Anomalies Detection on GRACE and GRACE-FO residuals Data Using AI/ML

Alexandre Belli, Bryant D Loomis, Eric Leon and Kenny E Rachlin

Science Systems and Applications, Inc. (SSAI), NASA GSFC, University of Michigan and KBRwyle

PRESENTED AT:



INTRODUCTION

The GRACE and GRACE Follow-On missions provide daily residual time series data from the Microwave Instruments (MWI), which are critical for gravity field modeling and other scientific studies. Detecting and removing anomalies in these time series, such as outliers, abrupt changes, or parasite signals, ensures that downstream analyses produce accurate results.

Traditional anomaly detection relies on human expertise or simple statistical thresholds, such as sigma cutoffs, which are often insufficient for complex or non-Gaussian data. In this study, we explore the use of Machine Learning (ML) to detect anomalies more efficiently and flexibly. By comparing unsupervised and supervised ML techniques, we aim to evaluate whether automated approaches can outperform human-driven methods while identifying potential limitations.

DATA AND OBJECTIVE

Dataset Details:

- GRACE and GRACE Follow-On MWI residuals (daily time series in mm/s).
- Data sampled nearly every second per day, yielding high-resolution, extensive datasets.
- Over thousands of days of data analyzed from both satellites, with variations in signal patterns due to instrument performance or external disturbances.

Research Objective:

Primary Goal:

Test the feasibility of ML algorithms for anomaly detection in GRACE/GRACE-FO residuals.

Secondary Goals

1. Assess whether automated anomaly detection performs better than manual editing or statistical thresholds.
2. Demonstrate the scalability of ML approaches to large, complex datasets.
3. Identify strengths and limitations of ML techniques and refine them for improved accuracy.

METHODS AND APPROACHES

Traditional Approach (Supervised Learning)

Process:

Manual pre-editing of data to remove known anomalies, labeled as "anomalies" versus "normal."

Use labeled datasets to train classification algorithms, such as Random Forests or Support Vector Machines.

Limitations:

Requires extensive manual effort to label data.

Computational cost of applying classification algorithms to years of daily data.

Prone to human errors and potential overfitting to pre-processed patterns.

Machine Learning Approaches

Isolation Forest

How It Works:

- Builds random decision trees to isolate data points based on their features.
- Anomalies are identified as data points that require fewer splits to isolate due to their rarity and dissimilarity from the majority.

Advantages:

- No assumptions about data distribution, making it robust to non-Gaussian or irregular patterns.
- Computationally lightweight and scalable to large datasets.
- Capable of identifying consecutive anomalies, useful for flagging extended problematic periods.

Drawbacks:

- Requires tuning of the "contamination factor," which assumes the proportion of anomalies in the dataset.
- May flag valid data as anomalies in datasets with minimal noise.

Autoencoder

How It Works:

- Neural network that compresses data into a low-dimensional representation (encoding) and reconstructs it (decoding).
- High reconstruction error indicates anomalies, as they deviate from the learned "normal" patterns.

Advantages:

- Effective for detecting non-linear and multi-dimensional anomalies.
- Learns complex data structures, adapting to evolving data patterns.

Drawbacks:

- Computationally heavier than Isolation Forest.
- Requires a large, well-labeled training dataset to prevent overfitting or underfitting.
- Reconstruction errors may be difficult to interpret.

Supervised Learning with Labeled Data

Approach:

- Use pre-edited datasets with anomalies flagged and removed to train classifiers.
- Compare predictions with human-edited results to validate model performance.

Challenges:

- Relies heavily on the accuracy of pre-edited data.
- Computationally expensive for large-scale datasets.
- Struggles to generalize to new types of anomalies not present in the training set.

RESULTS

Example Scenarios

Figure 1: A day with normal data distribution, showing potential outliers that are easily detected using simple statistical methods (e.g., sigma cutoff).

Observation: Traditional methods suffice in such cases but fail to detect anomalies with more complex structures.

Figure 2: A day with unexpected signal behavior.

Observation: The signal exhibits a pattern that is not entirely random, making it difficult to identify using simple methods. ML approaches like Isolation Forest and Autoencoder are more adept at detecting such anomalies without prior assumptions about data distribution.

Algorithm Performance

Isolation Forest:

Detected M anomalous days, significantly reducing the need for manual inspection.

Easily flagged days with consecutive anomalies (N anomalies in a row), enabling focused expert review.

Autoencoder:

Provided detailed insights into anomaly characteristics through reconstruction errors.

