OCO-2 v9 10-second-averaged X_{CO_2} Measurement Files

David Baker

16 Jan 2020

The OCO-2 Version 9 "Lite" files provide bias-corrected versions of the Level 2 raw X_{CO_2} retrievals found in the Standard product; additional quality filtering is performed as part of this calculation. We condense these corrected X_{CO_2} values into summary measurements, one every 10 seconds (or ∼67.5 km along-track), at a resolution more appropriate for the atmospheric transport models used in global-scale CO_2 flux inversions, which run with grid boxes 100s of km wide. These summary measurements can then be used more efficiently in the inversions, reducing the computations and I/O required while still capturing the relevent information contained in the original individual retrievals. We account for correlations between the individual retrievals when calculating the uncertainty assigned to the 10-second average X_{CO_2}. The 10-second grouped data described here could be assimilated as independent measurements if correlations at scales coarser than ∼67.5 km are neglected, or further correlations between 10-second averages could be modeled and accounted for in the inversions if not (we have not attempted to calculate such larger-scale correlations here). The choice of averaging interval could be changed from the 10-second value used here for problems with different goals and better knowledge of correlations.

The 10-second summary X_{CO_2} measurements, surface pressures, averaging kernels, and prior CO_2 profiles have been packaged in a netCDF file having a format similar to the "Lite" files produced by Chris O'Dell for the individual retrievals, although now there is a single file for the full span of the data rather than individual files for each day. In addition, a subset of the auxiliary parameters included in the "Lite" files are included here: these could be useful for a variety of reasons, for example allowing a check to be done of the X_{CO_2} bias correction at this 10-second resolution. The method for calculating the 10-second summary values for the X_{CO_2}, X_{CO_2} uncertainty, averaging kernel, and prior CO_2 profile needed in the inversions, as well as the extra parameters, is described below.

1 Notation:

X_{CO_2} – the pressure-weighted dry air CO_2 mixing ratio column average
σ_X – the uncertainty in X_{CO_2}
$\mathbf{a}_X$ – the averaging kernel vector associated with the retrieval of X_{CO_2}
v – any of various parameters associated with the retrieval of X_{CO_2} (e.g., aerosol optical depth, surface albedo, etc.)
J – the number of "good" X_{CO_2} retrievals used in the 10-second average

The names of variables provided in the 10-second average netCDF file are indicated in *bold-faced italic font*.

2 OCO-2 measurement averaging approach

A 10-second-average X_{CO_2} value is computed that is meant to capture the information content of the individual X_{CO_2} retrievals at scales relevent to global inversion models. An information-weighted average is used for all quantities, where the measurement "information" is taken as the inverse of the square of the X_{CO_2} uncertainty calculated in the retrieval, σ_X^{-2} (σ_X coming from variable *xco2_uncertainty* in the "Lite" files). Correlations between errors in the individual X_{CO_2} values are not accounted for in specifying the form of these weighted averages, but they are in determining the uncertainty placed on the final 10-second X_{CO_2} average. Finally, the "theoretical" single-shot uncertainties calculated by the retrieval are generally thought to be too low, since they do not capture the effect of certain systematic errors in the retrievals: we calculate an additional error term based on the spread of the retrieved X_{CO_2} values across the 10-second span that is added onto the "theoretical" errors in calculating the 10-second-average X_{CO_2} uncertainty.

2.1 Defining the weighted averages

OCO-2 makes three cross-scans per second, each of which yields measurements in eight separate fields of view; a maximum of 240 measurements per 10-second span are thus possible, but not all of these produce reliable retrievals due to clouds, high aerosol optical depths, or other problems that prevent the scene from passing the quality filters (*xco2_quality_flag*=0 indicates a "good" scene). Suppose we average up to 240 "good" measurements j in each 10-second span k as follows:

$$v_k = \sigma^{+2}_{X_k,uncorr} \sum_{j=1}^{J} \sigma^{-2}_{X_j} \; v_j \tag{1}$$

$$X_k = \sigma^{+2}_{X_k,uncorr} \sum_{j=1}^{J} \sigma^{-2}_{X_j} \; X_{CO_{2j}} \tag{2}$$

$$\mathbf{a}_{X_k} = \sigma^{+2}_{X_k,uncorr} \sum_{j=1}^{J} \sigma^{-2}_{X_j} \; \mathbf{a}_{X_j} \tag{3}$$

