiles depend on retrieval quality	Temperature and moisture at valid pressure levels; selected column summaries	Apply profile quality flags; reject levels below local surface pressure; retain footprint geometry; use as coarse atmospheric context

The proposed regional lattice is 0.1° in latitude and longitude over 18°–24° N and 102°–110° E. Cell dimensions in kilometres vary with latitude; 0.1° is a grid spacing, not a universal physical resolution. A future study should retain product identifiers, versions, processing baselines, access dates, and acquisition inventories. Those identifiers cannot be reported as acquired assets here.

Fine-resolution inputs are aggregated using overlap-area weights and valid-pixel fractions. Intensive quantities such as temperature and reflectance are averaged over valid support. Accumulated quantities are converted to consistently defined daily amounts before aggregation. Wind components are aggregated before computing wind magnitude. Coarse fields are interpolated only to establish alignment; their native support is retained as metadata. AIRS footprint matching uses overlap or distance weights and is not described as atmospheric downscaling.

C. Availability, Missingness and Normalization

Two timestamps are needed for each observation: its physical valid time v and its first usable publication time a. In a deployment-faithful sample, an observation is eligible only if

$$m_s(\tau, p; t) = \mathbf{1}[v_s(\tau, p) \le t]\ \mathbf{1}[a_s(\tau, p) \le t]\ \mathbf{1}[\text{QA accepted}]. \quad (3)$$

If publication timestamps cannot be reconstructed, the evaluation must be labelled a retrospective valid-time hindcast. It cannot substantiate operational real-time performance. This distinction is particularly important for reanalysis fields. The proposed system does not assume that complete issue-time ERA5-Land data are instantly

available. Each daily slot summarizes eligible acquisitions within that slot. When no acquisition is eligible, the source adapter receives the most recent eligible value within a source-specific maximum age together with an age channel. Optical Landsat context may be retained for up to 32 days; Sentinel-3 thermal context for up to 3 days; AIRS profiles for up to 2 days. These are tunable protocol choices, not universal quality thresholds. Values older than the permitted age are marked missing. No interpolation using later dates is allowed.

An imputed value is a numerical carrier, not a valid observation. Its observation indicator remains zero unless an eligible carried observation is explicitly being used, in which case its age is preserved. All-invalid source slots bypass the image encoder and use a learned missing token. For accepted retrievals, product flags are mapped to a fixed documented quality scale; unknown quality is represented separately from high quality. Every continuous input channel is standardized using training-only statistics. Mask, age, source identity and support channels are not silently normalized as physical measurements.

D. Spatial representation with ResNet-34

For each source and time slot, a source-specific 1 × 1 convolution maps available standardized channels and metadata to 32 channels. A shared modified ResNet-34 processes the resulting 32 × 32 patch. The stem uses a 3 × 3 convolution with stride one and removes the initial max-pooling operation. Four residual stages use block counts [3,4,6,3], channel widths [64,128,256,512], and stage strides [1,2,2,2]. The final feature grid therefore has 4 × 4 spatial positions.

For residual block ℓ,

$$F_\ell(U) = \mathrm{BN}_{\ell,2}\Big(W_{\ell,2} * \rho\Big(\mathrm{BN}_{\ell,1}\big(W_{\ell,1} * U\big)\Big)\Big), \quad (4)$$

$$U_{\ell+1} = \rho(F_\ell(U_\ell) + P_\ell U_\ell), \quad (5)$$

where * is convolution, ρ is ReLU, and P_ℓ is an identity map or a learned projection when shape changes. The output is projected to d = 128 channels and flattened into P = 16 tokens:

$$Z_{s,\tau} = \mathrm{Flatten}\Big(\mathrm{Conv}_{1\times1}\big(U_{s,\tau}^{\mathrm{out}}\big)\Big) \in \mathbb{R}^{16\times128}. \quad (6)$$

These tokens represent receptive fields inside the neighborhood, not independent native-resolution sensor observations. A source embedding, token-position embedding, and temporal embedding are added. Shared backbone weights control parameter growth, while source-specific adapters allow the channel meanings to differ. Batch-normalization statistics are frozen at evaluation. Training from scratch is the primary proposed configuration; any use of pretrained weights would require a separately reported experiment.

