Application of Running-Median Filtering for Ambient Meteorological and Eddy-Covariance Data

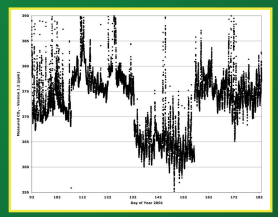
John Frank and Bill Massman, Rocky Mountain Research Station, Fort Collins, CO

The GLEES AmeriFlux Scaffold (2004-Present).

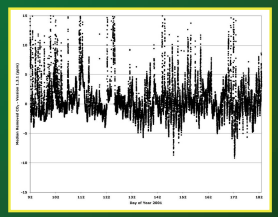
The GLEES-AmeriFlux Site

The GLEES-AmeriFlux site is located in the Snowy Range Mountains of the Medicine Bow National Forest, southeastern Wyoming. At 3140 m (10,300 ft), it is the highest elevation flux tower in the United States and one of the highest in the world. It measures energy and trace gas exchange between the Engelmann spruce/subalpine fir forest and the surrounding atmosphere. These measurements have been made at the site since November 1999.

The Snowy Range is notorious for severe weather, and it is a challenging environment in which to take micro-meteorological measurements. The quality assurance/quality control (QA/QC) processing of data can be difficult when data streams are corrupted with erroneous spikes, drifts, and offsets. A running median filter can be extremely useful for cleaning bad data out of contaminated data streams.



Original ambient CO_2 concentration measurement.

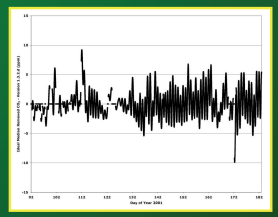


Ambient CO_2 concentration with 10-day running median removed.

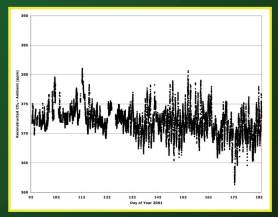
Ambient Meteorological Data

At times, ambient CO_2 concentration measurements were erroneously contaminated by sensor and output voltage drifts, offsets, and spikes due to bad weather, ground loops, lack of sensor maintenance, and spurious electrical activity. These problems were corrected by:

- Removing the 10-day running median
- Manually removing all sections of bad data longer than 0.44 day.



Ideal ambient CO_2 concentration used to identify bad data. The *ideal* was estimated from the 0.87-day running median and Fourier transform reconstruction.

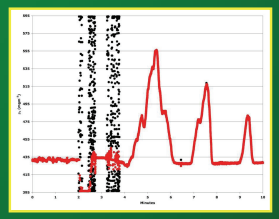


The reconstructed ambient CO_2 concentration includes the annually increasing global CO_2 average and the site specific annual and semi-annual sinusoidal variations.

- Estimating an *ideal* ambient CO_2 concentration by taking the 0.87-day running median of the data stream, transforming it to frequency domain, applying the approximate inverse 0.87-day running-median transfer function, and transforming it back to time domain (Note, this removed all frequencies > 1 day)
- Calculating the residuals between the real and *ideal* data streams
- Estimating an *ideal* σ for residuals from the 10-day running-median of the residuals.
- Removing all data with $p > 0.05$ based on their residual and *ideal* σ
- Adding lower-frequency terms for the annually increasing global CO_2 average along with the site specific annual and semi-annual sinusoidal variations.

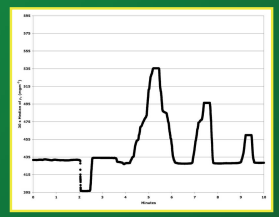
Eddy-Covariance Data

Energy and trace gas exchange were measured at the GLEES-AmeriFlux site using the eddy-covariance technique. Fast-response instruments were used to measure 3-dimensional turbulence, air temperature, water vapor and CO_2 density, and pressure. Most of theses were open-path instruments and were extremely susceptible to spiking and drifting due to rain, snow, ice, and dust. The eddy-covariance data streams were almost overwhelming (36,000 data points per sensor every 30 minutes) which made QA/QC processing even more of a computational challenge.

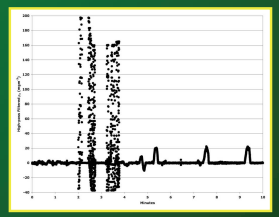


Fast-response CO_2 density during calibration to illustrate stability of the despiker. Data in black have been identified as spikes and removed.

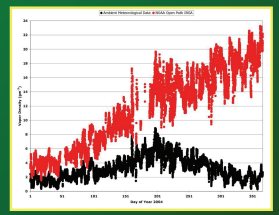
A common despiking algorithm for fast-response turbulence data (Højstrup 1993) was used on the data streams. This despiker continuously calculated the state statistics of a data stream (mean, variance, and correlation) and compared each data point to a forecast based mainly on the correlation to the previous data point. Data was removed when the difference from the forecast exceeded a threshold based on the variance. In some circumstances, though, this original despiker became unstable with the GLEES data. To improve the stability the 30-s running median was removed to high-pass filter all data before despiking, a second forecast and comparison was added based on the mean, and the threshold based on variance was capped by an empirical model for the variance of each measurement.



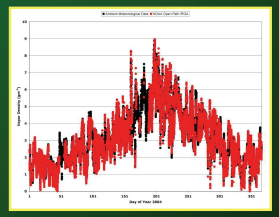
30-s running median of the fast-response CO_2 density.



High-pass filtered fast-response CO_2 density used for despiking.

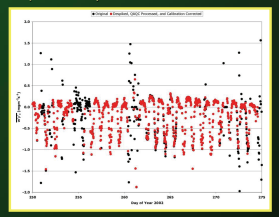


Despiked and QA/QC processed fast-response vapor density without calibration correction.



Despiked, QA/QC processed, and calibration corrected fast-response vapor density.

The despiked fast-response data streams were summarized for each half-hour and QA/QC processed and calibration corrected relative to the ambient meteorological data. Calibration correction was done by removing the 10-day running median from both data sets, multiplying by the coefficient of the 183-day running slope-only regression between both data sets, and adding the 10-day running median of the ambient metrological data. These data were processed to calculate the covariances only. Note, the covariances have not received WPL and other corrections, and therefore are not equivalent to fluxes. That step is yet to come...



Despiked, QA/QC processed, and calibration corrected 30-minute LI-7500 CO_2 covariance data. Original data are in black.

References: Højstrup, J. 1993. A statistical data screening procedure. Measurement Science and Technology. 4:153-157.

- Presented on October 17-19, 2007 at the AmeriFlux meeting in Boulder, CO.