Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 01| Issue- 05| 2022

Research Article **Received:** 18-04-2022 | **Accepted:** 11-05-2022 | **Published:** 28-05-2022

A Unified Data Warehouse Architecture for Multi-Source Forest Inventory Integration and Automated Remote Sensing Analysis

Mr. Rajesh Kumar Kanji
Independent Researcher, Plano, USA

Abstract: The integration of multi-source forest inventories with automated remote sensing workflows presents significant challenges for sustainable forest management and global climate mitigation initiatives. This review examines unified data warehouse architectures as critical frameworks for harmonizing disparate data streams—including satellite imagery, airborne LiDAR, UAV datasets, ground surveys, and historical records—into scalable analytical systems. We synthesize methodologies enabling near real-time inventory updates, focusing on spectral-index chronosequence analysis for disturbance detection, multi-sensor fusion strategies for canopy-penetrating observations, and dynamic growth-model recalibration. The paper assesses how optimized data warehousing supports machine learning and deep learning applications in automated attribute estimation, change detection, and error propagation management. Key challenges in data standardization, model transferability, and scalable cloud-based implementations are critically evaluated. Furthermore, the role of integrated architectures in advancing carbon monitoring, precision conservation, and long-term forest sustainability is analyzed. The synthesis concludes that unified data repositories are essential for operationalizing next-generation forest monitoring paradigms aligned with carbon neutrality objectives, providing robust foundations for policy-relevant decision-making through enhanced spatiotemporal analytics.

Keywords: multi-source forest inventory, remote sensing integration, data warehouse architecture, living inventory update, sustainable forestry.

INTRODUCTION

Forest inventory management is crucial to sustainable forestry, relying on exact data about individual tree attributes—such as species, height, and diameter at breast height (DBH). Traditionally, approaches depend on narrow datasets, often acquired from optical images or ground surveys, which fail to scale and retain accuracy in various situations. For example, traditional visual interpretation and conventional machine learning (ML) algorithms might face substantial hurdles in complicated circumstances like dense canopies or multi-tree crown tops, sometimes attaining accuracies as low as 59% [Coomes, D. A. *et al*., 2017]. These inadequacies hinder efficient decision-making in conservation, carbon sequestration monitoring, and urban planning—especially in vast wooded regions like Canada where raw data volumes are continuously expanding. Consequently, there is an urgent need for integrated systems that reconcile diverse data streams while automating analytical operations to support large-scale, real-time forest management.

Recent technological progress offers new potential. Remote sensing tools—satellites, unmanned aerial vehicles (UAVs), LiDAR, multispectral sensors—now allow extensive spatial and temporal data collection at high resolution [Chirici, G. *et al*., 2011]. This detail brings significant burdens: storing, processing, and interpreting vast, multi-source datasets. Data from a single location often arrives in disparate formats, resolutions, and timeframes. Models calibrated for one input type may falter when applied to another. While visual interpretation persists, dense canopies challenge its effectiveness. Machine learning shows promise but demands structured, validated inputs—requirements difficult to meet with fragmented or incomplete information.

The core difficulty often resides not in data scarcity, but in systems for meaningful integration [White, J. C. *et al*., 2016]. Diverse data sources are generally isolated. Merging them proves complex due to differing collection methods, inconsistent naming conventions, or missing metadata. Temporal variation adds complexity. Observations separated by months require standardization or historical context for valid comparison. Older paper records frequently contain meticulously verified coordinates exceeding newer digital sources in precision, yet rarely enter modern workflows. Without a coherent structure linking these elements, the longitudinal picture of forest dynamics fragments.

Focusing on individual tree classification and multi-source forestry data integration, this review undertakes a critical assessment of relevant methodologies, tools, and frameworks. It evaluates the operational deployment of available technologies, scrutinizing their inherent limitations and the persistent gaps that impede progress. Through a synthesis of historical precedents and contemporary advancements, the analysis aims to clarify the role of digital techniques in advancing

***Corresponding Author:** Mr. Rajesh Kumar Kanji

DOI- https://doi.org/ 10.5281/zenodo.15685056

tree-level comprehension of forest ecosystems and to propose how strategic data organization and collaborative integration might yield a more holistic perspective [Fassnacht, F. E. *et al*., 2016]. The review thus underscores the enduring difficulties associated with forest data processing, structuring, and validation, pointing toward specific domains where concentrated future effort holds promise for strengthening technical capacity and deepening ecological knowledge.

