We would like to thank reviewer #1 for his/her reviews. We addressed each comment below and highlighted our answers in red, the referee's comments are black.

1. Abstract, Conclusions, and Introduction paragraph 3. Recently a third global plumeheight
product has been created. It is a thermal technique, similar to one used for volcanic plumes in the past, and has been applied to MODIS. The reference is: Lyapustin, A., Y. Wang, S. Korkin, R.A. Kahn, and D. Winker, 2019. MAIAC thermal technique for smoke injection height from MODIS. IEEE Geosci. Remt. Sens. Lett., doi: 10.1109/LGRS.2019.2936332.

We were unable to implement the MODIS plume heights, since our paper was submitted before the Lyapustin et al. paper was published. We have now included references to it in the introduction.

The MODIS plume heights are not a truly global product, since only plume heights are only valid near hotspots and exclusively for fire plumes (others should be filtered, see Lyapustin et al. Conclusions: "To exclude
the transported smoke and ensure good quality of retrievals, we currently recommend to use Ha within ±75–150 km from the detected thermal hotspots as reported in the MAIAC quality assurance (QA) flag in the MCD19A2 product."
Thus, we felt the sentence in our abstract "Before the launch of TROPOMI, only two other satellite instruments were able to observe aerosol plume heights globally, MISR and CALIOP." and in the conclusions "The only satellites that could globally observe plume heights before the launch of TROPOMI were MISR and CALIOP." are still valid.

We included a short discussion of the MODIS plume height in the introduction:

"Very recently another plume height product has been created from MODIS observations, utilizing a thermal contrast technique (Lyapustin et al., 2019). These estimates are available globally, however, they are limited to plume heights near thermal hotspots."

2. Introduction, P2, lines 30-31. Also Section 6, P15, lines 4-6. MISR provides global coverage about once per week (about every 8 days near the equator, every 2 days near the poles). CALIPSO covers effectively 10-4 of the global surface, once every 16 days. This difference could be made clearer.

We have changed the sentence on p.2 accordingly.
From:
"However, these two instruments have the disadvantage of very limited coverage where most fires are missed […]."
To:
""However, these two instruments have the disadvantage of very limited coverage where most fires are missed […]; MISR provides global coverage about once per week

(8 days near the equator and every two days near the poles) and CALIPSO provides global coverage about every 16 days."

And on p. 15:
"those two satellites have a narrow-swath with a global coverage every week and 16 days, respectively."

3. Introduction, P3, lines 6-10. MISR stereo heights have also been validated against ground lidars.

We included the following sentence:
"Caliop and (standard) MISR plume heights have also been validated with ground-based lidars (e.g. Moroney et al., 2002; Naud et al., 2004; Kim et al, 2008; Tao et al.,2008)."

References:

Moroney, C., R. Davies, and J.-P. Muller (2002), MISR stereoscopic image matchers: Techniques and results, IEEE Trans. Geosci. Remote Sens., 40, 1547– 1559.

Naud, C., J. Muller, M. Haeffelin, Y. Morille, and A. Delaval (2004), Assessment of MISR and MODIS cloud top heights through intercomparison with a back-scattering lidar at SIRTA, Geophys. Res. Lett., 31, L04114, doi:10.1029/2003GL018976.

Kim, S.-W., Berthier, S., Raut, J.-C., Chazette, P., Dulac, F., and Yoon, S.-C.: Validation of aerosol and cloud layer structures from the space-borne lidar CALIOP using a ground-based lidar in Seoul, Korea, Atmos. Chem. Phys., 8, 3705–3720, https://doi.org/10.5194/acp-8-3705-2008, 2008.

Tao, Z., McCormick, M. & Wu, D. A comparison method for spaceborne and ground-based lidar and its application to the CALIPSO lidar. Appl. Phys. B 91, 639 (2008) doi:10.1007/s00340-008-3043-1

4. Section 2.1, P4, lines 22-23. I don't understand why a different (better) solar spectrum would be applied to the OFFL product than to the NRTI product. Once you have the better spectrum, can't it be used for the NRTI product too?

At the time of the NRTI processing the different (better) irradiance spectrum is not available.
NRTI retrievals are delivered within three hours of sensing, so only data available at that time can be used. In the OFFL processing more data are available, such as an

irradiance measurement closer to, but after, the radiance measurement. This is used precisely as the reviewer suggest (to use it in the retrieval then too), but is then called OFFL data. The differences between the data streams are not really important for this paper. The OFFL data was used, which is the best choice, when the time delay is not an issue.

5. Section 2.1, P4, lines 25-26. As described, the "quality flag" sounds more like a plume detection flag; if so, this might be a better description. Have you evaluated its actual quality, e.g., by using the MODIS FRP product?

