# PRACTICE AND EXPERIENCE IN USING PARALLEL AND SCALABLE MACHINE LEARNING IN REMOTE SENSING FROM HPC OVER CLOUD TO QUANTUM COMPUTING

*Morris Riedel*[1,2], *Gabriele Cavallaro*[2], *Jón Atli Benediktsson*[1]

[1] School of Engineering and Natural Sciences, University of Iceland, Iceland
[2] Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany

## ABSTRACT

Using computationally efficient techniques for transforming the massive amount of Remote Sensing (RS) data into scientific understanding is critical for Earth science. The utilization of efficient techniques through innovative computing systems in RS applications has become more widespread in recent years. The continuously increased use of Deep Learning (DL) as a specific type of Machine Learning (ML) for data-intensive problems (i.e., 'big data') requires powerful computing resources with equally increasing performance. This paper reviews recent advances in High-Performance Computing (HPC), Cloud Computing (CC), and Quantum Computing (QC) applied to RS problems. It thus represents a snapshot of the state-of-the-art in ML in the context of the most recent developments in those computing areas, including our lessons learned over the last years. Our paper also includes some recent challenges and good experiences by using Europeans fastest supercomputer for hyper-spectral and multi-spectral image analysis with state-of-the-art data analysis tools. It offers a thoughtful perspective of the potential and emerging challenges of applying innovative computing paradigms to RS problems.

***Index Terms***— High performance computing, cloud computing, quantum computing, machine learning, deep learning, parallel and distributed algorithms, remote sensing.

## 1. INTRODUCTION

Already ten years ago, Lee et al. [1] highlighted in a survey paper for RS a trend in the design of HPC systems for data-intensive problems is to utilize highly heterogeneous computing resources. Today, the use of HPC systems becomes more and more mainstream with many machines being heavily used by RS researchers, especially with new forms of multi-spectral or hyper-spectral image analysis techniques through DL. DL requires massive processing power and large quantities of data and open-source frameworks, whereby all of these three factors contributed to the high uptake of this unique ML technique in the RS community. CC evolved as an evolution of Grid computing to make parallel and distributed computing more straightforward to use than traditional rather complex HPC systems. In this context RS researchers often take advantage of Apache open-source tools with parallel and distributed algorithms (e.g., map-reduce [2] as a specific form of divide and conquer approach) based on Spark [3] or the larger Hadoop ecosystem [4]. Inherent in many ML and DL approaches are optimization techniques while many of them are incredibly fast solvable by QCs [5] that represent the most innovative type of computing today. Despite being in its infancy, Quantum Annealer (QA)s are specific forms of QC used by RS researchers [6, 7] to search for solutions to optimization problems already today.

In this experience and short review paper, we specifically focus on describing recent advances in the fields of HPC, CC, and QC applied to RS problems, covering innovative computing architectures utilizing multi-core processors, many-core processors, and specialized hardware components such as Tensor Processing Unit (TPU)s and QA chips. Relevant examples of using innovative architectures with parallel and scalable data science methods such as ML and DL techniques for RS show the community's uptake, including cutting-edge distributed training algorithms and tools. Also, we share lessons learned and thus add our own long year experience of using several of those innovative systems with ML and DL algorithms on many different RS datasets. This paper does not address the benefits of using such innovative computing systems or Field Programmable Gate Arrays (FPGA)s for on-board airborne or real-time processing on satellite sensor platforms due to the page restriction.

The remainder of the paper is structured as follows. Section 2 describes different innovative computing systems and their technological advances with examples in applying various ML and DL methods for RS datasets. Section 3 concludes the paper with some remarks and anticipates future directions and challenges in applying HPC, CC, or QC to RS problems. The main processing-intensive application areas and challenges for RS are briefly discussed in context.

This work was performed in the Center of Excellence (CoE) Research on AI- and Simulation-Based Engineering at Exascale (RAISE) receiving funding from EU's Horizon 2020 Research and Innovation Framework Programme H2020-INFRAEDI-2019-1 under grant agreement no. 951733.

