P3(10) Paper presented as poster during TECO-2016, published in: WMO, 2016, WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation (CIMO TECO 2016), IOM Report No. 125. Data Levels also published in: WMO, 2018, Solid Precipitation Intercomparison Experiment (SPICE) 2012-2015, Final Report, IOM Report No. 131

Description of the Quality Control and Event Selection Procedures used within the WMO SPICE project

Authors : Audrey Reverdin[(1)], Michael Earle[(2)], Andrew Gaydos[(3)], Mareile A. Wolff[(4)]

(1) Federal Office of Meteorology and Climatology (MeteoSwiss), Payerne, Switzerland, audrey.reverdin@meteoswiss.ch

(2) Environment and Climate Change Canada (EC), Halifax, Canada

(3) National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA

(4) Norwegian Meteorological Institute, Oslo, Norway

1. Introduction

The Solid Precipitation Intercomparison Experiment (SPICE) is a WMO/CIMO project aiming at assessing automatic instruments performance in the context of snow measurements. The measurement campaign took place during two winter seasons (2013-2015). Around 20 measuring sites got involved within SPICE, giving a unique opportunity to test different instruments on different climatic regimes over the two hemispheres. All the data collected from the different site reference systems, from the instruments under test and from instruments providing ancillary measurements have been archived, quality-controlled and led to a high computed level following the procedures described in this paper.

2. Data levels in SPICE

A system of data levels has been established to distinguish among datasets with different stages of processing and quality control. This system is built upon the existing framework used for satellite observations by the WMO and other organizations (WMO, 2014; NASA, 2010).

Level 0: The rawest output from an instrument, or instrument transducer, in native units (e.g. voltage).

Level 1: The time-stamped output from each individual instrument, or instrument and data logger, which has been converted into geophysical measurements (e.g. weight, mass, intensity). These data are generally recorded at the highest temporal resolution that is feasible for a particular instrument configuration at a particular site, and before any significant data quality control has been applied. These data are recorded and stored at each measurement site and transferred to the SPICE data archive at the National Centre for Atmospheric Research (NCAR), Boulder CO, USA.

Level 2: Quality-controlled datasets for one instrument, on one site.

> **Level 2a:** Level 1 data that have undergone both formatting and integrity checks to ensure the correct number of records per day (e.g. 1440 records/day for data with 1 minute sampling intervals) and the validity of field formats within a given record (e.g. number, text

string). These checks are performed automatically when data are ingested into the SPICE archive. Level 2a data are available for download from the SPICE archive.

Level 2b: Level 2a data after a basic (automatic and manual) data quality control procedure has been applied. The details of the procedure vary by gauge, and in some cases, by site, and have been developed through consultation with site managers. Basic data quality flags are added. For data with sampling intervals less than 1 minute, the output data and flags are aggregated to produce 1 minute values. Level 2b data are generated and made available for download at the SPICE archive. Details of the quality control procedures and flag criteria are provided in Chapter 3.

Level 3: 'Meteorologically-relevant' datasets derived from Level 2 data for single sensors and parameters. Processing is application-dependent, and may include aggregation to different temporal scales, or external calibration/adjustment to compensate for measurement artefacts or environmental factors. For example, weighing gauge data that have been aggregated to 'precipitation event' timescales (e.g. 30 minutes, 1 hour, or longer) and adjusted for wind-induced under-catch are Level 3 products.

Level 4: Integrated datasets derived using lower level datasets for multiple sensors and parameters. For example, the use of weighing gauge data (Level 2) in concert with data from a precipitation detector or disdrometer (Level 2) to identify and characterize precipitating periods would generate a Level 4 product. The event file described in Chapter 4 and comprising 30 min precipitation events dataset from all instruments on one site is a good example of level 3 product.

This paper focusses on the specific procedures used to generate data products from levels 1 to 4.

3. Quality Control procedure

The SPICE quality control (QC) consists of an automatic procedure applied to all data at the SPICE archive, followed by a manual procedure ensuring the best quality of the data for further analyses.