We have changed the description to the following to make the meaning of the quality flag a little clearer:
"In general, the OFFL product should perform better and is a better choice if timeliness is not an issue. Here, we evaluate the OFFL version only, as the NRTI version was not available for the time period that we investigated. As a first indication, the quality of each successful ALH retrieval is indicated by a quality assurance values (qa_value). If the input data or measurement configuration becomes close to a predefined limit, first the qa_value is lowered, if another limit is crossed, the pixel is filtered. E.g. all pixels with a solar zenith angle below 60º should have a good quality retrieval. However, for SZA > 60º the curvature of the Earth and the long photon path through the atmosphere may compromise a good retrieval. Above 75º, no retrievals are attempted. However, between 60º <SZA<75ºthe retrieval is performed, but the qa_values are lowered to 20%, to indicate to the user to use caution. This is done for all pixels with a (small) cloud fraction (qa lowered by 50%), small AAI (50%), high surface roughness (50%), and within sunglint and south Atlantic anomaly regions (50%).
Apart from the quantitative layer height, the quality flag provided alongside can be useful by itself, e.g. to locate and identify the presence of aerosol plumes and its vertical shape."

The quality flag cannot be evaluated with the MODIS FRP for several reasons:
A TROPOMI AER_LH may not be detected near a MODIS hotspot because:

- the plumes are typically downwind from the fires
- there might be clouds that interfere with the signal
- the SZA is too high to retrieve the TROPOMI AER_LH

A MODIS hotspot may not be near a TROPOMI AER_LH observation because:

- the TROPOMI AER_LH is not restricted to fire plumes, but can be from any source (volcanoes, dust, …)
- also some plume may be transported a long distance and can still be picked up by TROPOMI

6. Section 2.2, P5, line 2. Martonchik et al. (2004) did not evaluate the MISR plume height products. The main references for this product would be Muller et al. (2002) and Moroney et al. (2002).

We changed the references as suggested. Thank you for pointing this out and correcting this.

7. Section 2.2, P5, line 5. MISR actually has a standard stereo-height product, which is described in Muller et al. (2002) and Moroney et al. (2002). It runs on all the MISR data, and produces a reflectance-layer-reference-altitude, but does not call out aerosol plumes explicitly.

We have changed the following sentence in the manuscript, from:
"The plume height is not a standard product of MISR"
To:
"An operational MISR cloud-top product is available, however, the operational algorithm uses fixed-parameters that are applied to all scenes equally (Muller et al., 2002; Nelson et al., 2013). Instead, the plume height used here is not a standard product of MISR…"

8. Section 2.2, P5, line 17. The narrow MISR swath limits the frequency of global coverage.

We have changed the sentence as suggested, from:
"..(1) the swath limits the global coverage,.."
To:
"…(1) the narrow MISR swath limits the frequency of global coverage,…"

9. Section 2.3, P5, lines 27-28. The CALIOP "swath" is really a curtain, having a width of _100m, not several km. The data are usually averaged to several kilometers, but only along-track.

Thank you for pointing out this mistake.
We have changed the sentence from:
"…and has a very narrow swath width of just a few kilometres. In this study, we use the daytime aerosol layer product v4 (``Layer_Top_Altitude'', ``Layer_Base_Altitude'') (McGill et al., 2007; Vaughan et al., 2009) which provides the top and base height of aerosol layers detected (between the surface and 30 km) averaged to a 5 km horizontal resolution,…"
To:
"…and has a very narrow swath width of just a hundred meters. In this study, we use the daytime aerosol layer product v4 (``Layer_Top_Altitude'', ``Layer_Base_Altitude'') (McGill et al., 2007; Vaughan et al., 2009) which provides the top and base height of aerosol layers detected (between the surface and 30km) averaged over 5 km along the 100 m wide swath,…"

10. Section 2.3, P5, lines 30-31. Here you are using the CALIOP aerosol classification scheme, for which the key reference is: Omar, A.H., et al., 2009. The CALIPSO Automated
Aerosol Classification and Lidar Ratio Selection Algorithm. J. Atm. Oce. Tech. 26, pp1994-2014, doi: 10.1175/2009JTECHA1231.1.

Thank you we included the reference as suggested.

11. Section 2.4, P6, line 5. Small fires are also missed often by FRP, as well as those under heavy smoke plumes.

We modified the text to reflect this.
From:
"Note, that fires can potentially be missed due to cloud cover."
To:
"Note, that fires can potentially be missed for several reasons: due to cloud cover, under thick smoke plumes, as well as if the FRP signal is too low (e.g. small fires)."

12. Section 3.1, P7, line 23. For MISR, the contrast is assessed at a spatial scale of 1.1 km, which probably provides a lot more of the plume vertical structure than the model simulation – in particular, more extreme height maxima and minima.

