Reviewer #2

We thank Reviewer 2 for their comments and suggestions. The point-by-point responses are provided below, where comments from reviewers are in black, responses are in blue, and new text added in manuscript is in bold blue.

Rawat et all in their manuscript "Maximizing the Use of Pandora Data for Scientific Applications" present a methodology to increase the amount of "scientifically usable" columnar NO2 and HCHO data from Pandonia Global Network by applying different from PGN standard filtering criteria. The approach consists in using an independent uncertainty (detector photon noise propagated to slant columns) threshold to eliminate poor quality data. This threshold is derived from the independent uncertainty distribution for high-quality flagged data as $\mu + 3\sigma$. The data are further filtered by nrms (> 0.01) and maximum horizontal distance estimation for tropospheric columns (>20 km), and restoring measurements with < 10% relative error. The filtering results are verified by conducting linear regression analysis between different combinations of standard PGN quality flagged tropospheric column vs total column data of NO2 and HCHO. The main assumption is that the data are "scientifically useful" if correlation R2 is consistent for various flag combination of tropospheric column (scattered sky) vs direct sun measurements after filtering.

The focus of the paper is to better understand the quality of trace gas column measurements and "recover" PGN data potentially incorrectly labelled as low quality. This is a relevant topic for a publication in AMT considering the importance of PGN for satellite validation and air quality research. However, in current version this paper does not add any new knowledge about the quality of the measurements, and "physical" reasons for accepting more measurements.

We respectfully disagree with the reviewer. Comparing Pandora measurements from DS and SS that are fully independent of each other as well as Pandora measurements to surface in situ ozone that are also fully independent is an ideal way to demonstrate the value that can be gained from filtering Pandora data with our method. The independence of these measurements is fundamental 'physical evidence' of our approach.

We now better address the physical causes of high data loss in the standard PGN flags as overly stringent atmospheric variability and wrms thresholds which is fully explained in the new supplemental section 1.3.

Major comments:

The main assumption that the data are "scientifically useful" if linear correlation R^2 is consistent for various flag combination of tropospheric column from scattered sky vs total columns from direct sun measurements is not totally proven. While I agree that they have separate analysis "paths" they do not have to be correlated (e.g. sampling different air masses due to difference in observation geometries) and they can be correlated for wrong reasons (e.g effect of clouds and aerosols, observation geometry). Actually, the only times they could be correlated are under totally cloud free, homogeneous conditions and perfect instrument performance – the high-quality flagged data.

If two methods are measuring the same quantity in quick succession, the expectation of autocorrelation is fundamental. Therefore, these R^2 correlations are an indication that the measurements are valid.

The authors need to show how the parameter subset that goes into quality flag determination changes because of their filtering to convince that the resulting data are scientifically acceptable. There are certain metrics (e.g. wavelength shift) that have less impact on the DOAS fitting and air mass factor quality than others (e.g. clouds). The value of this paper would be to identify such main "drivers" of data quality based on a very detailed evaluation of instrumental and atmospheric uncertainties in PGN data.

This is now addressed in supplementary section S1.3, which was also requested by reviewer 1. The outcome that atmospheric variability (e.g., clouds) have no apparent bearing on data quality is a tremendous significance and raises the question of whether the atmospheric variability parameter is related to clouds at all.

In general, poor quality in direct sun DOAS fitting results can rise from instrumental problems (tracker pointing issues, coherent light interference, internal stray light, filter wheel issues,

spectrometer changes leading to wavelength and slit function drifts, etc, inaccurate location or time) and atmospheric (presence of the clouds, leading to change in photon path and spectral saturation, spatial stray light). Poor quality in scattered sky data is mainly due to presence of cumulus clouds at the higher scan angles, pointing at obstructions, presence of clouds in the reference spectrum and pointing close to the sun, changes in scattering conditions between the scan measurements etc. There are two parameters that reflect the data quality to the first order: nrms of the DOAS fitting residuals and relative column error. Nrms is instrument and fitting window dependent and thresholds can be determined from the fitting data. Also, nrms of 0.01 is a very large value for typical trace gas DOAS fits to be valid.

Our maximum limit of wrms < 0.01 was set empirically and our correlation between SS and DS indicates that these points are reasonable.

