NEON Algorithm Theoretical Basis Document

QA/QC Plausibility Testing

PREPARED BY	ORGANIZATION	DATE
Jeff Taylor	FIU	1/22/2013

APPROVALS (Name)	ORGANIZATION	APPROVAL DATE
Hank Loescher	FIU	9/12/2012
Hanne Buur	CCB DIR SE	9/20/2012

RELEASED BY (Name)	ORGANIZATION	RELEASE DATE
Anne Balsley	CCB ADMIN/DCS	1/22/2013

See Configuration Management System for approval history.
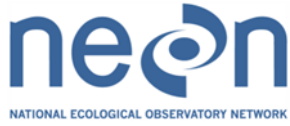
Change Record

REVISION	DATE	ECO #	DESCRIPTION OF CHANGE
A	1/22/2013	ECO-00246	Initial release
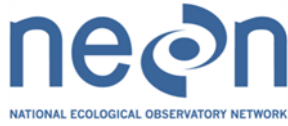
Title: NEON Algorithm Theoretical Basis Document – QA/QC Plausibility Testing	Author: J. Taylor	Date:1/22/2013
NEON Doc. #: NEON.DOC.011081		Revision: A

TABLE OF CONTENTS

	Title: NEON Algorithm Theoretical Basis Document – QA/QC Plausibility Testing	Author: J. Taylor	Date:1/22/2013
	NEON Doc. #: NEON.DOC.011081		Revision: A

1   DESCRIPTION

1.1   Purpose

This document specifies the plausibility algorithms as part of the automated Quality Control/Quality Assurance plan for TIS data [RD 03]. Specifically, this document describes the data flow, and the automated test routines for checking the plausibility of instrument observations. These plausibility tests will require site-specific parameters for many of the Level 0 Data Products, i.e., realistic high- and low value, some of which may rely on historical data available from public sources. This plausibility document includes several tests: range tests, persistence tests, stochastic tests, and variance tests, all of which will be used for most Level 0 Data Products, while others may only use one or two. Also described are some basic statistical properties and example equations/code to demonstrate implementation of these tests.

1.2   Scope

These algorithms are intended to be applied automatically to the L0 data (raw) to determine quality control in producing Provisional Level 1 data products. These tests will be used to automatically examine data over a short timescale (*e.g*., quasi-daily) and to determine the plausibility of each and every observation. The test quantities calculated in this document may be referenced by algorithms in other documents.

The data product algorithms can broadly be categorized as tests to check plausibility. Plausibility testing algorithms will focus on traditional range tests, stochastic tests, and persistence tests. Where possible, efficiency will be maximized through the use of combined tests.

2 RELATED DOCUMENTS AND ACRONYMS

2.1 Applicable Documents

AD[01]	NENO.DOC.000783	ATBD – De-Spiking and Time Series Analyses
AD[03]	NEON.DOC.005013	NEON Data Products Management Plan
AD[03]		
AD[04]		

2.2 Reference Documents

RD[01]	NEON.DOC.000243	NEON Glossary of Terms
RD[02]	NEON.DOC.000291	NEON Configured Sensor List
RD[03]	NEON.DOC.XXXXXX	TIS Data Quality Plan (modified NEON.FIU.011009.PLA.B)
RD[04]	NEON.DOC.000008	NEON Acronym List

2.3 Acronyms

Acronym	Explanation
ATBD	Algorithm Theoretical Basis Document
NOAA	National Oceanic Atmospheric Administration
QA/QC	Quality Assurance/Quality Control
USCRN	U.S. Climate Reference Network
WMO	World Meteorological Organization
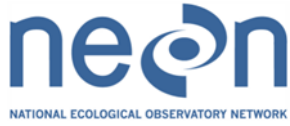
2.4   Verb convention

"Shall" is used whenever a specification expresses a provision that is binding. The verbs "should" and "may" express non-mandatory provisions. "Will" is used to express a declaration of purpose on the part of the design activity.

3   DESCRIPTION

The plausibility tests and outputs are described in detail below.

3.1   QA/QC Test Definitions

3.1.1   Range Tests

A Range Test checks that every recorded observation falls within reasonable minimum and maximum values for a given location and time of year. Ideally, these min/max range limits are determined from historical climate data and previously observed instrument data. For example, if the temperature in Hawaii was observed to be -30 °C, the range test would flag this as implausible because this is lower than the expected minimum value (*i.e.* out of range). The theoretical basis of how these values are determined is outlined in Section 4.1.

3.1.2   Persistence Tests

Persistence tests check that there is a realistic fluctuation of values over a designated period of time. Typically this involves two tests: a "sigma-test" and a "delta-test". The sigma-test uses the standard deviation or variance of the data over a given period of time and compares it to a given threshold value (threshold definition is discussed below). If the standard deviation is below this sigma threshold then the observations have not varied realistically and the test is failed. The delta-test examines the difference between pairs of subsequent observations over a given time period.   If the difference is less than the delta-threshold, then the observations have not varied realistically and the test is failed. By using both of these tests in tandem, instruments that are "stuck" at a constant value can be identified. For example, a radiation sensor that is completely covered with snow may report that there is adequate fluctuation between subsequent measurements (*i.e.* pass the delta-test) but the variance over a 24-hour period will be lower than expected because it is not able to view the daily change in solar radiation (*i.e.* fail the sigma-test), so the persistence test would flag the data over this 24-hour period as implausible. The theoretical basis of how the sigma and delta threshold values are determined is outlined in Section 4.1.

3.1.3   Stochastic Tests

Stochastic tests check to ensure that changes in a time series of data are realistic over a given period of time. They are similar to persistence tests in the sense that they check the plausibility of data based on temporal variation but, rather than be concerned with the minimum fluctuations, stochastic tests check for maximum fluctuations in data sets. A typical test is the "step test" in which successive data points are compared to determine if their difference exceeds a maximum threshold. Missing data points are also typically determined by a null test in which the number of dropped data points over a given period is monitored. A gap test is also necessary for data acquisition systems that don't report null values but rather just have a "jump" in sample times. For example, a compromised connection between a sensor and a data logger would result in realistic data variation (*i.e.* pass the step test) but have an increased number of dropped data points (*i.e.* fail the null test) so the stochastic tests would flag these data as implausible. The theoretical basis of how the step test and null test threshold values are determined is outlined in Section 4.1.

4   SCIENTIFIC CONTEXT

4.1   Theory of algorithm

4.1.1   Theoretical Basis – Statistical Distributions

The raw data received from every sensor must be scrutinized to determine if it is a reasonable value. This is most simply implemented by having a distribution of values that are considered "reasonable" for every sensor at every site. This range, persistence, and stochasticity can all be rigorously determined by constructing a statistical distribution based on existing data. For example, having a representative distribution of values is more effective than simply using historical minima and maxima as there is no way to ensure that these data, themselves, are reasonable.

Because the sensors monitor physical quantities that span numerous distributions, it is not possible to assume one fundamental distribution and calculate the desired quantities. However, as the point of interest is not with the distribution of the data but rather with a statistical quantity derived from these data, a sampling distribution of the statistic can be constructed. Since sampling distributions are constructed from randomly sampled data, the *Central Limit Theorem* states that the distribution will approach a Gaussian distribution as the number of samples approaches infinity. Therefore, regardless of the nature of the underlying data, a properly constructed sampling distribution of a dataflow (statistic) based on these data will always approach a Gaussian distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$ (1)

where, x is any random variable, μ is the population mean of the random variable, and σ is the population standard deviation of the random variable. For example, a statistic for the minimum temperature at a given location will have a Gaussian Distribution constructed from minimum temperature samples. From this sampling distribution, inferences about the population mean minimum temperature and population minimum temperature standard deviation can be used to define the minimum temperature value that will be used in plausibility tests.

