

Communication

SeaScope: A Transparent and Reproducible LLM-Assisted Framework for Maritime Earth Observation Analysis

Christos Sekas¹, Lydia Mavrofidopoulou², Ilias Agathangelidis^{2,*}, Constantinos Cartalis²,
Kostas Philippopoulos², Faidon Mavroudis², Stelios P. Neophytides³, Michalis Mavrovouniotis³,
Ioannis Yfantidis¹ and George Paterakis¹

- ¹ CTRL + Space Development E.E., 10682 Athens, Greece; csekas@ctrlspace.dev (C.S.); gyfantidis@ctrlspace.dev (I.Y.); gpaterakis@ctrlspace.dev (G.P.)
² Department of Physics, National and Kapodistrian University of Athens, 15784 Athens, Greece; lydimavr@phys.uoa.gr (L.M.); ckartali@phys.uoa.gr (C.C.); kphillip@phys.uoa.gr (K.P.); phaethon@phys.uoa.gr (F.M.)
³ Department of Big Earth Data Analytics, Eratosthenes Centre of Excellence, 3012 Limassol, Cyprus; stelios.neophytides@eratosthenes.org.cy (S.P.N.); michalis.mavrovouniotis@eratosthenes.org.cy (M.M.)
* Correspondence: iliasaga@phys.uoa.gr

Highlights

- SeaScope integrates LLMs, RAG, and Google Earth Engine to generate transparent and reproducible Earth Observation workflows from natural-language requests.
- The framework achieved high workflow-generation and event-detection success across diverse maritime monitoring applications.
- The results demonstrate the potential of AI-assisted workflows to make Earth Observation analysis more accessible while maintaining traceability, reproducibility, and human oversight.

Abstract

Earth Observation (EO) analysis increasingly relies on large and heterogeneous satellite datasets, yet developing EO workflows often requires specialized expertise in data selection, geospatial programming, and cloud-based processing. Recent advances in Large Language Models (LLMs) offer new opportunities for natural-language interaction with EO systems, although challenges related to transparency, reproducibility, and domain-specific reasoning remain. This study presents SeaScope, an explainable AI framework that integrates LLMs, Retrieval-Augmented Generation (RAG), scientific knowledge retrieval, and Google Earth Engine (GEE) to transform natural-language requests into transparent and executable EO workflows. The framework combines knowledge retrieval, code generation, cloud execution, provenance tracking, and interactive visualization within a unified environment. A pilot implementation is demonstrated through maritime and coastal monitoring applications, including oil spill detection, vessel monitoring, water quality assessment, floating debris detection, and air quality analysis. Multiple state-of-the-art LLMs are evaluated under both RAG and non-RAG configurations using representative EO case studies. The results indicate substantial differences among model families and show that retrieval augmentation can significantly improve workflow generation quality and reliability for capable models, while providing more limited benefits for smaller models. The proposed framework demonstrates the potential of explainable AI agents to support transparent, reproducible, and scalable EO analysis.



Academic Editor: Deepak R. Mishra

Received: 22 June 2026

Revised: 11 August 2026

Accepted: 19 August 2026

Published: 22 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: Earth Observation; Large Language Models; Retrieval-Augmented Generation; Google Earth Engine

1. Introduction

The rapid growth of Earth Observation (EO) data resources and cloud-based geospatial platforms has created new opportunities for environmental monitoring while simultaneously increasing the complexity of EO workflow development [1]. In maritime and coastal applications, analysts are often required to combine heterogeneous satellite datasets, domain-specific processing methods, and geospatial programming expertise to investigate phenomena such as oil spills, vessel activity, water quality, marine debris, and atmospheric pollution [2–6]. Although platforms such as Google Earth Engine (GEE) have substantially reduced computational barriers, the development of robust and reproducible EO workflows remains a technically demanding process.

Recent advances in Large Language Models (LLMs) have demonstrated strong capabilities in code generation and natural-language interaction [7]. In the geospatial domain, agentic programming approaches such as GeoAgent [8] and conversational systems such as ChatGeoAI [9] have highlighted the potential of natural-language interfaces for geospatial analysis. Similarly, remote-sensing vision-language models, including GeoChat [10] and EarthGPT [11], have demonstrated the ability of multimodal AI systems to interpret EO imagery and support EO-related reasoning tasks. Beyond image interpretation, LLM-based geospatial copilots, such as a multi-agent framework [12], have expanded these capabilities toward workflow orchestration and remote-sensing analysis.

Parallel developments have focused on geospatial code generation and retrieval-enhanced reasoning. AutoGEEval [13], Geo-FuB [14], and GeoCode-GPT [15] provide frameworks for evaluating and improving LLM-generated geospatial code, particularly in GEE environments [16]. More specifically, Hou et al. [17] demonstrated that Retrieval-Augmented Generation (RAG) can improve geospatial code generation through the integration of a GEE operator knowledge base, while GeoGraphRAG [18] highlighted the value of retrieval-based reasoning for automated geospatial modeling. In parallel, RS-Agent [19], EARTH-AGENT [20], and the EO-agent evaluation framework of Kao et al. [21] demonstrated the potential of tool-using agents and multi-step reasoning for EO applications. In addition, recent advancements include training-free multiagent LLM frameworks to reformulate the Earth Observation Satellite Task Scheduling Problem [22] and introducing architectural enhancements that improve conversational understanding, hierarchical tool composition, and the execution of complex Earth observation analyses [23]. Maritime-specific developments have also emerged [24] using artificial intelligence, often targeting ship detection imagery [25,26] and the integrated study of ocean systems [27].