## 2. TECHNOLOGICAL ADVANCES DRIVING INNOVATIVE COMPUTING SYSTEMS

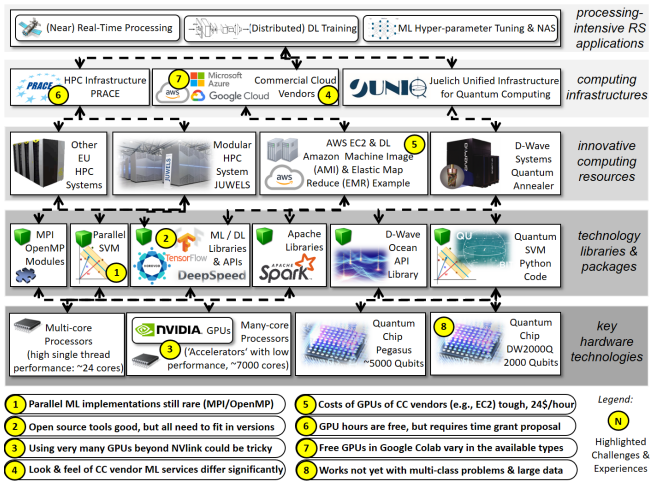


**Fig. 1**. Technology advances and our identified challenges.

Using multiple processors simultaneously with parallel computing and becoming a standard in applying HPC techniques to RS was already predicted ten years ago by Lee et al. [1]. Today, multi-core processors with high single-thread performance are broadly available via supercomputers for science and offer thousands to millions of cores[1] to be used by HPC techniques as shown in Figure 1. At the time of writing, for example, the Jülich Wizard for European Leadership Science (JUWELS)[2] system at the Jülich Supercomputing Centre (JSC) in Germany represents the fastest European supercomputer offering 122,768 CPU cores in its cluster module. While such systems and multi-core processors offer tremendous performance, the particular challenge to exploit this data analysis performance for ML is that those systems require specific parallel and scalable techniques. In other words, using such HPC systems with RS data effectively requires parallel and scalable algorithm implementations opposed to using plain sci-kit-learn[3], R[4], or different serial algorithms. Parallel ML algorithms for those systems are typically programmed using the Message Passing Interface (MPI) standard, and OpenMP that jointly leverage the power of shared memory and distributed memory via low latency interconnects (e.g., Infiniband) and parallel filesystems (e.g., Lustre).

Given our experience and frequent literature surveys, the availability of open-source parallel and scalable machine learning implementations that go beyond Artificial Neural Network (ANN)s or more recent DL networks is still relatively rare. The reason is the complexity of parallel programming and thus using HPC can be a challenge when the amount of data is relatively moderate (i.e., DL not always succesful). One example in this context is using a more robust classifier such as a parallel and scalable open-source Support Vector Machine (SVM) that we developed and used to speed up the classification of hyper-spectral RS images [8].

The many-core processor era with accelerators brought many advancements to both simulation sciences and data sciences, including many new approaches for analysing RS data with innovative DL techniques. The idea of using many numerous simpler processors with hundreds to thousands of independent processor cores enabled a high degree of parallel processing that fits very nicely to the demands of DL training whereby lots of matrix-matrix multiplications are performed. Today, hundreds to thousands of accelerators like Nvidia Graphics Processing Unit (GPU)s are used in large-scale HPC systems, offering unprecedented processing power for satellite image processing. For example, Europeans' fastest supercomputer JUWELS offers 3744 GPUs (booster module) of the most recent innovative type of Nvidia A100 cards. Over the years, our experience shows clearly that open-source DL packages such as TensorFlow[5] (now including Keras[6]) or pyTorch[7] are powerful tools that work very good for large-scale RS data analysis. Still, it can be challenging to have the right versions of python code matching the available tools and libraries versions on HPC systems with GPUs given the fast advancements of DL libraries, accelerators, and HPC systems. But the real challenge in using GPUs is not using one or a couple of GPUs connected by NVLlink or NVSwitches, but to scale beyond a large-scale HPC node setup using distributed DL training tools such as Horovod[8] or, more recently, DeepSpeed[9]. Our experience in using our supercomputer JUWELS with Horovod with a cutting-edge RESNET-50 DL network indicates a significant speed-up of training time without loosing accuracy [9]. In this initial study, we used 96 GPUs while in a later study driven by Sedona et al. [10], we achieved even a better speed-up on JUWELS using 128 interconnected GPUs after having more experience with Horovod. Using GPUs in the context of Unmanned Aerial Vehicle (UAV)s is shown in [11].

Beside HPC systems, also CC vendors offer multi-core and many-core processing power that is used by RS researchers with parallel and scalable tools such as Apache Spark [4] in the last years. For example, Haut et al. [3] uses Spark to develop a cloud implementation of a DL network for non-linear RS data compression known as AutoEncoder (AE). CC techniques such as Spark pipelines offer also the possibility to work in conjunction with DL techniques such as recently shown by Lunga et al. [12] for RS datasets. Over the years, we observed that CC makes parallel and distributed computing more straightforward to use than traditional HPC systems, such as using the very con-