$$\sigma^{-2}_{X_k,uncorr} = \sum_{j=1}^{J} \sigma^{-2}_{X_j} \tag{4}$$

where σ_{X_j} is the uncertainty in $X_{CO_{2j}}$ calculated by the retrieval and output in the "Lite" files (in *xco2_uncertainty*). Assimilating the summary value X_k with

an assumed uncertainty of $\sigma_{X_k,uncorr}$ would have the same effect as assimilating each retrieval $X_{CO_{2_j}}$ separately with uncertainty σ_{X_j}, assuming that each is independent of the other. [This is at least true at coarser scales at which the timing of each individual retrieval provides less information.]

We recognize, though, that the error in each individual retrieval is not independent from those nearby it in space and time. We initially derived an averaging strategy that tried to account for correlated errors when specifying the weights in the averages (see Appendix B). This approach had the undesirable feature of producing average values that often fell outside of the range of the input data, however. To avoid this, we decided to stay with the straight (uncorrelated) information averages specified in equations (1)-(3), and to account for error correlations only in the calculation of the uncertainty on X_{CO_2}.

2.2 Accounting for correlated errors in the X_{CO_2} uncertainty calculation

Across a 10-second span, errors in the individual retrievals $X_{CO_{2_j}}$ are likely to be positively correlated with each other due to radiative transfer modeling errors in the retrieval driven by changing aerosol distributions and changing surface conditions, for example. If we ignore these correlations and assume that the individual retrievals $X_{CO_{2_j}}$ are all independent, and assign an uncertainty of $\sigma_{X_k,uncorr}$ to X_k, we would give X_k too much weight in our flux inversion. Here we try to model the correlations and to come up with a more accurate uncertainty on X_k, instead. If we assume that the correlations, c, are constant across the 10-second span, then it may be shown (Appendix A) that the uncertainty in X_k (as defined in equation (2)) is given as:

$$\sigma^2_{X_k,corr} = \frac{1 - c + c\frac{(\sum \sigma_{X_j}^{-1})^2}{\sum \sigma_{X_j}^{-2}}}{\sum \sigma_{X_j}^{-2}} \tag{5}$$

Based on unpublished work done by Susan Kulawik, we use the following OCO-2 measurement correlation values:

$$c = \begin{cases} +0.3 & \text{over land} \\ +0.6 & \text{over water} \\ +0.6 & \text{for mixed land/water} \quad (data\_type = 9) \end{cases} \tag{6}$$

2.3 Inflating the theoretical X_{CO_2} uncertainties

Several previous studies have shown that the theoretical $X_{CO_{2_j}}$ uncertainties computed by the retrieval are too low (Kulawik *et al.*, 2019) and ought to be inflated when assimilating X_{CO_2} in flux inversions. Separating errors that may plausibly be treated as "random" from systematic across a 10-second span is challenging; here we are addressing the part that varies enough to be thought of as random. Kulawik *et al.* (2019) have quantified random errors in bias-corrected X_{CO_2} as being on the order of 0.9 and 0.5 ppm for data taken over

land and ocean (calculated as the portion of the error that can be reduced through averaging). Our form of such inflated errors will then be fed into the correlated error averaging equation (5) to get an improved uncertainty for the 10-second-averaged X_{CO_2} value.

An information-averaged error, $\sigma_{X_k,avg}$, on any single retrieval $X_{CO_2 j}$ in the 10-second span k may be calculated from the individual uncertainties produced by the retrieval as:

$$\sigma^{-2}{}_{X_k,avg} = \frac{1}{J}\sum_{j=1}^{J}\sigma_{X_j}^{-2} \tag{7}$$

This can be thought of as a sort of "theoretical" error or uncertainty on X_{CO_2} given by the linearized errors σ_{X_j} derived by the retrieval's covariance matrix. The subscript k in $\sigma_{X_k,avg}$ refers to the span across which the average was performed; we should keep in mind that this is really an average error on any individual retrieval j, however, and not an error associated with a summary X_{CO_2} value across the span.