Despite these advances, most existing approaches address individual components of the EO analysis pipeline rather than providing an integrated framework spanning knowledge retrieval, workflow generation, execution, and user interaction. While these developments have significantly expanded the capabilities of AI-assisted EO systems, most remain specialized toward a particular task, workflow generation, or knowledge retrieval. As a result, users often need to combine multiple tools and workflows to progress from a natural-language query to an executable and scientifically interpretable analysis. The integration of domain-grounded retrieval, executable GEE workflows, provenance tracking, explainability, and user validation within a unified EO framework remains comparatively limited. Furthermore, the impact of retrieval-enhanced reasoning on the generation of

operational EO workflows has received relatively limited evaluation across different LLM families and environmental monitoring tasks.

To address this gap, this paper presents SeaScope, an explainable AI framework that combines LLMs, Retrieval-Augmented Generation (RAG), scientific knowledge retrieval, and Google Earth Engine to transform natural-language requests into transparent and executable EO workflows. The framework integrates task interpretation, knowledge retrieval, code generation, cloud-based execution, provenance tracking, and interactive visualization into a human-in-the-loop environment where analytical decisions remain accessible throughout the workflow. SeaScope is demonstrated through a representative set of maritime applications, including vessel detection, oil spill monitoring, water quality assessment, floating debris detection, and air quality analysis, with a focus on the Mediterranean region and the Cyprus Exclusive Economic Zone (EEZ). Maritime environments present a diverse range of environmental monitoring challenges, often requiring the integration of multi-sensor EO datasets, domain-specific analytical methods, and cloud-based geospatial processing. As a result, they provide an appropriate and representative testbed for evaluating AI-assisted EO frameworks while maintaining broad applicability to other Earth Observation domains.

The proposed framework contributes to AI-assisted EO analysis in four key aspects. First, it integrates natural-language interaction, domain-specific knowledge retrieval, EO workflow generation, cloud execution, provenance tracking, and interactive visualization within a unified analytical environment. Second, it evaluates the impact of RAG on EO workflow generation through a comparative assessment of multiple state-of-the-art LLM families. Third, it implements a provenance-aware workflow paradigm in which retrieved sources, generated code, user requests, and visual outputs remain accessible throughout the analysis process, supporting transparency and reproducibility. Finally, it demonstrates the applicability of the framework through representative maritime EO case studies spanning multiple satellite sensors and environmental monitoring tasks.

2. Materials and Methods

2.1. SeaScope Framework

SeaScope is an explainable AI framework designed to support maritime and coastal EO analysis through natural-language interaction. The framework transforms user-defined objectives into executable GEE workflows, geospatial products, and traceable explanations. The current implementation supports applications including vessel detection, oil-spill monitoring, turbidity mapping, floating debris detection, water-quality assessment, and atmospheric pollution analysis using multi-source EO datasets.

The framework follows an interpret–retrieve–generate–execute–explain paradigm (Figure 1) implemented through a modular agent architecture. User requests are first translated into a structured task specification describing the spatial and temporal domain, phenomenon of interest, dataset constraints, analytical operations, and expected outputs. This representation provides a machine-interpretable description of the requested analysis and serves as the information structure exchanged between the different components of the framework. The task specification is processed by an orchestration layer that coordinates knowledge retrieval, workflow generation, cloud execution, result interpretation, and provenance tracking. To support scientifically grounded analyses, SeaScope employs RAG through a curated domain-specific corpus combining structured dataset metadata and processing specifications with unstructured scientific literature, EO guidelines, and validated workflow examples from authoritative and scientifically established sources. The retrieval component uses dense vector embeddings and Hypothetical Document Embeddings (HyDE) to identify relevant information, while the agent evaluates the retrieved

context and refines retrieval when additional domain knowledge is required. Retrieved content supports dataset selection, preprocessing, parameter configuration, threshold selection, and methodological validation throughout workflow generation.

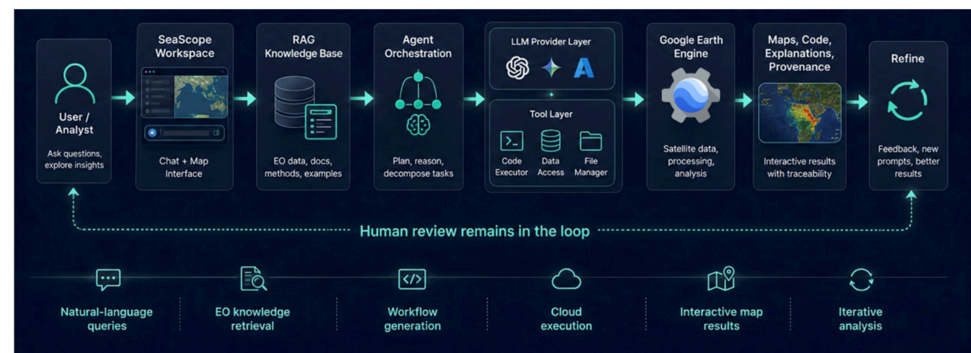


Figure 1. Architecture of the SeaScope framework for AI-assisted Earth Observation analysis.

Based on the task specification and retrieved context, the LLM synthesizes and reviews an executable GEE workflow. Rather than operating as a single-pass system, SeaScope follows an agentic process of iterative knowledge retrieval, workflow generation, tool execution, and refinement. Generated workflows may include dataset filtering, image compositing, cloud and land masking, radiometric preprocessing, spectral-index calculations, threshold-based classification, change detection, spatial statistics, temporal analysis, and map visualization. The resulting code is executed within the GEE cloud environment, enabling direct access to large EO archives.

When execution failures, dataset incompatibilities, or data-availability issues are encountered, the framework performs a bounded refinement procedure using execution feedback and retrieved knowledge. Corrective modifications are limited to parameter, dataset, or processing adjustments that preserve the original analytical objective. The analytical workflow remains fully documented by recording retrieved sources, workflow-generation steps, code revisions, execution feedback, and final outputs, enabling users to verify how the final result was produced. By integrating knowledge retrieval, agentic reasoning, EO workflow synthesis, cloud-based geospatial processing, provenance tracking, and interactive visualization, SeaScope supports transparency, reproducibility, and scientific accountability while maintaining users within the decision-making loop.