[1] Top500 list of supercomputers, online: https://www.top500.org/
[2] https://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUWELS
[3] https://scikit-learn.org/stable/
[4] https://www.r-project.org/
[5] https://www.tensorflow.org/
[6] https://keras.io/
[7] https://pytorch.org/
[8] https://horovod.ai/
[9] https://www.deepspeed.ai/

venient Jupyter[10] toolset that abstracts the complexities of underlying computing systems. Unfortunately, our experience revealed that the look and feel of seamlessly using CC services are very different when working with various vendors such as Amazon Web Services (AWS)[11], MS Azure[12], Google Collaboratory[13] or Google Cloud [14]. Also, the use of Jupyter for interactive supercomputing is also becoming more widespread as shown by Goebbert et al. for HPC systems of JSC in [13]. Even more challenging in our experience are the high cost when using CC services of vendors like AWS whereby using Nvidia V100 GPUs (i.e., p3.16xlarge) would cost more than 24 USD per hour. Our RESNET-50 studies mentioned above is using 128 GPUs for many hours, hence, we believe that for the time being RS researchers need to use still the cost-free HPC computational time grants to be feasible unless specific cooperations are formed with vendors. Such HPC grants are provided by e-infrastructures such as Partnership for Advanced Computing in Europe (PRACE)[15] in the EU (e.g., that includes free of charge A100 GPUs in JUWELS) or Extreme Science and Engineering Discovery Environment (XSEDE)[16] in the US. Our experience reveals further that free CC resources of commercial vendors typically have drawbacks like the Google Collaboratory example getting just different types of GPUs assigned that make it relatively hard to perform proper speed-up studies not even mentioning the missing possibility to interconnect GPUs for large-scale usage. Other examples of RS approaches of using Spark with CC are distributed parallel algorithms for anomaly detection in hyper-spectral images as shown in [14].

QA emerged as a promising and highly innovative computing approach used for simple RS data analysis problems to solve ML algorithms' optimisation problems [5]. Our experience with using quantum SVMs reveals that on QA architectures such as a D-Wave system[17] with 2000 qubits, RS researchers' possibilities are still limited by having only binary classification techniques or the requirement to subsample from large datasets and using ensemble methods [7]. Recent experience revealed that QA evolutions bears a lot of potentials since we are already using D-Wave Leap[18] with the QQ Advantage system that offers more than 5000 qubits and 35000 couplers enabling powerful RS data analysis. Another recent example of using QA for feature extraction and segmentation is shown by Otgonbaatar et al. in [15].

We observe a massive increase in complexity in the technological advances and an unprecedented heterogeneity in available HPC, CC, and QC systems that will raise challenges for their seamlessly use by domain-specific scientists such as in the RS community. Our studies on recent highly heterogenous Modular Supercomputing Architecture (MSA)s reveals quite some complexity for using cutting-edge HPC systems for RS researchers while still offering enormous benefits as shown by Erlingsson et al. in [16]. But not only is the pace of complex technologies fast (e.g., GPUs from K40s to A100s today), also new methods of Artificial Intelligence (AI) with innovative DL approaches or hyper-parameter optimization techniques like Neural Architecture Search (NAS) [17] are moving forward with an ever-increasing pace. Applying those cutting-edge computing systems with innovative AI technologies in complex RS application research questions requires more inter-disciplinary expertise than ever. One possible solution to deal with this complexity based on our experience at the JSC is establishing so-called domain-specific Simulation and Data Laboratories (SimDataLab)s[19] that consists of multi-disciplinary teams w.r.t. technologies but focus on one particular area of science (e.g., RS, neurosciences, etc.). Substantial multi-disciplinary efforts are needed in order to enable RS researchers to solve problems with high scientific impact through efficient use of HPC, CC, or QC resources today and even more in the future.

To help meet this challenge the JSC in Germany successfully operates since 2004 a wide variety of SimDataLabs as domain-specific research and support structure. A similar AI oriented research and support structure like SimDataLabs are the newly formed Helmholtz AI[20] local units in Germany at the German Aerospace Center and JSC that both support RS applications. Based on these succesful models, the University of Iceland has recently created the 'SimDataLab Remote Sensing'[21] with significant expertise to tackle computing challenges in RS funded by the Icelandic National Competence Center (NCC) under the umbrella of the EuroCC EC project [22]. Selected activities of this SimDataLab include supporting HPC, CC, and QC users in using parallel codes (e.g., maintaining the above mentioned parallel SVM code) or scaling their code on multiple GPUs by using tools (e.g., DeepSpeed or Horovod) in conjunction with deep learning packages (e.g., Keras or TensorFlow). Another benefit of SimDataLabs with domain-specific scientists is that the datasets become more manageable, leading to reduced storage usage by having large RS datasets centrally managed on computing resources instead of having them downloaded by each user separately. Even though the SimDataLab RS in Iceland is primarily supporting the Icelandic RS community[23], it also engages in strong international collaborations with many other countries (e.g., China, Germany, Italy, Spain, etc.) interested in overcoming computational challenges in RS.

