Answers to the Interactive comment on "Noise characteristics in Zenith Total Delay from homogeneously reprocessed GPS time series" by Anna Klos et al. written by Anonymous Referee #1

Thank you very much for your valuable feedback. We followed all suggestions you made and tried to address all comments. Hopefully, now the paper is clear. The answers are given below in italics.

This paper investigates the reprocessed ZTD time series of 120 permanent GPS stations spread among different climate zones. The seasonal and diurnal cycles present in the ZTD time series are first studied. But the main point of the paper is then to identify a better noise model for the ZTD time series than the widely used white-noise-only assumption. The authors come to the conclusion that an AR(4) + white noise model has to be preferred and finally discuss the implications of their refined noise model on the uncertainties of the trends of their ZTD series.

While the noise characteristics of GPS-derived tropospheric time series are indeed worthy of investigation, the paper however presents important defects in several methodological and formal aspects. A major revision therefore seems necessary to me. Specific major comments follow. A manuscript annotated with additional minor comments and corrections is also attached to this review.

A) Why using ZTD series rather than ZWD or IWV series?

At the end of the Introduction (page 3, lines 18-30), the authors recognize that studying ZWD or IWV series rather than ZTD series would be more climatologically meaningful, which is undoubtedly true, especially when focusing on trends. They then give two arguments to justify their choice of nevertheless sticking with ZTD time series:

* "The stochastic properties of the ZWD and ZTD time series are nearly identical" (supported by a single example in the supplementary material).

* "The preferred products for assimilation in NWP models are the ZTD estimates and not the IWV estimates".

I find the first argument quite questionable. First, even in the only provided example, the ZHD seems to have a comparable power as the ZWD in the monthly to semi-annual frequency band. So it seems likely that removing or not the a priori ZHD from the series (i.e., studying ZWD series rather than ZTD series) would have a noticeable impact on the noise analysis results. Then, what about other stations? Without evidence of the contrary, one could imagine that in some places, the variability of the ZHD may be even higher than that of the ZWD in specific frequency bands, which again would have an impact on the noise analysis results. And apart from the noise analysis itself, the aim of this study, as I understand it, is to eventually provide trends with realistic uncertainties for climate studies. But is there any climatological sense in providing ZTD trends?? I don't think that the authors' second

argument holds either, because although ZTD are assimilated in NWP models, they are currently not assimilated in climate models. At present, the long-term noise+trend analysis made by the authors can therefore not be performed on post-assimilation IWV series, but is only possible with "raw" GPS-derived ZWD series, possibly converted to IWV series.

The issues related to the ZWD -> IWV conversion and their potential impacts on the estimation of long-term trends are nicely summarized by the authors and can justify working with ZWD series rather than converted IWV series. However, as explained above, I can't see any reason to work with ZTD series rather than ZWD series, and I would encourage the authors to repeat their analyses based on ZWD series.

An alternative would be to provide strong enough evidence than using ZTD series rather than ZWD series has no impact on the noise analysis results *and the estimated trends*, for all stations. But this would require repeating the whole analysis on ZWD series anyway and complicate the paper unnecessarily. So it's probably best to show results for ZWD series only.

Thank you very much for this comment. Following the recommendations, we re-ran the entire analysis and focus now on the ZWD time series. Indeed, the ZWD data is a meaningful source and can be used in a climatological context. The point we wanted to address in this paper is that the ZTD/ZWD data are not characterized by pure white. This assumption is proven to be incorrect but is in fact widely used to assess trends with their uncertainties.

B) Homogenization

The procedure used by the authors to homogenize their ZTD time series is not very clearly described. The most disturbing aspect to me is that the authors seem to consider the homogenization of station position time series and ZTD time series as the very same problem (e.g. page 4, lines 41-42; page 5, lines 1-8). And it actually seems that the authors used discontinuities identified in their station position time series to homogenize their ZTD time series (page 5, lines 18-20).

I agree that equipment changes causing position discontinuities are likely to induce ZTD discontinuities as well, although this probably needs to be checked in every single case. On the other hand, I see absolutely no reason why earthquakes would cause discontinuities in ZTD time series. In brief, I don't think that the approach of homogenizing

ZTD time series based on discontinuities identified in station position time series is founded.

Another disturbing observation is that many of the estimated discontinuities provided in table S1 appear to be insignificant (or barely significant), even with the very optimistic white-noise-only assumption. This seems to confirm that the homogenization approach used by the authors is not well adapted to their ZTD series. I would therefore recommend that the authors use a different homogenization approach, based on their ZTD (or ZWD) series themselves. It is true that raw ZTD (or

ZWD) series are too scattered to allow visual identification of discontinuities, but the authors could use for that purpose the differences between their series and, e.g., ERAinterim series.

Besides the detection of discontinuities, another indispensable step in time series homogenization / modelling is not mentioned at all in the paper: the detection and removal of outliers. How did the authors "clean" their time series? Based on which assumptions and criteria? This information should appear in the paper.

