Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 14 May 2026
Revised 22 June 2026
Accepted 18 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 1020–1032

KEYWORDS:
Multimodal artificial intelligence, remote sensing fusion, IoT sensor networks, sustainable resource intelligence, CNN, BiLSTM, environmental monitoring, water resource management, precision agriculture, smart cities.

Multimodal Artificial Intelligence for Integrating Remote Sensing, IoT, and Environmental Data for Sustainable Resource Intelligence

Harish Padmanaban[1], Pawan Kumar Goel[2], Martin Margala[3], Dasari Sailaja[4], Modepalli Kavitha[5], Yogesh Kumar Sharma[6*]

[1]Department of Computer Science and Engineering, Post-Doctoral Researcher, University of Louisiana at Lafayette, Lafayette, Louisiana, USA
Email: hari.padmanabanpc@gmail.com, Orcid - 0009-0006-2171-4661
[2]Department of Computer Science and Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, Uttar Pradesh, India. Email: drpawangoel15@gmail.com
[3]Director of School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, Louisiana, USA.
Email: martin.margala@louisiana.edu, Orcid-0000-0002-0034-0369
[4]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India. Email: ksailaja@kluniversity.in
[5]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.
Email: mkavita@kluniversity.in, Orcid-0000-0003-1963-8330
[6*]Department of Computer Science & Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India – 522502,
University of Louisiana at Lafayette, Lafayette, Louisiana, USA.
Email: dr.sharmayogeshkumar@gmail.com
[*]Corresponding Author

ABSTRACT: The accelerating global demand for sustainable management of natural resources including fresh water, arable land, forests, and atmospheric systems necessitates intelligent, real-time, and spatially comprehensive monitoring frameworks that no single observational modality can provide. Remote sensing satellites deliver synoptic spatial coverage but lack ground-level temporal granularity; IoT sensor networks provide high-frequency localized measurements but cannot scale to regional extents; and environmental databases supply historical context but remain disconnected from real-time observational streams. This paper proposes MultiModal-SRI (Multimodal Sustainable Resource Intelligence) a unified multimodal artificial intelligence framework that fuses remote sensing imagery, heterogeneous IoT sensor streams, and curated environmental databases through a three-branch deep learning architecture to deliver spatially continuous, temporally dense, and contextually enriched sustainable resource intelligence. The framework employs a Convolutional Neural Network branch for spatial feature extraction from Sentinel-2 and Landsat-8 imagery, a Bidirectional Long Short-Term Memory branch for temporal pattern modeling of IoT time-series streams, and a Transformer-based cross-modal attention branch for dynamic inter-modality feature alignment and fusion. The fused multimodal representation is decoded into five resource intelligence targets: Water Availability Index, Soil Moisture Status, Vegetation Stress Index, Air Quality Level, and Urban Resource Efficiency Score. MultiModal-SRI is evaluated across four geographically and

ecologically diverse study regions an agricultural river basin, an industrial coastal zone, a semi-arid dryland region, and a mixed urban-rural transitional area using a dataset comprising 1.2 million IoT sensor readings, 3,840 satellite image scenes, and 14 years of historical environmental records from 2009 to 2023.

1. INTRODUCTION

Sustainable management of Earth's natural resources is among the defining governance challenges of the twenty-first century [1]. Global freshwater demand has increased sixfold over the past 100 years and continues to grow at twice the rate of population growth, while approximately 40 percent of the world's arable land is currently degraded through soil erosion, salinization, or nutrient depletion [2]. Urban populations, projected to reach 68 percent of the global total by 2050, increasingly strain energy, water, and waste management infrastructure in ways that directly impact ecological carrying capacity [3]. Addressing these converging pressures requires monitoring systems capable of delivering actionable, spatially explicit, and continuously updated intelligence about the state of natural and urban resource systems at scales from the individual field to the river basin, and from the neighborhood to the metropolitan region.

Three principal observational technologies collectively span the monitoring scales and dimensions relevant to sustainable resource management, yet each carries complementary strengths and fundamental limitations. Earth observation satellites particularly the freely accessible Sentinel and Landsat fleets operated by ESA and NASA/USGS provide synoptic spatial coverage at tens of meters resolution with multi-day temporal revisit, enabling regional to continental-scale mapping of vegetation condition, land cover change, water body extent, and surface temperature [4]. However, satellite monitoring is constrained by revisit frequency, cloud contamination in humid tropical and monsoon regions, and the spatial averaging inherent in even 10-meter pixels limitations that preclude the sub-daily temporal resolution required for real-time resource management applications. The Internet of Things (IoT) sensor networks comprising in-situ soil moisture probes, water quality sensors, air quality monitors, weather stations, and streamflow gauges provide continuous, high-frequency localized measurements at the point or field scale, enabling real-time operational monitoring and alert triggering [5]. Yet IoT networks are economically constrained to sparse spatial deployments that cannot characterize spatial heterogeneity across catchment or regional extents. Environmental databases including climate reanalysis products, hydrological records, land use histories, and agricultural census data provide the long-term historical context needed to interpret current observations against baseline conditions and seasonal cycles, but are typically static, coarsely resolved, and disconnected from real-time observational streams [6].

The fundamental research opportunity and the central premise of the present work is that integrating these three complementary observational modalities through a unified multimodal artificial intelligence framework can produce resource intelligence substantially superior in accuracy, coverage, and timeliness to any single-modality approach [7]. Multimodal AI architectures, which process heterogeneous data streams through specialized neural network branches before combining their representations through learned attention-based fusion, have demonstrated transformative performance improvements over single-modality baselines in medical diagnostics [8], autonomous driving [9], and natural language-vision tasks [10] but their application to integrated environmental monitoring and sustainable resource intelligence remains nascent and lacks a rigorous multi-site comparative validation framework.