2.2. Datasets and Earth Observation Workflows

The SeaScope framework integrates multiple Copernicus satellite products within Google Earth Engine (GEE) to support maritime and coastal monitoring applications. Sentinel-1 Synthetic Aperture Radar (SAR) Ground Range Detected (GRD) data were used for vessel monitoring and oil-spill analysis. Sentinel-1 provides all-weather, day-and-night microwave observations with spatial resolutions ranging from 10 to 40 m and a revisit frequency of approximately six days. Its sensitivity to surface roughness makes it particularly suitable for detecting vessels and identifying anomalous low-backscatter features associated with oil slicks and other marine surface films.

Sentinel-2 MultiSpectral Instrument (MSI) Level-2A Surface Reflectance data were used for water-quality assessment, turbidity mapping, and floating debris detection. Its multispectral capabilities support the derivation of spectral indicators related to water properties, suspended sediments, floating material, and coastal environmental conditions. Atmospheric monitoring was performed using Sentinel-5P NO₂ products, which provide observations of tropospheric nitrogen dioxide concentrations at regional scales. These data were used to assess atmospheric pollution patterns over maritime and coastal regions.

Accordingly, SeaScope supports multiple EO workflows corresponding to common maritime monitoring applications. Oil-spill and vessel-detection workflows primarily rely on Sentinel-1 SAR observations combined with spatial filtering, thresholding, and temporal comparison approaches. Water-quality and turbidity analyses are based on Sentinel-2 multispectral imagery, while floating debris detection combines spectral indicators designed to highlight floating material accumulations in coastal waters. Atmospheric monitoring workflows utilize Sentinel-5P observations to map nitrogen dioxide distributions and identify potential pollution hotspots. Several established spectral indicators were incorporated within the generated workflows, including the Normalized Difference Water Index (NDWI) for water delineation, the Normalized Difference Turbidity Index (NDTI) [28] for turbidity assessment, the Floating Debris Index (FDI) [29] for floating material detection, and the Normalized Difference Vegetation Index (NDVI) for discrimination between vegetated and non-vegetated floating targets.

2.3. Benchmarking Framework and Validation Design

The performance of SeaScope was evaluated using a combination of representative maritime EO case studies and a larger validation dataset spanning multiple environmental monitoring applications. The evaluation was designed to assess the ability of the framework to transform natural-language requests into scientifically meaningful and executable GEE workflows, while also examining the impact of RAG on workflow quality and reliability.

An initial benchmark was conducted using seven Mediterranean maritime EO case studies representative of common coastal and marine monitoring applications. These included oil spill detection, vessel detection, turbidity monitoring, floating debris detection, and air quality assessment. The oil-spill scenarios included historical events near Corsica (7 October 2018) (“Corsica 2018”), Baniyas, Syria (23 August 2021) (“Syria 2021”), and the Aléria-Solenzara region (13 June 2021) (“Corsica 2021”), while additional case studies examined vessel detection in the Port of Larnaca (“Vessel detection”), turbidity monitoring in the Gulf of Lion (“Turbidity”), floating debris detection in the Gulf of Trieste (“Debris”), and atmospheric pollution assessment in the Port of Limassol (“Air quality”). These applications were selected to cover diverse EO data sources, environmental conditions, and analytical requirements. For each application, methodological information, EO processing approaches, and indicative code examples were collected and incorporated into the SeaScope knowledge corpus. This material, together with dataset documentation and scientific references, was used to support the RAG configuration of the framework.

The benchmark included multiple state-of-the-art LLMs evaluated under both RAG and non-RAG configurations. For each case study, users interacted with SeaScope through natural-language prompts describing the requested EO analysis. The generated workflows were subsequently executed within GEE, and the resulting outputs were examined through both visual inspection and code review according to the evaluation criteria summarized in Table 1. Model performance was assessed using an expert-based evaluation score ranging from 0 to 10. The scoring framework considered the correctness and completeness of the generated result, the degree to which the workflow satisfied the user request, the number of interaction steps required to obtain a satisfactory solution, and the presence of implementation or execution errors. Additional consideration was given to workflow consistency across iterative interactions, particularly the ability of a model to preserve corrections and constraints introduced during previous exchanges. Generated code was also inspected to verify the correct use of datasets, temporal constraints, processing parameters, and application-specific methodological requirements. Higher scores therefore reflected more accurate, complete, and reliable EO workflows. To improve the robustness of the evaluation, the benchmark was independently assessed by two EO experts, each

evaluating approximately half of the benchmark cases according to the commonly defined criteria. Evaluator overlap was assessed across the 14 case-study–condition combinations by comparing the scores assigned independently by Evaluator 1 and Evaluator 2 using Gemini 3 Flash. Absolute agreement and consistency were estimated using the intraclass correlation coefficients ICC(2,1) and ICC(3,1), respectively, supplemented by Pearson and Spearman correlations, mean score difference, and Bland–Altman 95% limits of agreement. The evaluator-overlap analysis quantifies the consistency of the evaluators’ conclusions regarding the model’s typical performance across repeated conversations for the same case studies, under both RAG and non-RAG conditions.

Table 1. Evaluation framework used to assess the quality and reliability of generated EO workflows.

Score Range	Interpretation
0	No meaningful result was generated.
1–2	Incorrect or largely irrelevant output, indicating poor task understanding.
3–4	Partial task completion with major errors, omissions, or repeated failures.
5–6	Moderately acceptable result, requiring substantial user intervention or corrections.
7–8	Mostly correct output with minor errors and limited need for refinement.
9–10	Accurate, complete, and reliable result with correct implementation of task requirements and requested modifications.