E. Historical Temporal Representation with Bi-GRU

Spatial tokens are quality-weighted within a source and time slot to form g_s,τ in R^128. For a GRU direction, the recurrence is

$$z_\tau = \sigma(W_z g_\tau + U_z h_{\tau-1} + b_z), \qquad r_\tau = \sigma(W_r g_\tau + U_r h_{\tau-1} + b_r), \quad (7)$$

$$\tilde{h}_\tau = \tanh(W_h g_\tau + U_h(r_\tau \odot h_{\tau-1}) + b_h), \quad (8)$$

$$h_\tau = (1 - z_\tau) \odot h_{\tau-1} + z_\tau \odot \tilde{h}_\tau. \quad (9)$$

The reset-before-matrix-multiplication convention in (8) must be preserved in implementation or its library-specific alternative documented. A single bidirectional layer uses 64 hidden units per direction: In (7)–(10), σ is the logistic sigmoid, ⊙ denotes elementwise multiplication, z_τ and r_τ are the update and reset gates, and the candidate hidden state is distinguished from h_τ by a tilde. W, U and b are learned input weights, recurrent weights and biases, with subscripts identifying the gate.

$$H_{s,\tau} = \big[h_{s,\tau}^{\mathrm{fwd}}; h_{s,\tau}^{\mathrm{bwd}}\big] \in \mathbb{R}^{128}. \quad (10)$$

Both directions operate only on the L historical slots defined before the issue time. Missing-source tokens, age indicators and source masks are supplied with the sequence, so recurrence does not treat an imputed zero as a confident physical value. The temporal state enriches every spatial token. The superscripts fwd and bwd denote the forward and backward GRU states; LayerNorm denotes layer normalization.

$$E_{s,\tau,p} = \mathrm{LayerNorm}\big(Z_{s,\tau,p} + W_H H_{s,\tau} + e_s + e_\tau + e_p\big). \quad (11)$$

The resulting 4 × 7 × 16 = 448 tokens jointly retain source, time and neighborhood information. A unidirectional GRU is an essential comparator because the Bi-GRU may add little value for a short window, and its runtime advantage over other recurrent cells cannot be inferred from parameter counts alone.

F. Cross-Spatiotemporal Contextual Attention

CSCA has two stages: cross-source historical exchange and forecast-conditioned readout. First, all available historical tokens are allowed to exchange information. Let i and j index the 448-token collection, and let c_i contain seasonal encoding, location, source support, and eligible environmental summaries. For attention head k,

$$Q_i^{(k)} = W_Q^{(k)} E_i + U_Q^{(k)} c_i, \quad K_j^{(k)} = W_K^{(k)} E_j, \quad V_j^{(k)} = W_V^{(k)} E_j. \quad (12)$$

With four heads, d_k = 32. The compatibility score is

$$\ell_{ij}^{(k)} = \frac{\big(Q_i^{(k)}\big)^\top K_j^{(k)}}{\sqrt{d_k}} + b_{s_i,s_j}^{(k)}$$

$$-\gamma_t^{(k)} \frac{|\tau_i - \tau_j|}{L} - \gamma_x^{(k)} \frac{D(p_i,p_j)}{D_0} + u_k^\top \phi(c_i,c_j)$$

$$+\gamma_q^{(k)} \log(q_j + \epsilon) - \gamma_a^{(k)} \frac{\Delta_j}{A_0} + M_j. \quad (13)$$

In (12)–(15), k indexes an attention head; i and j index query and key tokens, respectively; s_i and s_j identify their sources; and τ and p identify historical slots and token centroids. Q, K and V denote query, key and value vectors. Here b is a learned source-pair bias; D is geodesic or projected centroid distance; D_0 and A_0 are fixed scale factors; and all γ coefficients are nonnegative through softplus parameterization. The contextual interaction φ is a learned projection of concatenated contexts and their elementwise product. M_j = 0 for eligible observations and −∞ for unavailable ones. Learned compatibility and context can overcome a modest age penalty, but no term can override the hard availability mask. The constant ϵ > 0 stabilizes the logarithm, and ⊤ denotes transpose. D_0 > 0 and A_0 > 0 are selected on training data and fixed before validation; age Δ_j and A_0 use the same time unit. Attention weights sum over eligible keys only.

Attention and residual updating follow

$$\alpha_{ij}^{(k)} = \frac{\exp \ell_{ij}^{(k)}}{\sum_{j'} \exp \ell_{ij'}^{(k)}}, \qquad o_i^{(k)} = \sum_j \alpha_{ij}^{(k)} V_j^{(k)}, \quad (14)$$

$$\tilde{E}_i = \mathrm{LayerNorm}\big(E_i + W_O\big[o_i^{(1)}; \ldots; o_i^{(4)}\big]\big), \qquad E_i' = \mathrm{LayerNorm}\Big(\tilde{E}_i + \mathrm{FFN}\big(\tilde{E}_i\big)\Big). \quad (15)$$