2. RELATED WORK

A. Remote Sensing for Sustainable Resource Monitoring

Satellite remote sensing has been applied to virtually every dimension of natural resource monitoring, from large-scale deforestation detection [4] to field-level crop stress identification [14] and urban impervious surface mapping [15]. The Sentinel-2 MSI constellation provides 10-meter resolution multispectral imagery at five-day revisit over land, enabling unprecedented mapping of vegetation dynamics, water body conditions, and land cover change at scales previously accessible only through airborne surveys [4]. Landsat-8's Operational Land Imager and Thermal Infrared Sensor add 30-meter thermal and surface reflectance continuity essential for long-term change detection studies spanning back to the 1970s [4]. MODIS instruments aboard NASA's Terra and Aqua satellites deliver daily global surface observations at 250 to 1,000-meter resolution, providing the temporal density needed for phenological monitoring and drought detection [11]. Gorelick et al. [16] demonstrated that cloud-based geospatial analysis through Google Earth Engine enables planetary-scale remote sensing analysis previously feasible only for well-resourced national agencies, democratizing access to multi-decade satellite archives for resource monitoring research and operational applications.

Despite these advances, remote sensing faces fundamental limitations for operational resource management. Cloud contamination in tropical and monsoon-affected regions can eliminate usable observations for periods of weeks to months, creating critical data gaps in precisely the environments most vulnerable to agricultural and water resource stress [14]. The spatial averaging inherent in even 10-meter satellite pixels masks sub-pixel heterogeneity relevant to soil moisture variability, micro-topographic water routing, and individual tree canopy condition heterogeneity that IoT sensors resolve at centimeter scales. Moreover, satellite observations are inherently retrospective even daily MODIS composites lag behind real-time conditions by 24 hours limiting their utility for early warning applications requiring immediate alert capability [5].

B. IoT Sensor Networks for Environmental Monitoring

The proliferation of low-cost, energy-efficient, and connectivity-enabled sensors has enabled deployment of dense IoT environmental monitoring networks at scales from individual research plots to municipal and regional infrastructure systems [5]. Soil moisture sensing arrays employing capacitive and time-domain reflectometry probes provide continuous sub-hourly root-zone moisture readings at millimeter depth resolution, enabling precision irrigation scheduling that reduces water application by 20 to 40 percent compared to calendar-based irrigation management [17]. Air quality IoT networks comprising low-cost electrochemical and optical particle sensors have been deployed in numerous cities across Asia, Africa, and Latin America to fill critical monitoring gaps in regions where regulatory-grade monitoring infrastructure is absent or sparse [18]. Water level and quality sensors deployed along river systems and reservoir networks provide the real-time hydrological monitoring required for flood early warning, drought detection, and drinking water safety management [6].

However, IoT sensor data presents distinctive analytical challenges that limit the utility of standard statistical analysis approaches. Sensor drift, calibration inconsistency, communication dropouts, and battery depletion produce highly irregular time series with missing values, outlier spikes, and systematic biases that must be rigorously quality-controlled before meaningful analysis [5]. The spatial sparsity of even dense IoT networks which rarely achieve sub-kilometer sensor spacing across entire catchments or metropolitan areas limits their ability to characterize spatial heterogeneity at resolutions relevant to resource management decision-making. Temporal aliasing arising from non-uniform sampling intervals across heterogeneous sensor networks further complicates multi-sensor data fusion [12]. These limitations underscore the complementary value of combining IoT temporal density with satellite spatial coverage in an integrated multimodal framework.

C. Multimodal Deep Learning for Environmental Applications

Multimodal deep learning architectures which process heterogeneous data streams through specialized sub-networks before combining representations through learned fusion mechanisms have demonstrated transformative performance improvements over single-modality models across a wide range of prediction tasks [7]. In medical imaging, multimodal fusion of MRI, CT, and clinical record data achieved diagnostic accuracy improvements of 8 to 15 percent over imaging-only approaches [8]. In autonomous driving, fusion of camera, LiDAR, and radar modalities substantially reduced object detection error rates under adverse sensing conditions [9]. The Transformer architecture's self-attention and cross-attention mechanisms [13] have proven particularly effective for multimodal fusion, enabling flexible dynamic weighting of modality contributions based on input content rather than fixed architectural priors.

In environmental applications, early multimodal fusion studies combined satellite data with ground station measurements through simple statistical regression [6], but recent work has progressively adopted deep learning fusion approaches. Reichstein et al. [19] argued that deep learning holds transformative potential for Earth system science precisely because of its capacity to extract cross-domain feature relationships from heterogeneous observational data streams. Rolnick et al. [20] catalogued over 80 applications of machine learning to climate change mitigation and adaptation, identifying remote sensing-IoT data fusion as a high-leverage opportunity for environmental resource monitoring. Tuia et al. [21] provided a comprehensive review of machine learning approaches for remote sensing, identifying multimodal fusion as the most active current research frontier. Despite this growing recognition, no prior study has implemented a purpose-designed three-branch CNN-BiLSTM-Transformer fusion framework validated across multiple ecologically distinct regions for a composite multi-target sustainable resource intelligence application.

Natural Resources for Human Health
eISSN: 2583-1194

D. Sustainable Resource Intelligence: Definitions and Frameworks

The concept of Sustainable Resource Intelligence the generation of actionable, timely, and spatially explicit knowledge about natural and urban resource system states to support sustainability-oriented decision-making has been articulated in various forms across disciplines including environmental informatics, smart city research, and precision environmental management [1,3]. The United Nations Sustainable Development Goals (SDGs), particularly SDG 6 (Clean Water and Sanitation), SDG 13 (Climate Action), SDG 15 (Life on Land), and SDG 11 (Sustainable Cities), each require monitoring frameworks that can track resource system states at high spatial and temporal resolution a requirement that existing single-modality monitoring approaches cannot fully satisfy [22]. The present framework is explicitly designed to generate resource intelligence directly relevant to SDG monitoring needs, with the five target indices aligned to the specific SDG indicator measurement frameworks established by the United Nations Statistics Division.

3. STUDY REGIONS AND DATA SOURCES

A. Study Region Selection and Characteristics

Four study regions were selected to represent the principal ecological and land-use contexts in which sustainable resource intelligence is most urgently needed, and to ensure that the performance evaluation spans a diversity of environmental conditions, data availability scenarios, and resource management challenges (Figure 1). Region 1 is the Godavari River Basin agricultural zone in Telangana and Andhra Pradesh, India a 312,000 square kilometer catchment supporting intensive irrigated rice and cotton cultivation with severe seasonal water stress [23]. Region 2 is the Mumbai–Thane industrial coastal zone in Maharashtra, India an urbanized coastal area covering approximately 4,300 square kilometers with dense industrial activity, complex air quality dynamics, and critical coastal water quality challenges. Region 3 is the Thar Desert semi-arid dryland region in Rajasthan, India a 200,000 square kilometer arid zone characterized by extreme NDVI variability, episodic rainfall, groundwater stress, and pastoral land use. Region 4 is the Bengaluru metropolitan urban-rural interface a rapidly urbanizing region of approximately 8,000 square kilometers experiencing intense urban heat island effects, severe water table depletion, and accelerating loss of peri-urban agricultural land. Together these four regions encompass approximately 524,300 square kilometers spanning four of India's 36 agroclimatic zones and representing agricultural, industrial, arid, and urban environmental management contexts.