To assess repeat-run consistency, Gemini 3 Flash was further evaluated on all seven case studies under both RAG and non-RAG conditions in five independent conversations per case study and condition, with each output scored on the same 0–10 evaluation scale. Consistency was quantified within each case-study and condition group using the mean, standard deviation (SD) and coefficient of variation (CV). RAG and non-RAG outcomes were compared across the seven paired case studies using Wilcoxon signed-rank tests.

To further assess the robustness and generalizability of the framework, an extended validation dataset comprising 105 documented environmental events was assembled. The dataset included 35 floating debris and marine litter cases, 35 turbidity cases, and 35 oil-spill events. Cases were compiled from scientific publications [30–40], the NOAA Incident Archive (<https://incidentnews.noaa.gov/>, accessed on 4 May 2026), and the MARIDA benchmark dataset [41], thereby covering a diverse range of environmental monitoring scenarios. The MARIDA-based cases utilized Sentinel-2 imagery and associated reference annotations, while the oil-spill dataset included events spanning different geographical regions, environmental conditions, and spill characteristics.

Gemini 3 Flash was selected as the representative model for all additional validation experiments due to its strong overall performance, fast response times, and relatively low computational cost. Specifically, it was used for the repeated-run consistency analysis, the inter-evaluator agreement assessment, and the extended validation involving 105 documented environmental events.

3. Results

Workflow quality and conversational efficiency exhibited a clear relationship across the evaluated models (Figure 2). Higher-performing systems generally required fewer user interactions, indicating a greater ability to generate satisfactory EO workflows with limited refinement. Gemini 3.1 PRO (RAG) achieved the highest overall performance, reaching a mean score of approximately 9.1/10 while requiring fewer than five messages on average. Gemini 3 Flash (RAG) also combined high evaluation scores with relatively efficient interactions. In contrast, lower-performing models typically required more con-

versational exchanges while producing less satisfactory outputs. Most models clustered between 7 and 9 messages, suggesting that iterative interaction remained necessary even for advanced systems. Overall, the distribution indicates that stronger models tended to achieve higher-quality results with fewer interactions.

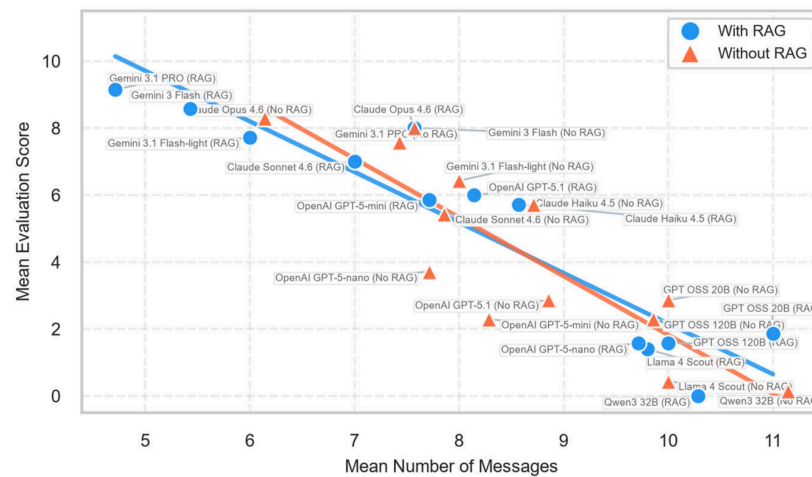


Figure 2. Relationship between workflow quality and conversational efficiency across the evaluated AI models, where higher evaluation scores and fewer user interactions indicate superior overall performance.

The overall evaluation scores are summarized in Figure 3. Retrieval-Augmented Generation (RAG) improved the performance of most evaluated models, although the magnitude of improvement varied considerably. The highest-performing systems were Gemini 3.1 PRO and Gemini 3.1 Flash, both of which achieved mean scores above 8/10 under the RAG configuration. Claude Sonnet 4.6, Gemini 3.1 Flash-light, OpenAI GPT-5.1, and OpenAI GPT-5-mini also demonstrated substantial improvements when retrieval support was available. In contrast, Claude Haiku 4.5 showed only marginal differences between configurations, while Claude Opus 4.6 exhibited slightly better performance without RAG. Models such as GPT OSS 20B, OpenAI GPT-5-nano, and Qwen3 32B consistently achieved lower scores regardless of configuration, suggesting that retrieval alone cannot compensate for limited reasoning or workflow-generation capabilities.

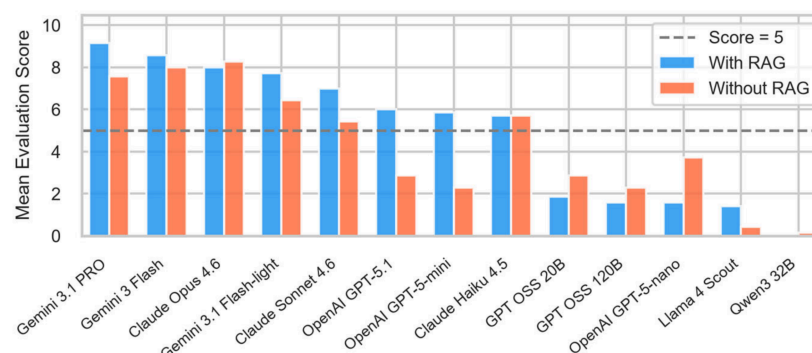


Figure 3. Overall mean evaluation scores of the evaluated AI models with and without Retrieval-Augmented Generation (RAG), with the dashed line indicating the satisfactory performance threshold (5/10).