Indeed, on a much finer resolution the minima and maxima would be more extreme. To reflect the resolution difference we averaged the MISR results to 10x10km (0.1x0.1deg) – the approximate resolution of the model - to be able to better compare the model and MISR plume heights. See text p. 11 l.16-18:
"To correct the impact of sensor resolution on the maximum plume height derived from a cluster of pixels in a given plume, the MISR pixels were averaged and binned on a 0.05x0.05 grid to approximately match the TROPOMI resolution."

13. Section 3.1, P8, line 2. Note that these are very large indices of refraction, both real and especially imaginary. Might apply to BC near source, but probably not hydrated or aged smoke particles.

The reviewer is correct that the real and imaginary parts of the refractive index will both decrease as the particle ages, but we are looking at fire plumes near the source. Kou (1996) (cited in the manuscript) found the value is 1.75 + 0.44i at 0% relative humidity [RH] for the complex refractive index of black carbon. This refractive index is unchanged up to 70% RH and is used by GEOS-CHEM. My value is for RH=99% is based on the assumption that there might be significant water from the combustion.

The thesis can be found here https://dalspace.library.dal.ca/handle/10222/55517; and further details can be found on p.12 of the thesis (p.32 if you use Adobe Reader's numbering).

Further, we realized that black carbon has a very high imaginary component in the refractive index. Other aerosols that might be part of a smoke plume is organic carbon (OC) which has a very low imaginary part – we used 1.36 + 0.001i (at RH=99%). Overall, we found there was not much difference between these two extreme cases of refractive index. The truth is probably a combination of BC and OC refractive index.

Many “MISR OSSE” plume heights were unchanged, and on average, we found the plume heights were 100m lower for OC than for BC.

We included the following in the manuscript to address the reviewer’s comments, p.x l.x:

“…at 99 % relative humidity (which is expected near the fire source): 1.68+0.36i…”

And p.8 l.23-25:

“We have also estimated the plume height assuming organic carbon (OC) with a refractive index of 1.36 + 0.001i (at RH=99 %), and found negligible differences between the plume heights obtained assuming BC and OC refractive indices for most cases (see Fig. S1 and S2).”

And in the supplement:

“**MISR OSSE with different refractive indices**

The black carbon (BC) reflective index (1.68+0.36i, RH: 99%) has an extremely large imaginary part, different to the refractive index of organic carbon (OC) where the refractive index is 1.36 + 0.001i (at RH=99%).

Looking at these two extreme cases of refractive indices, little difference was found for the MISR OSSE plume heights, most plume heights were identical, see Fig.S1. Only for plume profiles with a small plume above a large plume we found differences: the estimates assuming BC returned the plume height of the upper plume whereas the estimates assuming OC picked up the lower plume (see Fig. S2).

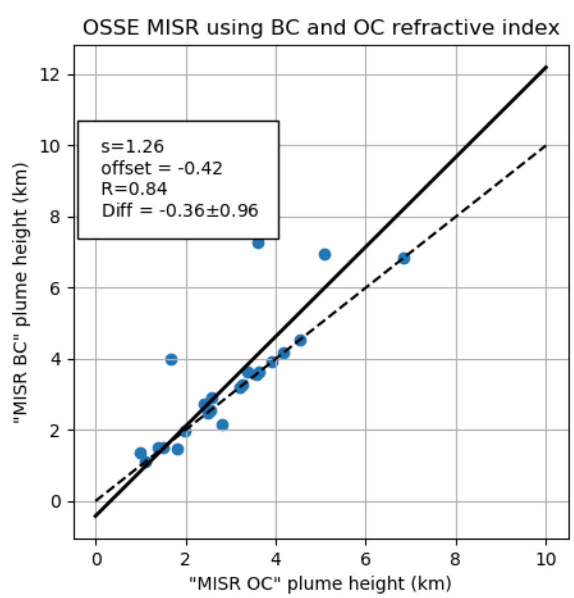

Fig. S1: MISR OSSE plume height estimates assuming a refractive index of BC and OC. The plume heights are identical (or very similar) except for three cases (shown in Fig. S2) where the plume height assuming BC is higher than the plume height assuming OC.