The homogenisation of ZWD data was re-run once again. We identified the discontinuities and verified the time series basing on (1) the International Terrestrial Reference Frame 2014 (ITRF2014) supplied discontinuity file, (2) earthquakes reported by the USGS Earthquake Hazards Program (https://earthquake.usgs.gov/) and (3) a manual inspection of all the position time series supported by epochs estimated with the Sequential t-test algorithm (STARS; Rodionov, 2004) using a segment length of 100 days and a confidence level of 95%. Our basic homogenisation strategy was to first identify the position offsets and then adopt the related epochs also for our discontinuities modelling of the ZWD time series. This is a very conservative approach as it will lead to the inclusion of many more offset epochs for the ZWD time series than necessary. To reduce the number of these epoch candidates in our basic homogenisation, we firstly excluded those epochs related to earthquakes as it is believed that these do not propagate into ZWD due to the loose constraints in our GPS processing strategy, and secondly applied a threshold criteria. For this we estimated the amplitudes of the ZWD discontinuities with a Least-Squares method and only retained those discontinuities with magnitudes of three times the estimated offset uncertainty. For simplicity during this process we assumed a white noise model but investigated any amplitudes suspected to be biased due to numerical artefacts. A third step in our basic homogenisation strategy included a manual inspection of the ZWD time series. In this way, the number of 530 offsets reported for the set of 120 stations was reduced to 333 discontinuities that were retained for the correction stage. The medium number of applied offsets was then equal to 3 offsets per 1 series. The amplitudes of offsets are juxtaposed in Table 1 in Supplementary materials.

Having corrected the ZWD time series, we estimated trends, their errors and the character of stochastic part once again. Indeed, as you drew our attention, now the values of trend do not differ between WH and AR+WH.

C) Deterministic model

The choice of the authors to estimate periodic signals at 1, 2, 3, 4 cpy and at 1, 2 cpd sounds quite arbitrary, as it is justified with just a figure (Figure 3) showing the PSD of a single station. I think that the least the authors could do would be to show *stacked* periodograms over their 120 stations rather than the periodogram of a single station.

This would either confirm that 4 annual harmonics are enough to remove seasonal variations from *all* the series, or might lead the authors to consider higher annual harmonics.

Regarding diurnal variations, the authors justify their choice of considering only two daily harmonics in the caption of Figure 3: "Remaining peaks in high frequencies were found to be non-significant." A first question that would need to be answered is: which criterion did the authors use to assess the significance of these peaks? But more importantly, does this conclusion hold for all stations? I seriously doubt so when looking at Figure S1, where high spectral peaks really jump out at every daily harmonic from the 1st to the 12th!

The choice of the considered annual and daily harmonics may therefore need to be revised. In any case, it has to be supported by stronger arguments than those presently given in the paper.

As suggested, the stacked PSD is now provided. The stacked power shows evident peaks of 1 and 2 cpy, which were modelled for the whole set of analysed stations. Additionally, periods of 3 and 4 cpy were added only for those stations which they were found significant for. Daily frequency and its overtones were also examined. We analysed the overtones of a day up to 12th harmonic. The medium amplitudes of 3rd-12th overtones was equal to 0.2 mm and was found not to be significant under the level of noise with a medium standard deviation of 40 mm. Therefore, we employed only daily and sub-daily frequencies to be modelled for the entire set of data. On the basis of that, the deterministic model we applied included f=365.25, 182.63, 1 and 0.5 days for all stations plus 121.75 and 91.31 days when significant. Also, this choice of frequencies was also validated with previous studies by e.g. Jin et al. (2009) who also focused on these periods.

Clear significant peaks of daily overtones seen in Figure 3 were found to be present due to the window we used to estimate the Welch periodogram, as their removal with deterministic model did not result in their decrease in PSDs.

D) Choice of the optimal noise model

Since the main aim of the paper is to identify an "optimal" noise model for ZTD time series, its main defect resides in my opinion in the way this "optimal" noise model is actually selected. It seems that the authors' initial intention is to explore the ARFIMA model class (page 9, lines 31-33). Why not (although this is a quite ambitious goal)? But then, only a very specific subset of ARFIMA models are investigated and compared, which are even not of successively increasing complexities. Talking only about the models that include additional white noise (which indeed seems necessary), the authors first consider two of the simplest possible ARFIMA models, namely AR(1)+WH and PL+WH=ARFIMA(0,d,0)+WH. But why isn't the third ARFIMA model with similar complexity (i.e. MA(1)+WH) considered? At the next level of complexity, the authors pick out only two specific models: ARMA(1,1)+WH and the socalled "ARFIMA(1,0)+WH" which I guess actually refers to ARFIMA(1,d,0)+WH with unknown d. But again, why aren't the other models of similar complexity (i.e. AR(2)+WH, MA(2)+WH and ARFIMA(0,d,1)+WH) tested? Last but not least, the authors then completely skip the next level of complexity, pick up a single model at the following level of

complexity (AR(4)+WH) and conclude, based on a sample of only 5 stations (Table 2) – in which AR(4)+WH is actually the preferred model for only 3 stations! – that AR(4)+WH is "the optimal model for ZTD series". I think it's an understatement to say that this conclusion is not supported by the results, for several reasons. First, if the aim is actually to explore the ARFIMA model class, then *all* ARFIMA models of *successively* increasing complexity levels should have been tested. Then, another methodological mistake is to select the most complex tested model as the "optimal" model, since nothing proves that more complex models (e.g. AR(5)+WH) would not have been preferred. The search for the optimal model should actually be made among increasingly complex models, *until the inflection point of the selection criterion (BIC) is reached*.