Next, the influence of retrieval augmentation is examined more directly (Figure 4), which presents the score difference between the RAG and non-RAG configurations. OpenAI GPT-5-mini and GPT-5.1 exhibited the largest average gains, followed by Gemini 3.1

PRO, Claude Sonnet 4.6, and Gemini 3.1 Flash-light. These findings indicate that advanced frontier and mid-sized models are generally capable of effectively integrating retrieved domain knowledge into the workflow-generation process. In contrast, several compact models showed limited improvements or even performance degradation when retrieval was enabled. Three of the four compact models achieved lower scores under the RAG configuration, suggesting that additional contextual information may not always be beneficial when the underlying model lacks sufficient reasoning capacity. Overall, the results indicate that the effectiveness of RAG depends not only on the quality of the retrieved information but also on the ability of the model to synthesize and utilize that information appropriately.

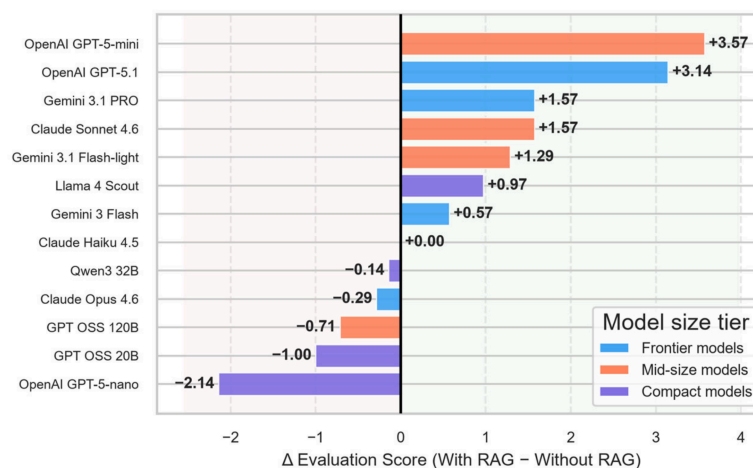


Figure 4. Impact of Retrieval-Augmented Generation (RAG) on model performance, expressed as the difference in mean evaluation score between the RAG and non-RAG configurations.

Furthermore, the impact of retrieval augmentation across the individual maritime EO case studies is assessed (Figure 5). Considerable variability was observed among both models and applications. Oil-spill scenarios generally exhibited the strongest positive response to RAG, particularly for models such as GPT-5.1, GPT-5-mini, Gemini 3.1 Flash-light, and Gemini 3.1 PRO. These applications often require domain-specific contextual information regarding sensor selection, preprocessing approaches, and interpretation of SAR signatures, making them particularly suitable for retrieval-enhanced reasoning. In contrast, air-quality assessment, vessel detection, and turbidity monitoring produced more moderate and model-dependent improvements. Several models also experienced performance degradation under the RAG configuration for specific case studies, indicating that retrieved information can occasionally introduce ambiguity or distract weaker models from the original analytical objective.

Across the five repeated runs, Gemini 3 Flash showed a mean within-group SD of 1.39 points on the 0–10 scale. Performance under the RAG configuration was slightly higher and less variable than the non-RAG configuration, with a mean score of 8.34 versus 8.0, a mean within-case SD of 1.35 versus 1.43, and a mean CV of 0.17 versus 0.25. However, none of these paired differences were statistically significant (Wilcoxon $p = 0.453$ for mean score and $p = 0.813$ for both SD and CV; seven paired case studies). Across the 14 matched case study-condition combinations, the evaluators demonstrated moderate agreement (ICC(2,1) = 0.47; ICC(3,1) = 0.46). Pearson ($r = 0.49$) and Spearman ($\rho = 0.44$) correlations further indicated that both evaluators generally identified the same case studies as exhibiting relatively stronger or weaker performance. Evaluator 1 assigned scores that were, on average, only 0.21 points higher than those of Evaluator 2, with no evidence of systematic bias (paired t -test: $p = 0.533$; Wilcoxon signed-rank test: $p = 0.635$). Furthermore, Evaluator 1's score fell within the range of Evaluator 2's scores in 13 of the 14 comparisons (92.9%)

and within the corresponding one-SD interval in 11 of the 14 comparisons (78.6%). All observed score differences lay within the Bland–Altman 95% limits of agreement (−2.2 to +2.7 points).

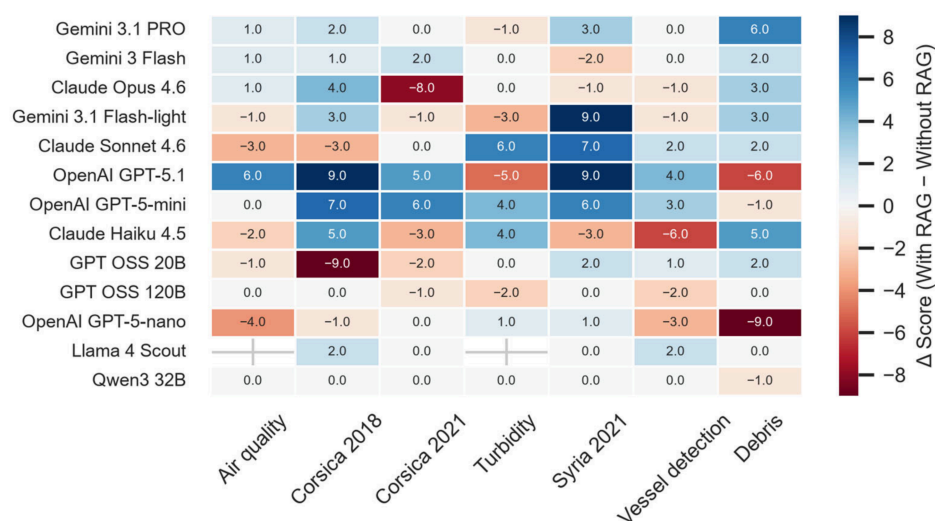


Figure 5. Impact of RAG across the evaluated maritime EO case studies, expressed as the difference in evaluation score between the RAG and non-RAG configurations.

