CHAPTER 8

DATA QUALITY TIERS OF SOIL MOISTURE DATA

Mike Cosh, Nandita Gaur

Learning Outcomes

The data quality tiering system described below can be used to categorize soil moisture data being produced by different networks, at a network or individual station level.

This tiering system provides an aspirational framework for network operators looking to improve data quality and a short-hand approach for data users to quickly assess whether a dataset's quality is likely to be appropriate for their intended use.

Networks benefit from setting goals and criteria for the products they produce. While networks may have individual goals and criteria outlined for themselves, standardized goals across networks can help coalesce soil moisture data from different networks in a more efficient way for stakeholders. This chapter outlines a tiering system for data quality self-assessment.

The approach described in this chapter for categorizing soil moisture networks into three tiers parallels a proposed tiering method for meteorological networks more broadly described by the World Meteorological Organization's Global Climate Observing System (WMO GCOS)[12]. Similar to the classification system described in this document, the WMO GCOS concept of a three-tiered structure is intended to support user guidance when selecting a dataset and is centered around data quality, data assurance, and documentation[13]. The network tiering approach described in this document differs in that it provides a set of guidelines or goals specific to soil moisture networks.

This chapter describes a method for soil moisture data quality tier assignment, which can be self-assessed for (1) a time series of soil moisture data from a certain sensor that is continually collecting long-term data, (2) a long-term station, and (3) a long-term network.

Broadly, the quality of soil moisture data and its utility to stakeholders can depend upon:

1. the numerical accuracy of the soil moisture dataset,
2. its spatial representativeness,
3. data latency,
4. ancillary information about the site, and
5. depth of soils that are represented.

There are a large variety of applications and purposes for network deployment, each with specific criteria and features. Based on available resources and purpose of the network, data quality objectives can vary greatly based on the five factors mentioned above. Any tiering system must

[12] Proposal for formalization and standardization of tiered network approach across domains and observing system programs. 2022. https://gcos.wmo.int/sites/default/files/2.3_c_concept_note_tiered_networks_v5_0.pdf?48eYWrX00RFgPm7j87Cle.PdX8grWXLo

[13] The tiers for the WMO GCOS networks are "Reference, Baseline, or Additional" networks. The GCOS proposed tiering approach was endorsed by the WMO in 2022, and specific criteria associated with the tiers are still in development as of 2024.

thus be comprehensive in defining the critical and common characteristics for networks, while also being flexible and applicable to the variety of conditions found among the many networks deployed in the past, present, and future. A system is therefore proposed with three criteria for determining the tier of a dataset: Error Analysis, Data Stream Density, and Metadata. Error analysis incorporates both errors arising due to choice of sensor and calibration (pre-data collection) and errors due to QA and QC issues (post data collection). Data stream density broadly refers to the spatial (depth-based) and temporal frequency of data collection and reporting. Metadata refers to the amount of standardized information per the Metadata Guidance document that the network provides. These criteria have been selected after discussions with network operators and data users that identified factors in selecting data and products for use. Broadly, each of these criteria will be evaluated for having 'None', 'Some', or 'All of the Ideal Criteria'.

THREE-TIER SYSTEM FOR DATA QUALITY

A three-tier system to categorize the quality of soil moisture data is provided in Table 8. The tiering system can be applied to a network, a station, or an individual time series of soil moisture data produced by a sensor within a station. It can vary over specific time spans for a specific station as well, because it is possible to have the quality of a station improve or degrade over time. For instance, stations in mountainous regions may have high latency in winter months because of access and data transmission logistics. Such a station could be classified as Tier I during the growing season or summer months, and Tier III during the winter. This will help users of the data understand the limitations of the network and data streams.

