# A Knowledge Graph Framework for Organizing Heterogeneous Datasets for Utilization in Classical and Quantum Computing: Current Challenges and Future Directions


1st Thilanka Munasinghe
*Department of Information Sciences and Technology*
*University at Albany*
Albany, New York, USA
tmunasinghe@albany.edu

2nd Kimberly A. Cornell
*Department of Information Sciences and Technology*
*University at Albany*
Albany, New York, USA
0000-0001-9551-9689

3rd Jennifer. C. Wei
*NASA Goddard Space Flight Center*
*GES-DISC*
Maryland, USA
jennifer.c.wei@nasa.gov

4th George Berg
*Department of Cybersecurity*
*University at Albany*
Albany, New York, USA
gberg@albany.edu

5th James Hendler
*Department of Computer Science*
*Rensselaer Polytechnic Institute*
Troy, New York, USA
hendler@cs.rpi.edu



***Abstract*—The lack of representation in interaction within environmental variables found in literature led to the development of a novel framework that reflects the true nature of the interconnectedness in our environment. We propose an Environmental Interaction Knowledge Graph (EIKG) framework. This general EIKG framework works as the basis for interconnected environmental events by knitting interrelated events such as hurricanes leading to storm surges, which lead to flood events that could cause events such as mudslides and landslides. The cascading nature of one event leading to another related event in the environment requires an adequate understanding of each event using contextual information before conducting any data-driven analytics. This vision paper showcases how the EIKG:floods, EIKG:wildfire EIKG:landslides, etc., can be derived from a base case framework of EIKG as those individual events are interconnected with some common denominator variables. As an example, the precipitation variable is used in the flood case study as well as in the wildfire or drought case study, as excessive precipitation levels lead to floods, and lack of precipitation leads to droughts and wildfires. We identify the precipitation variable as a "common-denominator-variable" in extreme weather events that play a key role in modeling the environment leading to different extreme weather events based on the variability of that variable (varying values where low precipitation leads to drought, and high values lead to floods). Insights from EIKG facilitate data analysis using both classical and Quantum Machine Learning (QML) techniques. The EIKG organizes heterogeneous datasets and integrates relationships to address extreme weather events. This study incorporates various datasets, including mobility data, socioeconomic data from the US Census Bureau, climate data from NASA, and critical infrastructure data.**



***Index Terms*—Knowledge Graphs, Quantum Computing, Heterogeneous Data, Environmental Interactions, Earth Observational Data**


## I. Introduction

Climate change and increasing extreme weather events, such as floods and wildfires, pose significant challenges to infrastructure resilience and population dynamics. In this paper, we introduce a novel framework called an Environmental Interaction Knowledge Graph (EIKG). This framework can be scalable and is designed to integrate and analyze big data on extreme weather and catastrophic environmental events, residential human mobility data, and critical infrastructure vulnerabilities. The use of high temporal and spatial resolution Earth Observational (EO) and environmental data is essential for climate change analysis, and they consist of the Big Data "Vs." Volume, Velocity, Verity, Veracity, etc. [1]. Studying the environmental, societal, and economic impact of extreme weather events requires using heterogeneous datasets that come from various sources. Therefore, complex problems and extreme weather events due to climate change inherently contain complexity that leads to data analytics with complex structures [2].

Studies on environmental systems modeling using Knowledge Graphs (KG)s have been conducted by the research community. The study conducted by Zafeiropoulos et al presents a socio-environmental system that enables scientists to easily extend their modeling work across different spatial and temporal scales when conducting data analysis [3]. The paper published by Islan et al discussed the "KnowUREnvironment: An Automated Knowledge Graph for Climate Change and Environmental Issues", a KG with related environmental issues with respect to climate change based on the findings from scientific literature [4].

What sets the EIKG apart is its ability to extend and adapt to specific extreme events, such as floods, wildfires, and heatwaves. We achieve this by leveraging graph theory in mathematics to represent interactions as heterogeneous nodes and relationships in a graph structure. This novel approach allows for a deeper understanding of how natural disasters affect human dynamics, taking into account the demographics and socioeconomic indicators of those affected regions. The EIKG also facilitates the integration of heterogeneous data sources, such as flood maps, socioeconomic indicators, and population migration patterns, into a unified model capable of performing complex reasoning and predictive analytics.