RELATED WORK

Recent advances in forest carbon monitoring emphasize multi-source remote sensing data fusion—integrating optical (e.g., Landsat, Sentinel-2), SAR (e.g., Sentinel-1, ALOS PALSAR), and LiDAR (e.g., GEDI, ICESat-2) platforms—within GIS-driven frameworks to enhance aboveground biomass (AGB) estimation across heterogeneous landscapes [Liang, X]. Feature-level fusion has become common, integrating spectral indices, radar backscatter, and structural metrics to alleviate sensor-specific limitations—particularly optical saturation in dense canopies and SAR sensitivity constraints—while cloud-computing platforms (e.g., Google Earth Engine) facilitate scalable wall-to-wall mapping. Machine learning approaches—particularly Random Forests and Gradient Boosting Machines—leverage these fused inputs to achieve superior accuracy over empirical regressions, though deep learning architectures (e.g., CNNs) show heightened proficiency in capturing complex spatial patterns from raw data, despite computational and interpretability challenges. Concurrently, process-based models (e.g., CASA, Biome-BGC) incorporate remote sensing-derived parameters to simulate carbon fluxes mechanistically, bridging data-driven and ecological perspectives for scenario-based forecasting. However, operational deployment faces barriers: persistent data heterogeneity, model transferability gaps across biomes, and sparse ground validation—especially in data-poor tropical regions—constrain policy-relevant applications like REDD+ MRV systems and carbon markets [Liang, X]. Emerging solutions advocate AI-augmented hybrid frameworks (e.g., physics-informed neural networks), uncertainty-aware workflows, and integrated sky-to-ground networks to unify satellite, UAV, and in-situ observations—collectively advancing toward real-time, scalable carbon monitoring aligned with global neutrality objectives.

Real-time forest inventory frameworks is distinguished by considerable breakthroughs in remote sensing data integration and computational innovation. Contemporary approaches, notably Enhanced Forest Inventories (EFIs), leverage airborne laser scanning (ALS) to generate spatially exhaustive, wall-to-wall raster estimates of structural attributes like height, volume, and canopy cover at fine resolutions (e.g., 20–30 m), surpassing conventional polygonal inventories in detail and objectivity [Reynolds, K. M. *et al*., 2003]. However, EFIs have limits in cost, update frequency (usually decadal), and inability to characterize non-stand-replacing disturbances or species composition without supplemental data. Consequently, research has focused toward incorporating regular, cost-effective satellite data to sustain currency. The advent of satellite constellations (e.g., Harmonized Landsat Sentinel (HLS), PlanetScope) provides unprecedented temporal resolutions (e.g., 2–5 days), permitting continuous monitoring through dense time-series analysis [Reynolds, K. M. *et al*., 2003]. For disturbance verification and attribute updating, targeted acquisitions using CubeSats, RPAS-derived digital aerial photogrammetry (DAP), or fine-resolution imagery (e.g., PlanetScope) provide critical fine-scale structural insights, though model transferability between ALS and DAP introduces potential bias [Reynolds, K. M. *et al*., 2003]. Growth integration leverages physiological or statistical models to boost features in undisturbed cells, yet error propagation from sequential updates and computational needs for near real-time processing remain important hurdles [Reynolds, K. M. *et al*., 2003].

Cormier, *et al*., developed a prototype data warehouse (DW) to integrate multi-source forestry data, including UAV imagery, LiDAR point clouds, survey records, and paper documents. Their DW transforms these datasets into inputs for machine learning and deep learning models, specifically YOLOv11, to classify individual tree species (ITS). The study pioneered the integration of paper records with UAV imagery, demonstrating that paper documents significantly improve ground-truth accuracy—particularly in dense canopies where visual interpretation alone achieved only 59% accuracy. The DW employs a star schema with three dimension tables (Date, Image, Species) and a fact table (Tree Metrics) to consolidate quantitative tree data. This design supports efficient querying, scalability, and temporal analysis while minimizing storage

Publisher: SARC Publisher

overhead. YOLOv11 handles real-time data ingestion, detecting trees and classifying species during ingestion, which streamlines the preprocessing pipeline. The architecture also accommodates future expansion to include video, text, and remote sensor data. For scalability, the authors leverage Canada's Digital Research Infrastructure (DRI), securing resources like GPU-accelerated HPC clusters and cloud storage to manage projected 10-year data growth. Their work establishes a centralized repository for long-term forest monitoring, addressing challenges in data harmonization and accessibility across government and industry stakeholders.

