Dear Reviewer #1,

We sincerely appreciate the time and effort you have devoted to reviewing our manuscript. The constructive comments and valuable suggestions have significantly helped us improve the clarity, rigor, and overall quality of our manuscript.

We have carefully revised the manuscript according to your comments. All changes in the revised manuscript have been highlighted in yellow for easy identification. A detailed point-by-point response letter is provided below, with the reply in blue.

Best regards,
The authors

Response to Reviewer #1 (egusphere-2026-5501)

The manuscript describes a new methodology to convert GNSS ZTD estimates to precipitable water vapour for real-time applications using AI weather forecast foundation models.

General comments

I really like the basic idea or concept of the study, which is also rather innovative. However, the presentation of the study is rather messy so that readers not familiar with GNSS PWV retrievals – which are to be expected for the ACP journal – will quickly loose the bottom line of your study. Now, you move back and forward quite frequently in the manuscript, and on top of that, use too often repetitions. To my opinion, a clear introduction/methodology section should include the following elements (in this order of appearance):

- The importance of PWV in meteorology (nowcasting) and Why is PWV nowcasting/forecasting important?
- How PWV is measured or obtained from models (your previous work Ding et al. 2024, 2025&, 2025b) in the context of real-time forecast (i.e. nowcasting).
- The role of GNSS in PWV retrieval and which auxiliary meteorological parameters are needed (basically surface pressure and humidity weighted mean temperature).
- The different approaches to calculate ZHD and Tm for nowcasting purposes
    1. Ideal situation: surface data and vertical PTU profiles available (from co-located observations or NWP forecast models)
    2. GPT3, Saastamoinen model
    3. Empirical relations between surface temperature and weighted mean temperature (e.g. Bevis et al., 1992).

Response: Thanks for your positive evaluation of the study concept and your detailed structural advice. We have reorganized the introduction and methodology following your proposed sequence, and cut redundant repetitions. The revised text sequentially introduces the importance of PWV nowcasting, existing PWV retrieval pathways including our prior work, the GNSS PWV principle and required auxiliary parameters, and the different ZHD/Tm) calculation approaches (ideal vertical PTU integration, GPT3, Saastamoinen, and Bevis). This restructure avoids confusing narrative jumps and makes the manuscript accessible to non-

expert readers as expected for ACP.

For instance, in your introduction, you are mentioning different errors or inaccuracies of those 3 approaches (the number of approaches is not necessarily fixed, just to give you an idea), but those 3 different approaches have not been explained before. You are speaking about "an empirical model", but without a clear meaning.
Response: In the revised introduction, we have clearly defined the PWV calculation approaches and clarify the exact definition of vertically empirical parameterizations before discussing their errors. We avoid ambiguous use of the term "empirical model" and distinguish them from climatological models like GPT3.

Also, to my opinion, in your study, the new element is that you are using vertical PTU profiles from AI weather forecast foundation models (instead those obtained from physical NWP models). So, you are somewhere in approach 1. from the here above summed up approaches. The most interesting question is if the AI weather forecast models give a more reliable PWV nowcast than physical NWP PWV nowcast? In this regard, does it then make sense to compare the PWV output of your approach with the cases (2. and 3.) where the vertical profile information is approximated roughly with only the knowledge of a lat-lon dependent and time-invariant climatology (e.g. GPT3) or only surface measurements at the site? They have been quite numerous studies that have shown that the accuracy of the PWV estimation is higher when using the vertical PTU information from physical NWP reanalyses or forecasts compared to using the other approaches. In this sense, the results (Fig. 3 and section 3.1) that the "integral" methods have smaller RMSEs for the ZHD and Tm than the Saastomoinen and Bevis approximations are quite trivial and not new at all. On the contrary, it is rather worrying that the Integral I13 method has a much higher ZHD RMSE than the Saastamoinen approach, and a clear explanation should be given here (13 vertical models should always be better than only surface values used, unless the 13 vertical models provide nonsense).
So, I would like to ask the authors to much clearer describe the state-of-the-art GNSS PWV nowcasting, better explain the novelty of this study in this context and focus on non-trivial analyses or results.
Response: Thanks for this detailed comment. We have rewritten the introduction to better describe GNSS PWV nowcasting, explicitly state the novelty of our study, and focus on non-trivial results.

We acknowledge the whether AI foundation models produce more reliable nowcasts than physical NWP models. Supporting references and discussion are now included in the revised introduction.

