Supplementary Material for Agentic AI in Remote Sensing: Foundations, Taxonomy, and Emerging Systems

1. RS Datasets Across Applications

In this section, we cover representative remote sensing (RS) datasets that ground the application taxonomy in the main paper. In Table 1, we group benchmarks across scene classification, semantic segmentation, object detection, change detection, building and road extraction, disaster and hazard mapping, text–image grounding, and earth observation (EO) foundation pretraining. The table lists each dataset's sensor modality, spatial resolution, and benchmark task. It groups together aerial RGB scene datasets [5, 29, 32], sentinel-based LULC collections [8, 10], disaster-focused resources such as xBD, FloodNet, and Sen1Floods11 [2, 9, 19], and large EO pretraining corpora including SSL4EO-S12 and EarthView [26, 28]. Collectively, these datasets offer a practical catalog for connecting specific RS tasks with suitable sensors and benchmarks when developing and evaluating new methods.

2. Datasets and Benchmarks for Agentic RS

In this section, we cover datasets and evaluation suites that explicitly target LLM-driven agentic methods in geospatial and RS. In Table 2, we summarize benchmarks for geospatial tool use and multi-step reasoning, including GeoBenchX [12] and GTChain-IT / CTChain-Eval [33], multi-turn multi-modal dialogue over SAR and infrared imagery in RS-VL3M [11], and realistic tool-augmented task suites in ThinkGeo and RescueADI [15, 20]. The table further includes ShapefileGPT for Shapefile-based spatial analysis [14] and generic tool-use evaluation frameworks like CORE [36]. Taken together, these benchmarks provide a focused basis for assessing agentic behavior in RS and for comparing emerging systems under consistent evaluation protocols.

References

[1] Yeshwanth Kumar Adimoolam, Bodhiswatta Chatterjee, Charalambos Poullis, and Melinos Averkiou. Efficient deduplication and leakage detection in large scale image datasets with a focus on the crowdai mapping challenge dataset. *arXiv preprint arXiv:2304.02296*, 2023. 2

[2] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 210–211, 2020. 1, 2

[3] Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific data*, 9(1):251, 2022. 2

[4] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote sensing*, 12(10):1662, 2020. 2

