1. ML Quality Control System Overview:

Welcome to the MVRE ML Quality Control System! In any machine learning task, the quality and reliability of the input data play a crucial role in achieving accurate and meaningful results. Our system is designed to ensure the integrity of your data through a comprehensive quality control process.
By integrating traditional quality control methods with the power of machine learning, our system enhances the accuracy and reliability of identifying outliers and ensuring high-quality data.

2. How to prepare your Dataset:

- **Required Features:**

The dataset should include the following features (order is not important):
- Prof_no *: Profile number
- year: Year of the data
- month: Month of the data
- Longitude_[deg]: Longitude in degrees
- Latitude_[deg]: Latitude in degrees
- Depth_[m]: Depth in meters (or Pressure in decibars)
- Temp_[°C]: Temperature in Celsius

- **Choosing Depth or Pressure:**

The user has the choice to include either the Depth or Pressure feature in their dataset. Only one of them is required.

- **Data Organization:**

The data should be organized profile by profile, with each profile's information listed below one another.

- **Units:**

- Longitude and Latitude should be specified in degrees.
- Depth should be specified in meters (m), or Pressure in decibars (dbar).

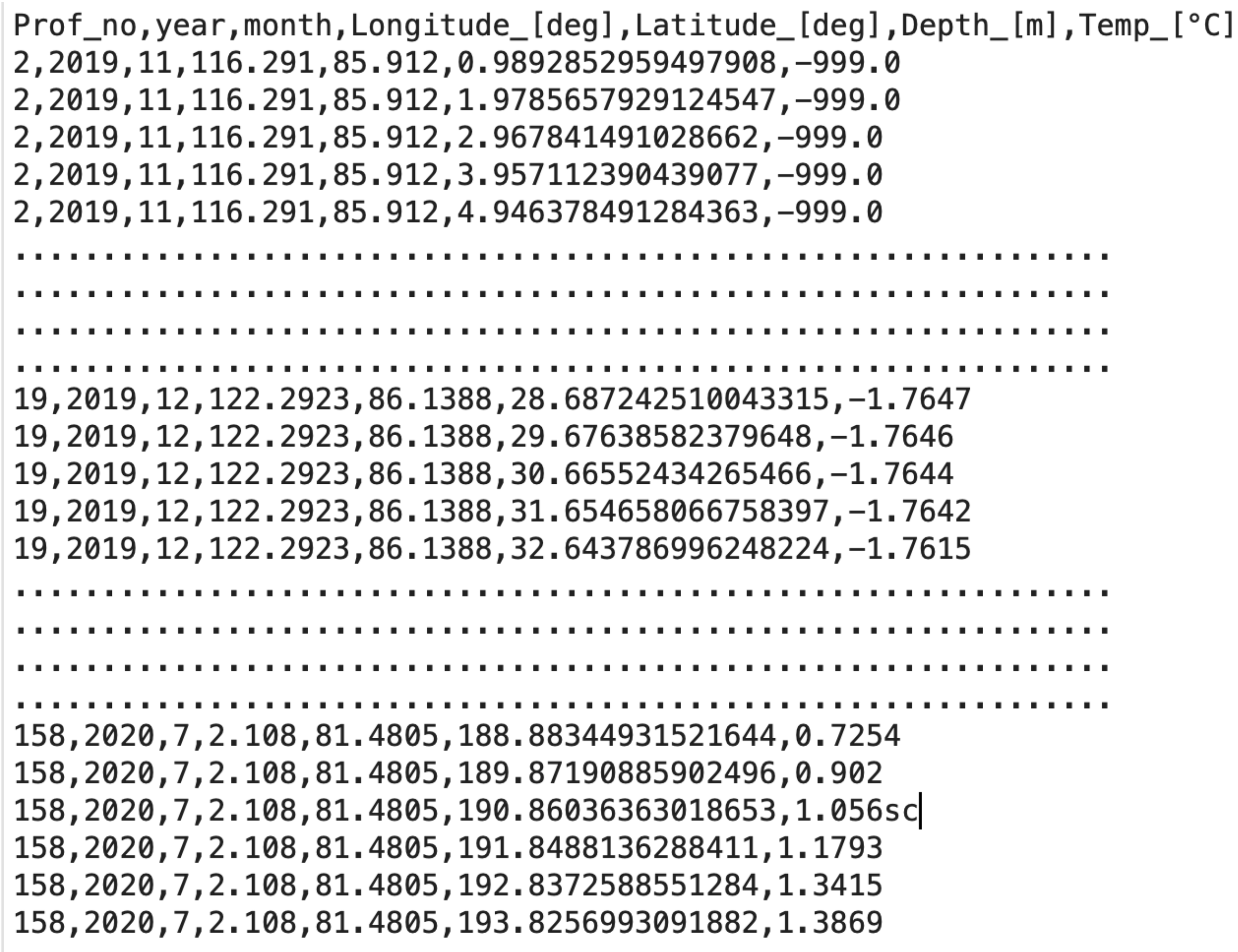
- **Handling Missing Values:**

Missing values should be filled with -999. In the provided example, missing values are denoted as -999 (which is very OK).