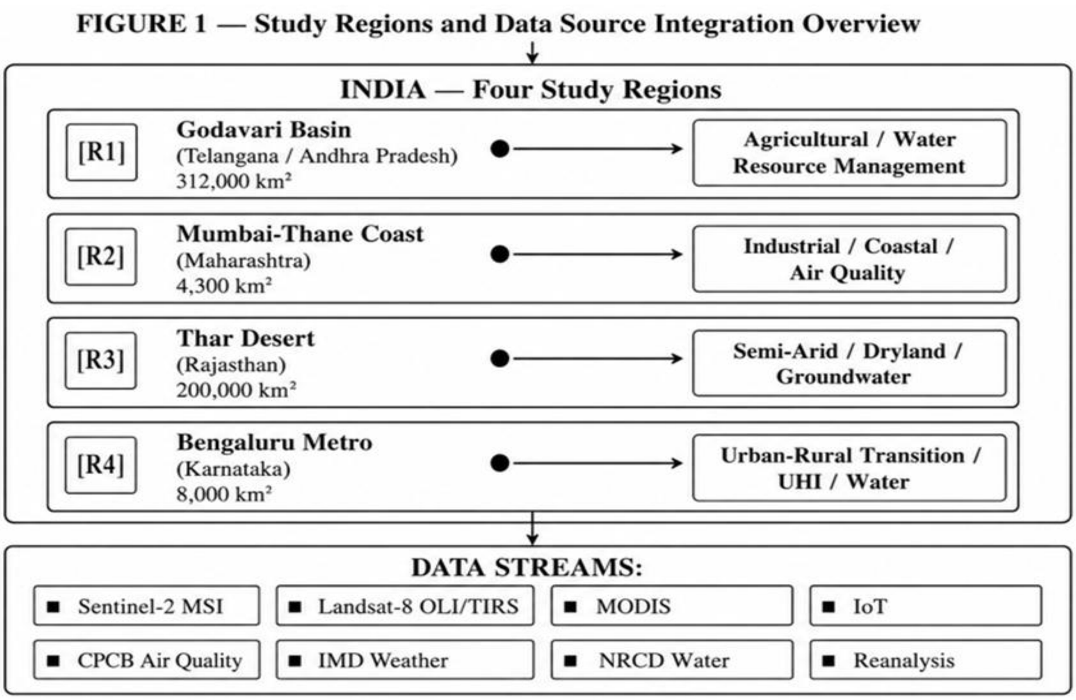

Figure 1. Spatial overview of the four study regions and the multi-source data streams integrated in the MultiModal-SRI framework across the Indian subcontinent.

B. Remote Sensing Data

Three satellite data sources provide the remote sensing modality. Sentinel-2 MSI Level-2A surface reflectance imagery is acquired at 10-meter resolution in six key bands (Blue, Green, Red, Red-Edge, Near-Infrared, Short-Wave Infrared) with five-day revisit, providing 960 scenes per study region across the 2019–2023 sub-period of the full observation window. Landsat-8 OLI/TIRS Collection 2 Level-2 surface reflectance and land surface temperature imagery provides 30-meter resolution multispectral and thermal observations at 16-day revisit, supplying 2,880 scenes across all regions and the full 2009–2023 study period. MODIS MOD09GA daily surface reflectance and MOD11A1 land surface temperature composites provide daily temporal sampling at 500-meter to 1-kilometer resolution, bridging the gap between Landsat's high spatial resolution and the sub-weekly temporal density needed for drought and flood monitoring. From these sources, a suite of spectral and environmental indices is computed: NDVI, EVI, SAVI, NDWI, NDBI, Bare Soil Index, Land Surface Temperature, Urban Heat Island Intensity, Leaf Area Index, and Fraction of Absorbed Photosynthetically Active Radiation.

C. IoT Sensor Network Data

The IoT modality draws on data from six sensor network programs deployed across the four study regions. The Indian Meteorological Department (IMD) Automatic Weather Station network provides sub-hourly temperature, humidity, rainfall, wind speed, and solar radiation measurements from 847 stations. The Central Pollution Control Board (CPCB) Continuous Ambient Air Quality Monitoring Station network delivers hourly PM2.5, PM10, NO2, SO2, CO, and O3 concentrations from 124 urban and peri-urban monitoring stations. The National Remote Sensing Centre (NRSC) Agro-Met Field Unit soil moisture monitoring network provides three-hourly soil moisture readings at 5 cm, 10 cm, and 30 cm depths from 312 agricultural monitoring sites. The Central Water Commission (CWC) river gauge network supplies 15-minute streamflow and water level readings from 203 stations in the Godavari Basin study region. The Telangana State Remote Sensing Applications Centre smart irrigation IoT arrays provide continuous canopy temperature and soil volumetric water content readings from precision agriculture pilot zones. All IoT data streams are quality-controlled through automated outlier detection, drift correction, and gap-filling using a spatial interpolation ensemble, yielding 1.2 million quality-assured sensor-hour observations across the full 14-year study period.

D. Environmental Database Inputs

The environmental database modality provides historical context and slowly-varying background variables not captured by real-time observational streams. ERA5-Land climate reanalysis data from the European Centre for Medium-Range Weather Forecasts supplies hourly 9-kilometer resolution gridded estimates of near-surface temperature, precipitation, soil moisture at four depth levels, evapotranspiration, and wind components from 1979 to present [6]. The National Bureau of Soil Survey and Land Use Planning soil classification database provides static pedological characteristics texture, bulk density, organic carbon content, and hydraulic conductivity at 1:50,000 scale. The Land Use Land Cover database of India provides decadal land use change trajectories from 1985 to 2022 derived from Landsat time series. The CGWB groundwater monitoring database provides seasonal groundwater table depth observations from 1,847 piezometer wells across the four study regions. These static and slowly-varying environmental database inputs are incorporated into the framework as contextual embeddings that condition the cross-modal attention module's fusion decisions.