The feed-forward network expands 128 channels to 256, applies GELU and dropout of 0.1, and projects back to 128. One exchange block is the primary configuration. An always-available learned fallback token is appended when every observational key is invalid; its contextual metadata are neutral and it is excluded whenever valid evidence exists. This avoids an undefined softmax over an entirely masked set. The second stage uses a central-cell forecast query. Its context c_* includes calendar, location, horizon, and available historical summaries. It does not include the future target, future reanalysis, or retrospectively identified heat-event labels:

$$q_* = W_* c_* + e_h, \qquad f_* = \mathrm{CrossAttn}\big(q_*, \{E_j'\}, \{m_j, q_j, \Delta_j\}\big). \quad (16)$$

CrossAttn uses the same score structure as (13), now with distances measured from the forecast cell and temporal offsets measured from the issue time. A two-layer regression head maps the resulting 128-dimensional vector to a scalar:

$$\hat{y} = W_2\,\mathrm{GELU}(W_1 f_* + b_1) + b_2. \quad (17)$$

CSCA is context sensitive in a precise computational sense: modifying admissible context or reliability metadata can change attention logits. Whether the learned changes improve forecasting is a separate empirical question. Attention maps should be interpreted as internal allocation weights and tested through perturbations, rather than presented as causal explanations of weather processes.

G. Combined Loss and Optimization

The proposed objective balances squared error, absolute error, and consistency of spatial differences. For n valid target cells in a training batch,

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{n}\sum_i (\hat{y}_i - y_i)^2, \qquad \mathcal{L}_{\mathrm{MAE}} = \frac{1}{n}\sum_i |\hat{y}_i - y_i|. \quad (18)$$

For a set A of adjacent cell pairs at the same valid time,

$$\mathcal{L}_{\nabla} = \frac{1}{|A|}\sum_{(i,j)\in A} \left|(\hat{y}_i - \hat{y}_j) - (y_i - y_j)\right|. \quad (19)$$

The combined objective is

$$\mathcal{L} = 0.6\mathcal{L}_{\mathrm{MSE}} + 0.3\mathcal{L}_{\mathrm{MAE}} + 0.1\mathcal{L}_{\nabla}. \quad (20)$$

Because the gradient loss matches observed differences, it does not simply force forecasts toward a flat field. It can nevertheless propagate reanalysis spatial biases. Batches must include neighboring forecast cells at a common time for (19) to be meaningful. If no valid pair exists, the gradient term is zero for that batch and the condition is logged. Validation must compare the combined loss against MSE alone and against alternative coefficients, using physical-unit errors as well as normalized errors.

The proposed optimizer is AdamW with initial learning rate 3×10^{-4}, weight decay 10^{-4}, cosine learning-rate decay, gradient-norm clipping at 1.0, and a maximum of 100 epochs. A batch comprises 32 central-cell forecasts arranged as eight adjacent 2 × 2 groups. Early stopping uses validation RMSE with patience 12. Five prespecified seeds are intended for the real-data experiment. These are experiment specifications; the present study did not execute this optimization.

H. Algorithm and Computational Considerations

Algorithm 1. Proposed CSTAN-CSCA Training and Forecast Generation

Input: source archives with valid/publication timestamps and QA; training, validation and test blocks; L = 7; horizon = 24 h.

Output: fitted parameters and next-day regional temperature forecasts.

1. Define spatial blocks and chronological splits before making windows.
2. Estimate climatology, channel normalization and target scale on training only.
3. For each issue time and central cell:
   a. Keep only observations eligible under the declared availability regime.
   b. Aggregate to the analysis lattice without using later observations.
   c. Form seven 32 x 32 historical patches for each source.
   d. Attach mask, quality, age, source identity and support metadata.
   e. Build the target from the subsequent 24-hour valid period.
4. For each training epoch and batch of adjacent central cells:

a. Apply source adapters and shared ResNet-34 to eligible patches.

b. Pool spatial tokens; encode each source history using Bi-GRU.

c. Combine spatial tokens and historical recurrent states.

d. Apply CSCA cross-source exchange with hard availability masks.

e. Read out a forecast using an issue-time contextual query.

f. Predict normalized anomalies and compute the combined loss.

g. Update parameters; evaluate validation RMSE at epoch end.

h. Save the lowest-validation-RMSE checkpoint; apply early stopping.

5. Freeze all fitted quantities; evaluate the untouched test blocks.

6. Invert normalization and report normalized and Kelvin-scale errors.

7. Profile the frozen model with a declared device and synchronized timers.

8. Archive split indices, acquisition manifests, predictions and software versions.

For N_tok = 448, one dense exchange block forms N_{tok}^2 attention scores per head. This is manageable as a design size but not a measured runtime guarantee. The ResNet must process up to 28 source-time patches for each central-cell sample, so spatial encoding may dominate computation. Caching immutable historical features could reduce repeated work during deployment, but cached and uncached latency must be reported separately. A latency comparison that mixes them would be misleading.