## II. Current Challenges

### A. Challenges in Temporal and Spatial Resolutions in Data Sources

We have identified that one of the main challenges in constructing an EIKG is the integration and harmonization of heterogeneous datasets from diverse sources. Datasets such as EO data, socioeconomic indicators, and residential mobility data are often disparate regarding resolution, temporal coverage, and format. These differences make it challenging to preprocess and integrate the data needed to construct a unified knowledge graph. For example, EO data is usually detailed and continuous, while socioeconomic data is generally summarized and reported at fixed intervals, such as annually or by census tracts. For example, US residential migration data is collected and available at four-year intervals, and socioeconomic data is available at five-year intervals, making it directly match the two data sources' temporal resolution is difficult for a given task [5], [6].

### B. Modeling Complex Interactions and Cascading Effects

Environmental interactions are complex and involve cascading events (e.g., a flood causing landslides and leading to infrastructure damages and failures in essential services, leading to population displacement). Accurately modeling these multi-level interactions requires expert domain knowledge, and sophisticated data structures such as graphs and complex algorithms are needed to capture the non-linear relationships and dependencies of extreme weather events.

Recognizing the critical need for proper data organization and computation techniques during an extreme weather event is essential. This vision paper investigates the integration of advanced computational methods with diverse but interconnected environmental datasets to address critical climate challenges, focusing on a chosen extreme weather event (flood) as a case study. This paper proposes a novel concept that integrates Knowledge Graph methods with the potential of Quantum Computing applications, alongside classical computing techniques. The goal is to address urgent environmental issues by leveraging NASA's extensive EO and weather datasets. This refined approach provides a robust framework for tackling the challenges of climate change.

By leveraging a Knowledge Graph to organize and interpret vast amounts of environmental data, traditional (classical) machine learning methods, and novel quantum computing methods such as Quantum Machine Learning (QML) to help understand the relevance of those technologies towards obtaining quantum advantage. We propose an Environmental Interaction Knowledge Graph for Quantum Computing (EIKG4QC). KGs have revolutionized managing and interpreting large, interconnected datasets, which are crucial for dissecting complex issues like climate change. This work helps gain deeper insights into the complex interplay of climate factors that contribute to extreme weather events caused by climate change, such as flooding, wildfires, etc, and how to conduct both classical and quantum computing to obtain reliable climate action solutions. Additionally, Our framework contains two components. As shown in the figure 2, Part 1: conducting classical (traditional) computing components, including knowledge graph implementation (EIKG:Flood) and conducting traditional AIML methods for floods (flooding as a use case), and Part 2: conducting quantum computing (novel) components using the IBM Qiskit environment on a real quantum computer.

## III. Common-Denominator-Environmental Variables

In this positioning paper, inspired by the "common denominator term in mathematics, we introduce a term called "Common-Denominator-Environmental Variables" (CDEV) in our study. The environmental-related variables, such as Precipitation (rainfall and snowfall), Land Surface Temperature (LST), Land Use, and Land Coverage (LULC), can be considered as common denominator variables. Those variables are typically monitored using EO satellite data and have been used in weather and environmental models. These EO datasets can be used to implement event-specific use cases such as floods (use case: floods) and build KGs. Once those individual KGs for case-specific events are implemented, we envision that they can be carefully integrated using contextual information based on how they relate to one another. Using their interrelations as the foundation, we propose our EIKG framework shown in Figure 1. The EIKG uses the EO dataset and other related datasets, such as socio-economics datasets, to build graph databases and conduct graph-based analytics using state-of-the-art Machine Learning (ML) techniques to conduct predictive analytics. These state-of-the-art ML techniques could be based on classical or Quantum Computing (QC) methods.

## IV. Proposed Solution and Future Directions

### A. Proposed EIKG Framework and Graph Data Model

In order to conduct proper graph data modeling, it is important to start with a clear use case describing the purpose and domain knowledge for implementing the graph data model. This process requires clearly identifying stakeholders, intended users of the model application, and domain experts who would test the implemented model against the use case and its validity. The data model clearly expresses the names of labels of the nodes, types of relationships, and properties of those nodes and relationships for the graph [7]–[9].