The automatic QC procedure developed for SPICE data includes the following steps: (1) a file formatting and integrity check, to ensure the uniformity of file formats to be used in subsequent analysis; (2) a plausibility check, to remove apparent outliers and identify potential baseline shifts; and (3) a noise filtering step by the mean of a Gaussian filter, to mitigate the contribution of high-frequency components to the measurement signal. The data are then aggregated in 1 min datasets and additionally manually quality-controlled to access to the level 2, as defined in previous section. Details for each step are outlined in subsequent sections.

3.1. File formatting and integrity check

Level 1 data files are checked to ensure that the correct number of records (time stamps) are present within a given time period, and that each record contains the correct number of fields.

Daily files for data with a 6 second sampling interval should have 14,400 records (10 records/minute x 60 minutes/hour x 24 hours), while those for data with a 1 minute sampling interval should have

1,440 records (1 record/minute x 60 minutes/hour x 24 hours). Any missing records are identified, the appropriate time stamps are inserted, and the corresponding fields are filled with null values (e.g. -999, NULL, or NaN). The missing records are tracked with the data presence flag (number 5), discussed in more detail in Section 3.6. Any duplicate records (repeated time stamps with the same data) are removed. In the event that a given record contains more or fewer fields/parameters than expected, the entire line is replaced with null values and flagged with the same presence flag. If a given field in a record does not match its expected format (e.g. text when a number is expected), all fields in the record are replaced with null values and again flagged. The resulting Level 2a data sets are used as inputs for the remaining steps of the QC procedure.

3.2. Plausibility check

Level 2a data undergo a series of checks to ensure that observations are within the range of expected/physically reasonable values, and to identify features which may confound subsequent data analysis.

3.2.1. Max/Min filter or Range Check

For each instrument parameter of interest for subsequent data analysis, a minimum and maximum expected value were defined according to mechanical constraints from the sensor or physical plausible values from that sensor (an accumulation from a weighing gauge, for instance, would be evaluated between 0 mm and the maximum capacity of the bucket, a temperature sensor will be evaluated between a range of -50 to +50 °C, a wind direction between 0 and 360°, etc.). If a given value lies above the maximum expected value or below the minimum expected value, it is replaced with a null value and flagged as 'erroneous' (number 4, see Table 1).

3.2.2. Jump and Baseline shift filter

Within a given dataset, 'jumps' may be observed. Jumps are intermittent deviations from the main data trend, or baseline, that fall within the maximum and minimum expected values, and hence, are not filtered out by the range check. A 'jump filter' is employed to identify points that differ from the preceding baseline values by more than a set threshold (the 'suspect' or 'erroneous' jump thresholds) and to flag them accordingly : suspect jumps are kept in the dataset and flagged as 'suspect' (number 3, see Table 1), while erroneous jumps are flagged as 'erroneous' (number 4), and replaced with null values. The parameter-specific value selected for the suspect and erroneous jump thresholds is meant to exceed the maximum expected increase of this parameter per 6 second or 1 minute (as defined by instrument operational limits and/or site climatology), in order to avoid filtering out values which may correspond to real measurements.

The jump filter tracks the number of points exceeding either jump thresholds relative to a baseline value. In certain instances, however, jumps are not intermittent, but correspond to an increase or decrease in the baseline. For weighing gauges, for instance, increases in the baseline may be associated with 'dumps' in which solid precipitation accumulated on the rim (a phenomenon referred to as 'capping') and falls into the bucket, resulting in an abrupt and sometimes significant increase in accumulation. Decreases in the baseline may be associated with emptying of buckets as part of regular gauge maintenance. Baseline shifts are identified by the jump filter when the number of consecutive jumps exceeds a 'plateau threshold,' set to correspond to a specified time period (e.g. 1

hour). When a new plateau has been identified, the associated data are not replaced with null values; rather, the first new plateau data is flagged (number 7, see Table 1) to indicate that manual assessment is required before the period can be considered for subsequent analysis. Gauge capping and related baseline shifts may impact data before or after the shift is observed. For example, a gauge may have been capped for an extended period before observing a dump, or a gauge may remain partially capped following a dump. Accordingly, the time periods (e.g. 1 hour) preceding and following baseline shifts are also flagged (number 8, see Table 1) for manual assessment, and possibly intervention.

