



## Digital Twin for Innovation in Oceanic Visualization and Exploration (DIVE)

### BENCHMARK PROCEDURES AND TOOLS

Prepared By

Nicolae Catalin Ristea (UNSTPB)

Date: 20.11.2025

Reviewed By

Cosmin Dănișor (UNSTPB)

Date: 30.11.2025

Approved By

Mihai Datcu (UNSTPB)

Date: 30.11.2025

## AUTHORS

| No. | Name                   | Role in the project                    |
|-----|------------------------|----------------------------------------|
| 1.  | Nicolae Catalin Ristea | CS III, Machine learning researcher    |
| 2.  | Cosmin Dănișor         | CS III, EO Image processing researcher |
| 3.  | Corin Văduva           | CS II, EO Data analytics               |
| 4.  | Andrei Anghel          | CS I, EO Signal processing researcher  |
| 5.  | Răzvan Mateescu        | CS I, Coastal engineering              |
| 6.  | Mihai Datcu            | CS I, Principal Investigator           |

## DOCUMENT CHANGE RECORD

| Issue | Date       | Chapter | Change              |
|-------|------------|---------|---------------------|
| 1.1   | 2025-10-30 | All     | Document created    |
| 1.2   | 2025-12-05 | All     | Version 1 completed |

## TABLE OF CONTENTS

|     |                            |   |
|-----|----------------------------|---|
| 1.  | Validation measures .....  | 4 |
| 1.1 | General description .....  | 4 |
| 1.2 | Confusion matrix .....     | 4 |
| 1.3 | Accuracy.....              | 5 |
| 1.4 | Precision and Recall ..... | 5 |
| 1.5 | F-score .....              | 5 |
| 1.6 | ROC curve.....             | 6 |

## 1. Validation measures

### 1.1 General description

Validation and verification of measurement methods are essential procedures designed to ensure that analytical techniques are suitable for their intended purposes and that the information derived from EO data accurately reflects real Earth processes and spatial coverages. These activities build both realistic expectations for analysts and confidence for end-users. Figure 1 provides an intuitive overview of the key concepts involved in method validation and verification:

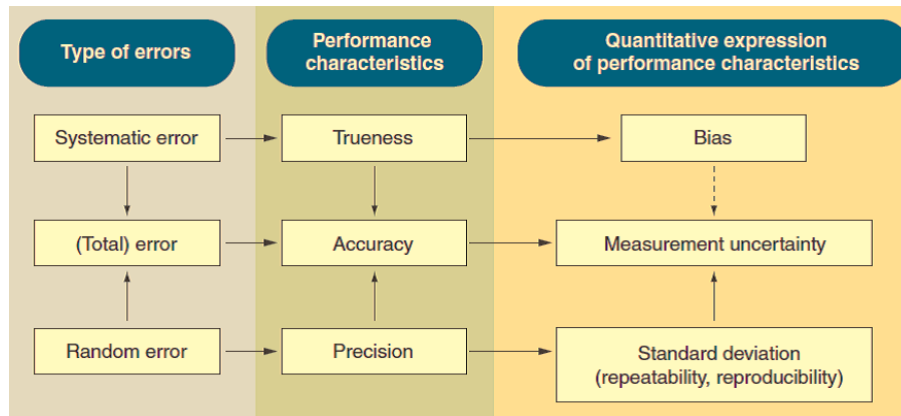


Figure 1. Qualitative and quantitative expressions of the type of error and their combination.

Errors in data processing may arise either randomly or systematically. These two categories are typically characterized qualitatively as precision (random error) and accuracy (overall closeness to the true value), and quantitatively as standard deviation and bias. Trueness refers to the qualitative degree to which the average of an infinite number of repeat measurements would align with a reference value; it is therefore inversely related to measurement bias and reflects the level of systematic error. Measurement accuracy describes how closely a processed result represents the actual state or value of the Earth's surface or processes, incorporating both random and systematic error components. A measurement is considered more accurate when total error is small, more true when bias is minimal, and more precise when random variability is low. Although precision is a qualitative notion, it can only be quantified indirectly through its opposite—imprecision.