Because the Gaussian Distribution is unimodal and symmetric, the random variable can be *normalized* by the standard deviation to yield a curve with the mean value centered at zero (see Figure 1). When the dataset is normalized, the integral between   and   represent 99.7% of all the data and the integral over   and   represent 95% of all the data. By exploiting these properties, we can define consistent parameters for all plausibility tests.

4.1.2   Theoretical Basis – Plausibility Tests

By utilizing sampling distributions for the appropriate test statistic, specified parameters for plausibility test values can be constructed for all measurements. The exact details of the sampling period and sample size will vary depending on the observation and sensor and shall be captured in the relevant ATBD. For example, precipitation will be observed on a much different time-scale than temperature. The derived specification parameters will ultimately be stored and maintained by NEON's CI team. Because some of the NEON Level 0 Data Products have not been extensively observed before, it may be challenging to adequately construct sampling distributions.   In these cases, best possible estimates of appropriate test parameters will be constructed for initial plausibility tests and, after a sufficient amount of data has been collected; these parameters will be updated based on the NEON observations.   The most up-to-date record of plausibility test parameters will be maintained by NEON's Cyber Infrastructure.   The schedule for regularly augmenting and updating this record will follow the standard protocol for managing TIS Data Quality [RD03].

4.1.2.1   Range Tests

As the range test relies on checking extreme values, it is necessary to construct sampling distributions of the minima and maxima observed for a given sample period. For many variables, this will occur on seasonal, diurnal or semi-diurnal time-scales (*e.g.,* temperature, radiation, humidity). It is also possible that an extreme value may not be statistically and/or quantitatively defined for some quantities (*e.g.*, minimum wind speed, maximum wind direction). From the distributions of these statistics, acceptable range
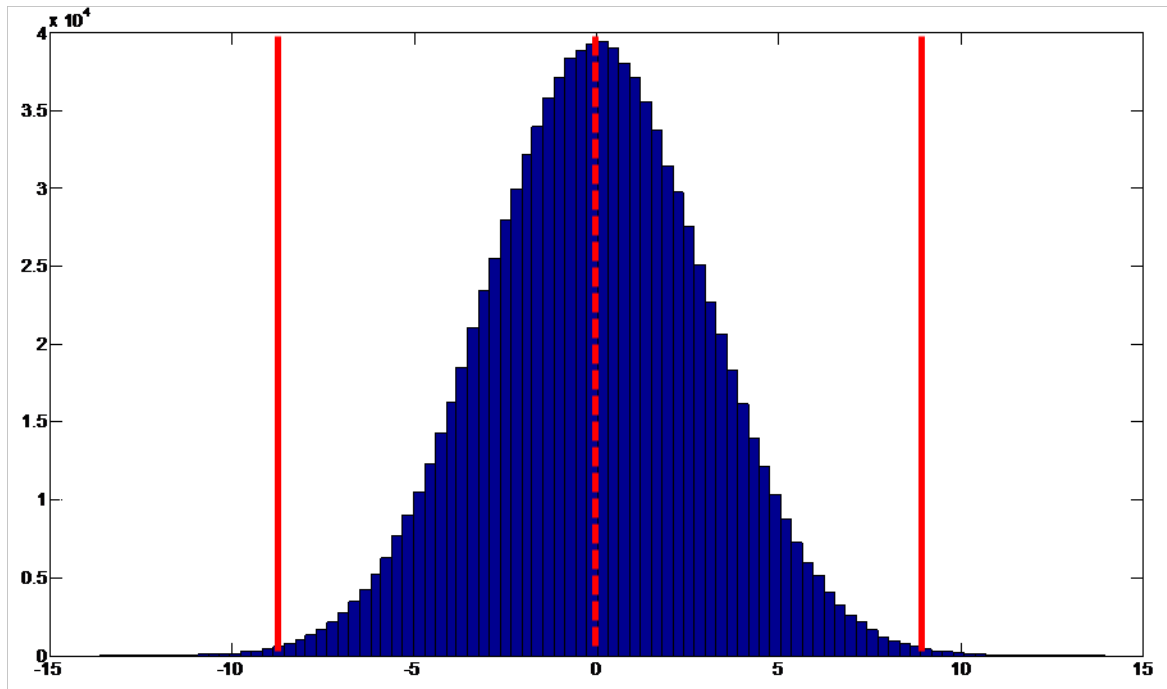
Figure 1: **Histogram of simulated data following a Gaussian distribution with normalized mean 0 (dotted red line) and standard deviation of 3. The range of values that lie between 3 standard deviations of the mean (solid red lines) represent 99.7% of all the data.**

thresholds are defined by   . By using the twice the standard deviation, 95% of all observed extreme values are considered acceptable, while the remaining 2.5% will be flagged as outliers. This threshold calculation will be applied to all incoming data streams unless explicitly stated otherwise.   It should be noted that the selection of statistical thresholds is always a subjective decision – it may be

that specific observations require a threshold based on 3σ or some other comparable metric.   If this is the case, it shall be explicitly detailed in the respective ATBD.

4.1.2.2   Sigma Tests

The sigma test relies on the variance or standard deviation of the data in a given sample. Consequently, a sampling distribution of all the sampled standard deviations of the data set will provide an inference of the minimum and maximum expected variability of a given parameter. A threshold of   (where   represents the mean standard deviation of the distribution and   is the standard deviation of the distribution of standard deviations) ensures that only the lowest/highest 2.3% variability is flagged. In many cases, care must be taken when scrutinizing the plausibility of this value and it will often need to be used in conjunction with other plausibility tests. For example, if there is no precipitation over a three day period (a very realistic case), the sigma test, alone, would reject this as having zero variability. This false-failure can be corrected by having two-stage tests where "zero variability conditions" are checked against other plausibility parameters.

4.1.2.3   Delta Tests

Similar to the sigma test the delta test scrutinizes the variability of a data set, but it focuses more on the observed noise rather than sampled variability over a specified period. The delta test utilizes the difference between subsequent observations to check the noise fluctuation. The mean and standard deviation of this sampling distribution represent how noise changes throughout the different observation samples. If this quantity changes less than the   threshold, it is flagged as being possibly "stuck" at a given value. Again, care must be taken with this test to ensure that observations that commonly read 0.0 are not being inadvertently flagged. In some cases, it may be advantageous to define the delta test threshold by the sampling precision of the instrument/data acquisition system, rather than statistically. For example, if the resolution of the instrument is 0.005, then the delta test may want to use a threshold of ~0.01 to see if values are "stuck" and only vary near the instrument's resolution.

4.1.2.4   Step Tests

The same distribution of subsequent observations is used to also define the threshold for the step test. Rather than scrutinize the smallest acceptable change in measurements, the step test seeks to ensure that there are no implausibly large jumps between measurements. The threshold is defined as   to ensure that only the largest 2.3% of all jumps are flagged.