Our method leverages the precise vertical profiles from AI foundation models to compute integrated ZHD and Tm, avoiding errors from vertically empirical parameterizations such as Bevis and Saastamoinen. Here "empirical models" denotes vertical parameterizations for profile integration, distinct from climatological models like GPT3, which apply DOY-dependent temporal modelling on pre-integrated Tm. This distinction is clarified in the manuscript.

Regarding Fig. 3 and Section 3.1, comparisons between the integration method and Saastamoinen/Bevis have indeed appeared in some existing studies, and the results in our

manuscript are largely expected. However, a ZHD/Tm comparison specifically targeting the number of pressure levels in AI foundation models, i.e., the 13-, 25-, and 37-level comparison in this manuscript has not yet been reported. Such a comparison is necessary for GNSS PWV research based on foundation models, and it provides an important reference for assessing whether newly emerging AI models are fundamentally suitable for real-time GNSS PWV retrieval.

For the unexpected finding that I13 integral ZHD RMSE exceeds Saastamoinen: the 13-level data are not directly integrated. The coarse profile is first interpolated/extrapolated to a finer grid before integration, so the comparison is between interpolation and vertical empirical model, not "with versus without vertical information". ZHD is highly sensitive to pressure (~2.3 mm/hPa). The 13-level configuration has a large gap between 1000 hPa and 925 hPa, and interpolation errors in different pressure levels lead to large ZHD biases. Saastamoinen only needs surface pressure and is anchored at the surface state. Therefore, the degraded I13 result stems from coarse vertical discretization failing to constrain surface pressure, rather than the uselessness of vertical T/RH data. This explanation has been added.

I'm also quite surprised that you could only identify 93 radiosondes sites collocated with a GNSS sites in 2022 that fulfill the rather loose criteria of (1) records available between the surface and 300 hPa, (2) at least five records of standard pressure levels <= 1000 hPa, (3) gaps between pressure levels being less than 200 hPa. Nowadays (and also already in 2022), radiosondes provide measurements every 1 or 2s (with an average ascent rate around 5-6 m/s), so these criteria are easily met for close to 100% of the radiosonde sites. Have you considered other radiosonde archives (e.g. the GTS, ECMWF) than IGRA for obtaining more high vertical resolution radiosonde data? How many radiosondes sites collocated with a GNSS receiver in 2022 actually failed to your radiosonde vertical resolution criteria?
Response: Thank you for this careful question. We clarify two key points here. First, the vertical quality criteria listed in Section 2.3 apply to individual radiosonde profiles, while the final 93 samples are GNSS-radiosonde station pairs for validation. Beyond the vertical screening, valid pairs require strict spatial-temporal matching. It is common for missing radiosonde records caused by equipment malfunctions or adverse weather conditions. While GNSS records are unaffected by weather, most stations still have varying degrees of missing data records. It is challenging to find stations with relatively complete, year-round synchronous observations and sufficiently small inter-station distance. We discard pairs with large temporal offsets between observation epochs or horizontal separation greater than 40 km, and this matching constraint is the dominant factor limiting the sample size, even if most individual radiosonde profiles may satisfy the vertical requirements.

Second, we have considered alternative archives including GTS and ECMWF radiosonde BUFR datasets. However, IGRA is the most widely adopted, quality-controlled unified global radiosonde archive; it integrates multiple data streams including GTS reports. We adopted IGRA for this work because of its consistent quality control and unified formatting, which facilitates automated profile parsing and pressure-level screening. The future work could incorporate raw BUFR archives to expand the sample size and access high-rate profile data.

Another general comment is that the English of the manuscript should really be improved.

On the other hands, it shows that the manuscript is not generated by AI to a large extent, for which I would really praise the authors. But, as there are co-authors on the manuscript which are quite fluent in writing scientific papers in English, there is no excuse for doing language editing yourself before submitting a next version.
Response: Thank you for this comment. The revised version has been thoroughly language-edited to improve grammar, clarity, and scientific style. We also appreciate your positive remark.

