R#	Reviewer Comment	Author's Response	Changes Made
1	The present work examines two years of handheld infrared thermometer observations and uses PWV calculated from a weighted average of the two nearest radiosonde sites. Remotely-sensed PWV observations are available from instruments that are situated more closely to the IR thermometer observing sites. The authors make use of these observations in a secondary sense to illustrate the overall annual evolution of PWV over the domain. These results are displayed in Figure B1, a welcome addition to the discussion. However, this figure also raises a significant question: are the radiosonde observations (especially EPZ) biasing the analysis? It seems to me that the most representative observation for the IR thermometer is going to be from the AERONET station: it's not the closest observation point, but it is the most similar in altitude and it is substantially closer than the radiosonde sites, and it also has observations that are much closer in time to the IR observations than the sondes do. Figure B1 indicates that the sondes tend to be more moist than the AERONET, with EPZ consistently more moist than the other observations. How is it known, therefore, that the spatially weighted average of the ABQ and EPZ sondes is the most representative observation and therefore the one around which this work should be built? The authors do not make primary use of the remotely sensed observations for two reasons, representativeness (due to the altitude difference between Suominet and the IR thermometer) and missing data (due to a year-long gap in the AERONET and Suominet observations). The first issue is a valid one, but I feel this second point deserves some more inspection. After all, the AERONET observations are still available for approximately half the observing period. It may be that two years of weighted-averaged radiosonde data that is 100s of km and approximately 6 h removed from the IR thermometer observations is better than one year of AERONET observations that is both spatially and temporally closer to the target. However, that point needs to be explicitly argued.	The primary reason for using radiosonde data is the availability of data. Over our two-year dataset (2019-2020), we have 522 days with clear-sky temperature measurements along with radiosonde PWV. The AERONET PWV data is only available for 142 of those days; if our analysis was built around AERONET, we would cut the size of our data by 73%. Similarly, the SuomiNet PWV data is only available for 270 days, reducing our dataset by 51%. Figure 3 shows a significant degree of variability in PWV, both seasonally and interannually. Therefore, data records of less than one year in duration will not adequately cover the range of values needed to obtain an accurate relationship between zenith clear sky temperature and PWV. However, both SuomiNet and AERONET have important roles to play in checking the robustness of our approach. Thus, we have incorporated Appendix B along with Figure B1 into the main body of the paper, and we included a more detailed discussion of the differences seen and the data availability issues. Furthermore, we added a new figure (Fig. 6) that is a two-panel scatter plot which compares our IR PWV product with both SoumiNet and AERONET PWV. The spatially-weighted average of the radiosonde measurements has good agreement with the SuomiNet and AERONET datasets and with significantly fewer missing days. We have also adjusted the weighting on the radiosonde data to better take into account the surface elevation difference. The result of this change is less weighting on EPZ data (25% compared with 40% in the previous version). Even with the EPZ values weighted less than the ABQ values that bias is not as prominent. All of these aspects are discussed in the revised manuscript.	Line 153 (markup): changed "nearby" to "within a 50-km radius" Line 155-159 (markup): Major additions to the discussion on usage of SuomiNet data. Line 161 (markup): changed "than NMT campus where the zenith sky temperatures are measured." to "in elevation than our measurement site." Line 162 (markup): changed "this" to "the difference in elevation" Line 162 (markup): changed "lead to a 20% systematic difference between South Knoll and NMT campus" to "load to a dry bias of 20% at South Knoll as compared to NMT campus." Line 165 (markup): changed "NMT campus. Unfortunately" to "Socorro, but" Line 166 (markup): changed ", which" to "that" Line 167 (markup): changed "that must be considered" to "in comparison to radiosonde PWV" Line 168 (markup): changed "utilizes" to "uses a weighted mean of" Line 176 (markup): Added "Figure 3 compares the radiosonde PWV data with both SuomiNet and AERONET observations for one year (2020)" Line 177 (markup): removed "(\DeltaPWV/PWV < 20%)" Line 181 (markup): changed "all of the analysis that follows, we use a weighted mean (inversely related to distance from Socorro)" to "the following analyses, we employ a 2-dimensional interpolation (which is linear in horizontal distance, and exponential in elevation)"

	In the end, this work relies on observations that are frequently 6 h old (or 6 h early), at least 100 km away, and much more moist than more local observations. Using the radiosonde observations may be the appropriate course of action, but it needs to be demonstrated that this set of decisions is the correct one.		Line 183 (markup): removed “Appendix B compares out derived PWV with both SuomiNet and AERONET observations for one year (2020)” Line 184 (markup): added “The corresponding linearized weighting factors are 0.75 for ABQ and 0.25 for EPZ” Line 186 (markup): changed “all three datasets are” to “PWV data from SuomiNet, AeroNet, and radiosonde means are all” Line 187 - 193 (markup): changes to discussion on the data gaps for SuomiNet and AERONET. Line 252 - 259 (markup): Added additional discussion on SuomiNet and AERONET analysis. Line 376 - 384 (markup): removed Appendix B. Figure B1: removed.