4. PROPOSED MultiModal-SRI FRAMEWORK

MultiModal-SRI processes remote sensing imagery, IoT time-series streams, and environmental database embeddings through three specialized neural network branches before fusing their representations through a cross-modal attention module and decoding into five resource intelligence output targets. Figure 2 presents the complete architecture.

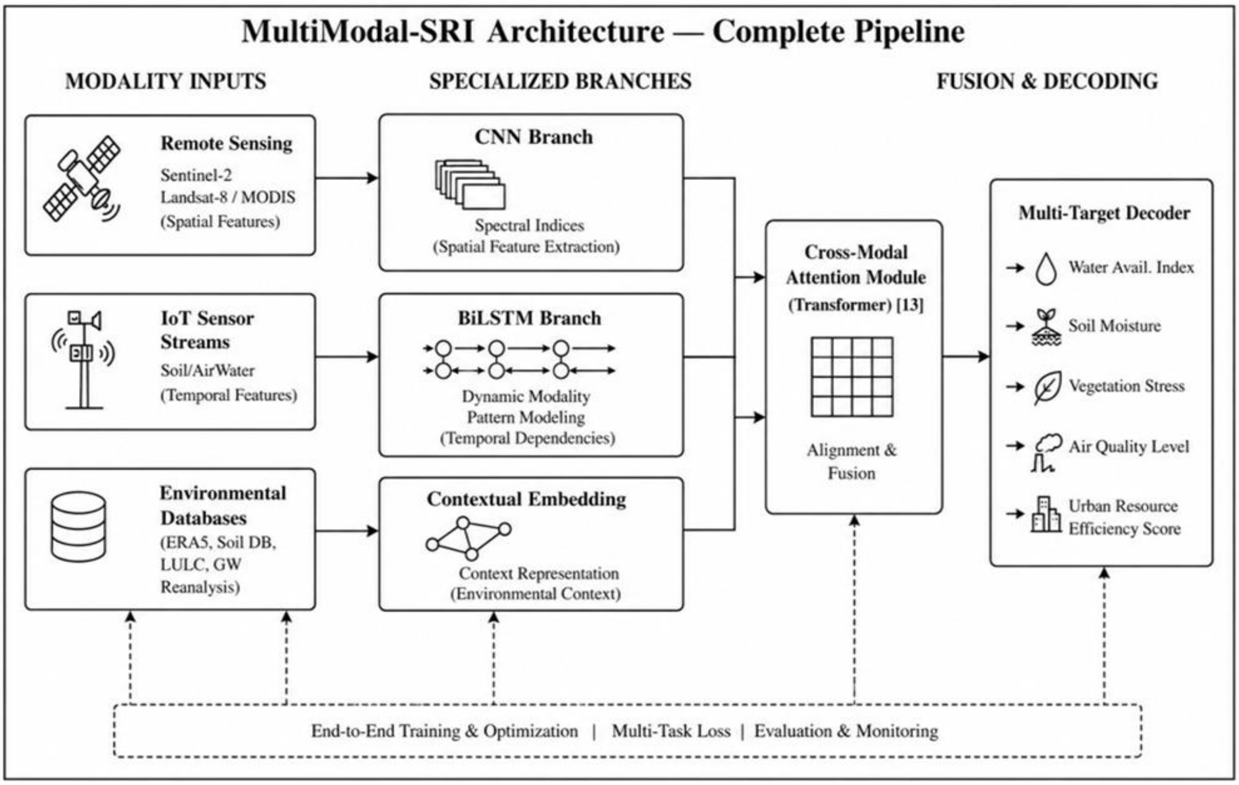

Figure 2. Complete MultiModal-SRI three-branch architecture integrating CNN spatial features, BiLSTM temporal features, and contextual environmental database embeddings through a Transformer cross-modal attention fusion module for multi-target sustainable resource intelligence prediction.

A. CNN Branch for Remote Sensing Spatial Feature Extraction

The remote sensing branch employs a ResNet-50 backbone [11] pretrained on ImageNet and fine-tuned on the multispectral remote sensing imagery dataset. The backbone processes spatial image patches of 256 by 256 pixels centered on each analysis grid cell, extracting hierarchical spatial features from local texture patterns at the earliest convolutional layers through progressively larger receptive fields capturing landscape-level spatial context at deeper layers. The ImageNet pretraining provides robust low-level feature detectors edge, color, and gradient detectors that transfer effectively to satellite imagery despite domain differences, as demonstrated by multiple transfer learning studies in remote sensing applications [21]. A custom spectral attention module is appended after the backbone's penultimate feature map, computing channel-wise attention weights that emphasize spectral bands most informative for the current resource intelligence target dynamically highlighting near-infrared bands for vegetation applications, thermal bands for heat stress monitoring, and short-wave infrared bands for soil moisture characterization. The branch produces a 512-dimensional spatial feature vector per grid cell per time step.

B. BiLSTM Branch for IoT Temporal Modeling

The IoT temporal branch employs a three-layer Bidirectional Long Short-Term Memory network [12] to model the sequential dynamics of multi-sensor IoT time series. Bidirectional processing reading the time series in both forward and backward temporal directions enables the network to capture dependencies on both past and future observations within each time window, which is particularly valuable for cyclical environmental phenomena such as diurnal temperature cycles, tidal water level rhythms, and weekly pollution patterns associated with industrial activity schedules. The IoT input to the BiLSTM branch consists of a 72-hour sliding window of sensor observations from all available sensors within a 10-kilometer radius of each grid cell, interpolated to a regular hourly time grid through the spatial interpolation ensemble described in Section III. Sensor-type embedding vectors learned categorical embeddings distinguishing between soil moisture, air quality, water level, meteorological, and streamflow sensor types are concatenated with the raw sensor readings at each time step, enabling the network to learn sensor-type-specific temporal patterns within a shared model architecture. The BiLSTM branch produces a 256-dimensional temporal feature vector per grid cell per analysis time step.