For those 10-second spans with more than one retrieval, the weighted standard deviation s_{v_k} across 10-second span k of any scalar quantity v may be calculated as:

$$s_{v_k}^2 = \frac{\frac{1}{J-1}\sum_{j=1}^{J}\frac{(v_j-\bar{v}_j)^2}{\sigma_{X_j}^2}}{\frac{1}{J}\sum_{j=1}^{J}\frac{1}{\sigma_{X_j}^2}} = \frac{\sigma_{X_k,avg}^2}{J-1}\left(\sum_{j=1}^{J}\frac{v_j^2}{\sigma_{X_j}^2} - \frac{\sigma_{X_k,avg}^2}{J}\left(\sum_{j=1}^{J}\frac{v_j}{\sigma_{X_j}^2}\right)^2\right) \tag{8}$$

where again the weights are specified as the inverse variance of the X_{CO_2} errors. Also, we should remember again the subscript k just indicates the span averaged over; s_{v_k} itself refers to the variability in parameter v for an individual scene, not in some average of that parameter across the span. The standard deviation s_{X_k} of $X_{CO_2 j}$, in particular, may then be defined as

$$s_{X_k}^2 = \frac{\sigma_{X_k,avg}^2}{J-1}\left(\sum_{j=1}^{J}\frac{X_{CO_2}{}_j^2}{\sigma_{X_j}^2} - \frac{\sigma_{X_k,avg}^2}{J}\left(\sum_{j=1}^{J}\frac{X_{CO_2 j}}{\sigma_{X_j}^2}\right)^2\right) \tag{9}$$

We use the raw (non-bias-corrected) X_{CO_2} values (*xco2_raw* from the "Lite" files) in this calculation; the standard deviation of the bias-corrected values is lower, especially over land, but we prefer to stay with the un-bias-corrected values to avoid having to assess the impact of the bias correction.

Figure 1 shows the distribution of s_{X_k} and $\sigma_{X_k,avg}$ side by side for five years of OCO-2 data as a function of viewing mode and the number of good retrievals across each 10-second span. While $\sigma_{X_k,avg}$ generally does a good job of approximating the actual errors s_{X_k} for data taken over the ocean, it underestimates then by a factor of two or more over land. The actual sampled errors tend to be higher for 10-second spans with fewer "good" scenes (J small) and lower for spans with more "good" scenes (J close to 240). This would make sense if, for example, the number of "good" scenes is inversely correlated with

how cloudy the span is, and if systematic retrieval errors are resulting from the impact of undetected clouds in the retrieval.

The random errors calculated by Kulawik *et al.* (2019) are somewhat lower than what we calculate on the top row of Figure 1 for land and about the same for ocean: the lower errors over land can be explained by the fact that we have chosen to examine the variability in raw instead of bias-corrected X_{CO_2}. In any case, these random errors are all larger than those given by the retrieval and some inflation approach is necessary. Instead of using some sort of global inflation factor, we choose here to use the actual spread in X_{CO_2} across each 10-second span, given by s_{X_k}, to do this inflation in a more local manner: this will increase the uncertainty more for cloudy scenes over land, and less for uncloudy conditions over the ocean, for example. This is done in an *ad hoc* manner, adding the sampled spread onto the theoretical spread in quadrature:

$$\sigma^2_{X_k} = \frac{1-c+c\frac{(\sum_j \sigma^{-1}_{X_j})^2}{\sum_j \sigma^{-2}_{X_j}}}{\sum_j \sigma^{-2}_{X_j}} + \frac{1-c+c\frac{(\sum_j s^{-1}_{X_k})^2}{\sum_j s^{-2}_{X_k}}}{\sum_j s^{-2}_{X_k}} \quad (10)$$

$$= \sigma^2_{X_k,avg}\left(\frac{1-c}{J} + c\frac{(\frac{1}{J}\sum_j \sigma^{-1}_{X_j})^2}{\sigma^{-2}_{X_k,avg}}\right) + s^2_{X_k}(c+(1-c)/J) \quad (11)$$

$$= c(s^2_{X_k} + \sigma^4_{X_k,avg}(\frac{1}{J}\sum_j \sigma^{-1}_{X_j})^2) + \frac{1-c}{J}(\sigma^2_{X_k,avg} + s^2_{X_k}) \quad (12)$$