3.3. Noise filtering

Precipitation accumulation measurements from weighing gauges are subject to noise, the magnitude of which increases with increasing wind speed and increasing accumulation in the bucket (Duchon, 2008). To mitigate the influence of noise, various types of filters can be employed. Several filter methods were tested within SPICE (see the SPICE Final Report for more details, soon available) and the Gaussian filter appeared to be the most effective one to remove noise. The Gaussian filter method applied a filter of a specified width (specified number of data points filtered in each step) to a moving window along the time series.
Consequently, over 1 min data, a Gaussian filter of a 4 minutes width and a 2 minutes standard deviation has been applied to the SPICE accumulation data coming from all kind of weighing gauges, but not on the other kind of parameters (ancillary measurements, 1min accumulations or intensities).

3.4. Aggregating data

3.4.1. Aggregating data from gauges with multiple transducers

The SPICE IOC has decided that the Geonor gauges, used as part of the Field Working Reference System (FWRS) and tested also on different SPICE sites, shall have three active transducers, working independently (Geonor configuration). The load of the bucket, including any accumulated precipitation, is shared among the three wires. While the precipitation amount could be determined from any of the individual transducers, an aggregate of the transducer outputs is typically used for two reasons: (1) averaging transducer outputs reduces the magnitude of noise due to wind effects relative to individual transducers (Duchon, 2008); and (2) there are often differences among the transducer outputs resulting from differences in orientation (with respect to the sun, with respect to vertical) and temperature, or aspects of the configuration (unbalanced load, vibration). Accordingly, the precipitation datasets for Geonor gauges with three transducers in SPICE were obtained by computing the arithmetic average of the outputs of the three transducers.

3.4.2. Aggregating to 1 min data

Quality-controlled datasets with sampling intervals of less than 1 minute are aggregated to generate 1 minute datasets, baseline frequency chosen for any SPICE analysis. For precipitation accumulation datasets, the aggregation step involves the selection of the last filtered data point from each minute.
For intensity and ancillary measurements (temperature, wind speed, humidity, etc.), the aggregation generally entails a simple block average, with some exceptions for specific parameters (e.g. vector average for wind direction). For precipitation type data measured by present weather sensors or

disdrometers, the highest numerical SYNOP code (Tab. 4680) of each minute represents the minutely value.

3.5. Manual Quality Control procedure

The automatic QC procedure drastically improved the raw SPICE dataset using a uniform and standardized approach, accounting for obvious failures of the sensor or unrealistic features in the data. However, some specific data issues originating from site maintenance, surrounding environment (e.g. birds, spiders, site visits, etc.), wrong installation or calibration, couldn't be tracked by the automatic process and needed a further step with a manual intervention. Moreover, the data flagged in previous automatic QC steps as 'suspect' (number 3, see Table 1), or as 'Baseline Shift' and 'Potential Capping' (numbers 7 and 9, respectively) were expected to be manually checked to confirm (or infirm) the poor quality of the data. Thus, in close collaboration with the Site Managers, the site logs of every site were shared and helped the analysts in assessing the pertinence of all suspect data features. The data identified as being wrong were manually removed from the dataset (replaced by null values) and flagged (number 6, see Table 1). Additionally, the baseline shifts in weighing gauge data confirmed as being due to maintenance (emptying, calibration check), were removed (and flagged with the same flag number) and baselines were leveled to provide continuous accumulated values for a better comparison with the reference (time series check). All the manual changes applied on the data were tracked in electronical form for further investigations, if needed.

The FWRS dataset have been particularly thoroughly manually QCed to ensure the best level of quality for the reference dataset used to derive all the subsequent SPICE results. The instruments under test, on the other hand, were manually checked by removing all the suspect data whose poor quality was identified to not be due to the sensor itself. Indeed, the intrinsic sensor failures were intended to be part of the evaluation of the sensor performance, and thus data were not removed.