Specific comments
lines 42-66: for non-expert readers, you are using terms that are not very familiar and should be explained before discussing their associated errors: ZHD, ZTD, ZWD, weighted mean temperature (Tm) --> which weights, an "empirical model", Global Pressure and Temperature 3 (GPT3), atmospheric vertical profile data (which variables?) for GNSS PWV conversion. A serious re-organization of sections 1 and 2.1 is necessary!
Response: Thanks for the comment. We have reorganized Sections 1 and 2.1 following the general and specific comments. Key terms including ZTD, ZHD, ZWD, Tm (with its weighting rule explained), empirical models, GPT3, and atmospheric vertical profile data (with relevant variables specified) are now defined before the discussion of their errors.

line 82: give the reference of the comprehensive review by Percy Liang of Stanford University
Response: The review we refer to is "On the Opportunities and Risks of Foundation Models" (Bommasani et al., 2021). This is a large collective report with >100 authors; Percy Liang led the report's framing and writing effort.

We have revised the sentence to remove ambiguity and clarify that Percy Liang led this collective report, while keeping the correct citation (Bommasani et al., 2021).

The revised text: "*The term foundation model describes any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks, as outlined in the landmark collective review (Bommasani et al., 2021), which was led by Percy Liang of Stanford University together with over 100 leading AI scholars.*".

line 96: give a reference for the < 1-2 mm ZHDs obtained with the Saastamoinen model. And is this number the uncertainty w.r.t. the ZHD obtained by integrating the refractive index (Eq. 2) in the zenith direction? Please explain.
Response: We appreciate this careful comment.

First, the value < 1-2 mm describes the estimation error of the Saastamoinen ZHD, not the absolute magnitude of ZHD output by the model. This error statistic is computed with respect to the reference ZHD derived from vertical integration of atmospheric refractivity along the zenith direction, exactly as the reviewer pointed out (the integral form equivalent to our Eq. 2).

Second, in Qin et al. (2026), the reference ZHD is calculated from ERA5 ray-tracing, which is implemented by integrating refractivity profiles. Their global assessment demonstrates that for altitudes 6-30 km near tropopause heights, the Saastamoinen model

achieves ZHD RMSE better than 1.1 mm, well within the <1–2 mm range. Mendes and Pavlis (2004) reported millimeter-level residuals of Saastamoinen against radiosonde ray-tracing (refractivity integration) for total zenith delay at optical wavelengths.

We have revised the sentence on line 96 to eliminate ambiguity and explicitly clarify that the 1–2 mm quantity denotes model estimation error relative to refractivity-integrated reference ZHD.

The revised text: "*The Saastamoinen model, which primarily uses surface pressure together with station latitude and height as inputs, achieves zenith hydrostatic delay (ZHD) errors smaller than 1–2 mm relative to the reference ZHD obtained by vertical integration of atmospheric refractivity, and is widely adopted for its simplicity and effectiveness (Qin et al., 2026; Mendes and Pavlis, 2004).*".

lines 107-121: you are only discussing the very simple "empirical" models based on a linear relationship between Tm and Ts. What about the empirical models based on a climatology and give a value for each DOY, lat and lon grid (e.g. GPT3). These should be mentioned here as well, together with the known uncertainties of those empirical models (linear Ts-Tm relationship + climatological based) for retrieving Tm as opposed to obtaining Tm by the complete integration over the vertical T and RH profiles (Eq. 6).
Response: Thank you for this thoughtful comment. The "empirical models" discussed in this manuscript refer to vertically empirical parameterizations (e.g., Bevis Tm and Saastamoinen ZHD), which are simplified alternatives to the full vertical integration of temperature and humidity profiles. This definition focuses on vertical approximation rather than temporal climatological modelling. Tm results from GPT3 and similar climatology models are built by further applying temporal modelling (DOY-dependent climatology) upon the integrated Tm values. To avoid confusion, we have clarified this distinction in the manuscript on climatology-based models such as GPT3 and their characteristics in the revised Introduction.

lines 162-164: this has already been mentioned in lines 142-144.
Response: Thank you for this careful check. We have removed the text in Lines 162–164.

line 170: this is the first time that you mention traditional empirical models that only need spatial coordinates and time. This should be mentioned and explained after Eq. 6.
Response: Thanks. We have added explanation for the empirical models in the revised Introduction as the suggestion of general comment.

lines 176-180: this should be highlighted more in the introduction, as I proposed in my general comment. The question to answer in the introduction is also: what is the added use of PWV forecast based on real GNSS ZTD data compared to pure NWM PWV forecasts?
Response: We have emphasized this point in the revised introduction following your general comment.