1	Finally, as I read through this work again, I’m left with one very fundamental question: how good is it? An analysis that shows the relationship for the IR PWV product to some kind of truth (be it the merged sondes, AERONET, etc.) seems to be lacking. The figures shown in the present work, such as the relationship between the sky temperature and PWV, are important but the relationship between the new product and the truth is critical. This could take the form of a scatterplot, histogram of the differences, box and whisker plot for the differences in various PWV bins, etc., but something should be in there. Crucially, I do not have a sense of how well the product performs as a function of different values.	For a journal that is focused on measurement techniques, it does not seem appropriate to refer to “truth” as opposed to actual measurements with their known uncertainties and possible biases. We believe that the integrated humidities from radiosondes likely provide the most accurate PWV compared to GPS or sun photometer measurements. However, the spatial and temporal interpolation to the Socorro location likely involves a higher degree of uncertainty for the weighted sonde means. In the end, there is simply not enough AERONET data with which to base our analysis, and SuomiNet has too large of an elevation offset. On the other hand, we appreciate the suggestion of adding scatter plots for comparison. We have added a figure (Figure 6) as discussed above.	
1	98. For the observations taken at 2300 UTC, are they matched to temporally averaged radiosonde observations or are they just matched to the nearest sonde time?	We consistently use the weighted average of PWV across all of our observation comparisons.	No changes
1	112. It is important to emphasize that the determination of clear or cloudy skies is a subjective observation by a human observer.	We have revised the paper to clarify that the observations are subjective.	Line 113 (markup): added subjective
1	115. The lack of brightness temperature observations below a given temperature threshold (resulting in NaN values) means that low PWV values cannot be observed with this method. What is the minimum PWV value that can be observed, and what is the seasonal distribution of missing data? This seems like an important issue that end users ought to be aware of. I assume that this is a more frequent occurrence in the high deserts of New Mexico than it is in the environment observed by Mims, and that wintertime values are more likely to be missing, but these points should be	As specified in the paper, the primary AMES sensor has a lower temperature threshold of -50 degrees C, and this results in NaN values less than 4% of the time (Table 1). As expected, the lowest temperatures occur during the coldest part of the winter. The low-temperature threshold limits minimum PWV to approximately 3 mm, as seen in Table 1 and in Figure 5.	No changes

	made explicit in the text.		
1	Figure 1: This is an extremely minor point, and you can address or ignore as you see fit, but I find figures easier to interpret when grid lines are present.	We have decided to keep the plots without gridlines to improve readability on data-heavy figures.	No changes
1	142. When you say ground temperature, do you specifically mean skin temperature as measured by the IR thermometer?	Ground temperature and skin temperature as measured by the IR thermometer are effectively the same, as now pointed out in the manuscript.	Line 105 (markup): added “(the effective IR skin temperature)”
1	165. If the Suominet and AERONET observations are going to be part of this analysis (even if only in the appendix), their locations should be noted on Fig. 2.	We have added the SuomiNet and AERONET locations to the map.	Figure 2 (revised): Changed to include Suominet and AERONET locations and is now colorblind friendly. The caption was also adjusted. Line 171 (markup): added “Socorro, ABQ, EPZ.” Line 171 (markup): added “, along with the locations of Socorro, SuomiNet, and Sevilleta AERONET sites.”