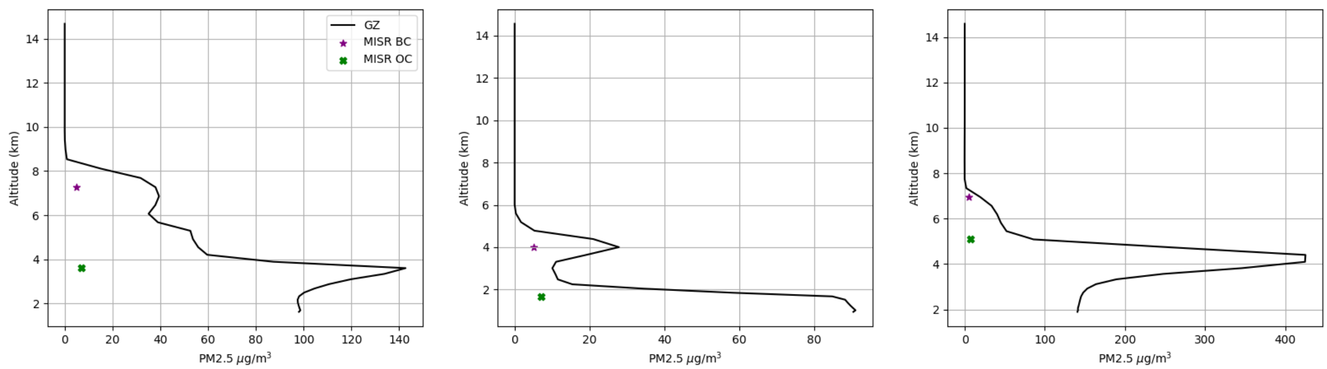

Fig. S2: The three profiles that lead to different plume heights when assuming a refractive index for OC versus BC. This happened when there is a secondary, smaller top plume that is more reflected for BC aerosols.

“

14. Section 4.2, P11, lines 29-30. CALIOP samples a curtain, so the data can be aggregated along-track to 5 km, but the cross-track width is still ~ 100m. There is nothing you can do about this, but it is worth noting that the sampling footprints of CALIOP and TROPOMI are still quite dissimilar.

We added the following sentence in the text to address this:
“Note that the Caliop data is averaged to 5 km, however along a narrow swath (~100 m), differences could arise due to the different sampling.”

15. Section 4.2, P12, lines 14-15 and Fig. 4. Here you might emphasize that by “thick,” you mean geometrically thick, and not optically thick. One would expect the differences in sampling among methods to be minimized for optically thick, geometrically thin plumes.

We have changed the sentences as suggested, and included “geometrically”.

We agree that probably geometrically thin but optically thick plumes should minimize the difference.

16. More generally, it might be helpful to identify explicitly the goal of the model and measurement comparisons in Sections 4 and 5. One would expect differences, due

to different spatial and temporal sampling, as well as sensitivity to optical depth and optical depth vertical distribution, among the measurements. The model assumptions contribute to differences among the simulations and with the measurements. So this is not really a "validation," as these could all be "correct" in the context of what they measure or simulate. Rather, I think you are exploring the sensitivity of the "plume height" result to different plume properties, measurement techniques, and modeling assumptions. As such, I find most useful the conclusions presented where you interpret the differences in terms of attributes of the derivation methods and plume properties.

We have changed the title to "The 2018 fire season in North America as seen by TROPOMI: aerosol layer height inter-comparisons and evaluation of model-derived plume heights" to remove the word "validation". The word "validation" is not mentioned in the manuscript.

The goal of the measurement comparisons in Sect. 4 and 5 can be found:
p. 1, l. 15-17
p. 3, l. 17-21

We further added a few sentences at the beginning of Sections 4 and 5 to highlight the purpose:
Satellite comparisons:

We added the following to Sect.4:
"As discussed in the previous section, there are fundamental differences between the plume heights observed by the different satellites. Here, the differences and correlation between the satellite plume height observations are discussed in terms of what is expected from the OSSE results and due to different observation times."

Model comparison:
The purpose of the model/satellite comparison is to evaluate if the model is "on the right track" or what it lacks. As discussed in the Sect. 5 and the conclusions, the modelled and observed plume heights correlate, however, especially over grassland, the model consistently overestimates the plume height, which is something that is helpful for the modelling community and something can be addressed in future releases of CFFEPS.

We added the following to Sect. 5:
"The modelled plume heights are compared to satellite observations with the aim to evaluate the modelled plume injection heights and to determine the strengths or weaknesses of the model."

17. Section 6, P15, lines 18-19. I'm wondering whether the "exact plume height" is really well defined when there are multiple layers.

We removed the word "exact" from the sentence.

18. Section 6, P15, lines 19-20. Actually, most aerosol plumes are not uniform in optical thickness, and when multiple layers are present, they rarely cover exactly the same area. As such, MISR will often pick up multiple layers, not in a single 1.1 km pixel, but over the plume area imaged by the instrument.

We have changed p.15, l.19-20, from:

“MISR on the other hand tends to respond to the upper aerosol layer, if there are any layers beneath MISR will not be able to pick this up.”

To:

“MISR on the other hand tends to respond to the upper aerosol layer if multiple plumes overlap the same pixel, if there are any layers beneath MISR will not be able to pick this up. However, often multiple layers of plumes do not overlap exactly the exact same area, so MISR will likely sense the lower plume heights over the plume area imaged by the instrument.”