The revised text: "*Another PWV source is direct integration of numerical weather prediction (NWP) fields, including AI-based NWP products such as those in our previous work (Ding et al., 2024, 2025a, 2025b). These are assimilated or forecast grid-point states that mix model physics, initial conditions,*

assimilation increments, and grid averaging. They are not observations and represent a different spatial volume, and are therefore treated separately from the observational PWV considered in this study."

Fig. 2 and lines 188-192: you should give much more details about this "simulation test". What are the global one-year average values of ZTD, ZHD, Tm used as initial and reference values? How representative are the -20 to 20 mm and -5 K to 5 K bias ranges adopted for ZHD and Tm respectively? Are these typical, realistic, uncertainty ranges for those variables? Also, be careful with the terminology: do you really want to discuss the impact of ZHD and Tm biases or rather the impact of the expected accuracies of these variables?

Response: Thanks for the comment. We have added more details for the simulation test, including the global one-year mean ZHD and Tm baseline values. The ±20 mm and ±5 K perturbation ranges are selected to cover typical realistic uncertainty bounds for ZHD and Tm. We keep the term "bias", as this experiment applies controlled artificial biases to investigate PWV sensitivity. The revised text is provided blow.

The revised text:

"*Figure 2 presents a simulation test to quantify how PWV responds to biases in ZHD and Tm. For this test, we adopt the global one-year mean ZHD and Tm (ZHD = 2.08 m, Tm = 270 K) as the baseline reference values and compute PWV following Equations (1), (4) and (5). We impose controlled biases ranging from −20 mm to 20 mm on ZHD and −5 K to 5 K on Tm; these perturbation bounds are selected to cover the typical realistic uncertainty ranges reported for ZHD and Tm. The left panel shows PWV bias under coexisting ZHD and Tm biases, while the right panel isolates their individual contributions by setting the bias of one variable to zero when the other is perturbed. The results show that a ZHD bias of ±6.12 mm or a Tm bias of ±8.97 K separately produces a 1 mm bias in PWV.*".

lines 194-201: as mentioned as a general comment: I am quite surprised that with those loose criteria on missing data values for radiosonde data, you only end up with 93 GNSS-RS co-locations in 2022.

Response: These criteria in Section 2.3 are for individual radiosonde (RS) stations, whereas the 93 records are GNSS-RS co-located pairs in Section 3.1. Valid pairs require further space-time matching; pairs with large gaps in common observation epoch or GNSS-RS distances >40 km are removed. We have clarified this difference in the revised text.

line 207-208: you should mention and refer explicitly the TU Wien grid model for ZHD and Tm here. To my opinion, this grid model should already have been described earlier (in the same paragraph as lines 107-121), see my earlier comment.

Response: Thank you for this suggestion. We consider this separation reasonable. The text in lines 107–121 introduces the general workflow for deriving PWV from GNSS ZTD, which is a generic approach applicable to different ZHD and Tm models. By contrast, lines 207–208 are placed in the data description of Section 2.3, to specify that the NGL GNSS products used in this study adopt the TU Wien grid model for ZHD and Tm. The two parts serve distinct purposes. We have confirmed that the TU Wien grid model and its reference are clearly stated

at lines 207–208.

lines 203-217: these paragraphs are a bit misleading. The most important message here is that you use the ZTD products from NGL. But you start this discussion by mentioning that NGL also provides ZTD, ZWD, gradients and PWV, and that ZHD and Tm from the TU Wien model, which may give the impression that you are using those as well. The last paragraph is also not very clear.
Response: Thank you for this careful comment. We have restructured this paragraph to state upfront that only NGL ZTD products are used in this study, while the other NGL tropospheric products (ZWD, PWV, gradients) are not utilized. We also revised the subsequent note to improve clarity. The revised text is shown below.

"*The Nevada Geodetic Laboratory (NGL) provides tropospheric delay products from over 20,000 GNSS stations worldwide, including ZTD, ZWD, gradients and PWV at 5-minute intervals. Only the NGL ZTD product is used in this study. The dataset spans 1994 to the present and its accuracy is comparable to IGS products, making it a reliable reference for evaluating tropospheric models (Ding and Chen, 2020). Among these more than 20,000 stations, only 8320 have a data completeness higher than 95% for 2022; hence, only these stations are adopted in our experiments. Note that NGL derives its ZHD and Tm from the TU Wien grid model, which introduces model errors into NGL's own ZWD and PWV products (Yuan et al., 2021; Ding et al., 2023). This is the key reason why we do not use NGL's ZWD or PWV.*