[5] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 1, 2

[6] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. Ieee, 2018. 2

[7] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 2

[8] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018. 1, 2

[9] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019. 1, 2

[10] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 1, 2

Dataset	Sensor / Modality	Resolution / Scale	Dataset Application
Scene / LULC classification			
UC Merced Land Use [32]	Aerial RGB	∼0.3 m, 256×256 patches	land-use scene classification (21 classes)
AID [29]	Aerial RGB	600×600 pixel patches	Aerial scene image classification (30 classes)
NWPU-RESISC45 [5]	Aerial RGB	256×256 pixel patches	Scene classification (45 classes)
EuroSAT [10]	Sentinel-2 multispectral	64 × 64 pixel patches	Land Use and Land Cover (LULC) classification (10 classes)
Million-AID [16]	Aerial RGB	Variable (0.5m to 153m)	Aerial scene classification (51 classes)
MLRSNet [18]	Optical Satellite/ Aerial RGB	256×256 pixel patches	Multi-label semantic scene understanding (46 scene categories, 60 labels)
Semantic segmentation (urban, LULC)			
Inria Aerial Image Labeling [17]	Aerial RGB	5000×5000 px, 0.3 m/pixel	Semantic segmentation
DeepGlobe Land Cover [8]	Satellite RGB	2448×2448 px, 0.5 m/pixel	Rural land cover semantic segmentation
LoveDA [27]	Spaceborne RGB satellite	1024×1024 px, 0.3 m/pixel	Land-cover segmentation under domain shift (rural/urban)
DynamicEarthNet [25]	Planet multi-spectral satellite	1024×1024 px, 3 m GSD	LULC semantic and change segmentation.
Dynamic World [3]	Sentinel-2 multi-spectral images	Global 10 m/pixel	Near real-time LULC mapping
Object detection / instance segmentation			
DOTA [30]	Optical aerial/satellite imagery (RGB/gray)	High-resolution, variable up to 20k	Oriented object detection in aerial images
xView [13]	WorldView-3 satellite imagery	0.3 m GSD, 1 km^2 chips	Overhead multi-class object detection
FAIR1M [24]	High-res optical satellite	0.3–0.8 m GSD, 1k–10k pixel	Fine-grained oriented object detection, classification
Change detection (bi-/multi-temporal)			
LEVIR-CD [4]	Google Earth VHR RGB	1024×1024 pixel, 0.5 m/pixel	Bitemporal building change segmentation
SYSU-CD [23]	0.5 m RGB aerial imagery	256×256 pixel, 0.5 m GSD	Bitemporal high-resolution change detection
S2Looking [21]	Side-looking RGB optical satellite imagery	1024×1024, 0.5–0.8 m GSD	Bitemporal building change detection
OSCD (Onera)[6]	Sentinel-2 multispectral optical imagery	600×600 at 10m resolution	Urban binary change detection.
Building / road extraction			
DeepGlobe [8]	Satellite Optical RGB	2448×2448 pixel, 0.5 m/pixel	Rural land cover segmentation.
DeepGlobe Road [7]	Satellite RGB	1024×1024 px tiles, 0.5 m/pixel	Road and street network extraction
CrowdAI [1]	RGB satellite imagery	300×300 pixel tiles, 0.3 m GSD	Building footprint detection / segmentation
Disaster, damage, hazard mapping			
xBD [9]	Multispectral satellite imagery	$\leq$ 0.8 m GSD	Building damage assessment, change detection
FloodNet [19]	UAV RGB	4000×3000 pixel, 1.5 cm GSD	Post-flood damage segmentation and VQA
Sen1Floods11 [2]	Sentinel-1 SAR imagery	512×512 chips, 120406 km^2 global	Flood and permanent water segmentation
UrbanSARFloods [34]	Sentinel-1 SAR	512×512 chips, 807500 km^2	Urban and open-area flood segmentation
FireRisk [22]	NAIP aerial RGB	270×270 px tiles, 1 m	Wildfire risk level classification
Text–image, captioning, VQA			
RSICD [31]	Aerial / satellite RGB	Patch-level	RS image captioning and text–image alignment
RSIVQA [35]	Multi-source aerial / satellite RGB imagery	Variable, 0.1-8 m GSD	VQA for RS scene understanding
FloodNet-VQA [19]	UAV RGB aerial	4000×3000 px, 1.5 cm GSD	Post-flood scene understanding, segmentation, VQA
Pretraining corpora / EO foundation			
SSL4EO-S12 [28]	Sentinel-1 SAR, Sentinel-2 multispectral	264× 264 pixel, 2640×2640 m	Self-supervised EO pretraining, downstream tasks Elib DLR +1
EarthView [26]	Multisource optical RS	Mixed 1-30 m GSD, global	Self-supervised pretraining for EO

Table 1. Representative benchmarks and datasets for remote sensing, grouped by application category (shown as section headers). The table highlights typical sensors, spatial scale, and primary benchmark tasks to support method selection and evaluation design.

	Dataset / Benchmark	Applications	Systems and Technologies
GeoBenchX [12]	Dataset and evaluation framework	Multi-step GIS reasoning	LangGraph ReAct agent, Python geospatial stack, and an LLM as Judge
GTChain-IT / CTChain-Eval [33]	Dataset and evaluation framework	Benchmarking LLMs on geospatial tool use tasks	Simulated tool-use environment and fixed GIS tool APIs
RS-VL3M [11]	Benchmark	Benchmark for multi turn dialogue over SAR/IR with joint perception	Infrared RS images with scene labels, combined with SAR-CLA and optical benchmarks in multi modality
ThinkGeo [20]	Benchmark	Benchmark to evaluate tool-augmented LLM agents on realistic remote sensing tasks	ReAct tool-calling with AgentLego tools, RGB/SAR imagery
RescueADI [15]	Benchmarks	Adaptive disaster interpretation	PSPNet, GroundingDINO, counting and area tools
Shapefile [14]	Benchmark	Benchmarking on 42 Shapefile spatial analysis tasks	27-function Shapefile GIS tool library
CORE [36]	Eval frameworks	Evaluation framework for tool-using agents	Simulated tool APIs with CORE path metrics

Table 2. Overview of datasets and evaluation benchmarks for LLM-driven agentic methods in geospatial and remote sensing.