4.1.2.5   Null and Gap Tests

The null test and gap test are used to monitor the loss of data.   The exact threshold for acceptable data loss will vary with instrument and sampling interval and, in some cases, may simply be defined as an arbitrary number (e.g. 0 or 1 maximum missing data value per day) or by a local calibration cycle.   For data that is sampled at higher frequencies, the statistical approach that has been used to define all plausibility thresholds should continue to be applied.   A sampling distribution of the number of missing data values over a given sampling period should be constructed. As with the other parameters, a threshold of   is chosen for flagging data with the null test. Because the data acquisition system does not report times with missing data, a gap test is used to explicitly check for a prolonged period of missing data. The threshold for this test must be chosen based on the sampling period and should be defined under optimal sampling conditions.

All of the plausibility test parameter definitions are summarized in Section 4.1.3.

4.1.3   Plausibility Test Parameter Definitions

Consistent testing parameters will be defined using the following approach:

Plausibility Test	Underlying Statistical Quantity Sampling Distribution	Calculation
Range Test	Extreme Values	Max: , Min:
Sigma Test	Standard Deviation	
Delta Test	Differences of Subsequent Pairs	(or defined by sampling)
Step Test	Differences of Subsequent Pairs	
Null Test	Missing Data	(or defined by sampling)
Gap Test	Large Gap of Missing Data	Defined by Sampling

5 ALGORITHM IMPLEMENTATION

5.1.1 Algorithms for Defining Test Parameters

Although a Gaussian probability distribution function (Eq. 1) can be constructed from historical climate data (e.g., form NCDC) for a few variables, this process is computationally expensive and inefficient for the amount of data NEON will be addressing. Without loss of generality, an algorithm that calculates the first two moments of a Gaussian Distribution (the mean and variance, respectively) can be constructed discretely to be:

$$\bar{x}(d) = \frac{\sum\limits_{y}^{H(d)} x(d, y)}{\sum\limits_{y}^{H(d)} 1} \qquad (2)$$

$$\sigma^2(d) = \frac{\sum\limits_{y}^{H(d)} \left[x(d, y) - \bar{x}(d) \right]^2}{\sum\limits_{y}^{H(d)} 1} \qquad (3)$$

where, x is a measurement statistic on a given day, d, , with a historical dataset of measurements on this day, H(d), and $\bar{x}$ and σ^2 are the derived mean and variance for this measurement statistic. For example, this could be a data set of daily maximum temperatures observed at a specific location for 30 years.

While this approach is computationally more efficient than manually constructing these parameters, it does not include additional information, such as temporally and spatially adjacent observations. To do this, a combined calculation incorporating spatial, temporal, and historical variation must be done. However, prior to entering the operational phase of the NEON Project, it is likely that historical data may be all that's available for a specific measured quantity at a specific location in which case, Equations (2) and (3) should be used to derive the appropriate plausibility parameters from these data.

5.1.2   Algorithms for Defining Test Parameters using Combined Calculations

Once the operational phase of the NEON Project has begun and there are more data representative of spatial and temporal variation available, algorithms utilizing a combined approach for defining plausibility parameters may be more appropriate. As the spatio-temporal correlation length scales are unique to each measurement statistic, a useful approach is to incorporate weighting factors for their respective influence. This modifies Equations (2) and (3) to become:

$$\overline{x_i}(d) = \frac{\sum_{j}^{N_i} \sum_{d'}^{D_d} \sum_{y}^{H(d')} w_1(j,i) \cdot w_2(d',d) \cdot x_j(d',y)}{\sum_{j}^{N_i} \sum_{d'}^{D_d} \sum_{y}^{H(d')} w_1(j,i) \cdot w_2(d',d)} \qquad (4)$$

$$\sigma_i^2(d) = \frac{\sum_{j}^{N_i} \sum_{d'}^{D_d} \sum_{y}^{H(d')} w_1(j,i) \cdot w_2(d',d) \cdot \left[x_j(d',y) - \overline{x_i}(d) \right]^2}{\sum_{j}^{N_i} \sum_{d'}^{D_d} \sum_{y}^{H(d')} w_1(j,i) \cdot w_2(d',d)} \qquad (5)$$

where, N_i is the set of neighboring sites measuring the same quantity (or alternatively, neighboring or redundant measurements at the same site), D_d is the set of adjacent dates upon which the quantity is measured, and w_1 and w_2 represent the spatial and temporal weighting factors, respectively. These are defined as:

$$w_1(j,i) = \begin{cases} 0, & j \notin N_i \\ 1, & j = i \\ \frac{1}{2} e^{-\left(\frac{|\Delta_{xy}|}{z} \right)^2}, & j \in N_i / \{i\} \end{cases}$$

$$w_2(d',d) = \begin{cases} 0, & d' \notin D_d \\ e^{-\left(\frac{|d'-d|}{t}\right)^2}, & d' \in D_d \end{cases}$$

where $|\Delta_{ij}|$ represents the distance between neighbouring sites in degrees (i.e. latitude and longitude), z represents the maximum allowable distance between neighbouring sites, and t represents the maximum time period over which adjacent dates of observation are considered. The temporal weighting is based on observations changing linearly with time and the spatial weighting is based on traditional Barnes Interpolation Analysis [Barnes, 1964].

5.1.3   Algorithm Framework

Once all of the plausibility test parameters have been defined, the tests are implemented in sequence for each observation at each site. It is anticipated that this will be an entirely automated testing procedure in which individual data streams are checked prior to any other data manipulation. Figure 2 shows an example process for this procedure.

The sequence shown here need not be followed in every case. Different observations and data products will require different sequences of tests that shall be detailed in the relevant ATBD. Furthermore, in the interest of computational efficiency, it may be sufficient to not subject data that has already been flagged to more plausibility testing. That is, if a "window" or subset of data has failed the sigma test, it may not need to be subsequently subjected to a delta test. All of these details can be found in the specific ATBD of interest.
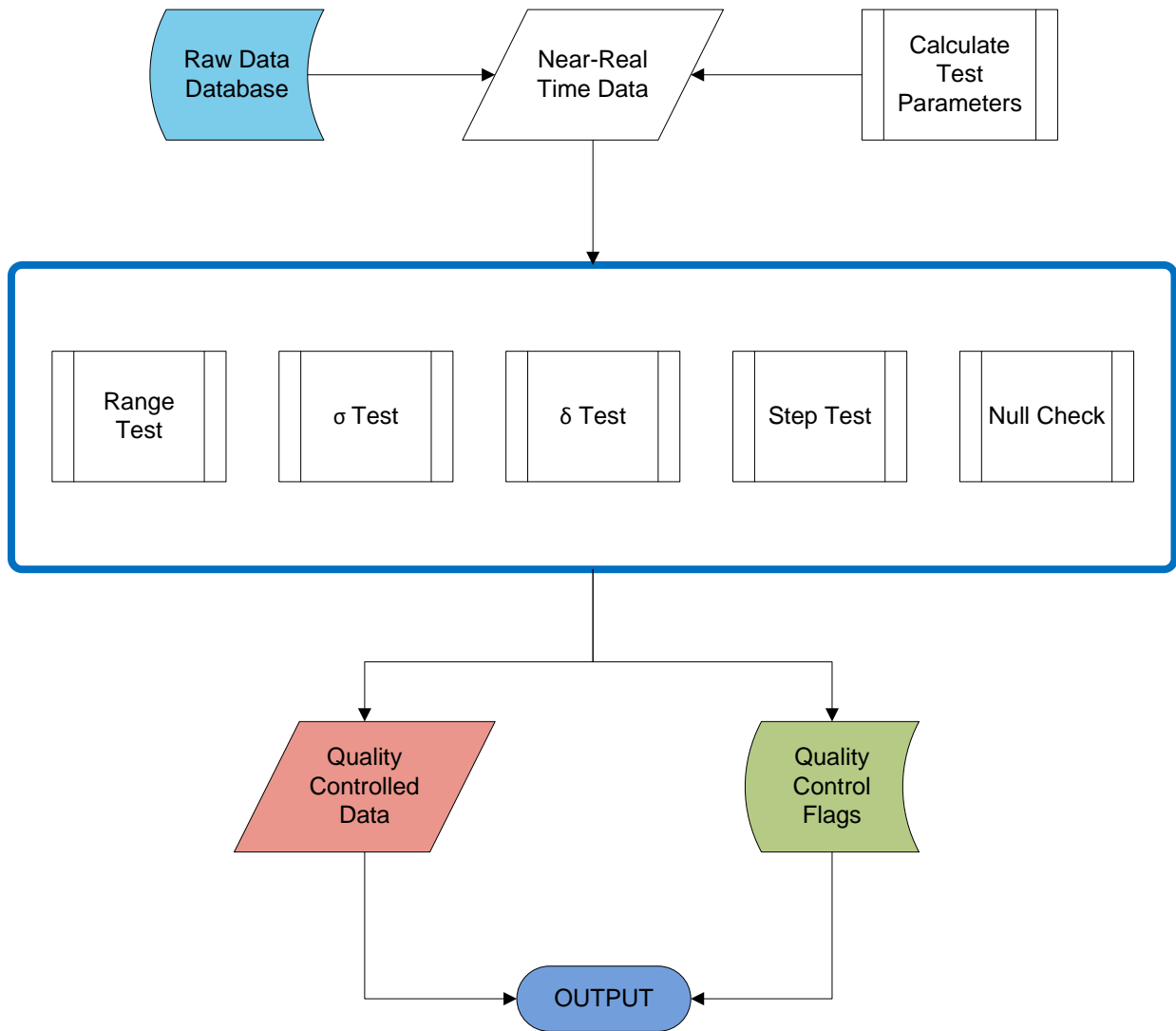
Figure 2: Data flow diagram for automated plausibility testing. The sequence of these tests need not be in the linear configuration shown here and relevant ATBDs will define the actual process.

