This paper from Sgheri et al. describes the End-to-End Simulator (E2ES) developed to fully investigate the potentiality of the future Earth Explorer 9 FORUM (Far-infrared Outgoing Radiation Understanding and Monitoring) mission that will be launched in the next years. The work explains in detail the architecture of the E2ES and then investigate the sensitivity of the acquired signal in different scenarios to atmospheric and surface geophysical variables through retrieval tests performed by the L2M_I module inverting the synthetic radiances. The E2ES retrieval module has been validated against the KLIMA and SACR independent codes, finally it has been tested for inhomogeneous scenes and in a real scenario built up from MODIS data.

The work is overall well written and structured, and it will represent reasonably a benchmark for the FORUM preparatory phase during the next years. However, some revisions have to be implemented on the manuscript before its final publication on the AMT journal.

**Major comments:**

The comparison of the retrieved products from L2M_I and KLIMA codes is fully investigated in Section 5.1, however I agree with point 2) of major comments and suggestions from Referee #1. Furthermore, the comparison between L2M_I and SACR codes has been poorly inspected, since relevant differences in the retrieved parameters reported in Table 9 have not been discussed. The cloud parameters retrieved by L2M_I are generally in agreement with true values, but L2M_I and SACR are generally not consistent (e.g. case 2.2, where difference on cloud top amount to 5.6 km). It is not clear at all if the SACR code tends to converge to local minima, however it should be clarified studying the retrieval quality through the chi-square reduction for - at least - two or three critical cases. Therefore, the authors should assess in Section 5.2.2 if this problem is affecting the SACR retrieval, if this may affect the L2M_I retrieval in cloudy sky conditions and how this drawback could be prevented.
Minor comments:

- LINE 92: when presenting FEI, it would be better to explain in detail its role in the detection of the FOV inhomogeneities, that’s not quite clear in this Section.
- LINE 100: is there any references confirming the 96.7% of the OLR being included in the range from 100 to 1600 cm$^{-1}$?
- LINE 179: Looking at the AER database, the gaseous species having absorption lines in the spectral range of FORUM are much more than 12. Which trace gases have been excluded from simulations? Did the authors perform any test to check if their spectral signatures on TOA radiances are negligible?
- LINE 193: Simulations in cloudy sky are generally more time consuming, so I guess that this choice was made to preserve computational efficiency. Did the authors test which is the effect on the high resolution cloudy spectra of a degradation of the spectral resolution from $10^{-4}$ to $10^{-2}$ cm$^{-1}$?
- LINE 200: can the authors specify from which database the 11 pre-defined surface types were taken?
- LINES 242-246: This is a little bit redundant, maybe it would be enough to explain that the safety margin on the FOV was considered to model the self-apodization effect.
- LINE 282: can the authors explain why is the resampling of the average spectrum required for the L2M_I? Which are the spectral resolutions of Level 1b and 1c respectively?
- LINE 339: the meaning of FWHM vertical resolution and how it should be calculated is not quite clear here.
- LINE 342: In my opinion, the first sentence of Section 4 is redundant, it could be removed.
- LINE 355: I suggest to re-write sentence as "we depart from real data using this approximation, however it does not have an impact on the assessment of the FE2ES performances".
- LINE 376: The vertical retrieval grid is considerably different from that used by Ridolfi et al. (2020) for FORUM. Can the authors justify the choice of this grid? Have any tests been performed to get an optimized grid?
- LINE 378: The authors could consider whether compare more in detail the information provided by the industrial consortia to the goal parameters (e.g. insert figure for comparison between G and industrial consortia noise, to show that this last is compliant to the goal).
- LINE 402-403: Are the spectroscopic database and continuum model in KLIMA the same of those used by the L2M_I module? If no, would the authors expect an even better consistency between the retrieved products from the two models if using the same spectroscopic data?
- LINE 417: Can the authors quantify the term “negligible” pointing out a threshold for the difference between retrieval errors provided by KLIMA and L2M_I?
- LINE 445: How is the sensitivity of the retrieval linked to the retrieval error? A reference to the information gain quantifier (Dinelli et al., 2009) would be useful to clarify this point.
- Figure 5 (bottom right panel): How can the authors explain the positive bias in the difference between retrieved and true water vapour above 100 hPa?
- LINE 470: How many emissivity values have been retrieved for the full spectral range and the FIR region only? The authors could consider whether to include (e.g. with a Table) a comparison between the number of retrieved values for the emissivity and the obtained DOFs. It would be useful to clarify the information content on surface emissivity in the different analysed scenario.
- LINE 503-506: It’s not quite clear from the text why the unretrieved atmospheric parameters are perturbed in this test.
- LINE 541: What’s the reduction factor on computational time obtained with the
application of the single layer approximation?

- **LINE 513**: This sentence is contradictory with results shown in Table 9. For example, differences in the retrieved cloud top are even higher than 5 km. As just mentioned in the major comments, the authors should explain at this point why it may occur
- **Figure 13**: The figure should include also some case tests for which the cloud is under the limit of detectability (e.g.: OD from 0.01 to 0.2 for CASE 1.2, from 0.01 to 0.15 for CASE 2.2)
- **LINE 621**: It is redundant to specify here the physical properties of the studied cloudy scenario, citing Table 5 would be enough
- **Figure 17**: It’s quite tricky to observe anti-correlation between retrieved emissivity and surface temperature from this graph, as left panel shows [retrieved-true] values and right panel shows [true-retrieved] values
- **LINE 654**: How can the authors explain this counterintuitive result on surface emissivity? I would have predicted a retrieved emissivity moving from water toward snow type with a larger coverage of the latter, as it is for surface temperature
- **LINE 729**: The authors should confirm in the conclusions that the CIC algorithm – in most cases - tends to prevent a clear sky classification even with low percentage of cloud contamination in the FSI FOV

**Technical revisions**

- **LINE 40**: Delete “respectively”. In the following sentence it’s better to shift “thus” at the beginning
- **LINE 41 and LINE 101**: “interannual” instead of “longer term”
- **LINE 52**: Please, specify the acronym of GCMs
- **Section 4**: Here is FEES acronym, I guess it should be FE2ES or E2ES
- **Table 4**: The right nomenclature is “cumulonimbus”, not “cumulus nimbus”
- **LINE 401**: I suggest “taking into account” instead of “takes”
- **LINE 405**: It would be better to specify the AER acronym at line 402, where AER is cited for the first time
- **LINE 431**: “Each column of the table” instead of “each row”
- **LINE 445**: “sensitivity” instead of “sensibility”
- **LINE 456**: “precision” instead of “performance”
- **LINE 489**: “top of the cloud” instead of “altitude of the top of the cloud”
- **LINE 498**: “and” is repeated twice
- **LINE 515**: the parenthesis is redundant, it could be removed