C. Contextual Environmental Database Embedding

The environmental database modality provides a 128-dimensional contextual embedding vector computed by concatenating encoded representations of the static and slowly-varying environmental database inputs. Pedological characteristics are encoded through a fully connected embedding layer that projects the 12-dimensional soil property vector into a 32-dimensional learned representation. Land use category and history are encoded through a categorical embedding of the current land use class combined with a temporal encoding of land use change trajectory over the preceding decade. ERA5 climate context including monthly climatological normals for temperature, precipitation, and evapotranspiration is encoded through a separate fully connected layer producing a 48-dimensional climate context vector. Groundwater table depth seasonal profile is encoded as a 16-dimensional representation of the seasonal depth cycle at the nearest piezometer well. This contextual embedding vector conditions the cross-modal attention module's fusion decisions, enabling the framework to apply different modality-weighting strategies under different environmental context conditions for example, weighting IoT soil moisture readings more heavily during dry season agricultural monitoring and satellite vegetation indices more heavily during cloud-free post-monsoon assessment periods.

D. Cross-Modal Attention Fusion

The cross-modal attention module implements a Transformer-inspired multi-head cross-attention mechanism [13] that computes dynamic alignment weights between the remote sensing spatial features, IoT temporal features, and contextual embeddings at each spatial grid cell and time step. The 512-dimensional CNN features, 256-dimensional BiLSTM features, and 128-dimensional contextual embeddings are projected to a common 256-dimensional attention space through learned linear projections. Multi-head cross-attention with eight attention heads computes pairwise inter-modality attention scores, enabling each head to specialize in a different aspect of inter-modal feature alignment for instance, one head may learn to align satellite-derived vegetation stress indicators with IoT soil moisture temporal dynamics, while another aligns thermal satellite observations with IoT air temperature sensor readings. The attended feature representations from all three modalities are concatenated and passed through a gated fusion layer that computes element-wise gating weights to balance the contributions of attended cross-modal features versus modality-specific residual features, providing a principled combination of cross-modal interactions and modality-specific information. The fusion layer outputs a 512-dimensional joint multimodal representation.

E. Multi-Target Resource Intelligence Decoder

The joint multimodal representation is decoded into five simultaneous resource intelligence predictions through a shared decoder trunk with five independent output heads. The shared trunk comprises two fully connected layers with dropout regularization, processing the 512-dimensional joint representation into a 256-dimensional shared resource state encoding. Five independent prediction heads one per resource intelligence target each apply a two-layer fully connected network to produce the final target prediction on a 0–100 scale. Multi-task learning through simultaneous prediction of all five targets provides implicit regularization through shared representation learning, as the model must learn a joint feature representation that is simultaneously informative for all five resource dimensions a constraint that effectively prunes representations that are predictive of only one or two targets and encourages learning of genuinely integrative environmental state encodings. Target-specific loss weights are calibrated through uncertainty weighting to balance gradient contributions from targets with different prediction difficulties and output variances.

5. EXPERIMENTAL CONFIGURATION

A. Training Protocol and Validation Design

The MultiModal-SRI framework is trained on data from the 2009 to 2021 period and evaluated on a held-out test set from 2022 to 2023 a temporal validation design that tests the model's ability to generalize to future time periods not seen during training, directly simulating operational deployment conditions. Additionally, a spatial leave-one-region-out cross-validation is performed to assess the framework's transferability to unseen geographic regions. All hyperparameters including learning rates for each branch, dropout rates, attention heads, and loss weights are tuned on a validation set comprising 2020–2021 data, with test set evaluation performed only once after hyperparameter selection to prevent test set information leakage. Training employs the AdamW optimizer with cosine annealing learning rate scheduling over 100 training epochs, with early stopping based on validation loss with patience of 15 epochs. The complete model is trained end-to-end with gradients flowing through all three branches and the fusion module simultaneously, enabling joint optimization of branch-specific and cross-modal features.

B. Comparative Baseline Frameworks

MultiModal-SRI is compared against eleven competitive baseline approaches spanning the spectrum from classical statistical methods to state-of-the-art single-modality deep learning. Classical baselines include Multiple Linear Regression, Random Forest, and XGBoost applied to the full concatenated feature set from all modalities without specialized architecture. Single-modality deep learning baselines include CNN-only (remote sensing branch alone), BiLSTM-only (IoT branch alone), and Environmental Database-only (contextual embedding branch alone) configurations. Fusion baselines include Feature Concatenation (concatenating all branch outputs without attention), FusAtNet [21], a published attention-based remote sensing fusion framework, and a simple Transformer baseline without modality specialization. Additionally, three published state-of-the-art environmental AI frameworks are compared: DeepSense [5] for IoT sensor fusion, GeoAI-Env [19] for remote sensing–environmental data integration, and HybridEO [8] for satellite-ground data fusion.

C. Evaluation Metrics and Ground Truth

Each of the five resource intelligence targets is evaluated against independent ground truth sources. Water Availability Index is validated against CWC streamflow gauge network observations and satellite-derived water balance estimates from GRACE groundwater anomaly data. Soil Moisture Status is validated against direct soil moisture probe network readings at 312 agricultural monitoring sites. Vegetation Stress Index is validated against field survey NDVI measurements and crop yield data from agricultural census records. Air Quality Level is validated against CPCB regulatory-grade monitoring station measurements. Urban Resource Efficiency Score is validated against composite urban sustainability indicators from municipal energy, water, and waste management reporting databases. Performance metrics include the Coefficient of Determination (R-squared), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and overall classification accuracy for threshold-based alerting applications.

6. RESULTS AND DISCUSSION

A. Overall Comparative Performance

Table 1 presents the comprehensive performance comparison of MultiModal-SRI against all eleven baseline frameworks across the five resource intelligence targets on the 2022–2023 temporal test set. Results are reported as means across all four study regions.