Table 8. Tiers of data quality

Element	Tier I	Tier II	Tier III	Uncategorized
1: Error analysis				
Sensor calibration	Soil-specific calibration with at least one post-deployment calibration activity.	Point scale and soil-specific calibration (Laboratory based)	Factory calibrated	Not defined
Quality assurance & quality control	Wide range of tests and data quality flags for Type I and Type II data errors[14]	Tests for Type I errors*	none	none
2: Data Stream Density				
Measurement frequency	Hourly	Hourly	> Hourly	>daily
Depths	3 depths or more	2 depths	1 depth	-
Temporal resolution	Near-real time	Daily	> Daily	Uncertain
Available data per quarter-year[15]	Reports 90% data/quarter	75%	50%	< 50%
3: Metadata	**Tier I**	**≥Tier II**	**≥ Tier III**	**No metadata**
Site characterization	Expert soil characterization	Map based estimates	Lat/Long	
Maintenance	Multiple times per year	Annual	Less than annual	

TIERS OF DATA QUALITY

A full description of metrics for tier classification can be found in the Appendices to this document. Summary descriptions of each category are provided below and in Table 8.

UNCATEGORIZED

The network or program collects data inconsistently or is lacking many parameters of quality assurance and control. An examples of soil moisture data that might be classified as "uncategorized" could be citizen science data that are collected on an irregular basis.

- Error analysis:
  - Sensor calibration is not defined by the network or is non-existent.
  - Data quality assurance and control protocols do not exist. Data are not flagged or quality checked following collection.
- Data stream density:
  - Data are collected on a less frequent basis than daily. Data collection may be sporadic.
  - Soil moisture is collected only at one depth or at different depths during different data collection events.

[14] *See Chapter 7 of this document for further information of Type I and Type II Data errors
[15] Under current operative conditions this may not possible; this element is currently only a recommendation.

  - Temporal resolution is not defined or irregular.
  - No more than 50% of the data collected within a 3-month time period are valid data (Chapter 7). Note: there may be some exceptions to this rule, for example, when frozen soils reduce sensor performance for a known, seasonal period.
- Metadata
  - No metadata are available.

TIER III DATA (BASIC/LOW QUALITY)

- Sensor calibration:
  - Only factory calibration has taken place. No soil-specific calibrations or in-lab tests have been conducted by the network operators.
  - No QA or QC is applied to data post-collection. Data are not flagged or checked.
- Data stream density:
  - Data are collected less frequently than on an hourly basis. Data may be collected only once daily.
  - Soil moisture and relevant parameters are measured at one or more depth per site.
  - Data are made available on a daily or less frequent basis.
  - At least 50% of the data collected within a 3-month time period are valid data (Chapter 7). Note: there may be some exceptions to this rule, for example, when frozen soils reduce sensor performance for a known, seasonal period.
- Metadata: (See NSCMMN Metadata Recommendations Guide for Tier Selection)
  - Latitude and longitude are provided (See NSCMMN Guidelines).
  - Soil and landscape characterization are not present or are incomplete (See NSCMMN Guidelines).
  - Maintenance does not occur on an annual basis or more frequently. Maintenance is sporadic.

TIER II DATA (MODERATE QUALITY)

- Error analysis:
  - Soil specific calibration in laboratory is complete for all installation locations for all deployed sensor makes and models (Chapter 2, 4).
  - Point scale calibration has taken place
  - Data-processing includes testing for and flagging Type I (visually observable) data errors (Chapter 7).
- Data Stream Density:
  - Data are collected at least hourly or more frequently.
  - Soil moisture and companion parameters are measured at two or more depths within the same soil column.
  - Temporal resolution is at least daily.
  - At least 75% of the data collected within a 3-month time period are valid data (Chapter 7). Note: there may be some exceptions to this rule, for example, when frozen soils reduce sensor performance for a known, seasonal period.

- Metadata: (See Metadata Guidance document for metadata criteria)
    - Site characterization (landscape cover, soil type, etc.) is conducted using estimates based on maps or is only partially available (Metadata Guidance document).
    - Latitude, longitude, and elevation are provided to a high degree of accuracy (Metadata Guidance document).
    - Site maintenance occurs annually (Chapter 5).