Detected subtle non-linear patterns missed by other methods but required more training and computational resources.

DISCUSSION AMD CONCLUSION

Discusion

Key Takeaways:

- Isolation Forest is well-suited for rapid anomaly detection in large datasets. Its simplicity and speed make it ideal for preprocessing.
- Autoencoders excel at identifying complex, high-dimensional anomalies but require significant computational resources.
- Combining these methods provides complementary strengths, enabling robust detection across diverse anomaly types.

Challenges and Future Work:

Fine-tuning contamination factors in Isolation Forest to minimize false positives and negatives.

Reducing the computational cost of Autoencoder training while maintaining accuracy.

Exploring hybrid approaches that combine unsupervised and supervised methods for greater precision.

Conclusion

ML algorithms like Isolation Forest and Autoencoders offer powerful tools for anomaly detection in satellite data, significantly reducing manual effort and improving detection efficiency. While human supervision remains essential, these techniques pave the way for scalable, automated data quality control for GRACE/GRACE-FO and similar missions. Future work will focus on refining these methods and extending their applicability to other datasets.

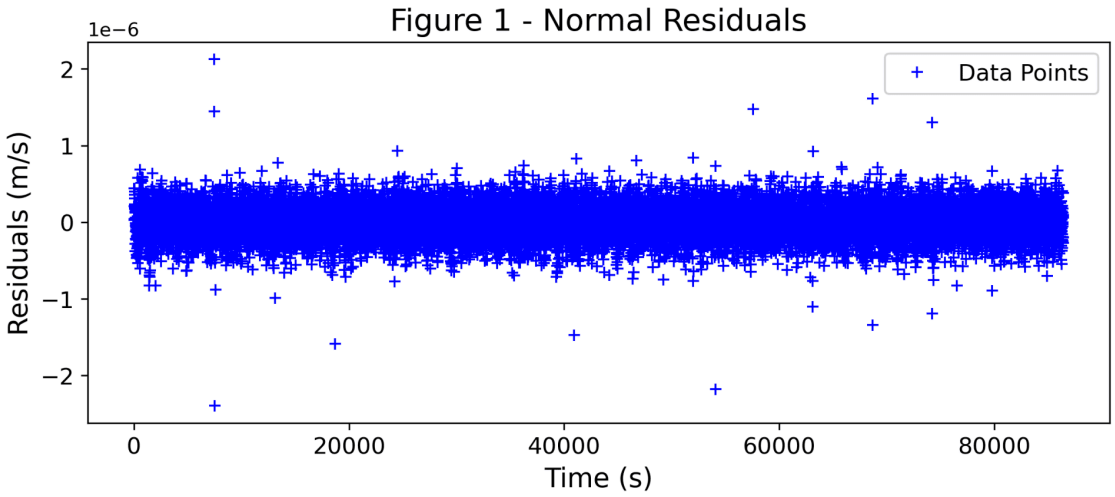


Figure 2 - Residuals with bad signal

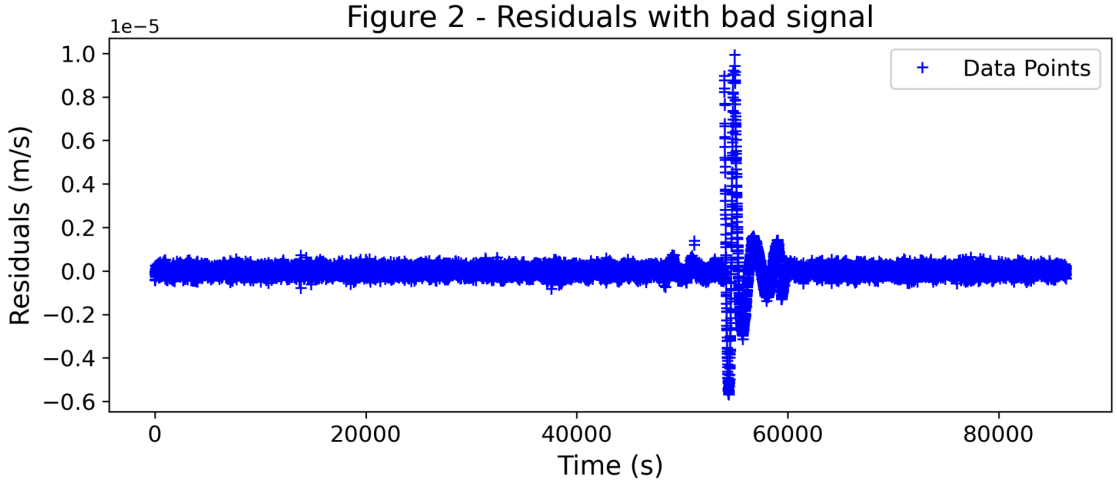


Figure 3 - Supervised Learning

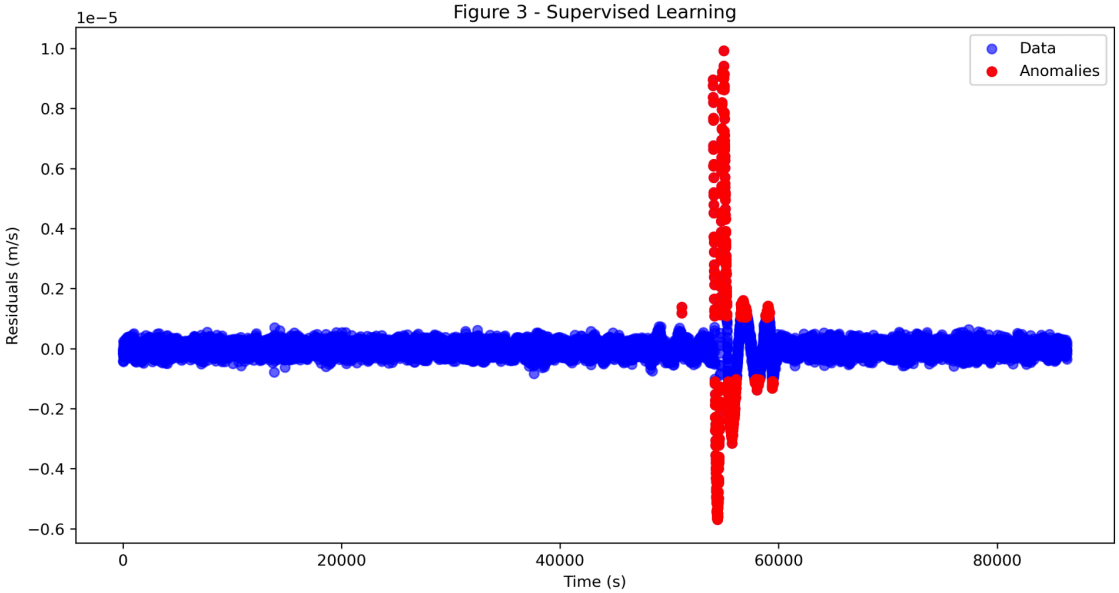


Figure 4 - Isolation Forest with contamination 0.0150

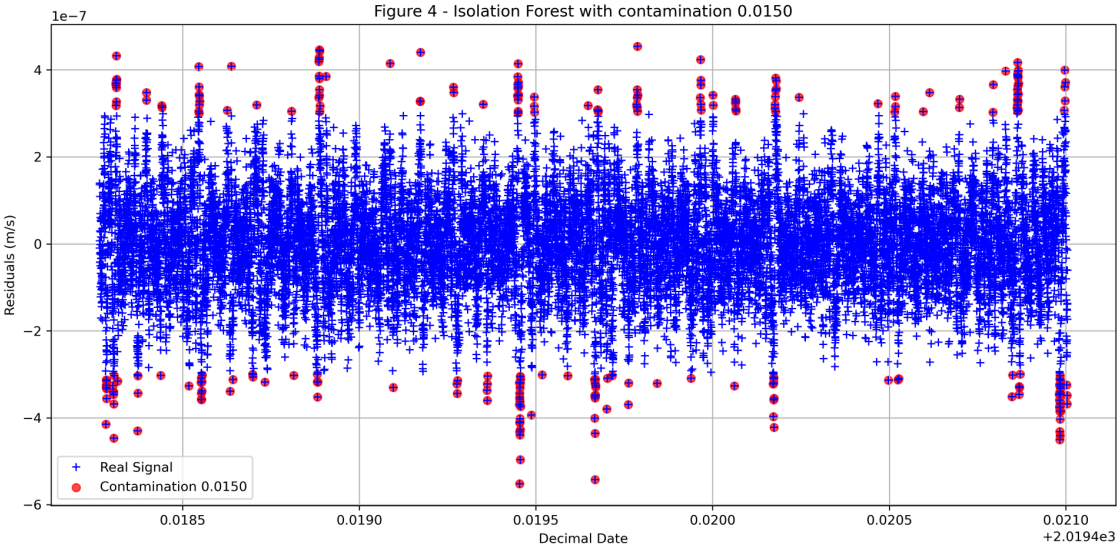


Figure 4 - Isolation Forest with contamination 0.0075