* In oceanographic data, a profile number refers to a unique identifier assigned to individual profiles or measurements taken during a specific oceanographic expedition or survey. A profile typically represents a vertical section of the ocean, providing measurements of various parameters such as temperature, salinity, depth, and other relevant variables at different depths or pressure levels. Each profile is assigned a profile number to distinguish it from other profiles collected during the same expedition or survey.

3. How your data should look like

Your dataset at the end should be looking as below:



```
Prof_no,year,month,Longitude_[deg],Latitude_[deg],Depth_[m],Temp_[°C]
2,2019,11,116.291,85.912,0.9892852959497908,-999.0
2,2019,11,116.291,85.912,1.9785657929124547,-999.0
2,2019,11,116.291,85.912,2.967841491028662,-999.0
2,2019,11,116.291,85.912,3.957112390439077,-999.0
2,2019,11,116.291,85.912,4.946378491284363,-999.0
...............................................................
...............................................................
...............................................................
...............................................................
19,2019,12,122.2923,86.1388,28.687242510043315,-1.7647
19,2019,12,122.2923,86.1388,29.67638582379648,-1.7646
19,2019,12,122.2923,86.1388,30.66552434265466,-1.7644
19,2019,12,122.2923,86.1388,31.654658066758397,-1.7642
19,2019,12,122.2923,86.1388,32.643786996248224,-1.7615
...............................................................
...............................................................
...............................................................
...............................................................
158,2020,7,2.108,81.4805,188.88344931521644,0.7254
158,2020,7,2.108,81.4805,189.87190885902496,0.902
158,2020,7,2.108,81.4805,190.86036363018653,1.056sc
158,2020,7,2.108,81.4805,191.8488136288411,1.1793
158,2020,7,2.108,81.4805,192.8372588551284,1.3415
158,2020,7,2.108,81.4805,193.8256993091882,1.3869
```

4. What will you download ?

After the user has downloaded the processed data, they will find three new columns added to their dataset:

- **' Trad_QF' (Quality Flag from Traditional Methods):**

This column contains flags generated by traditional quality control methods. The value '1' indicates that the corresponding feature includes a spike or suspect gradient, suggesting that it may be a questionable or erroneous sample. On the other hand, a value of '0' indicates that the feature is considered clean and does not exhibit any significant issues.

- **'ML_QF' (Merged Quality Flag)**

The 'merge_QF' column combines the Machine Learning Predictions and the Quality Flag from Traditional Methods. This column provides a comprehensive assessment by indicating the presence of a bad sample (flag 1) when both the traditional methods and the machine learning model confirm their anomalous nature. Conversely, if the sample is not flagged as bad by either method, it is indicated as zero, denoting a clean data point.

5. Recommendation:

We highly recommend giving significant consideration to the quality flags presented in the 'merge_QF' column. This column combines the insights derived from traditional quality control methods with the robust capabilities of machine learning. By leveraging this integrated approach, you can obtain a comprehensive and reliable assessment of the data quality, empowering you to make informed decisions based on the combined expertise of both methodologies.

6. Limitations of the System:

It is important to acknowledge that, despite its effectiveness, the ML Quality Control System has certain limitations that should be taken into consideration:

- **Potential for Mistakes:**

Like any machine learning system, the ML Quality Control System is not infallible. While it combines traditional methods with machine learning techniques to enhance accuracy, there is still a possibility of errors or misinterpretations. It is crucial to exercise caution and perform manual checks when interpreting the results generated by the system.

- **Performance with Large Datasets:**

When dealing with exceptionally large datasets, this Quality Control System may encounter performance issues or operational bugs. The system's efficiency can be affected by resource limitations, such as computational power or memory requirements. If you have a large dataset, we recommend reaching out to us so that we can work together to devise strategies to process and analyze your data smoothly.

- **Continuous Improvement:**

Our team is dedicated to continuously improving the ML Quality Control System. We are actively researching and developing new techniques to enhance the system's capabilities, overcome limitations, and address emerging challenges in data quality control. Your feedback and collaboration are essential in helping us refine and optimize the system further.

- **Data Usage Limitation:**

Please be aware that the ML Quality Control System processes data exclusively from regions located above the 60-degree north latitude line. Data situated below this latitude cannot be utilized within the system.

7. Concluding statements

Once the data is prepared following these guidelines, it can be used as input to your ML quality control system for further processing and analysis.

Now that you understand the key components of our ML Quality Control System and how to prepare your data, you are ready to take advantage of this robust framework to ensure the integrity of your datasets. Feel free to reach out if you have any questions or require further assistance.

We encourage users to exercise vigilance and maintain open communication with us, especially when encountering large datasets or facing challenges during the utilization of the ML Quality Control System. By working together, we can explore potential solutions and ensure a seamless and effective data processing experience.