If such an exhaustive search of the ARFIMA model class turns out not to be practically feasible, the authors may want to restrict their search for an "optimal" model within a smaller class (e.g. AR models of increasing orders). But even within the AR class, it may turn out that the inflection point of the BIC cannot be reached for computational reasons. If so, then no mention could be made of an "optimal" noise model anymore. Finally, another obvious reason why it cannot be concluded that AR(4)+WH is "the optimal model for ZTD series" is that in Table 2, it is the preferred model for only 3 stations out of 5! To be fair, and more informative, the authors should present results for their whole set of 120 stations. It would then likely appear that different stations have different preferred noise models, which might be an interesting result in itself. (Maybe the preferred noise model would depend on the climate zone?). But anyway, if a single preferred model needs to be chosen, then this choice should at least be made based on results for all 120 stations. A last question concerning the noise models tested by the authors is: did they try to consider variable white noise (VW) instead of constant white noise (WH)? Santamaria-Gomez et al. (2011) indeed found VW "significantly superior" to WH when modelling the noise of GPS station position time series.

In the new version of our paper we do not describe the "optimal" noise model. We state that the AR(1)+WH noise model is the preferred one over a pure white noise assumption. We justify this statement with a fraction of autoregressive part which constitutes into the AR(1)+WH combination. For most stations, the AR noise process contributes into AR(1)+WH combination more than 50%, which means that this noise gives a real advantage in description of stochastic properties of ZWD data over WH noise only. What is more, we also found a clear latitude and regional dependence of the fraction of the AR(1) part, which might prove that this noise contributes more into a character of stochastic part for some parts of the world than it does for another.

Naturally enough, one may go further and apply higher orders than 1 of autoregressive part. But, you have to be aware while implementing too high orders of AR part. You might go to infinity, but this is not the point. We added a simple AR(1) noise to white process to prove that even this noise model is preferred over a broadly applied pure white noise. We estimate the significance of trends when stochastic part is characterized by AR(1)+WH noise and show that the errors with WH only are

underestimated. Also, we prove that few trends which were previously analysed in terms of climate studies, should not be taken into account, until the value of trend is larger than the error itself.

E) Structure of the paper

Besides the presented results, the organization of the paper would also need to be revised. Different sections are in particular redundant or mixed with each other. More specifically:

* The first two paragraphs on page 3 (lines 1-16; details about GPS reprocessing and homogenization unnecessary in the introduction) should be merged into Sections 2.1 and 2.2. The beginning of the next paragraph (lines 17-20) should similarly be merged into Section 3.1.

* Instead of introducing trivial equations (Eq. 3 to 5), the beginning of Section 2.3 should rather focus on justifying the chosen deterministic model (cf. comment C).

Similarly, instead of defining different well-known noise models, the second part of Section 2.3 should rather introduce in a clear and precise way the approach that will then be followed to test different noised models and eventually select an "optimal" (or preferred) noise model (cf. comment D).

* The first paragraph on page 8 (lines 1-7; justification of the adopted deterministic model) should be revised according to comment C and moved into Section 2.3.

* Most of the results discussed in the rest of Section 3.1 seem to be repetitions of the findings of Jin et al. (2007; 2008). I don't think it would harm the paper to keep a short discussion about the estimated (semi-)annual and (semi-)diurnal signals, as well as Figures 4 to 7 (although they could be moved to the supplementary material). But since this is not the main subject of the paper and no new conclusions are reached compared to the results of Jin et al. (2007; 2008), I think that Section 3.1 could be shortened.

* The beginning of Section 3.2 (page 9, lines 9-37; details about the tested noise models) should be moved into Section 2.3. The rest of Section 2.3 will probably need to be entirely re-written according to comment D.

* Instead of real "Discussion" and "Conclusions", Sections 4 and 5 basically consist of repetitions of facts stated earlier in the paper and are almost entirely redundant with each other. Those two sections will also likely need to be entirely re-written (and merged into a single section) and should really focus on the main findings of the study, i.e. preferred noise model(s) and consequences for trends and their uncertainties.

Thank you very much for your valuable comments. We included them within the text. We also combined the Discussion and Conclusions parts into one. We have re-written the part you suggested, shortened the part about seasonal model and focused more on the noise properties itself. We added additional figure of fraction of AR(1) noise which means how much the AR(1) part contributes into

AR(1)+WH noise. This might help the reader to understand how important the AR(1) noise is to properly describe the character of ZWD data.

F) Language

The paper contains lots of English mistakes, but more importantly, many imprecise and unclear formulations. Some are marked in the attached annotated manuscript. But the paper will nevertheless require a very careful re-reading before being re-submitted.

We re-read and re-wrote the paper trying to make it more clear.