Lv, *et al*., addressed forest swamp classification challenges in the Changbai Mountain Ecological Function Protection Area using multi-source remote sensing data. Their study integrated 2019–2022 growing season datasets from Sentinel-1 C-SAR, ALOS-2 L-PALSAR, Sentinel-2 MSI, and Landsat-8 TIRS with environmental covariates. The methodology applied a two-stage framework: First, NDBI thresholding (NDBI > 0.12) excluded 94% of urban/agricultural areas. Second, an optimized Random Forest classifier (ntree = 1200, mtry = 28) with 10-fold cross-validation leveraged 42 features. Key discriminators included L-band HV backscatter (feature importance = 47), Sentinel-2 SWIR Band 12 (importance = 57), and land surface temperature gradients [Huang, C. *et al*., 2002]. This approach achieved 87.18% overall accuracy (Kappa = 0.84) at 10 m resolution. The study demonstrated L-band SAR's superior canopy penetration capability, improving classification accuracy by 4.2% over optical-only methods in ecotones (720–850 m elevation). Thermal-IR features further reduced spectral confusion between wetlands and forests. The resulting map quantified forest swamps at 229.95 km² (9% of protected areas), providing a transferable template for temperate mountain ecosystems [Huang, C. *et al*., 2002].

UNIFIED DATA WAREHOUSE ARCHITECTURES AND AUTOMATION FRAMEWORKS

The integration of diverse remote sensing (RS) and ground-based forest inventory data necessitates a unified data warehouse architecture capable of harmonizing disparate data types, scales, and temporal resolutions. The Forest Inventory and Analysis (FIA) program exemplifies this approach through its systematic aggregation of multi-source inputs—including satellite imagery (Landsat, MODIS, Sentinel), aerial photography (NAIP), lidar (airborne, spaceborne), and field plot measurements—into cohesive analytical frameworks. Central to this architecture is a scalable, cloud-based storage and processing infrastructure, such as Google Earth Engine (GEE) or the USFS-developed BIGMAP platform. These environments enable efficient handling of massive datasets (e.g., petabytes of Landsat time-series or high-resolution NAIP imagery) while providing tools for preprocessing, mosaicking, and algorithm application. Key to interoperability is the use of standardized geospatial protocols (e.g., WMS, WFS, WCS) and APIs, which facilitate seamless data streaming between storage systems and analytical tools. For instance, FIA leverages such protocols to dynamically link plot locations with auxiliary RS layers (e.g., intersecting FIA plots with LandTrendr-derived disturbance maps), ensuring spatial and temporal alignment crucial for model calibration and validation [Lister, A. J. *et al*., 2020]. The architecture further incorporates metadata schemas documenting data provenance, uncertainty, and preprocessing steps—critical for traceability and reproducibility in large-area estimates.

This unified structure directly supports advanced analytical workflows, such as model-assisted inference and small area estimation (SAE), by enabling on-demand access to synchronized multi-temporal datasets. For example, the integration of FIA plot data with Landsat time-series (LTS) via cloud platforms allows for near-real-time updates of forest attributes (e.g., biomass, canopy cover) across user-defined domains. The architecture's modular design accommodates specialized processing pipelines, such as the FIESTA software's estimation modules or the OBI-WAN system for GEDI-FIA biomass fusion, while maintaining compatibility with core inventory databases. Scalability is achieved through distributed computing frameworks that parallelize tasks like harmonic regression on LTS data or object-based image analysis (OBIA) of NAIP imagery. Crucially, the architecture mitigates historical challenges of data volume and computational latency, as evidenced by FIA's transition from manual photointerpretation to automated cloud-based workflows. This evolution underscores the architecture's role in transforming raw data into actionable insights—such as wall-to-wall carbon stock maps or county-level disturbance assessments—while ensuring consistency with FIA's rigorous statistical

Publisher: SARC Publisher

estimation protocols. By centralizing data access and processing, the architecture not only streamlines operational efficiency but also fosters innovation through reusable, interoperable components adaptable to emerging technologies like UAS or terrestrial lidar [Lister, A. J. *et al*., 2020].