5.1.4 Algorithm Framework

An example of potential code for testing the automated plausibility algorithms was implemented in MATLAB (i.e. it is written in syntax similar to traditional C and FORTRAN languages). All user Comments are preceded by "%" symbols.

5.1.4.1   Code for the Calculation of Test Parameters

```
function [Range, Step, Sigma, Delta, Null] = QC_PRT_define(filename)
%
% SUMMARY:
% MATLAB function to calculate thresholds for quality control of aquatics
% data based on historical datasets.
%
% DESCRIPTION:
% Function reads a set of data from a tab-delimited text file, constructs
% sampling distributions of the data based on predefined statistics, and
% determines acceptable threshold values with the intention that these values
% can later be applied to operational data quality control.   Figures are
% generated to show how these thresholds are applied to the dataset.
%
% INPUTS:
%   STRING filename: The name of the *.mat in which the data is stored.   The
%      file should have one array with two columns; the first is the date
%      and the second is the floating point data.   This version assumes that
%      the data is daily data only.
%
% OUTPUTS:
%   FLOAT Range: A 2 element vector describing the lower and upper range
%      value limits.
%   FLOAT Step: A 1 element variable defining the maximum step limit.
%   FLOAT Sigma: A 2 element vector describing the upper and lower sigma
%      limits.
%   FLOAT Delta: A 1 element variable defining the delta test limit.
%   FLOAT Null: A 1 element variable defining the null test limit.
%
% Jeff Taylor
% June, 2011


% Program parameter definition:

    path = 'H:\Sterling\prt';

    nsample = 1000; % The number of samples to use when sampling the data.
    ndatum = 100; % The number of data points to include in each sample.
    nullcheck = [0 0 0 0 0 2]; % Sampling rate of the data in [Y M D H M S]
```

```matlab
% Load the data and organize the array:

    cd(path);
    load(filename);

    nn = length(data);
    cereal = data(:,1);% - dateoffset;
    prt = data(:,2);


% Sort the data by date:

    [cereal,sortdex] = sort(cereal);
    prt = prt(sortdex);



% Now let's do some funky analysis!!!



% RANGE PARAMETERS:

% Sample the data to calculate means, sigmas, and "cutoffs"

    for i = 1:nsample
        r = randi(nn,ndatum,1);
        meaner(i) = mean(goodprt(r));
        sigmaer(i) = std(goodprt(r));
    end

    highcutoffer = meaner + 2*sigmaer; % This is the mean + 2 standard
deviations for each year.
    lowcutoffer = meaner - 2*sigmaer; % This is the mean - 2 standard
deviations for each year.
    highmeancutoffer = mean(highcutoffer); % This is the average of all
means+2sigmas.
    lowmeancutoffer = mean(lowcutoffer); % This is the average of all means-
2sigmas

    highcutoffthresh = highmeancutoffer + 2*std(highcutoffer);   % This is
the upper range threshold
    lowcutoffthresh = lowmeancutoffer - 2*std(lowcutoffer); % This is the
lower range threshold
```

```matlab
% SIGMA PARAMETERS:


    sigmacuthigh = mean(sigmaer) + 2*std(sigmaer); % This is the sigma mean
+ 2 sigma standard deviations for each sample.
    sigmacutlow = mean(sigmaer) - 2*std(sigmaer); % This is the sigma mean -
2 sigma standard deviations for each sample.



% STEP PARAMETERS:


    step = diff(goodprt); %Find the differences between all subsequent data
points.
    for i = 1:nsample
        r = randi(nn,ndatum,1);
        stepper(i) = mean(abs(step(r))); %Annual mean step difference
        sigmastep(i) = std(abs(step(r))); %Annual sigma step difference
    end

    stepcut = stepper + 2*sigmastep; %Annual mean step difference+2*sigma
(for maximum threshold)
    stepcuthresh = mean(stepcut) + 2*std(stepcut); %Step test threshold



% DELTA PARAMETERS:


    deltacut = stepper - 2*sigmastep; %Annual mean step difference-2*sigma
(for delta test)

    deltacuthresh = mean(deltacut) - 2*std(deltacut); %Delta test threshold
    if deltacuthresh < 0, deltacuthresh = 0; end %if the delta test threshold
is negative, set to 0



% NULL PARAMETERS:


    nullcereal = datenum(nullcheck);
    diffcereal = diff(goodcereal);

    for i = 1:nsample
        r = randi(nn,ndatum,1);
```

```matlab
        meannull(i) = nnz(find(diffcereal(r) > nullcereal));
    end


    nullcutoff = mean(meannull) +2*std(meannull);



% Let's output everything!

    Range = [lowcutoffthresh highcutoffthresh]
    Step = [stepcuthresh]
    Sigma = [sigmacutlow sigmacuthigh]
    Delta = [deltacuthresh]
    Null = [nullcutoff]
```