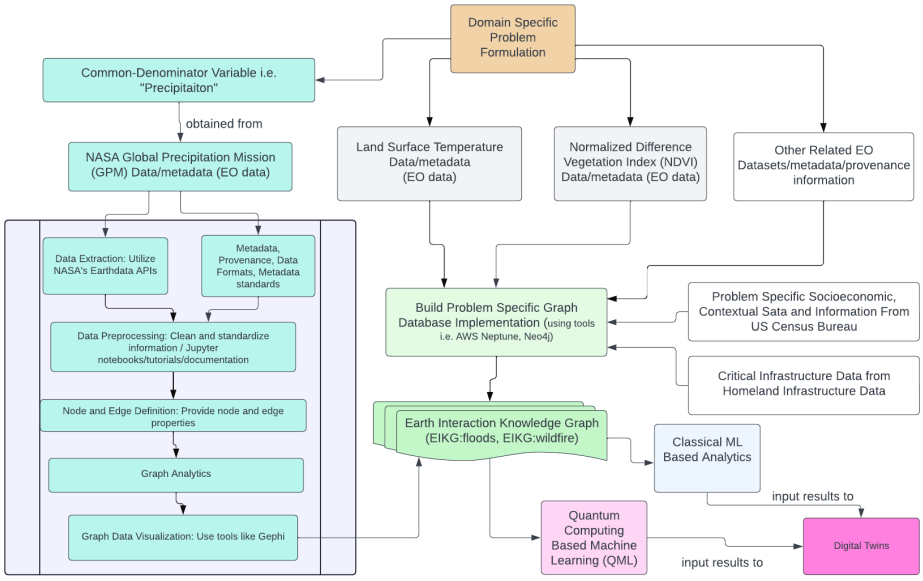


Fig. 1: Schematic View of the Environmental Interaction Knowledge Graph Workflow.

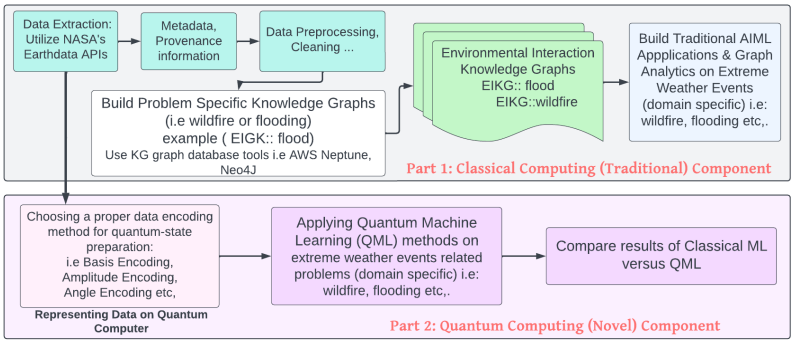


Fig. 2: Schematic View of the Classical Computing and Quantum Computing Workflow.

*1) Defining Node Labels and Properties:* In our use cases, we chose the dominant entities to be modeled as labels since the labels help us group the related nodes in our graph. Properties of the nodes will be used to identify the node and answer the relevant details in the application use cases by returning the property values from a query.

*2) Instance Model:* Using a setup sample data, we can evaluate if the use case is properly tested. For that, an instance model can be helpful in verifying that our data model meets the testing's intended requirements and satisfies the use case of the application. The following figure 3 shows an example of the instance model for the EIKG:floods, where we have created an instance of *Counties* and *FloodEvents* with their relevant relationships.

*3) Construction of the Knowledge Graph:* The EIKG will be the foundation for integrating heterogeneous datasets, including environmental and weather data, as well as socioeconomic and infrastructural datasets. The reason for using a knowledge graph is to capture the complex relationships between key entities, such as flood events, locations, infrastructure vulnerabilities, and human dynamics, such as population mobility.