3.6. Quality Control flags

The quality control and aggregation of datasets from reference gauges, systems under test, and ancillary gauges were accompanied by a flagging procedure to provide additional insight into gauge performance and data integrity, and to support the data analysis. A system of flags has been created that follows closely the approach implemented in the WMO Field Intercomparison of Rainfall Intensity (World Meteorological Organization, 2009). This approach is outlined in Table 1.

Flags are generated for each parameter of interest, for each gauge. Flags identified in datasets with sampling intervals of less than 1 minute will be carried forward when aggregating to 1 minute datasets. A threshold of 66% for flag carryover was set for the missing and erroneous flags (i.e. if 66% of points within a given minute are flagged, that flag is carried over to the 1 minute aggregate value). For the remaining flags, any instances of the flag being called within a given 1 minute period will result in that flag being carried forward to the 1 minute aggregate value.

A given data point can have only a single flag value within the defined system. For instances in which multiple flags are observed for a given data point, the following order of priority is applied:

Good (1) < Suspect (3) < Erroneous (4) < Manual intervention (8) < Baseline shift (7) < Site (6) < Missing (5)

As stated in Section 3.2, erroneous values are replaced with null values by the jump filter, while suspect values are only flagged. For datasets with sampling intervals less than 1 minute, the remaining suspect values may potentially impact the aggregate 1 minute values in cases where the 66% criterion is not met. To address this concern, if fewer than 66% of points within a given minute are flagged as suspect (3), and there are no higher priority flags (flags 4 or 5 meeting the 66% criterion, or any instances of flag 7 or 8 within that minute), the resulting aggregate value is flagged as suspect (3).

For instruments with multiple transducers, flags are generated separately for the aggregate 1 minute data from each transducer following the above criteria. The flags for the composite one minute instrument data (aggregating the contributions from each transducer) are aggregated such that the highest priority flag from a constituent transducer in a given minute is taken as the composite value.

Additional criteria have been proposed for the carryover of flags to identified precipitation events. This is discussed within the context of the event selection algorithm in Chapter 4.

Table 1: SPICE QC data flagging system.

Flag value	Data Classification	Data Characterization
1	‘’Good’	No issues detected
3	‘Suspect ’	Gauge diagnostic parameters indicate potential data issue
4	‘Erroneous’	Gauge diagnostic parameters indicate gauge or data error
5	‘Missing’	Missing data point (for datasets with sampling intervals of 1 min or longer) Insufficient number of samples used to compute minutely value (for datasets with sampling intervals less than 1 minute)
6	‘Site’	Adapted from site logs; data points manually flagged to reflect maintenance, malfunction, power outage, etc.
7	‘Baseline shift’	Baseline shift present; data should be checked manually
8	‘Potential capping’	Data within specified proximity of baseline shift, which may be impacted by gauge capping; data should be checked manually

4. Event Selection procedure

The identification of precipitation events is a key component in the analysis of performance of the instruments tested in SPICE, as well as for the derivation of transfer functions for each gauge type relative to gauges in reference configurations.

Precipitation events can vary extensively over the course of a season at any one site. The number of SPICE sites with different climatological conditions increases further the diversity of precipitation events that need to be taken into account for the analysis. In order to achieve comparable site data sets, a uniform method was required for defining and quantifying precipitation events, which could be applied to all SPICE data.

In the context of SPICE, a precipitation event is defined over a period of time when the accumulated precipitation detected by the reference system is positive, and meets certain criteria as outlined in the following sections. Selected events should be long enough to be reliably representative of snowfall events in a variety of climate and environmental conditions. For this reason, the baseline duration of a precipitation event is set at 30 minutes.

For any SPICE site, a Site Event DataSet (SEDS) is created. The SEDS contains data from all precipitation instruments operating on the site as well as from selected ancillary instruments for all 30 min intervals over which the FWRS reported a precipitation event.

The SEDS is derived from the 1 min quality-controlled datasets (level 2 data) and constitutes a level 4 data product, as defined in Chapter 2. The consistency of the approach allowed to have comparable SEDS among all sites.