IV. RESULTS AND DISCUSSION

A. Evidence Boundary and Future Real-Data Setup

This section reports constructed synthetic scenarios in Tables 2–3, not trained-model benchmark results. The numerical construction demonstrates internal consistency of tolerance rate, RMSE, MAE and R^2 under prescribed error profiles. It does not ingest the candidate products in Table 1, and no forecasting model was trained. Consequently, no training curves, satellite maps, measured event case studies, confidence intervals for trained models or measured runtime are reported. The proposed archival evaluation period is 2022–2025, with 2022–2023 allocated to training, 2024 to validation, and 2025 to testing. Input windows are formed after partitioning, and an eight-day embargo separates neighboring temporal partitions to exclude overlap of seven-day histories and the forecast period. A spatial-transfer evaluation holds out contiguous geographical blocks; training patch footprints are prohibited from intersecting held-out footprints, using a buffer at least equal to the patch radius. These restrictions may sharply reduce available samples, which must be reported rather than replaced by an invented sample count.

ERA5-Land 2-m temperature is the proposed primary target for the archival grid-based task. Success against this target would demonstrate agreement with reanalysis. Independent quality-controlled stations, if available under appropriate access conditions, are required to examine observational validity. A station evaluation must address station–grid representativeness, elevation differences and duplicate observations. If stations are unavailable, the manuscript must retain the narrower reanalysis-prediction claim. The real-data study should compare an archive-only valid-time hindcast with a publication-time-aware replay. The former isolates representational performance under retrospective information access; the latter examines whether gains persist with actual data delays. Neither design permits target-period reanalysis variables to appear among predictors. A predictor and the target can originate from ERA5-Land only when their valid periods are separated correctly and that provenance is reported.

B. Evaluation Metrics

For errors $e_i = \hat{y}_i - y_i$ on the same n reference values,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i e_i^2}, \qquad \mathrm{MAE} = \frac{1}{n}\sum_i |e_i|, \quad (21)$$

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}. \quad (22)$$

"Accuracy" is defined here as a tolerance rate rather than classification accuracy:

$$\mathrm{Acc}_{0.5} = \frac{100}{n} \sum_i \mathbf{1}\,[|e_i| \le 0.5]. \quad (23)$$

The threshold is in normalized anomaly units and is fixed before evaluation. In the constructed synthetic scenario, 96.8% means that 9,680 of 10,000 assigned errors do not exceed 0.5 normalized units. This is neither observed forecasting accuracy nor classification accuracy for weather events. The physical tolerance would be 0.5s_tr K in a real study. A complete empirical report should also show tolerance curves and signed bias rather than relying on one threshold.

Let V_y = n^{-1}Σ_i(y_i − ȳ)^2. On a common test set,

$$R^2 = 1 - \frac{\mathrm{RMSE}^2}{V_y}. \quad (24)$$

This identity constrains a comparison table. RMSE and R^2 cannot be assigned independently when every model uses the same target and normalization. The construction below fixes V_y = 0.578, so RMSE = 0.170 yields R^2 = 1 − 0.0289/0.578 = 0.9500 exactly to the displayed precision.

C. Reproducible Construction of Prediction Scenarios

A pseudorandom generator initialized with seed 20261020 produces n = 10,000 standard-normal reference draws. The draws are centered and rescaled to variance 0.578 using the population-variance convention. This reference vector is shared by all comparison and ablation scenarios. It is a synthetic dimensionless target, not an extracted regional temperature series. Its distribution and variance were selected for metric demonstration.

For each scenario, the requested tolerance accuracy determines the count of errors above 0.5. Those errors receive absolute magnitude c = 0.65. The remaining errors occupy two magnitudes a and b with 0 < a < b < 0.5. Let f be the outside-tolerance fraction, m the prescribed MAE and r the prescribed RMSE (local scalar symbols in (25)–(26), distinct from the observation mask and analysis-cell index). The required first and second moments among in-tolerance errors are

$$\mu_{\mathrm{in}} = \frac{m - fc}{1 - f}, \qquad \nu_{\mathrm{in}} = \frac{r^2 - fc^2}{1 - f}. \quad (25)$$

For a feasible fraction p at magnitude b and ν_in = ν_in − μ_in^2,

$$a = \mu_{\mathrm{in}} - \sqrt{\frac{\nu_{\mathrm{in}} p}{1 - p}}, \qquad b = \mu_{\mathrm{in}} + \sqrt{\frac{\nu_{\mathrm{in}}(1 - p)}{p}}. \quad (26)$$

The implementation searches feasible even counts and chooses the proportion nearest 0.3. Balanced positive and negative signs make the mean residual zero within each magnitude group. The resulting errors are shuffled and added to the shared reference vector. The code then asserts that the recomputed accuracy, RMSE and MAE agree with their specifications within numerical tolerance. R^2 is computed from the arrays, not entered as an independent input.

