Atmos. Meas. Tech. Discuss., referee comment RC1
https://doi.org/10.5194/amt-2021-130-RC1, 2021


# Comment on amt-2021-130

Anonymous Referee #1

Referee comment on "Atmospheric precipitable water vapor and its correlation with clear-sky infrared temperature observations" by Vicki Kelsey et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-130-RC1, 2021

## General Comments

Overall, I applaud the novel approach that the authors are taking to produce low-cost observations of an atmospheric parameter that has many applications, from climate studies to model assessment.  While they haven't developed the technique themselves, they are evaluating several different commercially available products to assess how well-suited they are to produce these observations and how a different location can impact the relationships used to obtain their targeted variable.

However, there are some fairly significant issues with the work that preclude publication at this time.  I believe that this will require the team to redo almost all of their analysis.  However, I do believe that ultimately the fundamental work should be published, and therefore I am recommending major revisions.

## Specific Comments

The first thing that struck me while reading this paper is that this is not a method to observe total precipitable water (TPW), but really a method to observe precipitable water vapor (PWV) in clear sky conditions. While one can argue that in clear skies the TPW is functionally equivalent to the PWV since there is no liquid or ice water present, this distinction is a valuable one: there are more sources of PWV data than TPW since measuring cloud characteristics is so challenging. There are several additional ways of measuring PWV that the authors do not address in the manuscript. This includes a direct retrieval from ground-based hyperspectral IR observations (Turner 2005 https://doi.org/10.1175/JAM2208.1), calculated from thermodynamic profiles retrieved from hyperspectral IR observations (Turner and Blumberg 2018 https://doi.org/

10.1109/JSTARS.2018.2874968), Raman lidar, aircraft, etc.

This leads into the most significant concern that I have about the present work: the training and validation dataset has significant drawbacks and better choices may be available. It may be true that in the desert southwest the temporal and spatial variability is not large, but it remains that the data being used is, at a minimum, located 110 km and 6 h away from the desired quantity. I am surprised that the authors did not utilize the Suominet observations of PWV from the Socorro area, especially since one of the authors is the contact for that particular observing site. This may be due to thinking that the present work describes a TPW product and not a PWV product. It is true that the observation site is located on a mountain while the IR observations are presumably taken at a lower altitude. This criticism is tempered somewhat by the fact that the two radiosonde sites used for validation differ in elevation by ~400 m and so altitude differences are going to be an issue regardless of the validation set used. That being said, a quick glance at a 14 day time series at Albuquerque (http://www.atmo.arizona.edu/products/gps/P034_14day.gif) and Socorro (http://www.atmo.arizona.edu/products/gps/SC01_14day.gif) doesn't really show a huge impact of the altitude (at least at the time of the writing of this review). Suominet has the advantage of a substantially better temporal resolution allowing a more direct comparison to the IR observations, and in fact, offering enough observations that it would be possible to average to reduce noise in the signal.

Even if they choose not to use Suominet observations, there are ways that the radiosonde dataset can be leveraged to create a more representative data sample.  Rather than using every single IR observation, it may be better to exclude from analysis the cases in which there is a substantial difference between the two sites, and/or between the 0000 and 1200 UTC launches. By focusing on cases in which the spatiotemporal variability is small, the authors can have greater confidence in the retrieved product. This will reduce the number of data points, but I fell will produce a stronger product overall.

The error analysis also seems to be somewhat lacking, as it tends to focus on the uncertainty of the regression while not addressing the influence of the uncertainty of the instrument or the measurement technique.  A monte carlo approach may prove useful here: by randomly perturbing the input brightness temperatures by a random value chosen from a gaussian distribution with a standard deviation equal to the instrument uncertainty, then repeating that over a set number of trials, it may provide a more realistic assessment of how the instrument itself may be contributing to the error bars of the retrieved value.  This doesn't include the uncertainty induced by the way the instrument is held, which may also expand the uncertainty of the retrieved value.

Finally, I'd like to see a greater exploration of the differences between Mims et al 2011 and the present work.  What is the RMSE of the current dataset, and how does that compare to the RMSE if you applied the Mims relationship to your data? In other words, how much are you improving the technique by tuning it for your specific location? Such an analysis would help increase the novelty of this paper.

Technical Comments

Line 50. Consider how PWV (not TPW) is also being measured by various systems, based on the discussion abovce.

Line 75. How are the observations actually being taken? Is a human pointing a hand-held system towards the sky and writing down the observed temperature, or is a more robust method being used? Many IR thermometers have adjustable emissivities, and the default isn't necessarily a blackbody. Were the emissivities set to the same value across all systems?

Line 77. Does the manufacturer note the wavelengths at which this instrument operates?

Line 99. This analysis of how to hand-hold a thermometer within 5 deg of zenith, and the fact that it results in less than 1 C uncertainty, is interesting, and the discussion of both points should be expanded.

Line 104. How are you screening for clouds? Observer judgement? Airport ceilometer? Satellite? IR thermometer threshold?

Line 111. I find it surprising that there is little dust in the middle of the high deserts of New Mexico. Why is the dust so low?

Fig 1. This figure is very confusing to me, and I apologize if there is something obvious that I'm missing. There are four categories: clear, cloudy, clear NaN, cloudy NaN. It seems like two separate things are going on. There is an instrument assessment to determine if the sky is clear or not (more detail on that is needed). But in the case of the NaNs, an external assessment of the clear our cloudy state has to be used because the instrument is not reporting anything. This is all coupled with the fact that the manuscript says that clouds were filtered out. Ultimately, I'm not sure what the figure is trying to tell me. A better approach may be a contingency table for each instrument that compares the external / instrument assessment in terms of clear/clear, clear/cloudy, cloudy/clear, and cloudy/cloudy, with special notes of the number of NaNs in each category.

Figure 2. By starting out the caption with (a,c) it is somewhat confusing to the reader (who may be more accustomed to going from a to b). It may be better to say something like "Comparisons between the AMES 1 and the FLIR i3 (left column) and the AMES 2 (right column) for clear sky (top row) and ground (bottom row)."

Line 140. This section would be greatly improved with a map showing the location of ABQ, EPZ, and Socorro, with elevation as the background color.

Line 156. The amount of data that is used in the analysis fits better in the methodology than in the results. I found myself using the values reported in Fig 1 to calculate the approximate number of datapoints for context before I got to this part of the paper.

Line 186. Is this R^2 for a linear correlation? If so, you may actually have a better fit than your numbers report, since the fit has an obvious non-linear shape.

Line 220: It doesn’t appear this way from the observations in Figure 4, but do the model studies show any evidence that the signal gets saturated (that is, is there a point where PWV is so high that any additional PWV can’t be detected from the brightness temperature observations)?

Line 257. This cost info is very important and should appear in the intro.