4.1. Description of Event Selection Algorithm

The event selection algorithm identifies precipitation events based on the quality-controlled data from two instruments that are part of the FWRS: the automatic weighing gauge, measuring accumulation, and the precipitation detector or disdrometer, reporting on the presence or absence of precipitation. The one minute datasets of these two reference instruments over one season are segregated in consecutive blocks of 30 minute intervals (i.e. 00h00, 00h30, 01h00, 01h30, etc.), over which the selection criteria are applied.

The flowchart in Figure 1 illustrates the two steps of the event selection algorithm, i.e. the event identification (step 1) and the event parameters processing (step 2). Over the first step, two algorithm options are considered: (1) when the precipitation detector is available on site and reports a valid output (column 1); and (2) when the precipitation detector is missing or outputs an invalid report (column 2). The third column in Figure 1 indicates when to proceed to the next step in the algorithm.

4.1.1. First step – Event Identification

In the first step, the data from the weighing gauge of the FWRS and, if available, the precipitation detector, are examined over the 30 minutes interval. To be selected as an event, the 30 minute window has to fulfill the following two conditions :

1) Net precipitation duration sufficiently long

The number of minutes during which precipitation is detected has to be more than 60% of the window time, i.e. more than 18 minutes. The precipitation duration is calculated based on precipitation detector data (first column in Figure 1) by looking at the number of "YES" cases that occurred during these 30 minutes. If the precipitation detector information is missing or unreliable (second column of Figure 1), the algorithm examines the data from the weighing gauge in the FWRS and identifies the number of minutes during which there is increasing accumulation. If this number exceeds 60% of the event duration, the net precipitation duration condition is considered to be met.

2) Accumulation of reference gauge sufficient

The total accumulation in the reference gauge during the 30 minutes has to be greater than a defined threshold. Based on previous experience, this threshold rate has been set to 0.25 mm over 30 minutes in the case where a reliable precipitation detector is available (first column in Figure 1), and to 0.5 mm over 30 minutes if it is not the case (second column).

A lower threshold was selected for the case where the event selection is based on the combination of the data from the precipitation detector and the weighing gauge, due to the higher degree of confidence given by the redundancy of the reports from two independently operating instruments. When only the weighing gauge report counts towards the decision for an event selection, the threshold is set to be more conservative.

Any 30 minute window during which these two conditions are fulfilled is considered to be a 30 min precipitation event, and is added to the SEDS. If these conditions are not fulfilled, the algorithm moves to the next 30 min interval.

To track which procedure of the two columns of Figure 1 was applied for the identification of each event, a flag was proposed to indicate if a precipitation detector was used or not, i.e. Flag = 0 or 1, respectively. This flag is reported in the SEDS, appended to the aggregated quality-control flag (see Section 4.2 for more details).

4.1.2. Second step – 30 min event parameters

For each 30 minute event identified, the algorithm outputs several parameters to characterize the event in detail for further analysis. The list of parameters in the output event file (i.e. the SEDS) was meant to be as consistent as possible for all sites to facilitate comparative analysis; however, since no two sites have identical equipment or sensor setups, some adaptation was required. Based on the general list outlined in second step of Figure 1, a site specific list of parameters for each individual SEDS was created.

The accumulation over the event was calculated by taking the difference between the last accumulation value and the first accumulation value over the 30 min interval. For Geonor weighing gauges, the accumulation of each individual transducers as well as the accumulation of the average of the three transducers were computed and reported in the SEDS. For Pluvio^2 weighing gauges, the 'Bucket RT' as well as the 'Accumulated NRT' accumulation were computed and reported in the SEDS.

For SUT outputting 1 min accumulation or intensity only, the sum, mean, minimum, maximum and standard deviation of the data over the 30 min event were computed and reported in the SEDS.

For all ancillary measurements, the mean, minimum, maximum and standard deviation of the data over the 30 min event were also computed and reported in the SEDS.

All together, these event parameters with their corresponding statistics and flag (see section **Error! Reference source not found.**) constitute the comprehensive Site Event DataSet (SEDS).