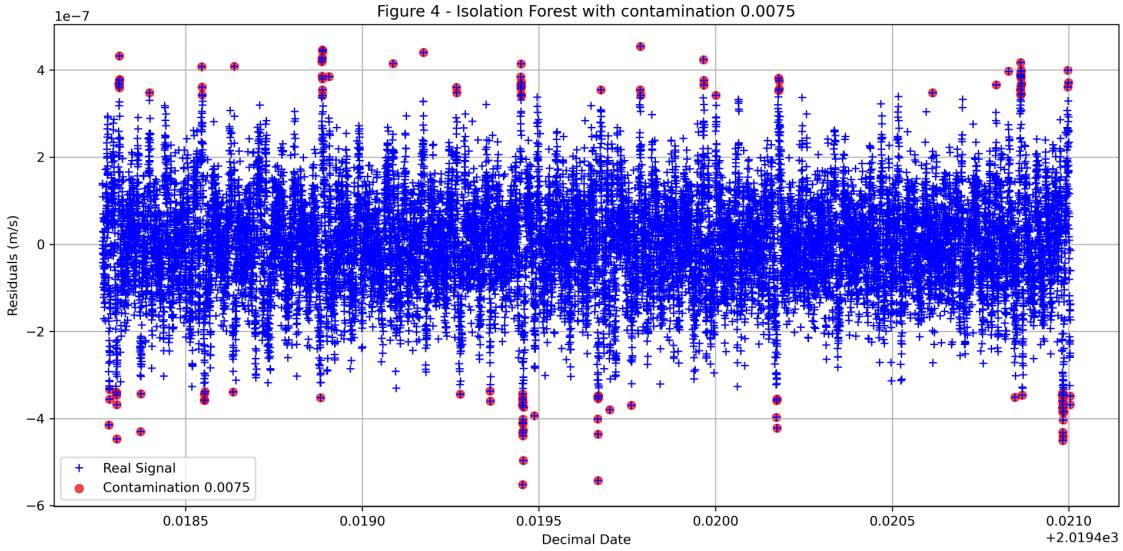


Figure 5 - mwi_resid_20220821.txt

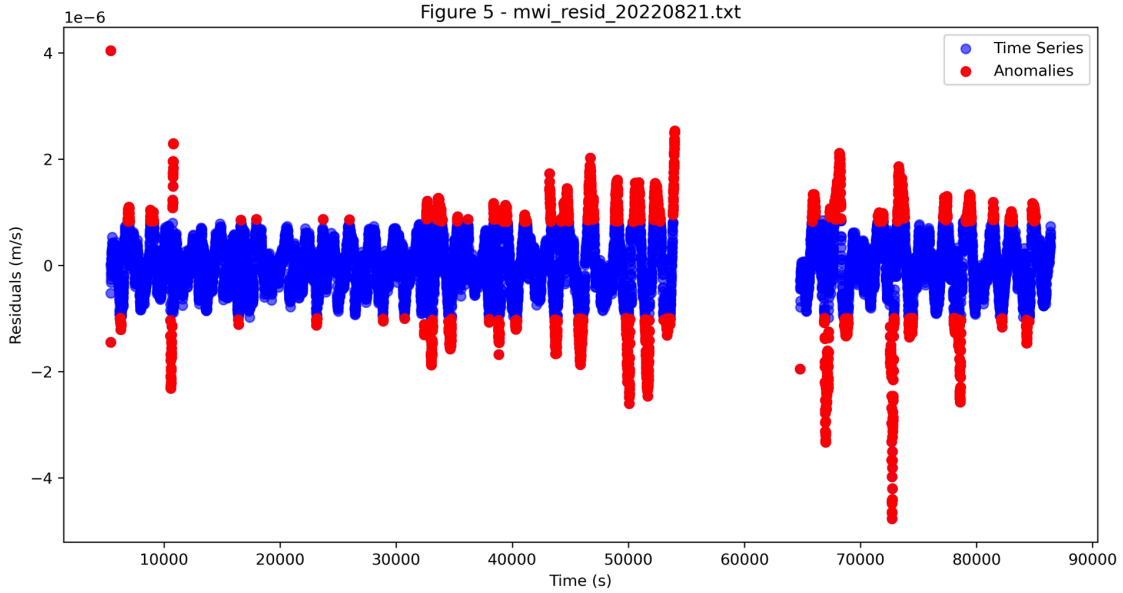


Figure 6 - Anomaly Detection with Autoencoder

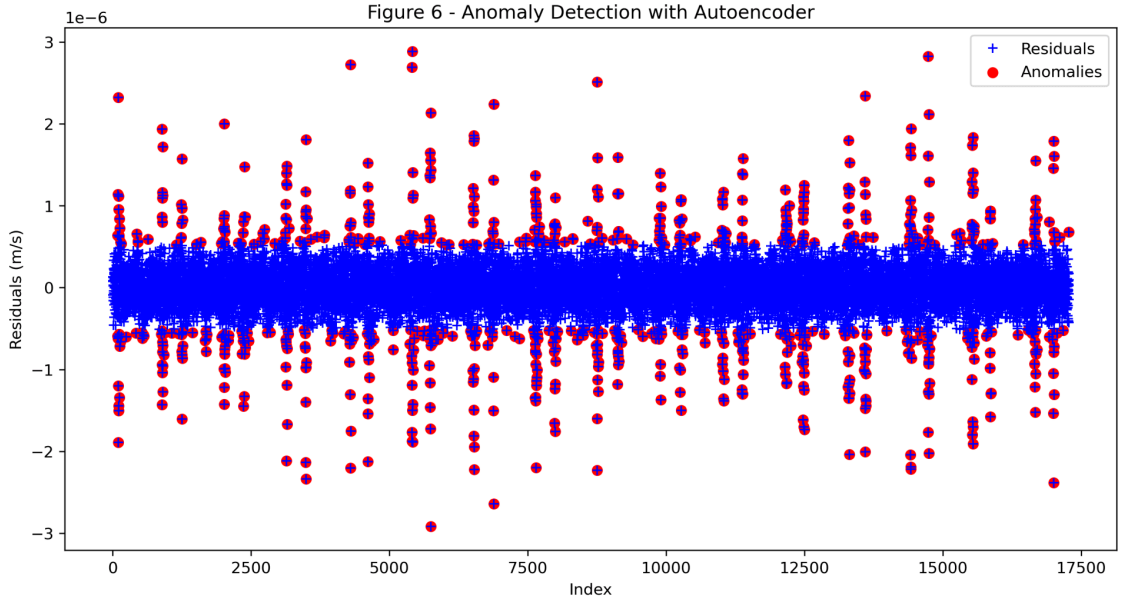


Figure 7 - Anomaly Detection with Autoencoder

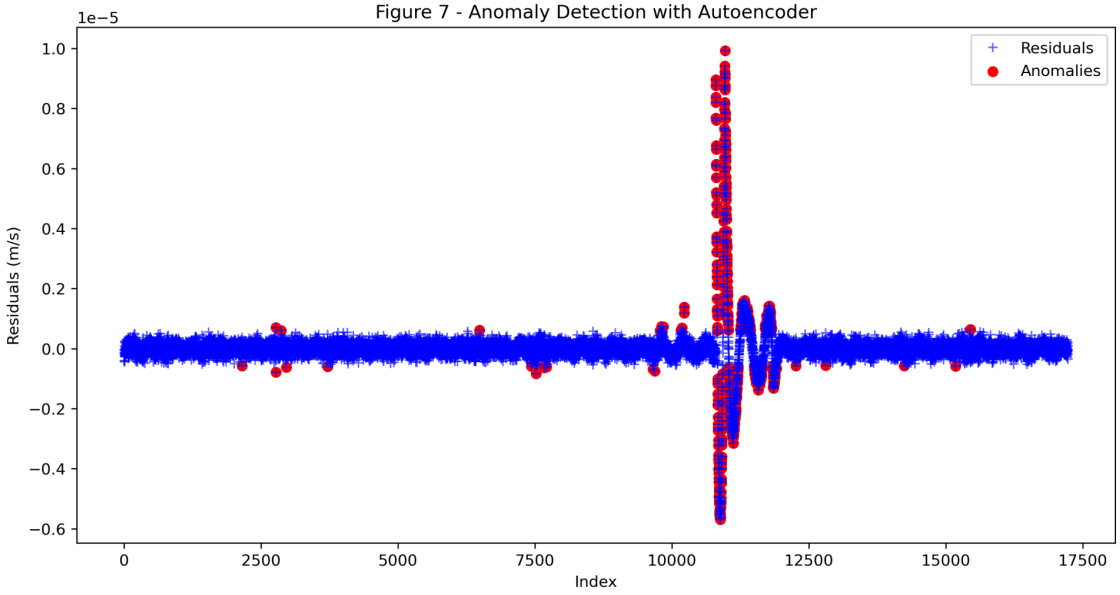
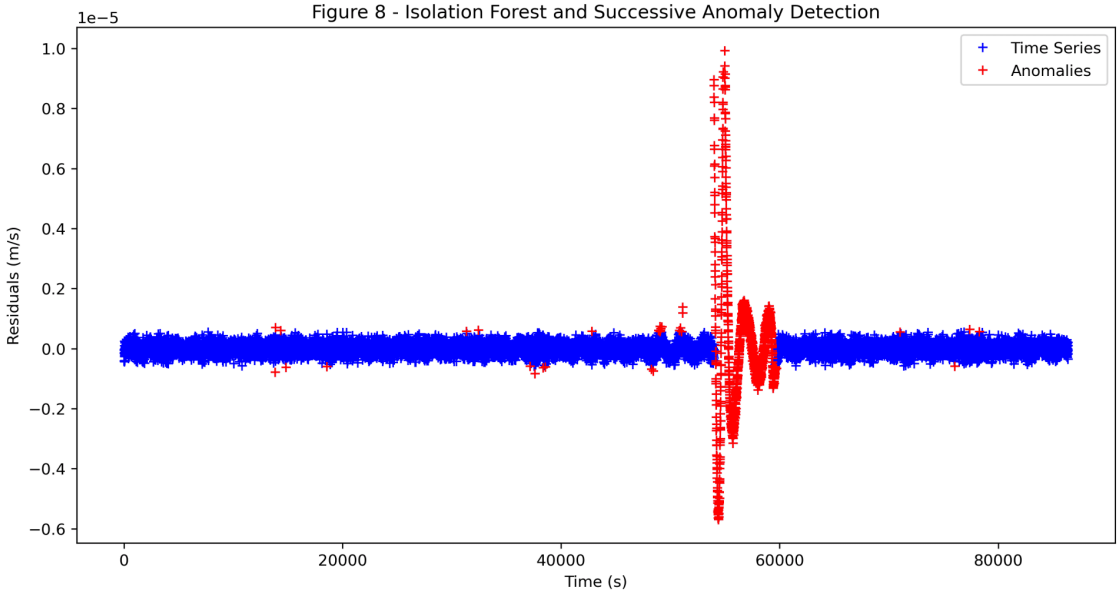


Figure 8 - Isolation Forest and Successive Anomaly Detection



AUTHOR INFO

Alexandre Belli - SSAI - alexandre.belli@ssaihq.com

Bryant D Loomis - NASA GSFC

Eric Leon - University of Michigan

Kenny E Rachlin - KBRwyle

TRANSCRIPT

ABSTRACT

Detecting anomalies is a crucial pre-processing step in data analysis to ensure the production of high-quality scientific outputs. Traditional statistical methods, such as standard deviation cutoffs, often fall short when dealing with real-world data