Overall, this is an interesting topic, but I found it hard to figure out what the paper is trying to accomplish. The purpose of the paper needs to be more clearly discussed in the first few pages.
The same comment has been raised by the first reviewer. We now clearly stated the two main research questions of the paper in the introduction:
"In this paper, we focus on answering the following questions.
1) How well are the spatial and temporal IWV variability represented by three different, independent, IWV datasets? As our primary dataset used here is a ground-based GPS IWV dataset covering a world-wide sample of 118 sites, we will assess the spatial variability only between those sites and the global IWV datasets will be sampled to the site locations. In this work, a first characterization of the spatial IWV variability is given by considering the geographical distribution of the IWV frequency distributions, an extension of the work by Foster et al. (2006). To assess the temporal IWV variability of the three datasets, we consider different time scales here: from the seasonal cycle to short inter-annual variability, and to trends over periods of around 15 years.
2) Can the spatial and temporal (inter-annual and trends) IWV variability be explained by changes of local meteorological variables (like e.g. surface temperature and pressure) and/or by low-frequency variability in atmosphere and ocean (on both global and regional scales), and if so, how? There are already a number of studies mentioning IWV patterns that result from interactions between the atmospheric circulation and the land and ocean surfaces (see e.g. Trenberth et al., 2005, Wagner et al., 2006, Shi et al., 2018, Wang et al., 2018). These studies are based on correlation studies between IWV and one such a teleconnection index (ENSO: all mentioned studies, Pacific Decadal Oscillation: see Shi et al., 2018). Here, we fit the monthly mean IWV time series by means of a stepwise multiple linear regression approach with proxies for the seasonal cycle and linear trend, and with local meteorological variables and teleconnection indices as explanatory variables. With this empirical approach, we aim at finding out which the most relevant variables are to explain the IWV variability for different regions, independent of the used IWV dataset. To our knowledge, it is the first time that such an analysis is done on the IWV time series of individual sites."

I find that the comparison between the GPS, GOMESCIA and reanalysis IPW data valuable and fairly well described in the paper. I would like to see a clearer discussion of the effect of clear-sky bias on the GOMESCIA data, and whether this could be an explanation for its larger discrepancy from the other data sources.
The scope of this paper is not the comparison of the different IWV datasets. This was the subject of our previous paper (Van Malderen et al., 2014) and has also been discussed in Beirle et al. (2018). However, we added some extra information from these papers in the manuscript, e.g.:
"The selection of cloud-free observations for GOMESCIA - unavoidable for water vapour retreivals from satellite measurements in the visible range, where cloudy scenes have to be masked out - corresponds to generally dryer atmospheric conditions, which likely results in low biased means. Comparisons to independent measurements in Beirle et al. (2018) result in relative biases of typically -5 to -10% for the total mean. In this context, it should be noted here that the GOMESCIA climate product was optimized for inter-instrumental consistency over time, not for accuracy (see Beirle et al., 2018). However, the impact of the dry bias on trend analysis is small, unless also the cloud properties themselves change over time."

The sorting of histograms of IPW at different sites is also valuable and clearly extends the earlier work in this area. I would like to see a more organized discussion of the impact of seasonal behavior on these histograms. The paper asserts that the seasonal behavior is important, but does not explicitly show that if the seasonal cycle is removed, the resulting distributions of the residuals are simpler (e.g gaussian or log-normal).
This suggestion has been taken into account by including a figure (new Fig. 5) of the classification after the seasonal cycle has been removed and a description of this analysis in the manuscript. Please see below (at your detailed comment on this issue) for more details.

I am less convinced by the "step-wise multiple linear regression." First, I don't understand the name – what is step-wise about it? Second, so many potential explanatory variables are used (which are said to be at least partly independent), that at least some of them are likely to have explanatory "power" by chance for the relatively short times series studied. I would be happier if the authors could clearly state a hypothesis, and then test it with a more limited set of explanatory variables consistent with the hypothesis. I am not a fan of a "throw everything at it and see what sticks" approach.
As also asked by the second reviewer, the step-wise multiple linear regression is explained in more detail in the paper. We indeed used a very large set of potential explanatory variables (ranging from 103 to 194, depending on the region), but the statistical test only retains on average 7 to 8 explanatory variables out of this list as contributing significantly to the correlation coefficient. We therefore believe in the explanatory power of this approach to select the most relevant variables for the IWV time series.
For information, we developed and applied the step-wise multiple linear regression initially to explain the time variability of the total ozone column at Uccle, Brussels. In this area, the relevant explanatory variables

(solar flux, QBO, EESC, stratospheric aerosols, EP flux) are well established by other studies and could be confirmed by our algorithm.