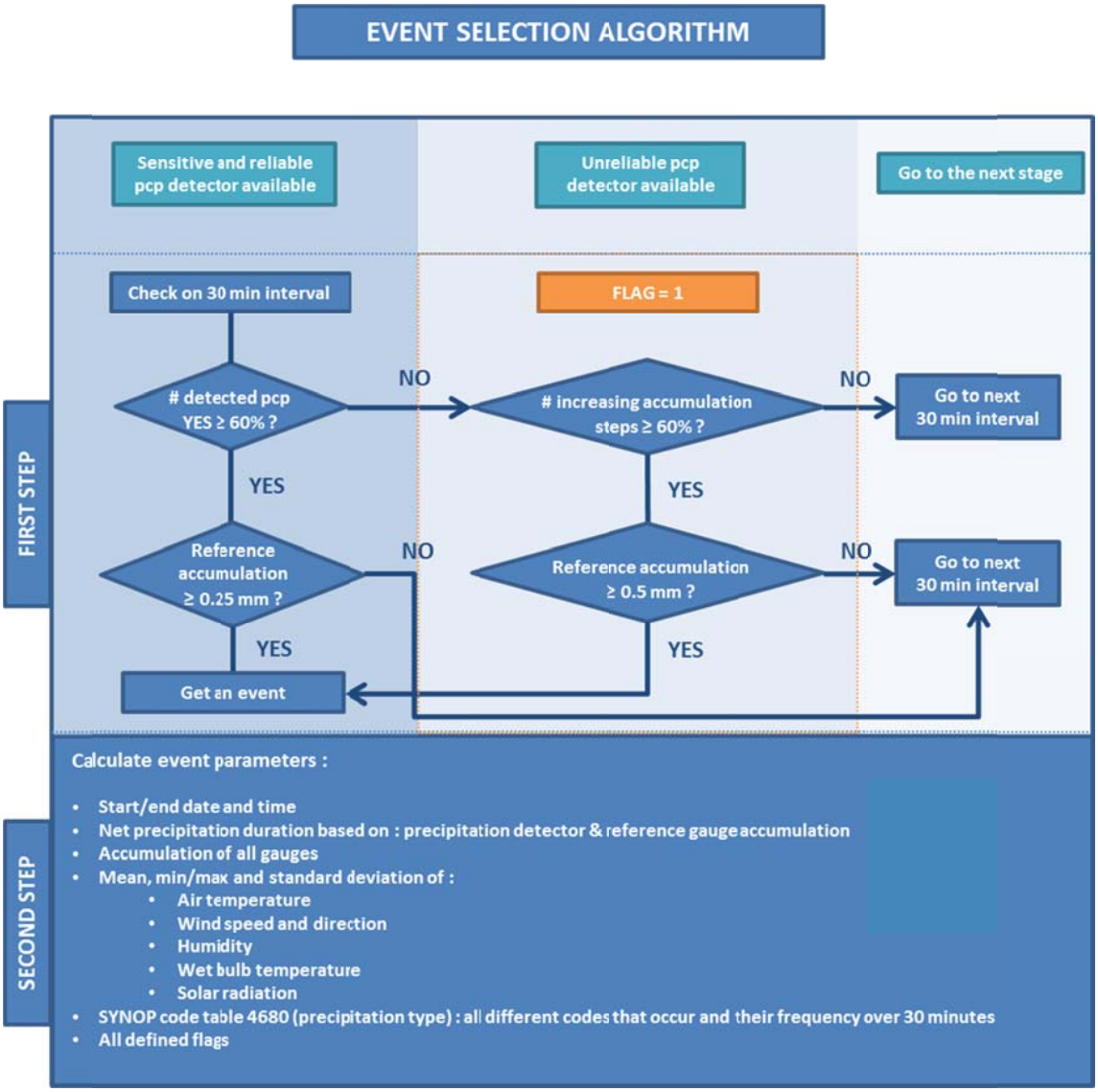

Figure 1: Flowchart representing the event selection algorithm and the list of event parameters for output.