The Forest Resource Assessment (FRA) of Nepal project exemplifies the necessity of a unified data warehouse architecture for integrating multi-source forest inventory data, combining conventional field measurements, visual interpretation plots, and satellite imagery into a cohesive analytical framework. This design acts as the core store and processing hub, enabling the smooth merging of diverse data types needed for large-scale mapping and estimation activities. Field data, acquired via a stratified systematic cluster sampling design, comprises extensive biometrical measures (e.g., diameter at breast height, species, tree height) from Concentric Circular Sample Plots (CCSPs). These ground observations are kept with plot-level information such as land-use class, forest type, and GPS coordinates. Simultaneously, remotely sensed data—specifically Landsat Thematic Mapper (TM) images and MODIS surface reflectance products—are ingested, pre-processed, and geometrically aligned within the system. Crucially, the architecture accommodates auxiliary data layers, such as visual interpretation plots produced from Google Earth and RapidEye imagery, classified according to FAO land-use guidelines [Muinonen, E. *et al*., 2012]. This integration supports the application of multi-source forest inventory (MSFI) methodologies, where the warehouse correlates spectral information from satellite bands (e.g., Landsat bands 1–7) to field-measured response variables (e.g., forest cover, stand volume). The use of Open Source tools (GDAL, GRASS GIS, Quantum GIS) for data processing ensures scalability and reproducibility, while structured database schemas maintain temporal and spatial consistency across repeated inventories—critical for REDD (Reducing Emissions from Deforestation and Forest Degradation) reporting and climate change mitigation strategies [Muinonen, E. *et al*., 2012].

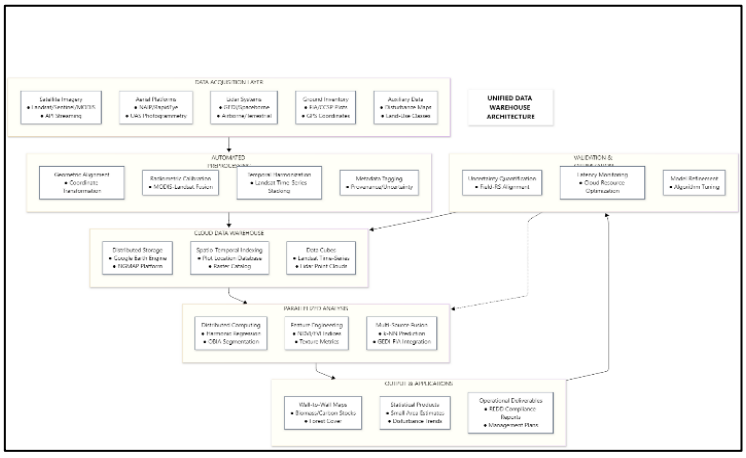

Fig 1: Integrated Cloud-Based Architecture for Automated Forest Inventory and Remote Sensing Analysis

Operationally, the architecture facilitates advanced analytical procedures, such as the *k*-nearest neighbor (*k*-NN) method, by structuring input data into distinct "reference" and "target" sets. The reference set comprises pixels overlapping field plots or visual interpretation points, holding observed variables (e.g., FAO land-use classes, stand volumes). In contrast, the target set comprises all pixels requiring prediction, leveraging feature variables (e.g., Landsat band values, NDVI ratios) derived from the image mosaic. The system automates essential pre-processing stages, including the relative radiometric calibration of Landsat images using MODIS data via local correction models, maintaining spectral consistency across mosaics. Furthermore, it manages the computation of spatial metrics (e.g., moving-window statistics for mean and standard deviation) and handles large-scale raster algebra for wall-to-wall prediction. Post-processing modules, integrated within the warehouse, apply spatial filters (e.g., 3×3 mode filters) to decrease classification noise and convert raster outputs into vector forms for practical applications like forest management planning [Muinonen, E. *et al*., 2012]. By centralizing data storage, quality control, and processing pipelines, the architecture eliminates silos, reduces latency in data retrieval, and enables the efficient reuse of field data for diverse remote sensing

Publisher: SARC Publisher

applications—exemplified by the FRA Nepal project's generation of forest cover and volume maps for the Terai region. This structured yet flexible approach assures that multi-source integration is not only additive but synergistic, boosting the accuracy and usability of forest resource evaluations [Muinonen, E. *et al*., 2012].

