Preprocessing for TIS Level 1 Data Products – QA/QC Algorithm Theoretical Basis Document

PREPARED BY	ORGANIZATION	DATE
David Durden	FIU	06/14/2016
Derek Smith	FIU	07/03/2013

APPROVALS	ORGANIZATION	DATE
Andrea Thorpe	SCI	08/30/2016
Rick Farnsworth	PSE	08/25/2016

RELEASED BY	ORGANIZATION	DATE
Judy Salazar	CM	08/31/2016

See configuration management system for approval history.

The National Ecological Observatory Network is a project solely funded by the National Science Foundation and managed under cooperative agreement by Battelle. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Change Record

REVISION	DATE	ECO #	DESCRIPTION OF CHANGE
A	12/05/2014	ECO-01052	Initial Release
B	08/31/2016	ECO-04040	Adding Regularization

TABLE OF CONTENTS

1 DESCRIPTION

This document specifies the preprocessing approach that will be used as part of the automated Quality Control/Quality Assurance (QA/QC) plan of observed instrument data [RD 02]. Specifically, this document outlines how data will be preprocessed, if required, to implement QA/QC processes that will be used to create TIS L1 DPs. Details on whether it is necessary to preprocess data prior to QA/QC and or other analyses will be specified either in a sensor's ATBD in the algorithm implementation section or be explicitly stated in a QA/QC ATBD.

2 RELATED DOCUMENTS AND ACRONYMS

2.1 Reference Documents

RD[01]	NEON.DOC.000008	NEON Acronym List
RD[02]	NEON.DOC.011009	NEON FIU Dataflow and QA Plan
RD[03]	NEON DOORS Requirements Database	

2.2 Acronyms

Acronym	**Explanation**
ATBD	Algorithm Theoretical Basis Document
DAS	Data Acquisition System
DP	Data Product
L0	Level 0
L1	Level 1
QA/QC	Quality Assurance/Quality Control
TIS	Terrestrial Instrument System

2.3 Variables

Variable	**Explanation**
f	Frequency (Hz)
n	Number of observations
t	Time (Seconds)
u_{DAS}	DAS Uncertainty (Seconds)

3 PREPROCESSING

The preprocessing to be applied will be identified in the sensor specific L0' or L1 transition ATBD. The two preprocessing procedures are identified below, creation of a regularized time series based on the sampling frequency of the sensor or a filling of missing timestamps if regularization is not necessary.

3.1 Regularized time series

Many time series analyses, such as frequency response corrections, lag corrections, and some despiking routines, necessitate a complete regularized time series to be performed accurately. Regularization of the time series consists of creating an equally spaced time series for a given dataset. Based on the range of the time period (T) (seconds) and the sensor's sampling frequency (f) (Hz), the number of observations (n) for the dataset can be calculated as follows:

$$n = T * f \quad (1),$$

Using this information a regularized time series can be generated for T by generating the a series with

$$t_{reg} = t_0 + \Delta t_{reg} * i \quad (2),$$

Where: t_{reg} = Regularized time series.

t_0 = Time of the predetermined initial timestamp for the dataset.

Δt_{reg} = Time interval between consecutive observations in the regularized time series ($\Delta t_{reg} = \frac{1}{f}$).

i = index operator ($i = 1, 2, \dots, n$)

After t_{reg} has been generated, the observed time series (t_{obs}) can be attributed to t_{reg} by binning the observations using a windowing function. Equations 3-5 define three windowing functions, centered, leading, and trailing, that will be used to bin the observations.

$$t_{reg} - \frac{\Delta t_{reg}}{2} < t_{obs} \leq t_{reg} + \frac{\Delta t_{reg}}{2} \quad (3)$$

$$t_{reg} - \Delta t_{reg} < t_{obs} \leq t_{reg} \quad (4)$$

$$t_{reg} < t_{obs} \leq t_{reg} + \Delta t_{reg} \quad (5)$$

The default windowing function will be the centered function described in Eq. (3); however, the specific windowing function to be applied may be specified in the sensor specific L0' or L1 transition ATBD.

Once the windowing function has been applied, the generated bins for each timestamp, t_{reg}, will be filled with either zero, one, or multiple observations. The value of expected datum ($x_{t_{reg}}$) at time t_{reg}

will be attributed as NA if no values fall within the bin. If a single value falls within the bin it will be the imputed $x_{t_{reg}}$. If multiple observations fall into a bin, $x_{t_{reg}}$ may be filled in one of three ways:

1. The closest t_{obs} to t_{reg} by minimum absolute deviation will be used to attribute $x_{t_{reg}}$
2. The first t_{obs} in the bin will be used to attribute $x_{t_{reg}}$
3. The last t_{obs} in the bin will be used to attribute $x_{t_{reg}}$

The default methodology to impute $x_{t_{reg}}$ will be to impute closest t_{obs} to t_{reg} by minimum absolute deviation (1. above); however, the methodology may be specified explicitly in the sensor specific L0' or L1 transition ATBD.

3.2 Filling missing timestamps

Only actual sensor observations will be output by the DAS. Therefore, any interruption in the data stream that results in missing observations is captured as a jump between sample timestamps. However, the QA/QC algorithms used to process TIS sensor data require that all data is complete with respect to a sensor's sampling frequency, f. Therefore, the number of samples, n, for a time period should be calculated following Eq. (1). In order to identify missing observation times, an observation's timestamp will first be compared to the timestamp of the next consecutive observation:

$$\Delta t = t_{x+1} - t_x \quad (7)$$

Where: Δt = Time interval between consecutive observations in the time series.
t_x = Time of the observation being assessed.
t_{x+1} = Time of the next consecutive observation in the time series.

Next, the difference that exists between the two timestamps is compared to the sampling frequency of the sensor as well as the timestamp uncertainty associated with data acquisition system (DAS), u_{DAS}. If $\Delta t \leq \frac{1}{f} + 2 * |u_{DAS}|$, then preprocessing is complete for the observation at t_x and preprocessing will progress to the next observation in the time series. If $\Delta t > \frac{1}{f} + 2 * |u_{DAS}|$, then Eq. (8) will be used to determine the number of number of missing samples (n) for a given time period (Δt).

Note: Time units for Δt and $\frac{1}{f}$ must be the same and information on the uncertainty associated with different sampling frequencies can be found in RD[03].

$$n = (\Delta t * f) - 1 \quad (8)$$

Where: n = Number of missing samples for the time interval being assessed.
Δt = Time interval between consecutive samples in the time series.

f = Nominal sample frequency

The results from Eq. (8) indicate the number of time-value pairs, where the value is NA and the time corresponds to the timestamp of the **expected** datum, that need to be inserted between t_x and t_{x+1}. Time-value pairs will be inserted starting with the next timestamp, according to the sensor's nominal sampling frequency, after t_x (i.e. $t_{x+1} = t_x + 1/f$). This will make the time series complete with respect to the sampling frequency of the sensor. The number of time-value pairs inserted in the time series, n, will always be *rounded half up* to the nearest integer. Once the time series has been preprocessed, it will proceed to the next step of the algorithm implementation process as specified in the sensor-specific ATBD.