1	165. Sometimes the text refers to Figure N, and other times it refers to Fig. N. This may be a stylistic choice, as it appears that the word is spelled out at the start of a sentence but not elsewhere, so I don’t know how much consistency you are going for here.	This style choice is based on the formatting guidelines laid out by the journal.	No changes
1	174-175: I’m not seeing where your product appears in Appendix B (unless you only mean the merged sondes). This goes back to the point I made in the major comments above about not really getting a sense of the skill or utility of the product.	Given that we have plotted 5 datasets in the time series plot (which now appears in the manuscript rather than the appendix), we have not added our PWV IR product to this plot. Instead, we now compare the PWV IR product to SuomiNet and AERONET in Fig. 6.	No changes
1	203. This seems like a counterintuitive way to approach the exceedance thresholding, as though the most important thing was to preserve 90% of the dataset instead of crafting a representative dataset. If the data are unrepresentative, they should not be used regardless of how many event dates must be removed. At a minimum, it is important to know how many standard deviations that 55% difference represents. (Also note: in the response to the reviewers, the authors stated this was a 75% threshold, so they should verify which value is the correct one.) It is easier to scientifically justify a standard-deviation-based filter than a filter designed to preserve a certain fraction of the	The 75% value was a typo and the correct value was 55%. We apologize for the confusion. To address this comment we have redesigned this feature to compare the standard deviation of the PWV measurements for the individual days with the average standard deviation over all days. The paper has been adjusted to reflect these changes.	Line 232 (markup): changed “A = 20.2 mm and B = 0.036” to “A = 18:48 mm and B = 0:034” Line 217 (markup): changed “relative difference” to “standard deviation” Line 218 (markup): changed “individual PWV observations to the daily mean of both ABQ and EPZ” to “the standard deviation of the PWV observations for a given day with the mean of the daily standard deviations over the entire dataset.” Line 220 (markup): changed “any

	total dataset, even if in the end you tune one filter to match the other.		difference exceeds a fixed 55%. This threshold value was determined so that no more than 10% of the days are rejected by this filter, while still ensuring" to "the standard deviation is more than twice the overall mean value" Line 223 (markup): changed "radiosondes do not bias" to "PWV, or between 00Z and 12Z observations, do not negatively impact" Line 224 (markup): Added "Approximently 12% of the days are rejected by this filter" Line 247 (markup): changed "mean" to "standard deviation" Line 250 (markup): changed "3.79" to "3.64" Line 260 (markup): changed "3.75" to "3.60" Line 262 (markup): changed "4.52" to "4.63"
1	226. This is close to what I was suggesting when I suggested a monte carlo simulation. My thinking was that you could take an IR-observed temperature, randomly perturb it by some value drawn from a gaussian, and plug that into your tool to obtain a new PWV. Do that a few thousand times, and you'll have an estimate on how the instrument uncertainties contribute to uncertainties in PWV. This doesn't take into account the uncertainties in the model, however, which your approach seems to do.	We appreciate your feedback.	No changes
1	230. Does this RMSE vary with the magnitude of the signal? Looking at Figure B1, a RMSE of 0.35 cm is very close to the observed value for the winter months. Do you expect that the error bars are very similar throughout the year, or do the larger PWV values in the summer have greater uncertainties associated with them?	Although it is possible that the RMSE contains a seasonal component, we feel that a detailed analysis of seasonal changes in RMSE in the paper is not not warranted at this point, given two years of available data.	No changes
1	285. This seems to imply that it may be possible to derive the appropriate relationships between PWV and the IR temp without	Based on the theory alone, it is not possible at this time to derive accurate fit coefficients for all locations, and it	No changes

	needing to take two years of manual observations to generate a testing dataset. Is this true? Or, rather, are there ways to arrive at the needed coefficients using existing data? (Earlier, when I said that I wanted to point my IR thermometer at the sky to get PWV, I meant it.) In all seriousness, you have done a good job demonstrating that the system needs to be trained to specific locations due to the large climatological variability in water vapor content. But are there ways to achieve acceptable results using a priori data? I think this is an important point for the issues raised in the conclusions, as substantial datasets will need to be collected by citizen scientists and school groups just to train the relationships. If an initial model can be implemented immediately from prior observations, NWP, etc., the adoption of such a program will likely increase.	will likely be necessary to build up a minimum database in order to carry out a widespread citizen science program. One step in this process is to explore the parameter space for a diverse set of locations and meteorological conditions. We feel that the current paper is an important step towards that goal.	
1	Figure B1. I keep coming back to this figure throughout reading and reviewing this paper. At times I wonder if this figure is important enough that it deserves promotion ot the main body of the paper.	We appreciate this suggestion, and have integrated Appendix B into the main paper along with additional discussion as noted above.	Changes discussed in the top row.