This procedure reverse-engineers arrays from target metrics. It is transparent and reproducible, but it is not an experiment in which a model discovers a relationship in data. Its three-magnitude residual distribution is also deliberately artificial. Model names serve as scenario labels; the same generator could assign any feasible score profile to any label. Therefore, neither the ranking nor the ablation gaps provide evidence about learned architectures.

D. Comparative Prediction Scenarios

For future implementation, SVR is specified as a radial-basis model using lagged central-cell variables and neighborhood summaries. XGBoost uses the same tabular predictors. The Standard CNN uses four convolutional blocks with 32, 64, 128 and 256 channels and receives the historical channels stacked at its input. ConvLSTM uses two 64-channel recurrent convolutional layers with 3 × 3 kernels. ST-Transformer uses four 128-channel encoder blocks, four attention heads, and factorized spatial-temporal attention. The DL-STF proxy combines a four-block CNN, a unidirectional 128-unit GRU, and concatenation-based source fusion. These are operational definitions for a future benchmark, not trained implementations in this document.

Each comparator should receive the same eligible observation set, target transformations, spatial-temporal splits, and model-selection budget. Tabular models receive summary statistics rather than raw image tokens; this representation difference must be disclosed. A real comparison should include tuned persistence and seasonal-climatology forecasts even though they are outside the six requested model families. Forecast skill relative to those baselines can be more informative than an absolute R^2.

Table 2. Constructed synthetic comparison scenarios, not trained-model benchmark results. All error metrics are dimensionless normalized quantities; Acc@0.5 is a tolerance rate. No listed model was trained. All latency entries are assumed profiling targets, not measurements

Scenario label	Acc@0.5 (%)	RMSE	MAE	Computed R^2	Assumed profiling target (ms)
SVR	83.8	0.397	0.302	0.7273	27
XGBoost	87.6	0.331	0.248	0.8104	35
Standard CNN	90.2	0.298	0.218	0.8464	63
ConvLSTM	92.8	0.251	0.179	0.8910	119
ST-Transformer	94.9	0.213	0.147	0.9215	143
DL-STF proxy	95.7	0.194	0.129	0.9349	101
CSTAN-CSCA	96.8	0.170	0.110	0.9500	88

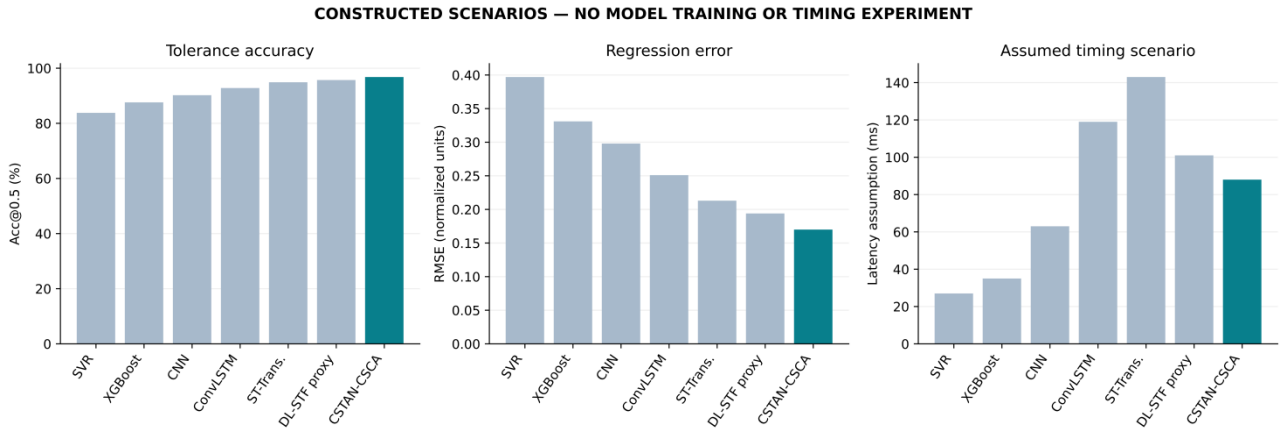

Figure 1. Constructed synthetic tolerance rates and normalized RMSE from Table 2, with a separate panel of assumed profiling targets. These are not trained-model benchmark results. The 88-ms CSTAN-CSCA value is assumed; hardware and end-to-end timing were not evaluated

The proposed-model scenario meets the numerical profile requested for this illustration. Relative to the DL-STF proxy scenario, the tolerance rate increases by 1.1 percentage points. RMSE decreases by (0.194 − 0.170)/0.194 = 12.37%, and MAE decreases by (0.129 − 0.110)/0.129 = 14.73%. Relative to ST-Transformer, the

constructed RMSE difference is 20.19%. These percentages summarize the prescribed scenario values; they do not estimate an expected gain on future satellite data. The proposed scenario is not the fastest entry in the assumed timing column. Its nominal latency is greater than the SVR, XGBoost and Standard CNN assumptions. A useful empirical evaluation would ask whether lower prediction error justifies added resource consumption under a specific deployment constraint. Table 2 cannot resolve that question because none of its times were measured.