5.1.4.2   Code for the Implementation of Plausibility Testing.

```matlab
Function QC_PRT_test(data,Range,Step,Sigma,Delta,Null,Gap)
%
% SUMMARY:
% MATLAB function to perform plausibility tests on a given dataset.
%
% DESCRIPTION:
% Function takes a given set of data and performs plausibility tests,
% specifically Range, Step, Sigma, Delta, and Null tests.   Resulting data
% flags are generated for output.
%
% INPUTS:
%   FLOAT data: An n by 2 element vector containing all of the data to be
%     analyzed.   The fisrt column should contain all of the date/time stamps
%     and the second column should contain the corresponding data.
%   FLOAT Range: A 2 element vector describing the lower and upper range
%     value limits.
%   FLOAT Step: A 1 element variable defining the maximum step limit.
%   FLOAT Sigma: A 2 element vector describing the upper and lower sigma
%     limits.
%   FLOAT Delta: A 1 element variable defining the delta test limit.
%   FLOAT Null: A 1 element variable defining the null test limit.
%   FLOAT Gap: A 1 by 6 element variable defining the maximum allowable gap
%     in time: [Y M D H M S].
%
%
% Jeff Taylor
```

```matlab
% June, 2011


% Program parameter definition:

    path = 'H:\Sterling\prt';
    dateoffset = 1721423.5; % To convert from Julian Date to MATLAB serial
dates.

    sigmaperiod = 500; % The time period over which the variance test is
applied in data points.
    deltaperiod = 100; % The time period over which the delta test is applied
in in data points.
    nullperiod = 50; % The time period over which the null test is applied.
    nullcheck = [0 0 0 0 0 2]; % The sampling rate of the data in [Y M D H M
S]

    obs = 'PRT Reading (mV)'; % The observed quantity that the data
represents -- will be used for plotting
    plottitle = 'Sterling Prototype Temperature'; % The title that will be
slapped onto plots.

% Organize the data

    n = length(data);
    cereal = data(:,1) - dateoffset;
    prt = data(:,2);

% Sort the data by date:

    [cereal,sortdex] = sort(cereal);
    prt = prt(sortdex);

% Now let's do some analysis!!!


% RANGE TEST:

    outrangedex = find((prt > Range(2)) | (prt < Range(1))); % Find the
outliers
    outrange = prt(outrangedex); % Remove them.
    outrangecereal = cereal(outrangedex); % Remove them.
```

```
% SIGMA TEST:
% Look at data in rolling periods to determine if it fails the sigma test.

    sigmatest = ones(n,1);
    for i = 1:n-sigmaperiod
        periodcheck = std(prt(i:i+sigmaperiod-1));
        if ((periodcheck > Sigma(2)))% || (periodcheck < sigmacutlow)) %for
PRT data, only check the high value
            sigmatest(i:i+sigmaperiod-1) = 0; % 1=pass, 0=fail.
        end
    end

    outsigmadex = find(sigmatest == 0); %Find the bad data.
    outsigma = prt(outsigmadex);
    outsigmacereal = cereal(outsigmadex); %Find the times that corresponds to
the bad data.


% STEP TEST:

    steps = diff(prt); %Find the differences between all subsequent data
points.
    badsteps = find(abs(steps) > Step); %test: is the step difference too
big?

    scount = 1;
    for i = 1:length(badsteps)
        bads(scount) = badsteps(i);
        bads(scount+1) = badsteps(i)+1; %flag both data points if they fail
the step test.
        scount = scount+2;
    end

    overstepdex = bads;
    overstep = prt(overstepdex); %Find the bad data
    overstepcereal = cereal(overstepdex); %Find the dates that correspond to
the bad data


% DETLA TEST:
```

```matlab
    deltatest = ones(n,1);
    for i = 1:n-deltaperiod
        deltacheck = std(steps(i:i+deltaperiod-1)); %test: are the step
chenges too small (i.e. is the sensor "stuck")?
        if deltacheck <= Delta
            deltatest(i:i+deltaperiod-1) = 0; % 1=pass, 0=fail
        end
    end


    stuckdex = (find(deltatest == 0)); %Find the bad data
    stuck = prt(stuckdex);
    stuckcereal = cereal(stuckdex); %Find the dates that correspond to the
bad data



% NULL TEST:

    nullcereal = datenum(nullcheck);
    diffcereal = diff(cereal);


    nulltest = ones(n,1);
    for i = 1:n-nullperiod
        nullchecker = nnz(find(diffcereal(i:i+nullperiod-1) > nullcereal));
        if nullchecker > Null
            nulltest(i:i+nullperiod-1) = 0; % 1=pass, 0=fail
        end
    end

    nulldex = find(nulltest == 0);
    nulls = prt(nulldex);
    nullscereal = cereal(nulldex);



% GAP TEST:

    gappycereal = datenum(Gap);
    missing = find(diffcereal > gappycereal);

    gcount = 1;
    for i = 1:length(missing)
        gaps(gcount) = missing(i);
```

```
        gaps(gcount+1) = missing(i)+1; %flag both data points if they fail
the step test.
        gcount = gcount+2;
    end

    gapdex = gaps;
    gapprt = prt(gapdex); %Find the bad data
    gapcereal = cereal(gapdex); %Find the dates that correspond to the bad
data

% Combine all of the bad data into one unique data set and separate from the
rest of the data.

    totalcrap = [outrangedex', outsigmadex', overstepdex, stuckdex',
nulldex', gapdex];
    totalcrap = unique(totalshit); %remove repeat offenders
    gooddex = setdiff(1:n,totalshit); %sort out the good data

    gooddata = prt(gooddex);
    goodcereal = cereal(gooddex);
```

5.2   Example

5.2.1   Test Parameter Calculation

For the purposes of illustration, prototype temperature observations in the form of voltage across a Platinum Resistance Thermometer at the North Sterling, CO site in Domain 10 are analyzed below. It should be noted that these data are preliminary and contain numerous known errors (which is useful for the purposes of this example).

A timeseries of approximately 1 month of data sampled at 1 second intervals are shown in Figure 3. These data will be considered the "historical" data for this site and observation such that the testing parameters can be defined from this timeseries (in practice, much more data will be used for this purpose).
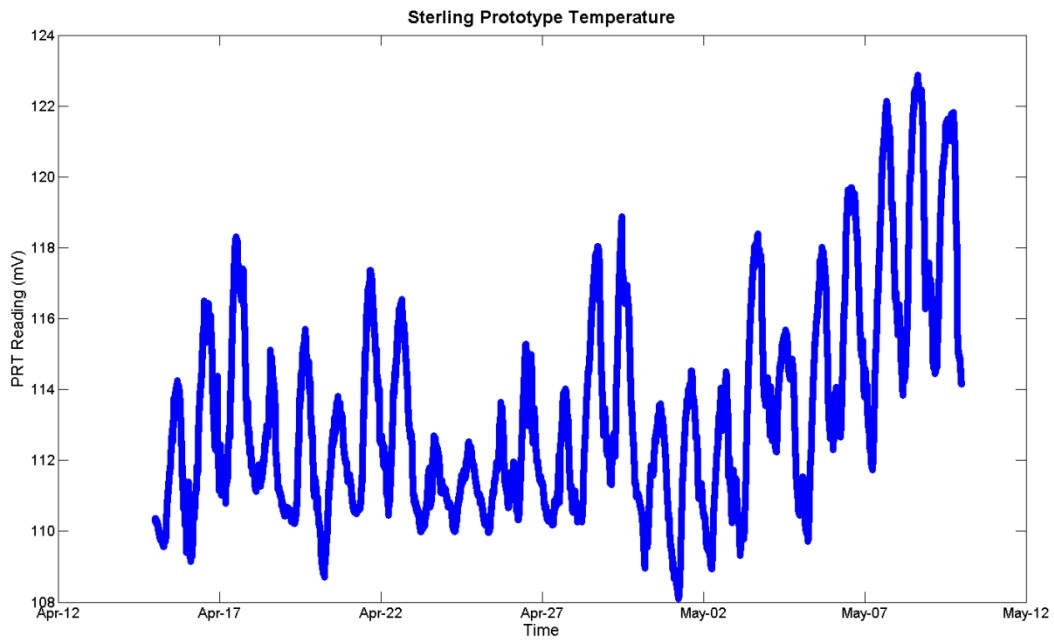
Figure 3: Timeseries of observations from Domain 10: North Sterling, Colorado.