We implement the EIKG in a manner to include the temporal aspect of the socioeconomic indicators and county-to-county residential mobility data in the EIKG, which represent the time-dependent changes of the indicators and their relationships over time [10]. For example, time-dependent variables such as socio-economic indicators (i.e., unemployment rate, which changes yearly for the same country), and it is important to capture those variations in our knowledge graph to show the level of change in the socioeconomic indicators before and after a major extreme weather event like flooding in a county. This can be done by introducing time-stamped entities and relationships with spatiotemporal attributes in the EIKG using the developed ontology as a guideline. Below is a high-level explanation of how to integrate the temporal aspects and suggestions for building the graph database.

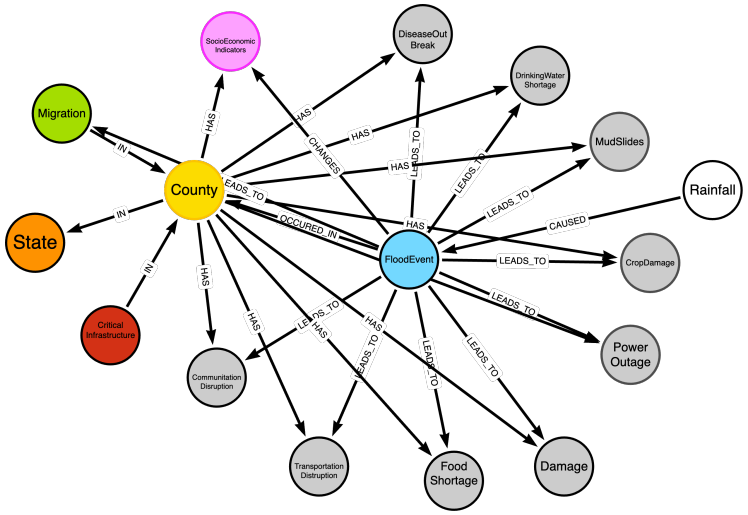


Fig. 3: High-level View of EIKG Graph Data Model Relating Floods, County and Relevant Flood Induced Events.

This development process will include the following:

1) **Data Integration**: Obtain relevant datasets from NASA, FEMA, U.S. Census Bureau, and other sources, generate necessary data frames, and save them as CSV files.
2) **Ontology Design**: Creating an ontology to define the classes and relationships between entities (e.g., `Flood`, `Infrastructure`, `Population`, `SocioeconomicIndicator`).
3) **Graph Database Implementation**: We plan to use the Neo4j academic research-licensed software to store, query, and visualize the EIKG knowledge graph.
4) **Use Case-Specific EIKG**: Specific EIKGs such as `EIKG::Flood` and `EIKG::Heatwave` will be implemented.

### B. Building the Ontology and the Knowledge Graph: EIKG

Prior to building the EIKG, our proposed research plan is to implement the ontology that helps create the KG and test its validity. We plan to conduct a thorough literature review to evaluate the existing ontologies published in papers.

*1) EIKG Ontology:* Before creating the ontology, it is important to define what it needs to achieve by defining the scope and purpose of the ontology for the EIKG. Identifying the key concepts and modeling them in classes and subclasses and defining their relationships between classes will help to model. This includes identifying the key questions the ontology will answer, the datasets used to integrate, and the domains it will cover (e.g., environmental, socioeconomic, mobility, infrastructure, etc.). Suggested questions include:

- What entities (e.g., floods, infrastructure, population) do we need to model?
- What relationships exist between these entities (e.g., floods damage critical infrastructure, flood affects human mobility)?
- What reasoning capabilities do we need (e.g., inferring population displacement from county-to-county residential mobility due to flood events)?

## V. Quantum Assisted Computing

### A. Representation of Data on a Quantum Computer

In order to use a quantum computer to conduct quantum computing-based machine learning tasks, the data in the Comma Separated Value (CSV) format needed to be encoded into quantum states where the Quantum Processing Unit (QPU) can use the encoded data. How classical data is encoded into a quantum state plays a crucial role. There are several techniques for encoding classical data, such as Basis Encoding, Amplitude Encoding, etc.; depending on the quantum encoding technique being used to encode the classical data, a suitable quantum algorithm will be used to conduct the quantum computing-based machine learning task [11]. We intend to follow proper data encoding techniques and useful quantum computing algorithms based on their usability as stated in the literature [11]–[14].