[11] Huiyang Hu, Peijin Wang, Yingchao Feng, Kaiwen Wei, Wenxin Yin, Wenhui Diao, Mengyu Wang, Hanbo Bi, Kaiyue Kang, Tong Ling, et al. RINGMO-Agent: A unified remote sensing foundation model for multi-platform and multi-modal reasoning. *arXiv preprint arXiv:2507.20776*, 2025. 1, 3

[12] Varvara Krechetova and Denis Kochedykov. GeoBenchX: Benchmarking LLMs in agent solving multistep geospatial tasks. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time Intelligence*, page 27–35, New York, NY, USA, 2025. Association for Computing Machinery. 1, 3

[13] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018. 2

[14] Qingming Lin, Rui Hu, Huaxia Li, Sensen Wu, Yadong Li, Kai Fang, Hailin Feng, Zhenhong Du, and Liuchang Xu. ShapefileGPT: A multi-agent large language model framework for automated shapefile processing. *International Journal of Digital Earth*, 18(2):2577884, 2025. 1, 3

[15] Zhuoran Liu, Danpei Zhao, Bo Yuan, and Zhiguo Jiang. RescueADI: adaptive disaster interpretation in remote sensing images with autonomous agents. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 1, 3

[16] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230, 2021. 2

[17] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International geoscience and remote sensing symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 2

[18] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020. 2

[19] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. 1, 2

[20] Akashah Shabbir, Muhammad Akhtar Munir, Akshay Dudhane, Muhammad Umer Sheikh, Muhammad Haris Khan, Paolo Fraccaro, Juan Bernabe Moreno, Fahad Shahbaz Khan, and Salman Khan. THINKGEO: Evaluating tool-augmented agents for remote sensing tasks. *arXiv preprint arXiv:2505.23752*, 2025. 1, 3

[21] Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021. 2

[22] Shuchang Shen, Sachith Seneviratne, Xinye Wanyan, and Michael Kirley. Firerisk: A remote sensing dataset for fire risk assessment with benchmarks using supervised and self-supervised learning. In *2023 international conference on digital image computing: techniques and applications (DICTA)*, pages 189–196. IEEE, 2023. 2

[23] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE transactions on geoscience and remote sensing*, 60:1–16, 2021. 2

[24] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 2

[25] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel

Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022. 2

[26] Diego Velazquez, Pau Rodriguez, Sergio Alonso, Josep M Gonfaus, Jordi Gonzalez, Gerardo Richarte, Javier Marin, Yoshua Bengio, and Alexandre Lacoste. Earthview: a large scale remote sensing dataset for self-supervision. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1228–1237, 2025. 1, 2

[27] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 2

[28] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. 1, 2

[29] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 1, 2

[30] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 2

[31] Bhavitha Yamani, Nikhil Medavarapu, and S Rakesh. Remote sensing image captioning using deep learning. In *2024 International Conference on Automation and Computation (AUTOCOM)*, pages 295–302. IEEE, 2024. 2

[32] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 1, 2

[33] Yifan Zhang, Jingxuan Li, Zhiyun Wang, Zhengting He, Qingfeng Guan, Jianfeng Lin, and Wenhao Yu. Geospatial large language model trained with a simulated environment for generating tool-use chains autonomously. *International Journal of Applied Earth Observation and Geoinformation*, 136:104312, 2025. 1, 3

[34] Jie Zhao, Zhitong Xiong, and Xiao Xiang Zhu. Urbansarfloods: Sentinel-1 slc-based benchmark dataset for urban and open-area flood mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 419–429, 2024. 2

[35] Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 2

[36] Yutong Zuo, Zirui Wang, Jiaxin Zhang, Yilun Wu, Bo Li, Erpeng Zhu, Lihong Jiang, Xifeng Zhang, Stanley K. S. Yau, Zhaoyuan Lin, et al. CORE: Full-path evaluation of LLM agents beyond final state. *arXiv preprint arXiv:2407.03728*, 2024. 1, 3