From this timeseries statistical sampling distributions were constructed by randomly sampling 100 data points, 1000 times.   From each sample of 100 data points, a mean and standard deviation were calculated.   The statistical sampling distribution of the mean values is shown in Figure 4.   Note that with only 1000 samples, the shape of the distribution approaches that of the Gaussian shown in Figure 1.   By applying the Central Limit Theorem to this distribution, the inferred population mean is 113.3 mV.
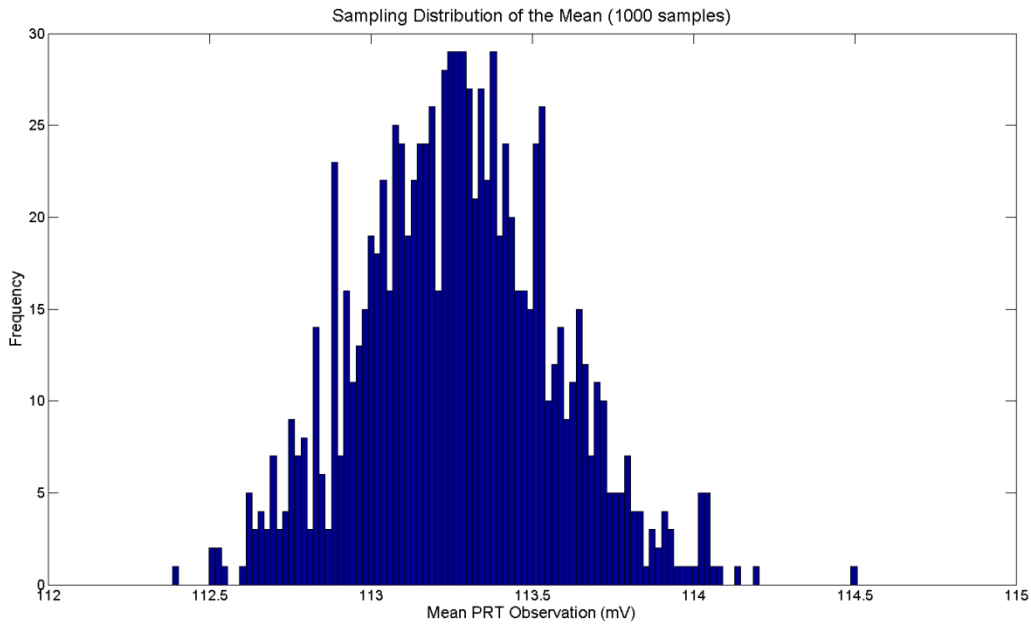
Figure 4: Statistical Sampling Distribution of the Sample Mean constructed from 1000 samples.

Using the same sampling characteristics, a statistical sampling distribution of the upper and lower range

limits ($\mu \pm 2\sigma$) can be constructed (see Figure 5).   From this distribution, the value of the upper

threshold range can be inferred to be: $\mu + 2\sigma = 119.2 + 2 * (0.74) = 120.7$.   Therefore, the Range
Test would consist of automatically checking all of the data to ensure that any values above this
threshold are flagged.

In a similar fashion, all parameters for step testing, sigma testing, delta testing, and null testing are
calculated by constructing sampling distributions (or, as mentioned in Section 4.2.2, may be defined by
the inherent data sampling/acquisition).

5.2.2   Plausibility Test Implementation

The same prototype temperature observations in the form of voltage across a Platinum Resistance
Thermometer at the North Sterling, CO site in Domain 10 are used for illustrating the efficacy of
plausibility testing.   It should be noted that these data are preliminary and contain numerous known
errors (which is useful for the purposes of this example).

A timeseries of approximately 2 months of data sampled at 1 second intervals are shown in Figure 6. These data will be considered the "test" data upon which all of the plausibility tests should be conducted. Using the test parameters derived from the procedure outlined in Section 4.4.1, these data were processed using all of the automated plausibility tests. The data that failed these tests were flagged and are shown in Figure 7.
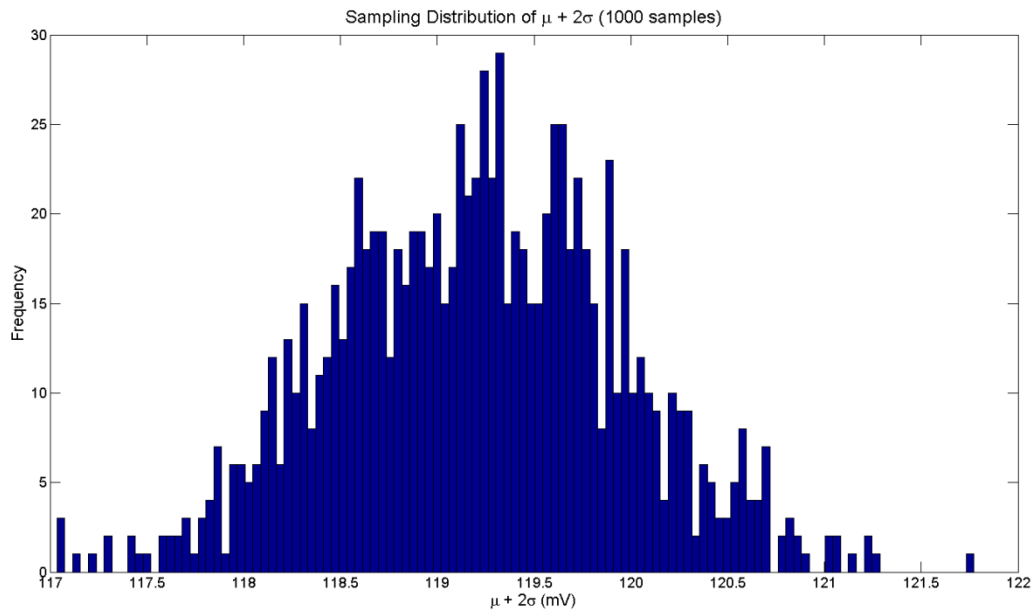