Finally, the extended validation dataset demonstrated consistently strong performance across all three application categories. Workflow generation success reached 94.3% for plastic detection, 97.1% for turbidity assessment, and 97.1% for oil-spill detection. Event detection success was 88.6%, 94.3%, and 91.4%, respectively, while the corresponding expert assessment scores were 7/10, 9/10, and 8/10. These results indicate that SeaScope can reliably generate executable EO workflows and successfully identify documented environmental phenomena across diverse monitoring applications. Performance was occasionally constrained by limitations inherent to the underlying EOs rather than the workflow-generation process itself. Cloud cover and reduced image quality affected the visibility of floating debris and turbidity features in a small number of Sentinel-2 cases, while several oil-spill events lacked suitable satellite acquisitions during the documented event period.

4. Discussion and Conclusions

The rapid emergence of Large Language Models has created new opportunities for simplifying access to EO data and cloud-based geospatial analysis. However, the results of this study demonstrate that the successful deployment of AI-assisted EO systems depends not only on advances in language models but also on the integration of domain knowledge, workflow traceability, and executable analytical environments. SeaScope represents one implementation of this emerging class of AI-assisted EO platforms, combining natural-language interaction, RAG, GEE workflow generation, provenance tracking, and human validation within a unified framework.

The benchmark results indicate that retrieval augmentation can substantially improve EO workflow generation when coupled with capable reasoning models. At the same time, the observed variability among models highlights that access to additional information alone does not guarantee improved performance. The effectiveness of retrieval-enhanced workflows ultimately depends on the ability of the underlying model to interpret, prioritize, and integrate external knowledge within the analytical process. These findings suggest

that future EO agents should consider not only retrieval quality but also adaptive retrieval strategies that account for model capabilities and task complexity.

Beyond the benchmark analysis, the extended validation dataset demonstrated that the proposed framework can successfully reproduce a broad range of documented environmental phenomena. Although not intended as a replacement for dedicated EO algorithms or expert interpretation, these results indicate that AI-assisted workflow generation can support the rapid exploration and analysis of diverse environmental events across different sensors, regions, and monitoring objectives. More broadly, the results support the ongoing transition toward integrated AI-assisted EO platforms that combine natural-language interaction, domain knowledge, and cloud-based geospatial processing within unified analytical environments. Traditionally, satellite data exploitation has required substantial expertise in sensor characteristics, geospatial programming, and cloud-based processing environments. SeaScope, together with other recently proposed AI-assisted EO frameworks, illustrates an emerging paradigm in which users interact with EO systems through analytical objectives rather than implementation details. In such environments, AI agents can assist with dataset selection, workflow construction, parameter configuration, execution, and result interpretation while maintaining transparency through provenance-aware reporting and user oversight.

Although the present evaluation focuses on maritime and coastal monitoring, SeaScope has also been piloted for urban heat and wildfire monitoring with limited changes to its domain knowledge base, supporting its transferability across EO domains. Future developments may incorporate larger knowledge bases, multimodal EO reasoning, autonomous workflow optimization, additional cloud-computing platforms, and tighter integration with scientific literature. As foundation models continue to evolve, retrieval-enhanced AI agents have the potential to become an increasingly important interface between users and the rapidly expanding EO data ecosystem. The combination of retrieval-enhanced reasoning and cloud-native EO processing provides a promising direction for the next generation of transparent and reproducible AI-assisted EO systems.

Author Contributions: Conceptualization, C.S., K.P., C.C., S.P.N. and M.M.; methodology, C.S., I.A. and K.P.; software, C.S. and I.Y.; validation, C.S., L.M., F.M. and G.P.; formal analysis, C.S., I.A., L.M. and F.M.; investigation, L.M.; resources, C.S.; data curation, C.S., L.M. and I.Y.; writing—original draft preparation, C.S., L.M., I.A. and F.M.; writing—review and editing, C.S., I.A., K.P., C.C., S.P.N. and M.M.; visualization, C.S. and I.Y.; supervision, C.S., I.A., C.C., S.P.N. and M.M.; project administration, C.S. and G.P.; funding acquisition, C.S., C.C., S.P.N. and M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union, via oc1-2025-TIS-01, issued and implemented by the ENFIELD project, under the grant agreement No. 101120657.

Data Availability Statement: The SeaScope source code, project documentation, and associated materials are publicly available at github.com/ctrl-space-labs/sea-scope (accessed on 10 August 2026).

Conflicts of Interest: Authors Christos Sekas, Ioannis Yfantidis and George Paterakis were employed by CTRL + Space Development E.E. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhao, Q.; Yu, L.; Du, Z.; Peng, D.; Hao, P.; Zhang, Y.; Gong, P. An Overview of the Applications of Earth Observation Satellite Data: Impacts and Future Trends. *Remote Sens.* **2022**, *14*, 1863. [\[CrossRef\]](#)
2. Brekke, C.; Solberg, A.H. Oil Spill Detection by Satellite Remote Sensing. *Remote Sens. Environ.* **2005**, *95*, 1–13. [\[CrossRef\]](#)