A lot of examples were provided on the data from Houston, Texas metropolitan area, a near coastal region with relatively high presence of partly cloudy conditions. The presented results of the proposed filtering suggest that more than 90% of scattered sky data were not impacted by clouds. This might be overly optimistic.

Sky-scan observations are almost never triggered to low or medium quality by atmospheric variability in the standard PGN flags (Table S4).

Direct sun HCHO depends on spectrometer stray light properties and/or some other potential optical interferences. As a result, caution should be taken when interpreting the measurements. For example, collocated instruments often will not produce the same HCHO total columns.

We appreciate the need for caution in interpreting the HCHO DS measurements. However, if the measurements are flawed, there should be no expectation of a correlation with surface in situ ozone. This relationship demands that we reassess the quality of the DS HCHO data. Nevertheless, we have discussed the residual stray light as a function of SZA in Figure S1, which showed it decreased or remained constant (~ 0.3%) with increasing SZA, which suggests stray light contribution to the column might not increase significantly at higher SZA. Also, differences between DS and SS in section 3.3 are analyzed along difference SZA, along different azimuth viewing in Figure S3 and for different seasons in Figure S2 and no specific behavior is observed.

It appears that there are interpolation errors in figures 8 and 9: constant Y values for changing X values.

There is no interpolation used in the present analysis. However, in figures 8 and 9 we see changing X for constant Y, as we match all the DS observations with the nearby SS observation within 5min.

To make it clearer now we have added a line in revised manuscript at page 9, line 226.

This enables an independent assessment of data quality by **taking advantage of the expected autocorrelation** of contemporaneous (within 5 min) DS and SS observations for different quality flag combinations.

I find the idea of using correlation improvement between column (HCHO) and surface (O3) measurements is a weak argument for selecting data quality of the measurements. The goal should be to derive this (surface to column) dependence based on the best quality data and not force it through selection of the data. Multiple studies have shown that column to surface ratios depend on a number of meteorological, emission and photochemistry conditions, so using it as an argument in favor of the new filtering might not convince the broader scientific community.

We respectfully disagree. Improved correlations do not happen by accident. Our filtering process clearly shows that there is improvement in the correlation between two entirely independent observables. Please explain how poor quality data could result in such an outcome.

GCAS measurements depend on surface reflectance, aerosol and trace gas profiles and need their own validation. These measurements are typically considered less accurate than ground-based measurements. Again, using them as a verification tool seems not appropriate.

We again respectfully disagree. Not only do we show correspondence between GCAS and Pandora, we show that the discrepancies between the two are entirely explainable. Comparing quantities attempting to measure the same variable with an eye for why they should or should not match is fundamental science. Good correspondence is unlikely to be accidental for two independent measurements.

It is important to note that we are not the first group to compare GCAS observations with another observational perspective. We have added a line in the revised manuscript at page 23, lines 458-460.

Additionally, the GCAS tropospheric column measurements have been shown to strongly correlate with in-situ aircraft observations for NO_2 (R^2=0.89) and HCHO (R^2=0.54), with columns differences in magnitude within 10% (Nowlan et al., 2018).

Missing detailed description of the standard PGN data flags and parameters that go into their determination

We have added the following paragraphs for data flagging in revised manuscript in page 8, lines 196-206 and detail Pandora data flagging propagation in supplementary section S1.3:

These quality flags are determined in various stages from L1 (raw data) to L2fit (spectral fit data) to L2 (processed data) for ensuring data quality through multiple checks at each stage. At the L1 stage, data is flagged into either low- or medium-quality based on instrument-related issues such as excessive dark counts, detector saturation, dark count differs significantly from the dark map for too many pixels, different effective temperature, and unsuccessful dark background fitting. In the L2fit stage, where spectral fitting is performed, data is further flagged based on factors including the quality of the fit, the wrms limit (normalized rms of fitting residuals weighted with independent uncertainty), and wavelength shift. Finally, at the final retrieval L2 stage, factors including retrieval error and atmospheric variability are used to categorize data into the medium or low-quality.

Need to provide more details for Figure S1: wavelengths, how this residual stray light was determined, etc

We discussed the residual stray light determination in the manuscript at page 27, line 543. We are not sure what else is needed.