[10] https://jupyter.org/
[11] https://aws.amazon.com/
[12] https://azure.microsoft.com/en-us/
[13] https://colab.research.google.com/
[14] https://cloud.google.com/
[15] https://prace-ri.eu/
[16] https://www.xsede.org/
[17] https://www.dwavesys.com/quantum-computing
[18] https://cloud.dwavesys.com/leap/
[19] https://www.fz-juelich.de/ias/jsc/EN/Expertise/SimLab/simlab_node.html
[20] https://helmholtz.at/
[21] https://ihpc.is/community/
[22] https://www.eurocc-project.eu/
[23] Icelandic Center for Remote Sensing, online: https://crs.hi.is/

## 3. CONCLUSIONS

We conclude that RS research entails many highly processing-intensive application areas for HPC, CC, and QC systems. Most notably, many current and future RS applications require real- or near real-time processing capabilities that those systems offer with significant speed-ups instead of just conventional desktop computers, small workstations, or laptops. Examples include environmental studies, military applications, tracking and monitoring hazards such as wildland and forest fires, oil spills and chemical/biological contaminations. Fast processing is also necessary for distributed training of DL networks and the computational-intensive process of hyper-parameter optimization using innovative approaches such as NAS. Future challenges include the broader use of transfer learning and intertwined usage of RS approaches with simulation sciences (i.e., using known physical laws with iterative numerical methods). Recently, the Center of Excellence (CoE) Research on AI- and Simulation-Based Engineering at Exascale (RAISE)[24] EU project started to identify such intertwined methodologies by using seismic imaging with remote sensing for oil and gas exploration and well maintenance in conjunction with simulations that raises high requirements for processing power.

## 4. REFERENCES


[1] C. Lee et al., "Recent Developments in High Performance Computing for Remote Sensing: A Review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 3, pp. 508–527, 2011.

[2] Q. Zou et al., "MapReduce Functions to Remote Sensing Distributed Data Processing - Global Vegetation Drought Monitoring as Example," *Journal of Software: Practice and Experience*, vol. 48, no. 7, pp. 1352–1367, 2018.

[3] J.M. Haut et al., "Cloud Deep Networks for Hyperspectral Image Analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 1–17, 2019.

[4] I. Chebbi et al., "A Comparison of Big Remote Sensing Data Processing with Hadoop MapReduce and Spark," in *4th International Conference on Advanced Technologies for Signal and Image Processing*, 2018, pp. 1–4.

[5] M. Henderson et al., "Methods for Accelerating Geospatial Data Processing Using Quantum Computers," 2020.

[6] R. Ayanzadeh et al., "An Ensemble Approach for Compressive Sensing with Quantum Annealers," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2020, to appear.

[7] G. Cavallaro et al., "Approaching Remote Sensing Image Classification with ensembles of Support Vector Machines on the D-Wave Quantum Annealer," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2020, to appear.

[8] G. Cavallaro et al., "On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4634–4646, 2015.

[9] R. Sedona et al., "Remote Sensing Big Data Classification with High Performance Distributed Deep Learning," *MDPI Remote Sensing*, vol. 11, no. 24, 2019.

[10] R. Sedona et al., "Scaling up a Multispectral RESNET-50 to 128 GPUs," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2020, to appear.

[11] R. Wang et al., "An Effective Image Denoising Method for UAV Images via Improved Generative Adversarial Networks," *Sensors*, vol. 18, no. 7, 2018.

[12] D. Lunga et al., "Apache Spark Accelerated Deep Learning Inference for Large Scale Satellite Image Analytics," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1–17, 2020.

[13] J.H. Goebbert et al., "Enabling Interactive Supercomputing at JSC Lessons Learned," in *Proceedings of International Conference on High Performance Computing*, 2019, Lecture Notes in Computer Science Vol. 11203.

[14] Y. Zhang et al., "A Distributed Parallel Algorithm Based on Low-Rank and Sparse Representation for Anomaly Detection in Hyperspectral Images," *MDPI Sensors*, vol. 18, no. 11, 2018.

[15] S. Otgonbaatar et al., "Quantum Annealing Approach: Feature Extraction and Segmentation of Synthetic Aperture Radar Image," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2020, to appear.

[16] E. Erlingsson et al., "Scalable Workflows for Remote Sensing Data Processing with the Deep-Est Modular Supercomputing Architecture," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2019, pp. 5905–5908.

[17] C. Peng et al., "Efficient Convolutional Neural Architecture Search for Remote Sensing Image Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2020.


[24] Center of Excellence RAISE, online: coe-raise.eu