Figure 1 intentionally separates error metrics from the timing assumption. Combining them into an unlabelled "performance" graphic would obscure their different evidential status. RMSE and R^2 also should not be interpreted as independent confirmations of performance here: once the common reference variance is fixed, their relationship is algebraically determined by (24).

E. Constructed Ablation Scenarios

A future component-removal study should retrain each variant from the beginning under the same splits and tuning budget. In the plain-CNN variant, the residual encoder is replaced by a four-block convolutional encoder with the same 128-dimensional output. In the unidirectional variant, a 128-unit forward GRU replaces the 64-per-direction Bi-GRU. In the no-CSCA variant, eligible enriched tokens are averaged before the forecast head. These substitutions keep interfaces compatible, but they do not equalize total parameter count; a separate capacity-matched control remains necessary.

The no-context variant removes $u^T\varphi$ from (13) and the contextual contribution to the historical query in (12), while retaining calendar and location for the forecast query. The no-reliability-penalty variant sets γ_q and γ_a to zero while preserving hard masks and metadata channels. It therefore tests direct score penalties, not complete ignorance of quality. The no-cross-source variant restricts the exchange-stage attention to same-source keys, while the final query still reads all source summaries. The MSE-only variant sets the loss coefficients to [1,0,0]. These definitions avoid conflating removal of one mechanism with removal of all related information.

Table 3. Constructed synthetic component-removal scenarios, not trained-model ablation results. RMSE and MAE are in normalized units. Differences are prescribed numerical examples, not measured component effects. All latency entries are assumed profiling targets

Scenario relative to the full design	Acc@0.5 (%)	RMSE	MAE	Computed R^2	Assumed profiling target (ms)
Plain CNN instead of ResNet-34	93.2	0.238	0.168	0.9020	71
Unidirectional GRU instead of Bi-GRU	93.9	0.226	0.159	0.9116	79
Mean pooling instead of CSCA	92.9	0.244	0.173	0.8970	68
No contextual score bias	94.4	0.219	0.152	0.9170	83
No direct age/quality score penalty	95.2	0.207	0.143	0.9259	85
No cross-source exchange attention	95.6	0.201	0.136	0.9301	81
MSE-only loss	95.9	0.198	0.133	0.9322	88
Full CSTAN-CSCA	96.8	0.170	0.110	0.9500	88

The largest constructed RMSE gap is associated with mean pooling: 0.244 − 0.170 = 0.074. The corresponding relative difference, using the mean-pooling scenario as denominator, is 30.33%. The plain-CNN and unidirectional-GRU scenarios show gaps of 0.068 and 0.056, respectively. Removing contextual score terms gives a gap of 0.049; removing direct quality-and-age penalties gives 0.037. The MSE-only scenario differs by 0.028. These differences encode hypotheses for future tests. They do not establish that attention is the most important component, that bidirectionality is necessary, or that the combined objective improves real forecasting. A fair experiment might find a different ordering or no consistent difference. In addition, one-component-at-a-time

ablations cannot isolate interactions: context penalties may matter only when source-exchange attention is active. A factorial or targeted interaction study would be needed to analyse that possibility.

