Summary for the Synthetic Data Breakout Group

Below is a summary of the main points of discussion.

(1) Level 1 Data

The synthetic data will be used to test both the SDPC pipeline, and validate the accuracy of the operational algorithms. The main conclusion of the breakout group was that these uses may have different requirements, which are discussed below.

1.1 Considerations for testing the SDPC pipeline

Performing a second set of radiative transfer (RT) simulations that will allow easier assessment of operational algorithms
The current simulation has the capability of simulating more realistic scenarios than what are assumed by the operational algorithms (e.g. anisotropic surface reflectivity, scattering clouds, temperature dependent cross sections). This makes it difficult to assess whether differences between the truth and retrieved atmospheric quantities are related to algorithm assumptions of software implementation errors.

Explore implementation of faster numerical methods
The current line-by-line simulations are computationally expensive. Implementation of the fast-PCA VLIDORT model could produce spectra of sufficient accuracy for testing the SDPC, with significantly lower computational burden. Broader use of the fast-PCA model for algorithm validation can be assessed by comparing results from algorithms applied to the fast-PCA spectra and the full line-by-line model.

1.2 Considerations for algorithm validation

Inclusion of o2 and n2 rotational raman scattering (RRS)
The synthetic spectra are generated using VLIDORT, as light polarization must be considered for level 0 to 1 processing. VLIDORT does not include RRS, but an estimate is needed for the RRS-based operational cloud retrieval algorithm. A solution for postprocessing the VLIDORT output to include RRS will be developed, which may be via lookup tables, or simplified RT calculations.

Additional processes
Plant fluorescence was suggested for inclusion in future simulations, to test the solar induced fluorescence retrieval.

There is a need to carefully assess the individual test requirements for each operational algorithm. How many test cases are required? What time resolution.
To supplement a limited number of full scans, a carefully selected additional subset of simulations can be performed. These should be designed to test specific hypotheses on sources of bias in each operational algorithm.

(2) Level 2 Data

A limited number of scans of L2 data will be produced through the full SDPC pipeline due to the computational expense of the full line-by-line simulations. However there are alternate approaches to generating L2 data that avoid simulation of full spectra, that could satisfy the requirements of the L2 user base. There are four potential sets of synthetic L2 data:

1. **Realistic L2**: synthetic observations generated through the SDPC
2. **Truth L2**: the chemical transport model output sampled to the TEMPO observation grid
3. **Truth+Random L2**: Truth L2 with noise added based on the instrument specified precision
4. **Semi-realistic L2**: L2 data including noise and algorithm-specific biases, which are generated by sidestepping full spectra calculations. This may be most appropriate for users that require higher temporal resolution than may be afforded by option (1) (e.g. data assimilation)

Options (2) and (3) require little computational expense, and should satisfy users whose main interest is in visualization.