Both the theoretical and sampled single-retrieval errors are assumed to have the same positive correlations discussed in Section 2.2, and the impact of both on the 10-second average are calculated with the same correlated error equation (5), then added in quadrature to calculate a 10-second summary uncertainty. It can be shown that this gives a similar result to adding the two errors together in quadrature first, then feeding the combined error through the correlated error equation (5): the result is exactly the same in the limit of when the uncertainties for all the individual shots are the same. For spans in which s_{X_k} is much larger than $\sigma_{X_k,avg}$, this gives an uncertainty of about s_{X_k}. For scenes in which the sampled variability s_{X_k} underestimates the actual variability, as happens occasionally when J is small, the theoretical uncertainty $\sigma_{X_k,avg}$ provides a floor. The inflated uncertainty, σ_{X_k}, for the 10-second X_{CO_2} average is placed in variable *xco2_uncertainty* in the netCDF file.

2.4 Model error

When the OCO-2 data are assimilated in flux inversion models, we suggest that an additional error related to the transport model itself be added to the measurement error:

$$\sigma^2_{k,assim} = \sigma^2_{X_k} + \sigma^2_{k,model} \quad (13)$$

While $\sigma_{k,model}$ should ideally be calculated by each modeler to capture the pecidillos of their own individual transport model, we provide an example of

this quantity in variable *model_error*. The value in *model_error* was calculated from the difference of the GEOS-Chem and TM5 modeled CO_2, with annual-mean biases subtracted off.

3 Data Selection

Each 10-sec average value is assigned a sounding identification number, *sounding_id*, with format YYYYMMDDHHMMSo, where YYYY=year, MM=month, DD=day, HH=hour (00-23), MM=minute (00-59), S=10-sec range (0-5), and "o"=a data type flag. Summary measurements with the 10-sec range given by the "S" field (0-5) are computed from measurements with the seconds variable (*date*(6)) from the "Lite" file in the ranges 00-09, 10-19, 20-29, 30-39, 40-49, and 50-59. The data selection flag "o" is computed as in Table 1; it is also output in the *data_type* variable for convenience. The data falling into each *data_type* category are summed seperately inside each 10-second span, with separate values being output to the netCDF files.

10-sec avg *data_type*	"Lite" file variables				10-sec avg mode descriptor
	xco2_quality_flag	*operation_mode*	*surface_type*	*land_fraction*	
1	0	0	1	80-100	land nadir
2	0	1	1	80-100	land glint
3	0	2	1	80-100	land target
4	0	3	1	80-100	land transition
5	0	0	0	0-20	water nadir
6	0	1	0	0-20	water glint
7	0	2	0	0-20	water target
8	0	3	0	0-20	water transition
9	0	all combinations		20-80	mixed land/water

Table 1: How the *data_type* variable used in the 10-second average files is defined, based on variables in the "Lite" files.

Only "good" data (those with L2 flag *xco2_quality_flag*=0) are included in the averages. Data over the ocean (*land_water_indicator*=0) and over inland water (*land_water_indicator*=2) are lumped together in the water (*surface_type*=0) category. Data with mixed water and land scenes (*land_fraction*=20-80%) are relegated to *data_type*=9. Restricting the *land_fraction* range to 80-100% for land retrievals and 0-20% for water retrievals is normally done as part of the Level 2 processing as part of determining which surface physics mode to use for the retrieval. Here, we have added this check to our processing, since the pointing correction done as part of the v9 processing has resulted in some of the scenes that fell into these ranges in the v8 processing to now be in the middle

20-80% range (that is, no new data rejection was done based on this flag in the v9 processing): we reject these scenes here to stay true to the original criteria.

For the netCDF 10-second average file to be used in the v9 OCO-2 MIP study ("OCO2_b91_10sec_GOOD_r24.nc4"), only *data_type* values 1, 2, and 6, and averages with $J \geq 10$ have been included in the file, since the other data are not used.

Occasionally, files are made with "bad" data (those with L2 flag *xco2_quality_flag*=1) also included. In this case, the "bad" data is averaged together separately from the "good" data, and the *sounding_id* variable is no longer a unique identifier: the *xco2_quality_flag* variable must also be referenced to distinguish the "good" from the "bad" data.

4 Recipe for using the 10-second summary measurements in the inversion inter-comparison experiments

1. Obtain the 10-second summary files via anonymous ftp:

   ftp ftp.cira.colostate.edu

   cd ftp/BAKER

   get OCO2_b91_10sec_GOOD_r24.nc4

2. Use the summary X_{CO_2} measurements, X_k (variable *xco2*), computed across the 10 second spans ($\sim$67.5 km along-track swaths) in place of the individual retrievals.