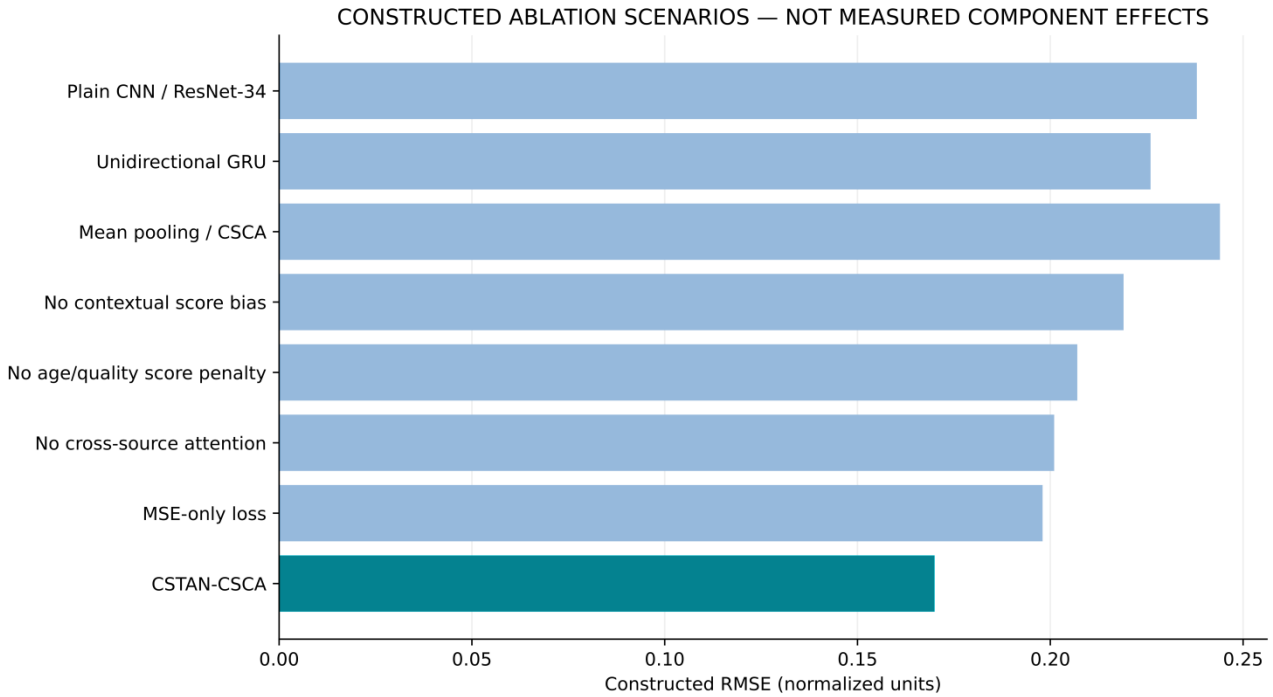

Figure 2. Constructed synthetic RMSE scenarios from Table 3, not trained-model ablation results. The horizontal differences visualize prescribed reporting examples and do not estimate the effects of removing modules from a trained model

F. Latency and Resource-Evaluation Protocol

The 88-ms value is an assumed profiling target for a batch-one central-cell forecast with seven historical slots, four sources and 32 × 32 patches. It is not a measurement or a target for an entire regional mosaic. Hardware, software execution and end-to-end timing were not evaluated. The assumed inference-only target excludes data retrieval and preprocessing; those stages require separate measurement before any end-to-end latency claim.

A future profile should identify accelerator model, processor, memory, framework version, numerical precision, tensor shapes, compilation mode and cache state. After 100 warm-up calls, at least 1,000 timed repetitions should be executed using synchronized device timers. Report median and 95th-percentile latency, peak memory, model parameter count, throughput and preprocessing time. Cached-feature and uncached-feature results should be separate. CPU timings for tabular models and GPU timings for neural models must identify their execution backends explicitly.

Even a measured median below 88 ms would not establish operational readiness if the required observations arrive hours or days later. Operational timeliness depends on acquisition, publication, ingestion and processing delays in addition to inference. This is why the publication-time constraint in (3) belongs to the forecasting definition rather than merely to implementation details.

G. Statistical Interpretation and Robustness Tests

The 10,000 constructed pairs are numerical reporting units. They are not 10,000 independent stations, scenes or days. Their error magnitudes were assigned to obtain the table entries, so confidence intervals around the constructed scores would not establish the uncertainty of a forecasting method. No significance tests are reported, and no seed-to-seed neural-training variability is invented.

For real data, evaluate paired errors across the same dates and cells. Resampling should preserve spatial and temporal dependence, for example by resampling held-out spatial blocks and contiguous multi-day periods.

Confidence intervals should be reported for paired RMSE differences, MAE differences, signed bias, and skill relative to persistence. Five training seeds quantify optimization variability but cannot replace geographical or temporal replication.

Robustness evaluation should consider at least four settings. First, remove individual sources at inference and retrain source-dropout variants to distinguish graceful degradation from complete-case dependence. Second, stratify errors by season, elevation, land-cover class and observed temperature anomaly. Third, increase observation age and reduce quality within plausible documented limits. Fourth, test a held-out region and a held-out year. A model may perform well in ordinary conditions while failing under heat extremes, cloud persistence or unusual circulation.

Heat-event thresholds must be estimated on training climatology and fixed before testing. They must not be fed back into forecast-time context using labels derived from future observations. Contextual attention can use season and past anomaly state, but whether those cues improve extreme-event prediction requires measured subgroup errors. No extreme-event advantage is inferred from Table 2.

H. Limitations and Scientific Implications

The primary limitation is evidential: this paper specifies an architecture and demonstrates arithmetic consistency, but it has not established predictive validity. No synthetic sensor cubes were passed through a trained network. The residual construction bypasses optimization, missing-data processing, sensor fusion and physical dynamics. Consequently, Table 1 describes candidate inputs rather than a processed dataset, and Tables 2–3 describe prescribed prediction scenarios rather than benchmark results.