Figure 5: Statistical Sampling Distribution of the Sample Mean added to twice the sample standard deviation.
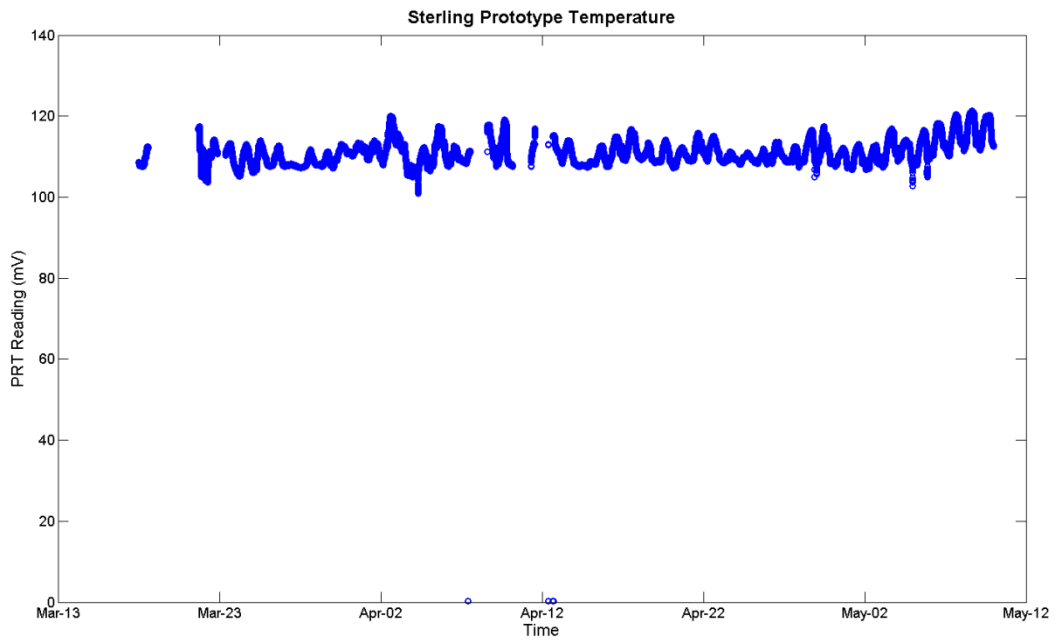
Figure 6: Timeseries of observations from Domain 10: North Sterling, Colorado.

Figure 7: Timeseries of observations from Domain 10: North Sterling, Colorado. The different coloured symbols represent the different flags that have been applied by automated plausibility testing.

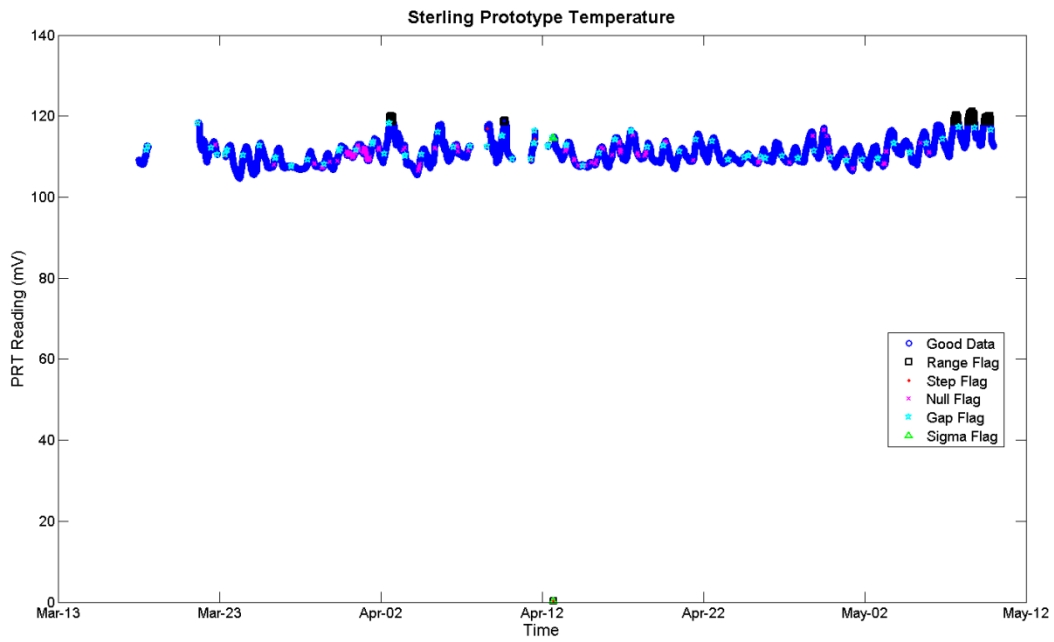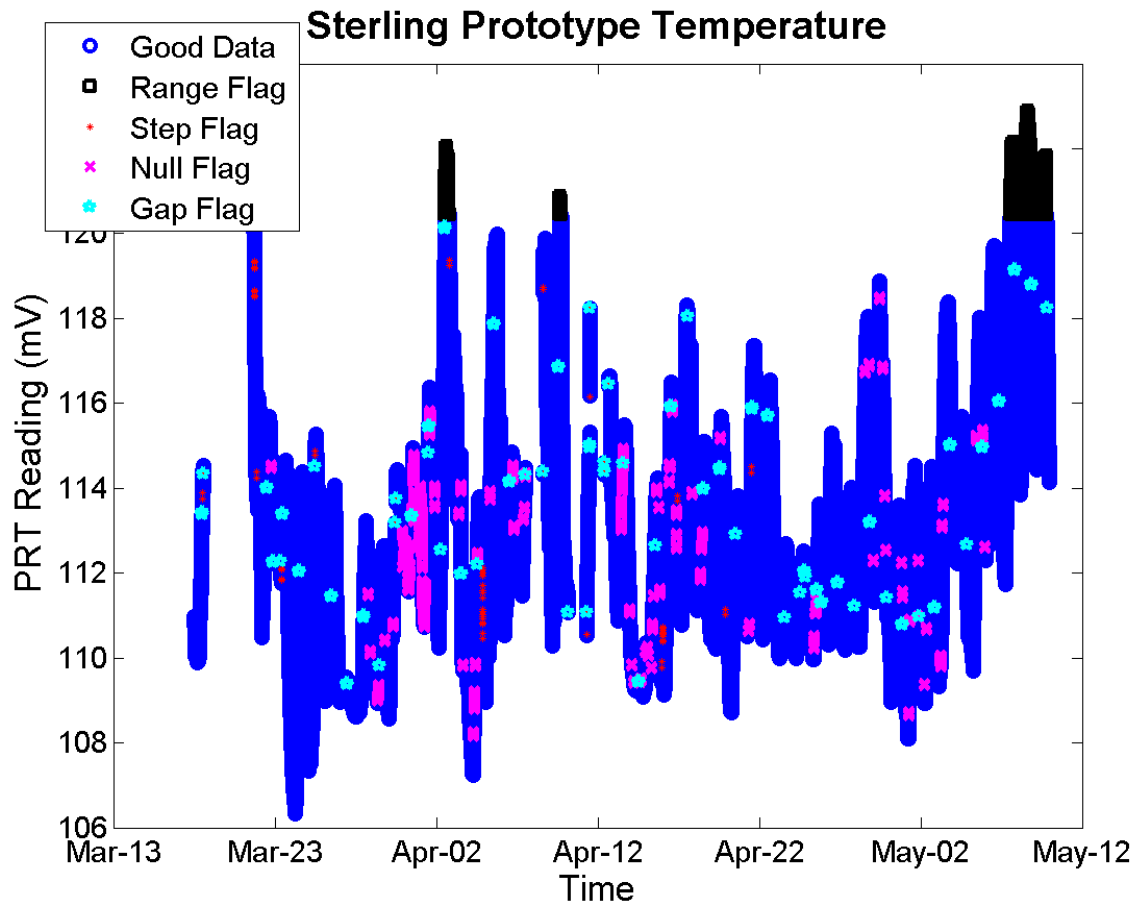Figure 8: Close-up of the timeseries displayed in Figure 7. The different coloured symbols represent the different flags that have been applied by automated plausibility testing.
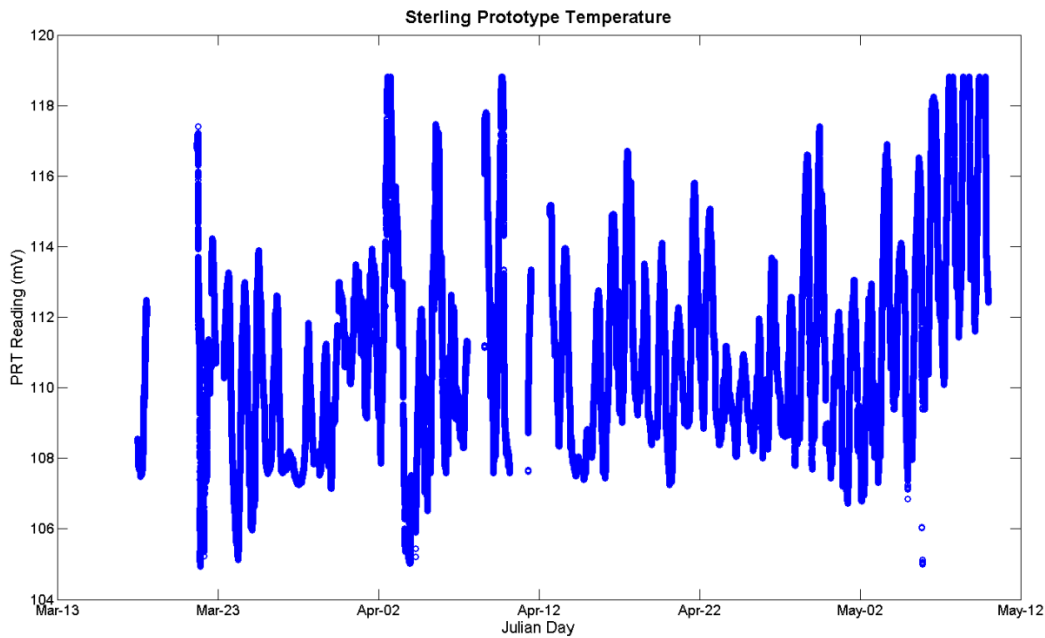
Figure 9: Timeseries of observations from Domain 10: North Sterling, Colorado. All of the flagged data has been removed, leaving only observations that passed all automated plausibility tests (~4.1% of the raw data was flagged).