APPLICATIONS AND CHALLENGES

The integration of high-density UAV-LiDAR data with machine learning techniques allows for precise estimation of individual tree attributes such as diameter at breast height (DBH), total height, and timber volume, all of which are critical for modern forest inventories. This method performs well across a variety of machine learning models (Support Vector Regression, Random Forest, Artificial Neural Networks, Extreme Gradient Boosting), with relative Root Mean Square Errors (rRMSE) consistently less than 15% for DBH, 9% for height, and 29% for volume [Dalla Corte, A. P. *et al*., 2020]. This level of accuracy makes detailed stand-level assessments possible, promotes sustainable timber management, and improves carbon stock quantification without the need for labor-intensive field surveys. Furthermore, the scalability of UAV-LiDAR systems enables rapid data acquisition over difficult terrains, making them ideal for operational monitoring in integrated crop-livestock-forest systems like the *Eucalyptus benthamii* plantations investigated in this study [Dalla Corte, A. P. *et al*., 2020].

Despite its advantages, this methodology has significant limitations, owing to the high dimensionality and multicollinearity of LiDAR-derived predictor variables (such as height percentiles and canopy metrics). Correlated predictors complicate traditional statistical approaches and increase the risk of overfitting in machine learning workflows, as demonstrated by stepwise regression models with severe Variance Inflation Factors (VIF > 10). Processing high-density point clouds (1,500-2,500 points/m²) and tuning hyperparameters across multiple algorithms can be computationally demanding [Dalla Corte, A. P. *et al*., 2020]. Furthermore, while machine learning techniques like SVR reduce multicollinearity effects, their "black-box" nature limits their interpretability for forest managers. Model transferability across forest types is still uncertain, necessitating species-specific calibration and validation to maintain accuracy—a challenge exacerbated by variability in sensor specifications and flight parameters. Future research should address these limitations by employing dimensionality reduction techniques and standardized data collection procedures [Dalla Corte, A. P. *et al*., 2020].

Satellite-based wildfire detection and prediction systems enable critical disaster management applications that require real-time data. These systems help with evacuation planning by mapping active fire perimeters and spread trajectories, which improves emergency responders' situational awareness. Deep learning models enable rapid damage assessment via automated burn severity classification (e.g., unburned to completely destroyed), which speeds up post-fire recovery efforts. Furthermore, predictive algorithms use historical fire data, meteorological variables, and terrain metrics to forecast fire behavior 24-48 hours in advance, allowing for better resource allocation for firefighting operations. The fusion of multi-sensor satellite data (thermal, optical, and radar) within a unified architecture enables continuous monitoring under a variety of conditions, including cloud cover, smoke, and darkness, ensuring persistent surveillance of high-risk areas.

The heterogeneity of data and the computational demands present significant challenges. Satellite datasets vary significantly in spatial resolution (3m-1km), temporal frequency (minutes to 16 days), and preprocessing requirements (atmospheric correction, cloud masking). Training robust models requires large, labeled datasets, but public repositories are scarce due to the sensitivity of fire locations and damage metrics. Noise from sensor artifacts, cloud obstruction, and spectral ambiguities (such as smoke/cloud confusion) all reduce detection accuracy. Real-time processing imposes latency constraints, especially for high-resolution data streams that require extensive computation. Model interoperability is hampered by inconsistencies in evaluation metrics across studies (F1-scores, IoU, and RMSE), which prevent standardized benchmarking. Furthermore, model transfer across ecosystems is limited due to regional differences in fuel types, topography, and climate patterns.