TIER I DATA (HIGH QUALITY)

- Error analysis:
    - Soil specific calibration in laboratory must be complete for all installation locations for all deployed sensor makes and models (Chapter 2, 4).
    - At least one post-deployment field calibration or validation activity must have been completed for each sensor deployment location (Chapter 6).
    - Data post-processing includes a wide range of tests and associated flags for both Type I (visually observable) and Type II (complex) data errors (Chapter 7). A key is provided for all error flags.
- Data stream density:
    - Measurements are taken least hourly, if not more frequently.
    - Soil moisture and accompanying parameters are measured at three or more depths within the same soil pit/trench.
    - Temporal resolution is near real time. (Data are transmitted multiple times daily.)
    - At least 90% of the data collected within a 3-month time period are valid data (Chapter 7). Note: there may be some exceptions to this rule, for example, when frozen soils reduce sensor performance for a known, seasonal period.
- Metadata: (See Metadata Guidance document for metadata classifications)
    - Soil characteristics, including soil texture, salinity, pore size, etc., have been characterized by a soils expert.
    - Latitude, longitude, and elevation are provided to a high degree of accuracy
    - Site description (landscape type, slope, etc.) are collected in-person.
    - Station maintenance is conducted multiple times per year (Chapter 5).

OTHER TIERING CONSIDERATIONS

For representing soil moisture in certain units other than volumetric soil moisture, Tier I metadata is a pre-requisite. These include fraction available water and % field capacity. Hence, to support some stakeholder uses it may be beneficial to maintain Tier I *metadata* (per the Metadata Guidance document) even if the other parameters do not conform to that tiering.

To address seasonal impacts to data collection and data quality, data availability should be measured per 3-month period. It is possible that a station or network meet different tier criteria during different seasons of the year: therefore, data users might choose to utilize the network for only the growing season, or some other time frame.

A network, station, or time series under consideration will be classified based on the lowest tier it conforms to, based on all three elements of data tiering (error analysis, data stream density, and metadata). A few examples are provided below.

1. For determining the tier for a network of 10 stations, if eight stations are Tier I while two stations are Tier II, the classification of the network would be Tier II. However, this network can advertise that 80% of their stations conform to Tier I while 20% correspond to Tier II.
2. For determining the tier for a station with five soil moisture sensors, if two sensors are Tier II but the remaining are Tier II, the station tier would be classified as a Tier III.
3. For a time-series of soil moisture data from a sensor, if the measurements correspond to Tier I for two elements per Table 8 but Tier III for the remaining element, the time series will be classified as a Tier III.

WHO WILL DETERMINE THE TIERING LEVEL OF A STATION OR NETWORK?

These elements are intended for network self-evaluation but may also be subject to peer review, as usually occurs in scientific reviews and publications. Generally, a station would be classified only as high as their lowest tier class among metrics for evaluation. However, there may be some situations, such as performance during periods of frozen soils, where temporal caveats are reasonable.

WHY SHOULD YOU PARTICIPATE IN THE TIERING EXERCISE?

There are various reasons for using the tiering system.

1. The system is designed with both network operators and stakeholders in mind.
2. It enables better integration of soil moisture data from different sources and networks, which can increase the large-scale usability of soil moisture data as is often required by stakeholders.
3. At a network level, it provides a standardized baseline for different networks to compare themselves with other networks. This information can be used to identify areas of improvement for the network and identify areas that require investment of additional resources.
4. Finally, a network may use the tiering to support its intra-network management decision-making. For example, if a network characterizes themselves as having 80% Tier I stations and 20% Tier II stations, that information might be used to create aspirations for station improvements and selective investment of resources.
5. The tiering system is transparent and allows stakeholders to identify the tier of data they require. As such, stakeholders can quickly identify networks that provide data that is of interest to them, while network operators can identify additional stakeholders for their dataset creating the potential for additional sources of funding for maintenance.