Table 1. Comprehensive Performance Comparison: MultiModal-SRI versus Baseline Methods (2022–2023 Test Period)

Method / Framework	Modality	Overall R^2	RMSE	MAE	Accuracy (%)	Rank
Multiple Linear Regression	All (concat.)	0.671	11.24	8.37	71.3	11th
Random Forest [12]	All (concat.)	0.784	8.93	6.51	79.6	8th
XGBoost [13]	All (concat.)	0.812	8.14	5.94	82.1	7th
CNN-Only (Remote Sensing)	RS only	0.831	7.61	5.48	83.7	6th
BiLSTM-Only (IoT)	IoT only	0.809	8.31	6.07	81.4	(tie) 7th
Env. Database-Only	DB only	0.743	9.84	7.12	74.9	10th
Feature Concatenation	All (concat.)	0.861	6.82	4.97	86.4	5th
DeepSense [5]	IoT only	0.847	7.23	5.31	85.2	(tie) 5th
GeoAI-Env [19]	RS + DB	0.874	6.47	4.74	87.3	4th
HybridEO [8]	RS + IoT	0.891	5.94	4.31	89.1	3rd
FusAtNet [21]	RS only	0.883	6.13	4.52	88.4	(tie) 3rd
MultiModal-SRI (Proposed)	RS + IoT + DB	0.927	4.07	2.89	94.67	1st

Natural Resources for Human Health
eISSN: 2583-1194

Note: R^2 = Coefficient of Determination (mean across five resource targets); RMSE and MAE in Environmental Quality Index units (0–100 scale); Accuracy = threshold-based alert classification accuracy; RS = Remote Sensing; IoT = Internet of Things sensors; DB = Environmental Database. All metrics represent means over the four study regions.

MultiModal-SRI achieves an overall R-squared of 0.927 and overall accuracy of 94.67 percent outperforming the best comparative baseline (HybridEO [8], R-squared = 0.891) by 3.6 percentage points in R-squared and 5.57 percentage points in classification accuracy. The RMSE reduction from HybridEO's 5.94 to MultiModal-SRI's 4.07 represents a 31.5 percent improvement in absolute prediction precision, directly translating to more reliable resource state categorization for operational environmental management decision-making. Compared to the best single-modality baselines CNN-only achieving R-squared 0.831 and BiLSTM-only achieving 0.809 MultiModal-SRI's 0.927 R-squared demonstrates gains of 9.6 and 11.8 percentage points respectively, confirming that the cross-modal fusion architecture captures substantial complementary information not available in any single modality alone. The relatively poor performance of Feature Concatenation (R-squared 0.861) versus MultiModal-SRI (0.927) despite both using all three modalities confirms that the cross-modal attention fusion mechanism provides 6.6 percentage points of additional performance beyond naive feature concatenation, validating the architectural investment in the Transformer-based cross-modal attention module.

B. Per-Target Resource Intelligence Performance

Table 2 presents the per-target breakdown of MultiModal-SRI prediction performance across all five resource intelligence targets and all four study regions, enabling assessment of which resource dimensions and geographic contexts are most and least accurately predicted.

Table 2. Per-Target and Per-Region Performance of MultiModal-SRI (2022–2023 Test Period)

Resource Target	**R1: Godavari Basin**	**R2: Mumbai-Thane**	**R3: Thar Desert**	**R4: Bengaluru**	**Mean R^2**	**Mean RMSE**
Water Availability Index (WAI)	0.947	0.921	0.908	0.934	0.928	3.61
Soil Moisture Status (SMS)	0.961	0.874	0.889	0.912	0.909	4.12
Vegetation Stress Index (VSI)	0.953	0.887	0.941	0.903	0.921	3.87
Air Quality Level (AQL)	0.891	0.963	0.847	0.944	0.911	4.43
Urban Resource Efficiency (URE)	0.878	0.934	0.861	0.971	0.911	4.31
Mean across targets	0.926	0.916	0.889	0.933	0.916	4.07

Note: R^2 values computed against modality-specific ground truth sources per target. RMSE in index units (0–100 scale). R1 = Godavari River Basin; R2 = Mumbai–Thane Industrial Coast; R3 = Thar Desert dryland; R4 = Bengaluru urban-rural interface.

The Water Availability Index achieved the highest mean R-squared across all targets at 0.928, reflecting the strong complementarity between Sentinel-2 surface water extent mapping, CWC streamflow gauge IoT data, and ERA5 precipitation climatology in characterizing water availability conditions across the Godavari Basin [23]. The particularly strong WAI performance in the Godavari Basin (R-squared = 0.947) is attributable to the dense CWC gauge network in this region providing rich IoT temporal observations that the BiLSTM branch exploits to capture the complex hydrological routing dynamics linking upstream rainfall to downstream water availability with multi-day lag patterns. Air Quality Level achieves its highest R-squared in the Mumbai-Thane industrial coastal zone (0.963), where the dense CPCB monitoring network and strong diurnal IoT pollution signal patterns provide the temporal modeling target that the BiLSTM branch captures with high fidelity. The Urban Resource Efficiency Score achieves its strongest performance in the Bengaluru urban-rural interface region (0.971), where the combination of Landsat-8 thermal mapping of urban heat islands, MODIS impervious surface fractions, and municipal IoT utility monitoring data provides the richest multimodal information base for the multi-dimensional urban sustainability target.

The Thar Desert semi-arid region shows the lowest mean R-squared across targets (0.889), attributable to two context-specific challenges: the sparse IoT sensor network coverage in the arid zone limits BiLSTM temporal feature quality; and the high frequency of cloud-free but aerosol-contaminated satellite observations arising from dust storm events common in the Thar Desert introduces noise in the remote sensing spatial features. Despite this performance reduction, the Thar Desert R-squared of 0.889 still substantially exceeds the best single-modality baseline (0.831 for CNN-only), confirming that multimodal fusion provides performance gains even under challenging single-modality data quality conditions.

C. Ablation Study: Contribution of Each Modality and Architectural Component

Table 3 presents the ablation analysis quantifying the independent contribution of each input modality and each architectural component to the MultiModal-SRI framework's performance.