CONCLUSION

This review demonstrates that a unified data warehouse architecture is critical for implementing multi-source forest inventory integration and automated remote sensing analysis. Such architectures provide the framework for integrating disparate, high-volume data streams—including

Publisher: SARC Publisher

satellite imagery, LiDAR, UAV datasets, ground surveys, and historical records—into scalable, query-efficient repositories. The synthesis shows that these integrated systems enable critical capabilities such as near real-time inventory updates via spectral chronosequence analysis, improved attribute estimation via multi-sensor fusion (particularly canopy-penetrating L-band SAR and LiDAR), and dynamic model recalibration using machine learning and deep learning. Cloud-based implementations, aided by distributed computing, efficiently manage computational demands and data volume, allowing for applications such as continuous disturbance detection, precision conservation, and carbon stock monitoring. However, significant challenges remain. Data heterogeneity in format, resolution, and temporal frequency necessitates strict standardization protocols and consistent metadata schemas. Model transferability across various forest biomes remains limited, necessitating site-specific calibration. Furthermore, managing error propagation during sequential updates and ensuring the interpretability of complex "black-box" algorithms remain ongoing issues. Despite these challenges, integration achieved through unified data warehouses improves the accuracy and timeliness of forest resource assessments. This structured approach provides the robust, policy-relevant analytics required for sustainable forest management and meeting global carbon neutrality goals. Future work must prioritize improving interoperability, creating transferable models, and refining uncertainty quantification within these integrated systems.

REFERENCES

1. Coomes, D. A., Dalponte, M., Jucker, T., Asner, G. P., Banin, L. F., Burslem, D. F., Lewis, S. L., Nilus, R., Phillips, O. L., Phua, M. H. & Qie, L. "Area-based vs tree-centric approaches to mapping forest carbon in Southeast Asian forests from airborne laser scanning data." *Remote Sensing of Environment*, 194 (2017): 77–88.
2. Chirici, G., Winter, S. & McRoberts, R. E. (Eds.). *National forest inventories: contributions to forest biodiversity assessments.* Springer Science & Business Media, 2011.
3. White, J. C., Coops, N. C., Wulder, M. A., Vastaranta, M., Hilker, T. & Tompalski, P. "Remote sensing technologies for enhancing forest inventories: A review." *Canadian Journal of Remote Sensing*, 42.5 (2016): 619–641.
4. Fassnacht, F. E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L. T., Straub, C. & Ghosh, A. "Review of studies on tree species classification from remotely sensed data." *Remote Sensing of Environment*, 186 (2016): 64–87.
5. Liang, X., Yu, S., Meng, B., Wang, X., Yang, C., Shi, C. & Ding, J. "Multi-source remote sensing and GIS-driven forest carbon monitoring for carbon neutrality: Integrating data, modeling, and policy applications."
6. Reynolds, K. M., Hessburg, P. F. & Bourgeron, P. S. "Making ecosystem-based management operational: A case study of the Okanogan Basin, Washington." *Forest Ecology and Management*, 175.1–3 (2003): 29–52.
7. Huang, C., Davis, L. S. & Townshend, J. R. "An assessment of support vector machines for land cover classification." *International Journal of Remote Sensing*, 23.4 (2002): 725–749.
8. Lister, A. J., Andersen, H., Frescino, T., Gatziolis, D., Healey, S., Heath, L. S., Liknes, G. C., McRoberts, R., Moisen, G. G., Nelson, M. & Riemann, R. "Use of remote sensing data to improve the efficiency of national forest inventories: A case study from the United States national forest inventory." *Forests*, 11.12 (2020): 1364.
9. Muinonen, E., Parikka, H., Pokharel, Y. P., Shrestha, S. M. & Eerikäinen, K. "Utilizing a multi-source forest inventory technique, MODIS data and Landsat TM images in the production of forest cover and volume maps for the Terai Physiographic Zone in Nepal." *Remote Sensing*, 4.12 (2012): 3920–3947.
10. Dalla Corte, A. P., Souza, D. V., Rex, F. E., Sanquetta, C. R., Mohan, M., Silva, C. A., Zambrano, A. M., Prata, G., de Almeida, D. R., Trautenmüller, J. W. & Klauberg, C. "Forest inventory with high-density UAV-Lidar: Machine learning approaches for predicting individual tree attributes." *Computers and Electronics in Agriculture*, 179 (2020): 105815.

Publisher: SARC Publisher

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Kanji, R.K. "A Unified Data Warehouse Architecture for Multi-Source Forest Inventory Integration and Automated Remote Sensing Analysis." *Sarcouncil Journal of Engineering and Computer Sciences* 1.5 (2022): pp 10-16.

Publisher: SARC Publisher