In this paper, we do not claim with our multiple linear regression approach that explanatory variable X can explain part of the IWV time variability of dataset X at station Z. Instead, we group the stations in regions (guided by the analyses of the frequency distribution and the seasonal variation) and discuss only the dominant explanatory variables for that specific region, for the three datasets used. By doing this, we are confident to cancel out to a certain extent the uncertainty related to the identification of specific explanatory variables for specific IWV time series.

From the literature, some obvious candidates for IWV explanatory variables could be determined, like the surface temperature, surface pressure, NAO, ENSO (see e.g. Wagner et al., 2006, Shi et al., 2018, Wang et al., 2018), Pacific Decadal Oscillation (PDO, see Shi et al., 2018). Rather than trying to confirm these links between IWV and those explanatory variables, we aim at finding out which are the most relevant variables by our empirical approach here. In this context, it should be noted as well that indices like ENSO have been defined rather arbitrarily (depending on the geographic location of the available weather stations) and are purely empirical as well. So there is already some arbitrariness in the whole matter. These last lines have also been added to the manuscript.

There are numerous cases of strange English usage/wording, some of which I mention below, but there are far more than I can explicitly call out. I recommend that the paper be edited by a native English speaker.
We went through the entire manuscript again, and try to avoid alternative English usage/wording than the usual formulations in scientific papers.

Some more detailed comments below:
Page 3, line 20-22. Strange Wording (“vastly”). Also, what does neutral mean in this context?
Neutral means here: nonionized component of the atmosphere (so, not the ionosphere). We explained this in the manuscript and changed “vastly” by “located mostly”.

Page 3, line 30. Strange wording (“disposed of”). Maybe change to “This process results in a world-wide. . ..”
Done.

Page 4, line 29. “downsized”? How was the conversion from 5-minute observations to 6 hourly observations performed? And why is this needed/appropriate considering that the reanalysis can be considered a snapshot at the synoptic times? (I’m not saying that the downsizing is wrong, I just want it explained better.
We changed the sentence into “We did not apply any time interpolation, so that, as the ERA-Interim reanalysis is only available at 0, 6, 12, and 18h UTC, the resulting GPS IWV dataset is also to the utmost available at these mentioned times.”

Page 8. I can’t follow the discussion of the clear sky/cloudy biases. Is the GOMESCIA only available in clear-sky conditions? If so, wouldn’t it be expected that the GOMESCIA is biased low compared to measurements available in all-sky conditions?
On page 8, we wrote that “Looking at the biases, we found that 70% of the GPS stations have a negative IWV bias with respect to ERA-Interim, while 60% of the stations have a positive bias compared to GOMESCIA. Especially for sites in Europe, Southeast Canada and East USA, GOMESCIA shows a dry bias with respect to GPS, which in turn is dry-biased compared to ERA-Interim.” So, indeed, GOMESCIA is biased low (dry bias) compared to both ERA-Interim and GPS. We tried to be clearer about this by linking “positive bias” to “wet bias” and “negative bias” to “dry bias” immediately.

Page 9 and 10. Discussion of different distribution classes at different locations. It is unclear how the sorting into classes was performed. Are all types of fits tried, and then some criterion applied? Please explain more clearly.
Indeed, we first fitted the different distribution classes separately (Gaussian, single lognormal, 2 lognormals) by means of a non-linear least squares fit to the frequency distributions of a station. A first, automated sorting into the different classes has been applied, based on the value of the chi-square goodness-of-fit statistic. Therefore, those values are compared for the Gaussian and single lognormal fit and the distribution with the lowest value is chosen. To sort between the different lognormal classes (single, shouldered, and bimodal), we defined some range intervals for the chi square of the standard lognormal fit (resp. $X^2 < 0.003$ for lognormal, between 0.003 and 0.01 for shouldered lognormal and above 0.01 for bimodal). These limiting values have been determined beforehand by comparing the chi square values and the quality of the fits (plots) by eye. This automated procedure is then used as a guidance (proposal) for the interactive procedure, in which we show all the frequency distribution fits of the different classes (and their chi squares) in one plot. In resp. 80% and 75% of the cases for respectively ERA-Interim and GPS, the