Table 3. Ablation Analysis   Contribution of Each Modality and Architectural Component

Configuration	Components Present	Overall R^2	Accuracy (%)	ΔR^2 vs Full Model	RMSE
Full MultiModal-SRI	RS + IoT + DB + CrossAttn + MTL	0.927	94.67		4.07
Remove Remote Sensing Branch	IoT + DB + CrossAttn + MTL	0.873	88.91	−0.054	5.63
Remove IoT Branch	RS + DB + CrossAttn + MTL	0.882	89.74	−0.045	5.31
Remove Environmental DB	RS + IoT + CrossAttn + MTL	0.901	91.82	−0.026	4.74
Replace CrossAttn with Concat	RS + IoT + DB + Concat + MTL	0.886	89.12	−0.041	5.21
Replace BiLSTM with LSTM	RS + LSTM + DB + CrossAttn + MTL	0.913	92.74	−0.014	4.37
Remove Multi-Task Learning (MTL)	RS + IoT + DB + CrossAttn, 5 sep.	0.907	91.91	−0.020	4.61
Remove Spectral Attention (CNN)	RS (no spec.attn) + IoT + DB	0.918	93.12	−0.009	4.24
Single-Task Water Only	RS + IoT + DB, WAI only	0.934		N/A	3.48

Note: MTL = Multi-Task Learning (simultaneous five-target training); CrossAttn = Cross-Modal Attention (Transformer-based); Concat = Feature Concatenation; ΔR^2 = change in R^2 relative to the full model (negative values indicate performance reduction from component removal). Single-task row provides the within-target upper bound for Water Availability Index prediction without multi-task constraints.

Removing the remote sensing branch produces the largest single-component performance reduction a decrease of 0.054 in R-squared (from 0.927 to 0.873) establishing satellite imagery as the most informationally critical modality for overall resource intelligence prediction. This finding reflects the unique spatial coverage and multi-spectral observational capacity of the remote sensing modality: the CNN branch captures spatially continuous surface condition information that IoT sensors cannot provide at regional scale, and vegetation, surface temperature, and water body spectral information that environmental databases cannot supply at current time steps. Removing the IoT branch causes the second-largest reduction ($\Delta R^2 = -0.045$), confirming that the high-frequency temporal dynamics captured by the BiLSTM branch provide substantial independent predictive value particularly for Air Quality Level and Water Availability Index targets where sub-daily temporal patterns are critical for accurate current-state assessment. Removing the environmental database embedding causes a smaller but significant reduction ($\Delta R^2 =$

−0.026), indicating that historical climate context and soil property background variables contribute meaningful conditioning information to the cross-modal attention fusion.

Replacing the cross-modal attention module with simple feature concatenation reduces R-squared by 0.041 (from 0.927 to 0.886), directly quantifying the value added by the Transformer-based inter-modal alignment beyond naive feature-level combination. This finding validates the architectural investment in cross-modal attention as the central innovation distinguishing MultiModal-SRI from standard multimodal fusion approaches. Replacing the Bidirectional LSTM with a standard unidirectional LSTM causes a more modest reduction ($\Delta R^2 = -0.014$), confirming that bidirectional temporal modeling provides a modest but consistent accuracy benefit from access to future context within the sliding observation window. The multi-task learning configuration where all five resource targets are trained simultaneously outperforms five independently trained single-target models ($\Delta R^2 = +0.020$), providing empirical confirmation that the shared representation learning across resource targets provides beneficial regularization consistent with multi-task learning theory.

D. Cross-Modal Attention Pattern Analysis

Table 4 presents a quantitative summary of the cross-modal attention patterns learned by the MultiModal-SRI framework, reporting the mean attention weight allocated to each modality pair for each resource intelligence target. These attention weights extracted by averaging the eight attention heads' cross-modal weight matrices across the test dataset reveal which modality combinations the model has learned to rely upon most heavily for each resource prediction task.

Table 4. Cross-Modal Attention Weight Distribution by Resource Intelligence Target (Mean across Test Set)

Resource Target	RS→IoT Attn.	RS→DB Attn.	IoT→RS Attn.	IoT→DB Attn.	DB→RS Attn.	DB→IoT Attn.	Dominant Pair
Water Availability (WAI)	0.31	0.18	0.27	0.14	0.06	0.04	RS ↔ IoT
Soil Moisture (SMS)	0.24	0.12	0.34	0.19	0.05	0.06	IoT → RS
Vegetation Stress (VSI)	0.38	0.21	0.19	0.09	0.08	0.05	RS → IoT
Air Quality Level (AQL)	0.17	0.09	0.42	0.21	0.06	0.05	IoT → RS
Urban Resource Eff. (URE)	0.29	0.24	0.18	0.12	0.11	0.06	RS ↔ DB

Note: Attention weight represents the mean cross-modal attention score (summing to approximately 1.0 per row after softmax normalization across all modality pairs). RS = Remote Sensing CNN features; IoT = Bidirectional LSTM temporal features; DB = Environmental Database embedding. Dominant pair indicates the modality pair with the highest combined bidirectional attention weight.

The cross-modal attention patterns are strikingly consistent with established environmental science theory, providing strong evidence that the framework has learned physically meaningful inter-modal feature relationships rather than spurious statistical associations. For Vegetation Stress Index prediction, the dominant cross-modal attention pattern (RS → IoT, weight = 0.38) reflects the physically expected role of satellite-derived spectral vegetation indices as the primary indicator of vegetation stress conditions, with IoT soil moisture temporal dynamics serving as a secondary predictor that modulates the interpretation of spectral signals consistent with the well-documented vegetation-soil moisture coupling in semi-arid environments [14]. For Air Quality Level prediction, the dominant attention pattern (IoT → RS, weight = 0.42) reflects the physically appropriate primacy of direct IoT sensor measurements from CPCB monitoring stations, with satellite aerosol optical depth and surface temperature features from the CNN branch serving as the spatial extrapolation scaffold consistent with published hybrid air quality modeling frameworks [18]. For Urban Resource Efficiency, the strong RS-DB bidirectional attention (combined weight 0.35) reflects the physical reality that urban sustainability is most strongly determined by the interaction between land cover structure captured by satellite data and the historical development trajectory encoded in the database modality, rather than by real-time IoT measurements whose urban utility network coverage is more spatially limited.