### B. Quantum Machine Learning Applicability

Over the recent years, QML methods have been developed and compared with their classical ML methods [15], [16]. Some of the prominent QML methods are Variational Quantum Classifier (VQC) and Variational Quantum Regressor (VQR), both derived from the framework of Variational Quantum Eigensolvers (VQE) [15], [17]. Practical applications of the VQC have been tested on well-known ML datasets from the University of Irvine ML dataset repository [18]–[20]. Among the QML methods, Variational Eigenvalue solver-based approaches have advanced over the past decade since Peruzzo et al. introduced them in 2014 [21]. Our proposed work focuses on using QML techniques such as Quantum Support Vector Classifier (QSVC) and Variational Quantum Algorithms (VQA) [16], [19]. Inspired by studies in the literature, we plan to explore how QSVC, VQC, and VQR could be implemented to predict the county-to-county migration impact based on floods by classifying the intensity level. We will also

TABLE I: Datasets used in the Analysis

| Dataset | Source | Description | Resolution | Temporal Coverage |
|---|---|---|---|---|
| **GPM IMERG Daily Precipitation** | NASA | Global precipitation data used to measure rainfall intensity and patterns. | 0.1° x 0.1° | Daily, ongoing |
| **SMAP L4 Global Daily Soil Moisture** | NASA | Soil moisture data used to assess flood potential and water retention capacity of the soil. | 9 km | Daily, ongoing |
| **MODIS Land Cover Type Yearly L3 Global** | NASA | Land cover data to evaluate land use and its impact on flood susceptibility. | 500 m | Yearly, ongoing |
| **American Community Survey (ACS) 5-Year Estimates** | U.S. Census Bureau | Socioeconomic data including income levels, unemployment rates, and educational attainment. | Census Block Group (CBG) and county level | 5-year estimates, ongoing |
| **Residential Mobility Data 4-Year Estimates** | U.S. Census Bureau | County-to-County residential mobility. | County level | 4-year estimates, ongoing |
| **Hospitals and Power Plants Critical Infrastructure Location Information** | Homeland Infrastructure Foundation-Level Data (HIFLD) | Locations and attributes of critical infrastructure to assess flood vulnerability. | Point of Interest (PoI) data with Lat/Lon | Updated periodically |
| **National Hydrography Dataset (NHD)** | U.S. Environmental Protection Agency (EPA) | Detailed hydrography data to map water bodies and assess flood risks in relation to infrastructure. | Various, high resolution | Updated periodically |
| **FEMA Flood Hazard Maps** | Federal Emergency Management Agency (FEMA) | Flood hazard data used to assess flood risk areas, including flood maps and flood zones. | Varies by region | Updated periodically |
| **NOAA Sea Level Rise Data** | National Oceanic and Atmospheric Administration (NOAA) | Data on sea level rise and predictions to evaluate the impact of rising sea levels on flood risks. | Point data with Lat/Lon | Ongoing, updated periodically |
| **FEMA National Risk Index (NRI)** | Federal Emergency Management Agency (FEMA) | Comprehensive data on risk, resilience, and vulnerability to natural hazards, including flooding. | County level | Updated periodically |
| **Social Vulnerability Index (SVI)** | Center for Disease Control (CDC) | Index measuring the social vulnerability of communities, based on factors like socioeconomic status, household composition, and access to transportation. | Census tract level | Updated periodically |

conduct other relevant prediction tasks and compare the results with those of classical ML models found in the literature.

## VI. Conclusion

The proposed EIKG is an innovative framework for integrating heterogeneous datasets to model complex environmental interactions and support classical and QML analytics. This vision paper helps address challenges such as integrating diverse datasets and modeling cascading environmental events. The proposed framework demonstrates the potential to enhance understanding of extreme weather and its impact on human dynamics, such as human mobility. The introduced common-denominator-environmental variables and spatiotemporal considerations in data modeling emphasize the framework's adaptability and relevance to various use cases by identifying the shared environmental factors that lead to floods, wildfires, and droughts. Through the implementation of case-specific EIKG and its attempt to integrate with advanced quantum computing techniques, this work pushes the boundaries of current machine learning methods. The insights gained from implementing an EIKG will help to better understand extreme events and find data-driven solutions to mitigate the adverse effects of climate change.