class proposed by the automated procedure have been adopted. The largest difference between the automated and interactive procedure took place for the Gaussian distribution, which was largely overestimated by the automated procedure. This procedure has now also been explained in more detail in the manuscript:
"In this work, we used a non-linear least squares fit to compute (separately) Gaussian, lognormal, and bimodal distribution functions for the IWV distributions observed at each GPS sites, making use of the same formula for the lognormal distribution function as in Foster et al. (2006). The sorting into the different categories was based for 75-80% of the sites on the values of the chi-square goodness-of-fit statistic (e.g. Gaussian if this statistic is lower than the lognormal one, a limiting value for the lognormal fit statistic to discriminate between a single and bimodal lognormal distribution) and determined interactively for the remaining sites by checking by eye the different frequency distribution fits. Examples of each category are given in Fig. 3a-b-d. We added an extra category, in between a lognormal and bimodal distribution, also in terms of the range of the chi-square goodness-of-fit statistic for the single lognormal distribution fit for these sites. For this category, there is one clear lognormal distribution which characterises the majority of the distribution, but an additional, secondary lognormal distribution (most often at the higher IWV side, an upper mode) is needed to explain the overall frequency distribution "satisfactorily", i.e.in terms of the chi-square goodness-of-fit statistic and/or by eye. We call it a "shouldered" lognormal distribution, see Fig. 3c."
Also, if the various distributions mostly occur because of seasonal cycles, maybe it would be good to show analysis after the seasonal cycle is removed??
We included an extra figure in our manuscript and extended/grouped this analysis: "To illustrate the impact of the seasonal variability on the shapes of the frequency distributions, we show in Fig. 5 the classification of the sites after the seasonal cycle has been removed from their time series (by subtracting the overall monthly means), both for GPS and ERA-Interim. A first thing to note is that no more bimodal distributions are present, but are replaced (mostly) by single lognormal or even Gaussian distributions. In particular, it is now clear that the dominance of the original bimodal distribution for the Asian sites (see Fig. 4a) is linked to the seasonal behaviour due to the monsoon, which is responsible for the reverse lognormal distribution with high median value, the strong upper mode, with the lower mode being caused by the dry season. The bimodality for the (sub)tropical sites is caused by the seasonal IWV variation too, and the reverse lognormal distribution is now very prominent for the deseasonalized IWV time series of (sub)tropical coastal or island sites (see Fig. 5). Another striking difference between the classification of the frequency distributions from the original and deseasonalized time series is situated in Europe: after removing the seasonal cycle, the dominating shouldered lognormal distributions in Fig. 4a are turned into standard lognormal distributions in Fig. 5a. So, in Europe, the shouldered lognormal distribution originates from the seasonal IWV variation. A remarkable feature for North America is the very similar geographical distinction in the continent in terms of the GSD values in Fig. 4b and the classes of frequency distributions of the deseasonalized IWV times series (Fig. 5b): whereas the sites in western part of North America (low GSD values) have standard lognormal distributions, the central and east North American stations (higher GSD values) are best fitted by shouldered lognormal distributions. For those latter stations, the IWV variability caused by weather or inter-annual variability seems to be more complex (multimodal) than if the seasonal variability is added."

Page 11, lines 19-20. This could be tested by subsetting the reanalysis data so that it too has gaps at the same time as the GPS data.
Thank you for this suggestion. We investigated this, and, as a matter of fact, the explained variabilities, the correlation coefficients, and the percentage of sites with a biannual cycle are now 64%, 0.776 and 68% respectively for ERA-Interim, and 55%, 0.710 and 60% for GOMESCIA. These numbers lie in the same range as for the GPS dataset, GOMESCIA being even lower. This has been adapted in the manuscript as follows:"These values are 81% and 0.895 for ERA-Interim respectively, and 71% and 0.831 for GOMESCIA. In this context, it should be noted that a slightly higher number of sites were found to have a statistically significant contribution from a biannual cycle for ERA-Interim (79%) and GOMESCIA (75%) than for GPS (69%), which also contribute to a better linear regression representation. The worse parameterisation by harmonics for the GPS dataset can be explained by the presence of gaps in the IWV monthly mean time series: if we subset the ERA-Interim and GOMESCIA monthly means to the GPS time series, the explained variabilities, correlation coefficients, and percentage of sites with a biannual cycle now lie in the same range as for the GPS dataset (even lower for GOMESCIA)."

Page 16. I wonder if the reason for the poor fitting for the west coast of North America sites it due to the fact that for much of this region, the rainy season is in the winter, where the temperatures are low. In the summer, it is dry but warmer, so there is not that much of a change in IPW. Other regions, such as eastern NA or Europe, there is still significant rainfall in the summer season, leading to a much larger correlation with surface temperature.
Thank you for this very valuable suggestion. Our data indeed shows the seasonal variations in surface temperature, IWV and precipitation for the majority of the sites of the west coast of North America. We included your suggestion in the manuscript, but we must confess that we have no idea how we might check it with our approach. Your suggestion is also consistent with the amplitudes of the seasonal cycle being

larger for the central and eastern sites in Northern America than for the western coast sites. So we also mentioned it in Sect. 5, where we included a discussion of the geographical pattern of the seasonal cycle, as requested by the first reviewer.