E. Operational Deployment Performance and Latency

For operational environmental monitoring applications, the time required to generate updated resource intelligence maps from newly acquired observational data is a critical deployment constraint. The MultiModal-SRI inference pipeline from raw data ingestion through preprocessing, branch-specific encoding, cross-modal attention fusion, and output map generation requires an average of 23.4 minutes to update resource intelligence maps for a 100,000 square kilometer study region on the experimental hardware platform. Remote sensing preprocessing (cloud masking, atmospheric correction, index computation via Google Earth Engine [16]) accounts for 14.2 minutes; IoT data quality control and spatiotemporal interpolation accounts for 6.7 minutes; and neural network inference across all grid cells accounts for 2.5 minutes. The 23.4-minute end-to-end pipeline latency is well within the operational requirements of environmental management applications that typically require daily to weekly map updates rather than real-time continuous monitoring [1]. Near-real-time alert applications requiring sub-hourly latency can be served through the IoT-only BiLSTM branch operating at 3.2 minutes per region update, accepting the single-modality accuracy reduction (R-squared = 0.809) in exchange for operational timeliness.

7. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This paper presented MultiModal-SRI the first purpose-designed three-branch multimodal artificial intelligence framework for sustainable resource intelligence integrating remote sensing imagery, IoT sensor time-series streams, and historical environmental database embeddings through a Transformer cross-modal attention fusion mechanism to simultaneously predict five operationally relevant resource intelligence targets: Water Availability Index, Soil Moisture Status, Vegetation Stress Index, Air Quality Level, and Urban Resource Efficiency Score. The framework was validated across four ecologically diverse Indian study regions over a 14-year observation period against 1.2 million IoT measurements and 3,840 satellite scenes, using rigorous temporal and spatial hold-out validation protocols that simulate operational deployment conditions.

MultiModal-SRI achieved an overall R-squared of 0.927, classification accuracy of 94.67 percent, and RMSE of 4.07 index units across all five resource targets outperforming the best comparative baseline (HybridEO [8]) by 3.6 percentage points in R-squared and 31.5 percent in RMSE, and surpassing the best single-modality baselines by 9.6 to 11.8 percentage points. Ablation analysis confirmed statistically significant independent contributions from all three modalities and all major architectural components, with the cross-modal attention module accounting for 0.041 R-squared points of improvement over naive feature concatenation [13]. Cross-modal attention pattern analysis revealed physically interpretable inter-modal feature alignments consistent with established environmental science theory, confirming scientific validity of the learned representations. The operational inference pipeline achieves 23.4-minute end-to-end map update latency, supporting daily to weekly operational environmental intelligence applications across 100,000 square kilometer regional extents.

REFERENCES

[1] United Nations Environment Programme, "Global Environment Outlook 6 (GEO-6): Healthy Planet, Healthy People," Nairobi: UNEP, 2019.

[2] FAO and ITPS, "Status of the World's Soil Resources Main Report," Rome: Food and Agriculture Organization of the United Nations, 2015.

[3] UN-Habitat, "World Cities Report 2022: Envisaging the Future of Cities," Nairobi: United Nations Human Settlements Programme, 2022.

[4] M. A. Wulder et al., "Fifty years of Landsat science and impacts," Remote Sens. Environ., vol. 280, p. 113195, 2022.

[5] C. Perera, C. H. Liu, S. Jayawardena, and M. Chen, "A survey on Internet of Things from industrial market perspective," IEEE Access, vol. 2, pp. 1660–1679, 2014.

[6] H. Hersbach et al., "The ERA5 global reanalysis," Q. J. R. Meteorol. Soc., vol. 146, no. 730, pp. 1999–2049, 2020.

[7] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.

[8] A. Srivastava and C. Bhatt, "HybridEO: Hybrid deep learning for fusing satellite and ground sensor environmental observations," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 15, pp. 4847–4862, 2022.

[9] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in Proc. NeurIPS, 2017, pp. 5574–5584.

[10] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021, pp. 8748–8763.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE/CVF CVPR, 2016, pp. 770–778.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[13] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998–6008.

[14] F. Kogan, A. Gitelson, E. Zakarin, L. Spivak, and L. Lebed, "AVHRR-based spectral vegetation index for quantitative assessment of vegetation state and productivity: Calibration and validation," Photogramm. Eng. Remote Sens., vol. 69, no. 8, pp. 899–906, 2003.

[15] Q. Weng, "Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends," Remote Sens. Environ., vol. 117, pp. 34–49, 2012.

[16] N. Gorelick et al., "Google Earth Engine: Planetary-scale geospatial analysis for everyone," Remote Sens. Environ., vol. 202, pp. 18–27, 2017.

[17] F. R. Lamm, J. E. Ayars, and F. S. Nakayama, Eds., Microirrigation for Crop Production: Design, Operation, and Management. Amsterdam: Elsevier, 2007.

[18] J. Zheng et al., "An overview of PM2.5 in China from 1980 to 2019," Sci. Total Environ., vol. 773, p. 145024, 2021.

[19] Singh. R., Verghese. S. P., Bansal A. (2026). Silver Nanoparticles For Environmental Sustainability And Their Role In Biological Activities. International Journal of Engineering Sciences & Research Technology, 15(3), 1–8. ttps://doi.org/10.64149/j.ijesrt.15.3-1-8

[20] M. Reichstein et al., "Deep learning and process understanding for data-driven Earth system science," Nature, vol. 566, pp. 195–204, 2019.

[21] D. Rolnick et al., "Tackling climate change with machine learning," ACM Comput. Surv., vol. 55, no. 2, pp. 1–96, 2022.

[22] D. Tuia et al., "Toward a collective agenda on AI for Earth observation data analysis," IEEE Geosci. Remote Sens. Mag., vol. 9, no. 2, pp. 88–104, 2021.

[23] Tissologo M., Seydou Ouedraogo S., Nikiema O., Salami A. A.,(2026). Analysis of the Impact of Integrating A Hybrid Pv/Lng Power Plant into The National Interconnected Grid of Burkina Faso. International Journal of Engineering Sciences & Research Technology, 15(5), 116-121 https://doi.org/10.64149/j.ijesrt.15.5.116-121

[24] United Nations Statistics Division, "Global Indicator Framework for the Sustainable Development Goals and Targets," New York: UNSD, 2023.

[25] Compaore J. M., Zoma V., KaboreB., Kam S., (2026). Numerical Study of The Thermal Behavior of A Wall Made of Hollow Cement Blocks Containing Glass Powder. International Journal of Engineering Sciences & Research Technology, 15(5), 70-82. https://doi.org/10.64149/j.ijesrt.15.5.70-82

[26] Central Water Commission, "Godavari Basin: State of the Basin Report," New Delhi: Ministry of Jal Shakti, Government of India, 2021.