3. Assimilate the 10-second summary X_{CO_2} values, sampling CO_2 in the model with the 10-second summary averaging kernels (*xco2_averaging_kernel*) and prior CO_2 profiles (*co2_profile_apriori*) provided in the file, at the time, latitude and longitude provided.

4. Assimilate each 10-second grouped measurement as if it were independent of all the rest.

5. Do NOT add additional uncertainty to account for systematic errors to the 10-second summary measurements (these are built into the 10-second uncertainties given by (14)).

Note: To choose only certain types of OCO-2 data (e.g. for use in the OCO-2 MIP experiments), use the *data_type* variable (or the last digit of *sounding_id*): 1=land nadir, 2=land glint, 3=land target, 4=land transition, 5=water nadir, 6=water glint, 7=water target, 8=water transition, or 9=mixed land/water. If there are multiple types of measurements in each 10 second span, these will be grouped into separate 10-second summary measurements: note that these different measurements may not be in time order inside of each 10-second segment.

5 References

Kulawik *et al.*, Validation of OCO-2 and ACOS-GOSAT using HIPPO, TCCON, and surface sites, (AGU poster), 2019.

6 Appendix A: Computation of the uncertainty on 10-second-averaged X_{CO_2} considering error correlations for an uncorrelated information-weighted average

If we form a vector $\mathbf{x} \equiv [X_{CO_2\,1}, X_{CO_2\,2}, \ldots, X_{CO_2\,J}]^T$ of J retrieved X_{CO_2} values to be averaged in our 10-second span k, then their weighted average is calculated as:

$$X_k = \mathbf{w}^T\mathbf{x}/\mathbf{w}^T\mathbf{1} = \frac{\sum_{j=1}^{J} w_j X_{CO_2\,j}}{\sum_{j=1}^{J} w_j} \tag{14}$$

with $\mathbf{1}$ being a vector of ones. Choosing $w_j = \sigma_{X_j}^{-2}$, where σ_{X_j} is the uncertainty in $X_{CO_2\,j}$ calculated by the retrieval, makes it an information-weighted average. If we allow there to be correlations between the errors of the individual elements of $\mathbf{x}$, then the uncertainty in X_k is given by

$$\sigma_{X_k}^2 = E[\mathbf{w}^T d\mathbf{x} d\mathbf{x}^T \mathbf{w}]/(\mathbf{w}^T\mathbf{1})^2 = [\mathbf{w}^T[\mathbf{S}^T\mathbf{C}\mathbf{S}]\mathbf{w}]/(\mathbf{w}^T\mathbf{1})^2 = [(\mathbf{S}\mathbf{w})^T\mathbf{C}(\mathbf{S}\mathbf{w})]/(\mathbf{w}^T\mathbf{1})^2 \tag{15}$$

where the covariance matrix for errors in $\mathbf{x}$, $\mathbf{P} = \mathbf{S}^T\mathbf{C}\mathbf{S}$ is specified in terms of a correlation matrix $\mathbf{C}$ and diagonal matrix $\mathbf{S}$ with $[\sigma_{X_1}, \sigma_{X_2}, \ldots, \sigma_{X_J}]^T$ on the main diagonal. Vector $\mathbf{S}\mathbf{w}$ is then given as $\mathbf{S}\mathbf{w} = [\sigma_{X_1}^{-1}, \sigma_{X_2}^{-1}, \ldots, \sigma_{X_J}^{-1}]^T$. Suppose we further assume that the errors in X_{CO_2} for all the individual retrievals j in the span are correlated with each other with the same positive correlation coefficent c, so that the correlation matrix $\mathbf{C}$ may be specified as

$$\mathbf{C} = \begin{bmatrix} 1 & c & \ldots & c \\ c & 1 & \ldots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \ldots & 1 \end{bmatrix} \tag{16}$$