4.2. Event flags

In addition to data parameters, the SEDS file includes event flags to inform on events which may be less reliable and could impact subsequent analysis. The event flag is composed of an aggregation (on the 30 min event) of the 1 minute quality control flags (as defined in Section 3.6) and of the flag coming from the event selection algorithm, indicating which options was applied to get the event, i.e. with or without a precipitation detector (see Section 4.1.1 and Figure 1). The aggregation of 1 min QC flags is based on the percentage number of 'Good' data flags (number 1, see Table 1) reported during the 30 min event.

The event flag approach is outlined below in Table 2. As a consequence, each parameter in the SEDS file is accompanied by its corresponding 30 min event flag.

Table 2: SPICE data quality flagging system for precipitation event files, using QC flagging system described in Table 1.

Flag value	Data Classification	Data Characterization
1	'Good'	Number of 1 minute data points with QC flag = 1 > 80%
11	'Good/no precip detector'	Same as 1, but event selected without precipitation detector
2	'Suspect'	60% < number of 1 minute data points with QC flag = 1 < 80%
21	'Suspect/no precip detector'	Same as 2, but event selected without precipitation detector
3	'Doubtful'	Number of 1 minute data points with QC flag = 1 < 60%
31	'Doubtful/no precip detector'	Same as 3, but event selected without precipitation detector

4.3. SLEDS and SNEDS

The SEDS gives a good opportunity to analyze precipitation events with a high level of confidence. However, the 30 min intervals that weren't selected by this process still contain light precipitation events, or have no precipitation at all. Other studies in SPICE were interested on these remaining 30 min intervals for different purposes. Snow On the Ground (SOG) analysis for instance, needed to focus on non-precipitation events to better interpret the data of SOG sensors, while another analysis required the remaining cases where light events occurred. Following these needs, two additional event files were produced for each site: the Site Non-Event DataSet (SNEDS), accounting for 30 min intervals over which no precipitation occurred, and the Site Light-Event DataSet (SLEDS), comprising all the remaining 30 min intervals. The format for these two files was exactly the same as the SEDS, since they are also based on 30 min intervals. The criteria used to create SEDS, SLEDS and SNEDS are summarized in Table 3. The three files were computed for every SPICE site, for every season, and cover, all together, the whole 1 min datasets available from the site.

Table 3: Criteria used to compute the three different event files : the precipitation event file (SEDS), the light precipitation event file (SLEDS) and the non-precipitation event file (SNEDS). 'Ref Acc' refers to the accumulation of the FWRS weighing gauge, 'PrecipDet_Y' to the number of minutes with precipitation detected by the precipitation detector, and '# Ref_Acc_Y_min' to the number of minutes of increasing accumulation from the FWRS weighing gauge.

	SEDS Site Event DataSet	SLEDS Site Light Event DataSet	SNEDS Site Non Event DataSet
Conditions to fulfill over a 30 min interval to get an event	Not flagged : Ref Acc ≥ 0.25 mm PrecipDet Y ≥ 18 min Flagged : Ref Acc ≥ 0.5 mm # Ref_Acc_Y_min ≥ 18 min	Not flagged : 0 < Ref Acc < 0.25 mm PrecipDet_Y ≥ 1 min Flagged : 0 < Ref Acc < 0.25 mm # Ref_Acc_Y_min ≥ 1 min	Not flagged : Ref Acc ≤ 0.05 mm PrecipDet_Y = 0 min Flagged : Ref Acc ≤ 0.05 mm # Ref_Acc_Y_min = 0 min

5. References

Duchon, C.E. (2008). Using vibrating-wire technology for precipitation measurements. In S.C. Michaelides (Ed.), Precipitation: Advances in measurement, estimation and prediction (pp. 33-58). Berlin: Springer.

National Aeronautics and Space Administration (2010). Data processing levels for EOSDIS data products. Retrieved from http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/

World Meteorological Organization (2015). WMO Space Programme: from data to products. Retrieved from http://www.wmo.int/pages/prog/sat/dataandproducts_en.php