3. Dong, S.; Feng, J.; Gu, Z.; Yin, K.; Long, Y. A Review of Artificial Intelligence and Remote Sensing for Marine Oil Spill Detection, Classification, and Thickness Estimation. *Remote Sens.* **2025**, *17*, 3681. [CrossRef]
4. Kalogirou, E.; Christofi, K.; Makri, D.; Iqbal, M.A.; La Pegna, V.; Tzouvaras, M.; Mettas, C.; Hadjimitsis, D. Oil Spill Detection Using Convolutional Neural Networks and Sentinel-1 SAR Imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2025**, *XLVIII-G-2025*, 757–764. [CrossRef]
5. Evagorou, E.; Hasiotis, T.; Petsimeris, I.T.; Monioudi, I.N.; Andreadis, O.P.; Chatzipavlis, A.; Christofi, D.; Kountouri, J.; Stylianou, N.; Mettas, C.; et al. A Holistic High-Resolution Remote Sensing Approach for Mapping Coastal Geomorphology and Marine Habitats. *Remote Sens.* **2025**, *17*, 1437. [CrossRef]
6. Batista, T.; Jamal, S.A.; Isbaex, C. Sentinel Data for Monitoring of Pollutant Emissions by Maritime Transport—A Literature Review. *Remote Sens.* **2025**, *17*, 2202. [CrossRef]
7. Zheng, Z.; Ning, K.; Wang, Y.; Zhang, J.; Zheng, D.; Ye, M.; Chen, J. A Survey of Large Language Models for Code: Evolution, Benchmarking, and Future Trends. *arXiv* **2023**, arXiv:2311.10372.
8. Chen, Y.; Wang, W.; Lobry, S.; Kurtz, C. An LLM Agent for Automatic Geospatial Data Analysis. *arXiv* **2024**, arXiv:2410.18792.
9. Mansourian, A.; Oucheikh, R. ChatGeoAI: Enabling Geospatial Analysis for Public through Natural Language, with Large Language Models. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 348. [CrossRef]
10. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; Khan, F.S. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 27831–27840.
11. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Mao, X. EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5917820. [CrossRef]
12. Lee, C.; Paramanayakam, V.; Karatzas, A.; Jian, Y.; Fore, M.; Liao, H.; Yu, F.; Li, R.; Anagnostopoulos, I.; Stamoulis, D. Multi-Agent Geospatial Copilots for Remote Sensing Workflows. In *Proceedings of the IGARSS 2025—2025 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: New York, NY, USA, 2025; pp. 1084–1089.
13. Hou, S.; Shen, Z.; Wu, H.; Liang, J.; Jiao, H.; Qing, Y.; Zhang, X.; Li, X.; Gui, Z.; Guan, X.; et al. AutoGEEval: A Multimodal and Automated Framework for Geospatial Code Generation on GEE with Large Language Models. *arXiv* **2025**, arXiv:2505.12900.
14. Hou, S.; Zhao, A.; Liang, J.; Shen, Z.; Wu, H. Geo-FuB: A Method. for Constructing an Operator-Function Knowledge Base for Geospatial Code Generation with Large Language Models. *Knowl. Based Syst.* **2025**, *319*, 113624. [CrossRef]
15. Hou, S.; Shen, Z.; Zhao, A.; Liang, J.; Gui, Z.; Guan, X.; Li, R.; Wu, H. GeoCode-GPT: A Large Language Model for Geospatial Code Generation. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *138*, 104456. [CrossRef]
16. Zhao, A.; Gui, Z.; Liang, J.; Liu, Y.; Peng, D.; Wei, W.; Hou, S.; Wu, H. Extraction of Geoprocessing Modeling Knowledge from Crowdsourced Google Earth Engine Scripts by Coordinating Large and Small Language Models. *Int. J. Geogr. Inf. Sci.* **2026**, *40*, 808–841. [CrossRef]
17. Hou, S.; Liang, J.; Zhao, A.; Wu, H. GEE-OPs: An Operator Knowledge Base for Geospatial Code Generation on the Google Earth Engine Platform Powered by Large Language Models. *Geo-Spat. Inf. Sci.* **2026**, *29*, 429–450. [CrossRef]
18. Liang, J.; Hou, S.; Jiao, H.; Qing, Y.; Zhao, A.; Shen, Z.; Xiang, L.; Wu, H. GeoGraphRAG: A Graph-Based Retrieval-Augmented Generation Approach for Empowering Large Language Models in Automated Geospatial Modeling. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *142*, 104712. [CrossRef]
19. Xu, W.; Yu, Z.; Mu, B.; Wei, Z.; Zhang, Y.; Li, G.; Wang, J.; Peng, M. RS-Agent: Automating Remote Sensing Tasks through Intelligent Agent. Available online: <https://arxiv.org/abs/2406.07089v3> (accessed on 28 May 2026).
20. Feng, P.; Lv, Z.; Ye, J.; Wang, X.; Huo, X.; Yu, J.; Xu, W.; Zhang, W.; Bai, L.; He, C.; et al. Earth-Agent: Unlocking the Full Landscape of Earth Observation with Agents. Available online: <https://arxiv.org/abs/2509.23141v3> (accessed on 28 May 2026).
21. Kao, C.H.; Zhao, W.; Lam, C.; Umap, A.; Revankar, S.; Speas, S.; Bhagat, S.; Datta, R.; Phoo, C.P.; Mall, U.; et al. Towards LLM Agents for Earth Observation. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026*; Liakata, M., Moreira, V.P., Zhang, J., Jurgens, D., Eds.; Association for Computational Linguistics: San Diego, CA, USA, 2026; pp. 2597–2611.
22. Chen, J.; Pedrycz, W.; Wang, F.; Wu, J.; Wu, G.; Xing, L.; Chen, Y.; Song, Y. A Tri-Stage LLM-Coordinated Framework for Order-Driven Scheduling of Earth Observation Satellite Tasks. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 3002015. [CrossRef]
23. Terry, B.; Wagner, O.; Yetkin, O.; Gordon, J.; Baptist, J.R.; Borges, D.; Gowda, S. Improving EO-GPT: An LLM Orchestration Framework for Earth Observation Analysis. In *Proceedings of the IGARSS 2025—2025 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: New York, NY, USA, 2025; pp. 6286–6290.
24. Zhang, W.; Cai, M.; Zhang, T.; Lei, G.; Zhuang, Y.; Mao, X. Popeye: A Unified Visual-Language Model for Multi-Source Ship Detection from Remote Sensing Imagery. Available online: <https://arxiv.org/abs/2403.03790v2> (accessed on 28 May 2026).
25. Xue, W.; Ai, J.; Zhu, Y.; Sun, X.; Zhang, Y.; Gao, G. LMCNet: Lightweight Modality Compensation Network via Knowledge Distillation for Salient Ship Detection under Missing Modality Conditions. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 6547–6560. [CrossRef]