Then the uncertainty in X_k, $\sigma_{X_k,corr}$, is then given by

$$\sigma^2_{X_k,corr} = [\sigma^{-1}_{X_1}, \sigma^{-1}_{X_2}, \ldots, \sigma^{-1}_{X_J}] \begin{bmatrix} 1 & c & \ldots & c \\ c & 1 & \ldots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \ldots & 1 \end{bmatrix} \begin{bmatrix} \sigma^{-1}_{X_1} \\ \sigma^{-1}_{X_2} \\ \vdots \\ \sigma^{-1}_{X_J} \end{bmatrix} / (\mathbf{w}^T \mathbf{1})^2 \quad (17)$$

$$= \left[(1-c) \sum_{j=1}^{J} \sigma^{-2}_{X_j} + c \left(\sum_{j=1}^{J} \sigma^{-1}_{X_j} \right)^2 \right] / \left(\sum_{j=1}^{J} \sigma^{-2}_{X_j} \right)^2 \quad (18)$$

$$= \frac{1 - c + c \frac{(\sum \sigma^{-1}_{X_j})^2}{\sum \sigma^{-2}_{X_j}}}{\sum \sigma^{-2}_{X_j}} \quad (19)$$

$$= \sigma^2_{X_k,avg} \left[\frac{1-c}{J} + c \frac{\left(\frac{1}{J} \sum \sigma^{-1}_{X_j} \right)^2}{\sigma^{-2}_{X_k,avg}} \right] \quad (20)$$

where $\sigma^{-2}_{X_k,avg} \equiv \frac{1}{J} \sum_{j=1}^{J} \sigma^{-2}_{X_j}$.

7 Appendix B: Computation of the uncertainty on 10-second-averaged X_{CO_2} considering error correlations for a correlated information-weighted average

[Math for what you get if you do not specify the form of the X_{CO_2} average as in (2), but rather calculate what it should be if correlations are considered in calculating the average. The equation one obtains in this case frequently gives average values that fall outside the range of the inputs; because of this, we choose instead to specify the form of the average with (2) and then consider correlations when computing the error it. The form for the error is different from (5), as well.]

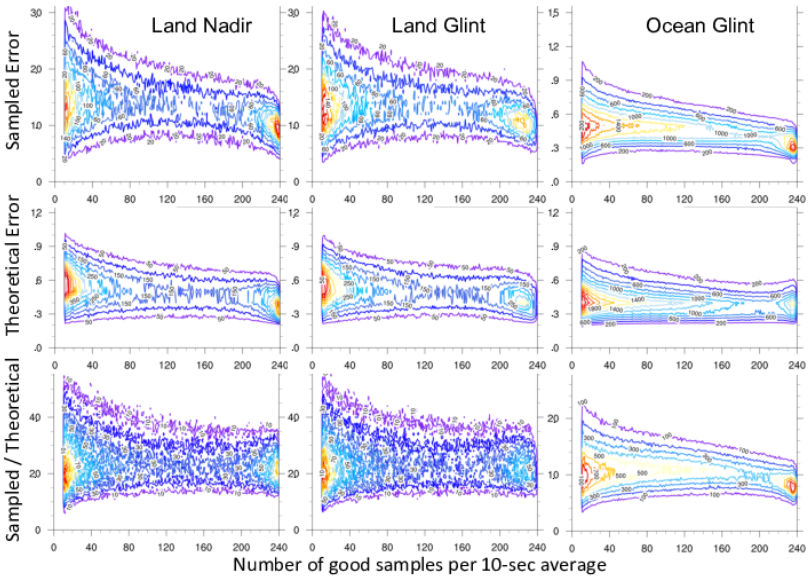

Figure 1: The sampled error in X_{CO_2} computed as the weighted standard deviation of the "good" X_{CO_2} values used in the average for each 10-second span (top row) is compared to its expected value as calculated using the "theoretical" X_{CO_2} uncertainties given by the retrieval (middle row); the ratio of the sampled error over the theoretical error is also given (bottom row). The plots are 2-D histograms representing the frequency that a single 10-second span across the full 2014-2019 OCO-2 data span falls within each range of error magnitude ([ppm], y-axis) and number of "good" shots in the average (x-axis). The theoretical errors tend to represent the actual sampled errors quite well for ocean scenes, but are at least a factor of two too low for land scenes. Scenes tend to be either cloudy (small J, left side of plots) or relatively cloud-free (J close to 240, right-hand side of plots), with fewer cases in between; the nearly-cloud-free scenes are affected by noticeably lower errors than the cloudy scenes, a dip that is only partially represented by the theoretical errors.