The automated plausibility tests resulted in the following data quality report (with additional annotations for explanation):

A. Range Test: The range thresholds were found to be [104.04 to 118.56] mV. There were 150 643 values outside of this range resulting in 3.2% being flagged.

B. Step Test: The step threshold was found to be 0.2015 mV. There were 36 values outside greater than this step resulting in 7.5 ppm being flagged.

C. Sigma Test: The sigma thresholds were found to be [2.57 to 3.56] mV. Because the temperature observations in this data set have considerable bias and variation, the lower sigma threshold is much larger than the anticipated noise in the baseline observations. For this reason, the lower variance test was not applied and the plausibility of the variation over small time scales was monitored solely by the step tests and delta tests. While this is not nominally optimal, it is appropriate for this, relatively noisy, dataset. Utilizing the test for only the upper sigma range and applied over a rolling domain sequence of 500 data points, there were 999 instances where the variance is greater than the acceptable sigma range resulting in 0.02% being flagged.

D. <u>Delta Test</u>: Due to the narrow range of variation in the observations, the delta threshold was found to be negative and, consequently, set to 0 for this test.   This will happen with observational datesets of this nature and should typically have a threshold set at the precision resolution of the sensor.   For this particular prototype dataset, this value was not available and the delta test was not applied.   Nominally, the delta threshold would be applied over a rolling domain sequence of ~100 data points, or similar.

E. <u>Null Test</u>: The null threshold was found to be 12.6 missing data points.   This was applied over a rolling domain sequence of 50 data points resulting in 42 804 instances where there were more missing values than the threshold causing 0.9% to be flagged.

F. <u>Gap Test</u>: The gap threshold was chosen to be 5 mins.   There were 116 time gaps greater than this threshold resulting in 24 ppm being flagged.

By combining all of the plausibility tests together, this resulted in 194 581 data points being flagged, or 4.1% of all the data in question.   It should be noted that many poor observations are flagged by multiple tests, so the total number of flagged data points is not simply the linear addition of flagged data points from individual failed tests. The details of the code used to implement these tests are provided below.

6   SAMPLE CODE

7   UNCERTAINTY

While the metrics calculated here will be linked to the overall uncertainty estimations as described in the TIS Data Quality Plan (RD[03]), there are no associated uncertainty estimations for these metrics. These data quality metrics are intended to provide high-level, quantitative scrutiny for the performance of sensors rather than provide direct confidence estimates for the data products themselves.   Individual data product ATBDs should inform how the uncertainty parameters are calculated.

8   VALIDATION AND VERIFICATION

8.1   Algorithm validation
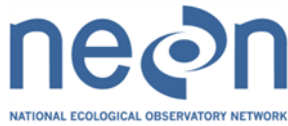
Verification and validation of algorithm implementation shall be managed by DPS, in consultation with FIU and CI in accordance with TIS objectives.   Additional process detail may be found in the NEON Data Products Management Plan (AD[02]).

It is anticipated that testing will be done through construction of a "unit test harness", where trusted input data is used in tests of individual functional components of the algorithm against a set of expected test outputs. Test data sets will be generated synthetically to simulate plausibility failure cases in which each of the aforementioned plausibility tests can be verified.   Output results should be similar to those demonstrated in the example above.   The exact use cases for which plausibility tests will pass/fail are highly data-product specific and, for this reason, care must be taken to ensure that the application of these tests and their associated parameters are documented in each data product's respective ATBD.
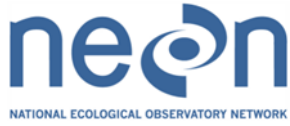
9   SCIENTIFIC AND EDUCATIONAL APPLICATIONS

Upon completion of all plausibility testing for a given set of observations at a given location, <u>all of the data</u> and associated quality information shall be made available for the next sequence of automated quality control testing as defined in the TIS Data Quality Plan [RD 03].   Records of the flagged data should be detailed in the quality control flag report for later consideration and general statistics of the data flags should be output for regular scrutiny.

10   FUTURE PLANS AND MODIFICATIONS

It will be necessary to maintain regularly updated records of all test parameters associated with plausibility testing as they will vary by measurement subsystem and by site location.   All of these parameters will ultimately be calculated and stored by CI according to the TIS Data Quality Plan [RD03]. The tables below are merely intended to serve as an example of the parameters that will need to be calculated for each sensor at each location.   Table A illustrates the details of an example data product, while Table B summarizes example parameters for the Sterling site that were shown in previous Sections.

Table A: Instrumentation, Observations, and Sampling

Observation (L0)	Units (L0/L1)	Instrument	Sampling	Period of Variation
Temperature	milliVolts/Celsius	PRT air temperature sensor	1 Hz	24 Hours

Tables B: Sample Plausibility Parameters by Variable and Location

Table B.1: Temperature (1 meter) Plausibility Parameters by Location

Domain	Location	Range	Sigma	Delta	Step	Null	Gap

10	Sterling	104.04-118.56	0-3.56	0	0.2015	12.6	5 mins

11   BIBLIOGRAPHY

Barnes, S. L., A Technique for Maximizing Details in Numerical Weather Map Analysis, *J. Appl. Meteor.*, **3**: 396-409, 1964.

12   CHANGELOG

DATE	REVISION	CHANGES
9-4-2012	A	Initial Release