## Acknowledgment

We thank the University at Albany and its College of Emergency Preparedness, Homeland Security, and Cybersecurity, as well as the Rensselaer Polytechnic Institute, for their support.

## References

[1] J. Anuradha *et al.*, "A brief introduction on big data 5vs characteristics and hadoop technology," *Procedia computer science*, vol. 48, pp. 319–324, 2015.

[2] F. Hadzic, H. Tan, and T. S. Dillon, *Mining of Data with Complex Structures*. Springer, 2011, vol. 333.

[3] A. Zafeiropoulos, E. Fotopoulou, and S. Papavassiliou, "Participatory socio-environmental systems modeling over knowledge graphs," in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.

[4] M. S. Islam, A. Proma, Y. Zhou, S. N. Akter, C. Wohn, and E. Hoque, "KnowUREnvironment: An Automated Knowledge Graph for Climate Change and Environmental Issues," in *AAAI 2022 Fall Symposium: the Role of AI in Responding to Climate Challenges*, 2022.

[5] B. Bhaduri, E. Bright, P. Coleman, and M. L. Urban, "Landscan usa: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics," *GeoJournal*, vol. 69, pp. 103–117, 2007.

[6] J. Mennis and J. W. Liu, "Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change," *Transactions in GIS*, vol. 9, no. 1, pp. 5–17, 2005.

[7] J. Barrasa and J. Webber, *Building Knowledge Graphs.* " O'Reilly Media, Inc.", 2023.

[8] P. Liu, Y. Huang, P. Wang, Q. Zhao, J. Nie, Y. Tang, L. Sun, H. Wang, X. Wu, and W. Li, "Construction of typhoon disaster knowledge graph based on graph database neo4j," in *2020 Chinese Control And Decision Conference (CCDC)*. IEEE, 2020, pp. 3612–3616.

[9] T. Bratanic, *Graph Algorithms for Data Science: With Examples in Neo4j*. Simon and Schuster, 2024.

[10] X. Li and C. Zhang, "Spatiotemporal trends of poverty in the united states, 2006–2021," *Applied Spatial Analysis and Policy*, vol. 18, no. 1, p. 3, 2025.

[11] M. Schuld and F. Petruccione, *Machine learning with quantum computers*. Springer, 2021, vol. 676.

[12] S. Ganguly, *Quantum Machine Learning: An Applied Approach*. Springer, 2021.

[13] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill *et al.*, "Quantum advantage in learning from experiments," *Science*, vol. 376, no. 6598, pp. 1182–1186, 2022.

[14] M. Schuld, *Supervised learning with quantum computers*. Springer, 2018.

[15] J. Tilly, H. Chen, S. Cao, D. Picozzi, K. Setia, Y. Li, E. Grant, L. Wossnig, I. Rungger, G. H. Booth *et al.*, "The variational quantum eigensolver: a review of methods and best practices," *Physics Reports*, vol. 986, pp. 1–128, 2022.

[16] M. Schuld and N. Killoran, "Quantum machine learning in feature hilbert spaces," *Physical review letters*, vol. 122, no. 4, p. 040504, 2019.

[17] J.-G. Liu, Y.-H. Zhang, Y. Wan, and L. Wang, "Variational quantum eigensolver with fewer qubits," *Physical Review Research*, vol. 1, no. 2, p. 023025, 2019.

[18] A. Asuncion, D. Newman *et al.*, "Uci machine learning repository," 2007.

[19] P. Sen, A. S. Bhatia, K. S. Bhangu, and A. Elbeltagi, "Variational quantum classifiers through the lens of the hessian," *Plos one*, vol. 17, no. 1, p. e0262346, 2022.

[20] D. Maheshwari, D. Sierra-Sosa, and B. Garcia-Zapirain, "Variational quantum classifier for binary classification: Real vs synthetic dataset," *IEEE access*, vol. 10, pp. 3705–3715, 2021.

[21] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'brien, "A variational eigenvalue solver on a photonic quantum processor," *Nature communications*, vol. 5, no. 1, p. 4213, 2014.