### 1.2 Confusion matrix

A confusion matrix is a valuable tool for evaluating the performance of classification models. In Earth observation, these models are often used for tasks like land cover classification, object detection, and change detection. A confusion matrix breaks down the model's predictions and actual class labels into four categories:

- True Positives (TP)
- True Negatives (TN)
- False Positives (FP)
- False Negatives (FN)

Assuming a task focused on land cover classification using EO Data as an example, the four categories can be explained as follows:

- TP: These are the instances where the model correctly predicted a certain land cover class, and the ground truth also identifies it as that class.
- TN: These are the instances where the model correctly predicts a class other than the one in question, and the ground truth also identifies it as a different class. For example, the model predicts "Water," and the area is actually "Urban."
- FP: These are instances where the model incorrectly predicts a class, but the ground truth classifies it differently. For example, the model predicts "Water," but the area is "Urban." This represents a classification error where the model falsely identifies a certain class.

- **FN:** These are instances where the model incorrectly predicts a different class, but the ground truth identifies it as the class in question. For instance, the model predicts "Forest," but it is actually "Water." This represents a missed classification of the desired class.

The confusion matrix enables the calculation of various performance metrics such as precision, recall, F1 score, accuracy, and more (see next sections).

### 1.3 Accuracy

Accuracy is a simple and widely used metric that measures the overall correctness of your classification model. It calculates the ratio of correctly classified instances to the total number of instances in the dataset. It can be computed using the formula below:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Accuracy might not be the best metric for imbalanced datasets.

### 1.4 Precision and Recall

Precision and recall are validation measures commonly used in machine learning, especially in binary classification tasks, to evaluate the performance of a model. They provide insights into how well a model is identifying and classifying positive instances (True Positives) and avoiding false classifications (False Positives and False Negatives). These metrics are particularly useful when dealing with imbalanced datasets, where one class is much more prevalent than the other.

- **Precision** is a measure of how many of the predicted positive instances are true positives. It answers the question, "Of all the instances the model classified as positive, how many were correctly classified?"

$$\text{Precision} = TP / (TP + FP)$$

A high precision value indicates that the model makes accurate positive predictions and has a low rate of false positives.

- **Recall**, also known as sensitivity or true positive rate, measures how many of the actual positive instances were correctly classified by the model. It answers the question, "Of all the actual positive instances, how many did the model correctly identify?"

$$\text{Recall} = TP / (TP + FN)$$

A high recall value indicates that the model is effective at identifying most of the positive instances in the dataset, and it has a low rate of false negatives.

In classification tasks, precision and recall are often in tension, and improving one may reduce the other. The choice of which metric to prioritize depends on the specific objectives and associated costs of the task. Prioritizing high precision—minimizing false positives—typically requires a higher decision threshold, at the risk of lowering recall. Conversely, if maximizing recall is more important—capturing as many true positives as possible—then a lower threshold is appropriate, even though this may decrease precision. Evaluating precision and recall across multiple thresholds allows the construction of a precision–recall curve, which visualizes this trade-off and supports selecting the threshold that best aligns with the application's requirements.

### 1.5 F-score

The F1 score is a single metric that combines precision and recall into one value. It can be useful when a balance between precision and recall is required. It is calculated as:

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

A high F1 score indicates a good balance between precision and recall.

## **1.6 ROC curve**

A useful graphical representation to assess the performance of the model in binary classifications is a Receiver Operating Characteristic (ROC) . It illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (specificity) as the discrimination threshold of the model is varied.

The ROC curve considers:

- True Positive Rate (TPR) or Sensitivity: TPR is the proportion of actual positive cases (true positives) that the model correctly identifies as positive. It is calculated as follows:  
$$\text{TPR} = \text{TP} / (\text{TP} + \text{FN})$$
- False Positive Rate (FPR) or Specificity: FPR is the proportion of actual negative cases (true negatives) that the model incorrectly identifies as positive. It is calculated as follows:  
$$\text{FPR} = \text{FP} / (\text{FP} + \text{TN})$$

Where TP, TN and FN are defined in Section 5.2. A good classifier would have a curve that leans toward the top-left corner, indicating high TPR and low FPR across various threshold values. We may quantify the overall performance by calculating the Area Under the ROC Curve (AUC-ROC), where a higher AUC-ROC indicates a better classifier.