characterized by high variability, discontinuities, and unexpected behaviors. These methods require frequent adjustments and human supervision to maintain performance, making them less reliable and efficient.

Machine Learning (ML) and Artificial Intelligence (AI) provide a suite of algorithms that address many of these limitations, enhancing the accuracy and efficiency of anomaly detection in satellite data. Although advanced algorithms like neural networks may increase computational demands and necessitate thorough dataset preparation, the outcomes are highly promising.

In this study, we applied ML/AI techniques to the GARCE/GRACE-FO intersatellite range-rate residuals to detect outliers and other anomalies, such as abrupt signal changes and unexpected behaviors. We quantified the impact of these anomalies on the gravity field. Our approach involved experimenting with various supervised and unsupervised algorithms, including autoencoders, isolation forests, and random forest classification. We will present an overview of the advantages and disadvantages of each technique and discuss which method proved most effective in qualifying and quantifying the anomalies' impacts on the gravity field. Our analysis leveraged the entire GRACE data span from 2003 to 2016 for training purposes, utilizing manually edited residuals for supervised learning.

Additionally, we will briefly outline our forward-looking strategy for integrating ML/AI techniques in the Geophysics and Geodesy Lab at NASA GSFC, emphasizing the importance of continued advancement in this field.

EVALUATIONS

#	Average Score
There are currently no completed evaluations for this presentation	

+