26. Galdelli, A.; Narang, G.; Pietrini, R.; Zazzarini, M.; Fiorani, A.; Tasseti, A.N. Multimodal AI-Enhanced Ship Detection for Mapping Fishing Vessels and Informing on Suspicious Activities. *Pattern Recognit. Lett.* **2025**, *191*, 15–22. [[CrossRef](#)]
27. Song, N.; Lv, C.; Nie, J.; Ye, M.; Zhao, E.; Ma, J.; Liu, X.; Wei, Z. LLM-DSK: A Domain-Specific Semantic Knowledge-Guided Ocean Environment Prediction Method Based on Large Language Models. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 19453–19469. [[CrossRef](#)]
28. Lacaux, J.P.; Tourre, Y.M.; Vignolles, C.; Ndione, J.A.; Lafaye, M. Classification of Ponds from High-Spatial Resolution Remote Sensing: Application to Rift Valley Fever Epidemics in Senegal. *Remote Sens. Environ.* **2007**, *106*, 66–74. [[CrossRef](#)]
29. Biermann, L.; Clewley, D.; Martinez-Vicente, V.; Topouzelis, K. Finding Plastic Patches in Coastal Waters Using Optical Satellite Data. *Sci. Rep.* **2020**, *10*, 5364. [[CrossRef](#)] [[PubMed](#)]
30. Yananto, A.; Iswari, M.Y.; Yulianto, F.; Radityo, S.H.H.; Nugroho, S.; Fachrudin, I. Spatial and Temporal Oil Spill Distribution Monitoring in the Java Sea in 2019 with Multi-Temporal Sentinel-1 SAR Data. *AIP Conf. Proc.* **2024**, *3069*, 020133. [[CrossRef](#)]
31. Rajendran, S.; Vethamony, P.; Sadooni, F.N.; Al-Kuwari, H.A.-S.; Al-Khayat, J.A.; Seegobin, V.O.; Govil, H.; Nasir, S. Detection of Wakashio Oil Spill off Mauritius Using Sentinel-1 and 2 Data: Capability of Sensors, Image Transformation Methods and Mapping. *Environ. Pollut.* **2021**, *274*, 116618. [[CrossRef](#)] [[PubMed](#)]
32. Prastyani, R.; Basith, A. Utilisation of Sentinel-1 SAR Imagery for Oil Spill Mapping: A Case Study of Balikpapan Bay Oil Spill. *JGISE J. Geospat. Inf. Sci. Eng.* **2018**, *1*, 22–26. [[CrossRef](#)]
33. Chaturvedi, S.K.; Banerjee, S.; Lele, S. An Assessment of Oil Spill Detection Using Sentinel 1 SAR-C Images. *J. Ocean. Eng. Sci.* **2020**, *5*, 116–135. [[CrossRef](#)]
34. Nandika, M.R.; Rachman, H.A.; Eguchi, T.; Setiawati, M.D.; Chatterjee, U.; Adi, N.S.; Husain, B.H.; Muslim, A.M. Sentinel-1 SAR Imaging for Detection and Analysis of Oil Spills in the Coastal Waters of Bintan: A Case Study and Environmental Monitoring Application. In *Proceedings of the 2023 8th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR)*; IEEE: New York, NY, USA, 2023; pp. 1–5.
35. Yang, Y.-J.; Singha, S.; Goldman, R.; Schütte, F. Dataset of Oil Slicks, Look-Alikes and Remarkable SAR Signatures Obtained from Sentinel-1 Data in the Eastern Mediterranean Sea. *Earth Syst. Sci. Data* **2025**, *17*, 6807–6837. [[CrossRef](#)]
36. Cheng, P.; Qi, Y.; Zhao, J. Automated Oil Spill Detection Using Deep Learning and SAR Satellite Data: A Case Study of the Lower Congo Basin. *Front. Earth Sci.* **2026**, *13*, 1667450. [[CrossRef](#)]
37. La, T.V.; Pelich, R.-M.; Li, Y.; Matgen, P.; Chini, M. Monitoring of Spatio-Temporal Variations of Oil Slicks via the Collocation of Multi-Source Satellite Images. *Remote Sens.* **2024**, *16*, 3110. [[CrossRef](#)]
38. Okba, Z.; Fernández-Rodríguez, E.; Rowland, Z.; Ligonio, L. Analysis of the Spread of the Callao Oil Spill off the Coast of Lima, Peru Using Earth Observation Data. In *Proceedings of the 73rd International Astronautical Congress (IAC)*, Paris, France, 19 September 2022.
39. Li, C.; Kim, D.; Park, S.; Kim, J.; Song, J. A Self-Evolving Deep Learning Algorithm for Automatic Oil Spill Detection in Sentinel-1 SAR Images. *Remote Sens. Environ.* **2023**, *299*, 113872. [[CrossRef](#)]
40. Yang, Y.-J.; Singha, S.; Goldman, R. A near Real-Time Automated Oil Spill Detection and Early Warning System Using Sentinel-1 SAR Imagery for the Southeastern Mediterranean Sea. *Int. J. Remote Sens.* **2024**, *45*, 1997–2027. [[CrossRef](#)]
41. Kikaki, K.; Kakogeorgiou, I.; Mikeli, P.; Raitos, D.E.; Karantzas, K. MARIDA: A Benchmark for Marine Debris Detection from Sentinel-2 Remote Sensing Data. *PLoS ONE* **2022**, *17*, e0262247. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.