Several scientific limitations would persist even after real-data implementation. Training on ERA5-Land targets may reward agreement with reanalysis biases. A 0.1° grid cannot validate fine urban thermal contrasts. AIRS profiles may offer only limited additional information once reanalysis history is included. Landsat context can become stale during rapid land-surface change. The ResNet backbone may be computationally expensive relative to simpler neighborhood statistics, especially for a short historical window. Each limitation suggests a specific empirical control rather than a reason to assume benefit.

The architecture also contains design assumptions that could fail. Nonnegative age penalties presume that freshness is generally desirable, yet older high-quality observations may be more useful than recent noisy observations. Context-dependent compatibility can partly compensate, but calibration must be examined. Direct quality scores may not be comparable across products. The proposed scalar quality mapping should therefore be tested against source-specific embeddings and uncertainty estimates. Attention weights alone cannot determine which sensor is physically responsible for a prediction.

The methodological value of the study lies in making these assumptions inspectable. The target variable, information boundary, spatial support, normalization, comparator definitions and runtime unit are declared before claiming performance. This framework permits a future empirical study to fail informatively: a negative result could reveal redundant sources, ineffective context penalties, or a mismatch between architectural complexity and available data.

V. CONCLUSION AND FUTURE WORK

This paper presented CSTAN-CSCA, a proposed reliability-aware architecture for next-day regional temperature forecasting using Sentinel-3, Landsat-9, ERA5-Land and AIRS information. The design combines modified ResNet-34 spatial encoding, historical-window Bi-GRU sequence representation, and a two-stage attention module that exchanges information across source, space and time. Its scoring rule explicitly incorporates environmental context, spatial distance, observation age, retrieval quality and hard availability constraints. A combined regression objective and a leakage-conscious evaluation protocol complete the specification. Tables 2–3 report constructed synthetic scenarios, not trained-model benchmark results. For the scenario labelled CSTAN-CSCA, the prescribed tolerance rate is 96.8%, normalized RMSE is 0.170, normalized MAE is 0.110, and the corresponding computed R^2 is 0.9500. The 88-ms latency is an assumed profiling target; hardware and end-to-end timing were not evaluated. These values illustrate metric consistency and do not

establish empirical superiority, sensor-fusion effectiveness or operational forecasting capability.

Future work should implement the architecture, assemble versioned product manifests, and evaluate chronological and geographical holdouts under both retrospective and publication-time-aware access. ERA5-Land is the proposed primary target for the archival evaluation; success against it demonstrates agreement with reanalysis, not independent observational accuracy. Independent quality-controlled station validation remains necessary. Subsequent experiments should test capacity-matched encoders, unidirectional recurrence, source removal, contextual-score alternatives and loss interactions. Uncertainty calibration, physically meaningful error stratification, and measured end-to-end timing should accompany those comparisons. Only after those tests should the illustrative tables be replaced by observed results and the manuscript be presented as an empirical forecasting paper.

Conflicts of Interest

The authors declare that there is no conflict of interest concerning the publication of this paper.

Funding Statement

The authors received no financial support for the research or publication of this paper.

Acknowledgements

None.

VI. REFERENCES

1. K. He, et al., "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. Google Scholar | Publisher Link
2. K. Cho, et al., "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. Google Scholar | Publisher Link
3. A. Vaswani, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017. Google Scholar | Publisher Link
4. European Space Agency, "Instruments," *Copernicus Sentinel-3*, 2026. Online: http://esa.int/Science_Exploration/Space_Science/Instruments
5. U.S. Geological Survey, "Landsat 9," *Landsat Missions*, 2026. Online: https://www.usgs.gov/landsat-missions/landsat-9
6. J. Muñoz-Sabater, et al., "ERA5-Land: A State-of-the-Art Global Reanalysis Dataset for Land Applications," *Earth System Science Data*, vol. 13, pp. 4349–4383, 2021. Google Scholar | Publisher Link
7. AIRS Project, "Aqua/AIRS L2 Standard Physical Retrieval (AIRS-Only) V7.0 (AIRS2RET)," *Goddard Earth Sciences Data and Information Services Center*, Greenbelt, MD, USA, 2019. Google Scholar | Publisher Link
8. X. Shi, et al., "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," *Advances in Neural Information Processing Systems*, vol. 28, pp. 802–810, 2015. Google Scholar | Publisher Link
9. T. Nguyen, et al., "ClimaX: A Foundation Model for Weather and Climate," in *Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol. 202, pp. 25904–25938, 2023. Google Scholar | Publisher Link
10. T. Chen, and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. Google Scholar | Publisher Link