As an additional note: NGL ZTD itself is not produced for real-time applications. We select it for our experiments because our objective is to develop a global method to convert GNSS ZTD into high-accuracy real-time PWV, rather than generating operational real-time GNSS PWV products. Thus, the non-real-time nature of the NGL ZTD does not affect our conclusions. The accuracy metrics reported here are relative values, referenced against retrospective calculations using ERA5. We also note that users may adopt forecast ZHD during GNSS processing, which we recommend. Since ZHD error is absorbed into ZWD and has negligible influence on ZTD (Ding et al., 2023), this alternative workflow is functionally equivalent to the scheme illustrated in Figure 1.".

as mentioned in the general comment: I'm not a big fan of your section 3.1. You should first specify much clearer what RMSEs for ZHD and Tm are considered here, by making the link with or reference to the equations (e.g. 2 and 6) earlier in the manuscript. I have several other concerns here.
Response: Thanks for the comment. We have revised Introduction and Section 2.1 according to your general comments. We now clearly define the ZHD and Tm RMSE metrics at the start of this section and connect them for better clarity. We welcome your additional specific concerns.

line 225: the upper panels of Fig. 3 are not scatter plots, but maps. And here you talking about the 2600 radiosondes sites worldwide, but apparently you only have 93 GNSS-radiosonde co-locations.
Response: Thanks for the comment. We have updated Fig. 3 by replacing scatter plots with map plots. We would like to clarify the difference between the two groups of panels. The upper

panels evaluate the **precision** of empirical models using 13-, 25-, and 37-level ERA5 data, with the integral method taken as the reference. This assessment does not rely on radiosonde observations, so it uses the global set of 2600 radiosonde sites and is not constrained by the availability of valid GNSS-RS paired records. In contrast, the bottom panels assess **accuracy** against real radiosonde measurements, which is why we apply strict matching criteria and only retain the 93 qualified GNSS-radiosonde co-located pairs. We have clarified this difference in the manuscript.

line 226 and lines 232-233: how have the pressure levels of the 13, 25, and 37-level ERA5 data been selected? Randomly? Of course, the lower levels will contain the most important information for the Tm and ZHD retrieval.
Response: No, they were not selected randomly. Each configuration follows a documented rationale: (1) 13 levels are the standard WeatherBench subset (50–1000 hPa), established as the most impactful levels for almost all data-driven weather prediction and used by Pangu-Weather, GraphCast, FengWu and other weather forecast foundation models. (2) 25 levels correspond to the ECMWF HRES 25-level set, adopted by the VMF3_FC products. (3) 37 levels are the native ERA5 pressure-level set, used as the high-fidelity reference.

We agree that the lower levels carry the most information for Tm and ZHD. The 13/25/37 configurations were chosen precisely to quantify how much vertical resolution is needed before ZHD/Tm accuracy saturates, and the evidence indicates that 25 levels are sufficient, while 13 levels may not.

lines 230-237: aren't those rather trivial and quite logical results? Give references to earlier work that have obtained the same results. And shouldn't you include more sophisticated Tm models than the simple Tm-Ts linear regression models in your analysis here? I think you should.
Response: Thanks for the comment. While the underlying physical reasoning is intuitive, we are not aware of prior studies that specifically compare the performance of ZHD and Tm derived from 13-, 25-, and 37-level ERA5 inputs, which correspond to the native vertical grids of Pangu-Weather, VMF3-FC and GraphCast/FengWu. This specific vertical-level comparison has not been reported previously, which is why we do not have directly matching references.

Regarding the selection of Tm models: we have already explained in Section 2.1 that Bevis is chosen as a representative simple Tm–Ts linear model. More sophisticated regional, seasonal, gridded or neural-network-based Tm models (Yao et al., 2012, 2013, 2014; Huang et al., 2019, 2023; Ding, 2018) can achieve higher accuracy, yet many of them are not publicly available or cannot be applied under the time window of our experiment. Our goal here is to examine how vertical resolution (number of PTU pressure levels) affects the error propagation of a widely used, easy to implement empirical formulation, rather than to carry out a comprehensive benchmark of all advanced Tm models. We have reinforced this explanation in the manuscript to clarify our experimental scope.

lines 240-242: I really cannot understand that making use of T and RH data at 13 pressure levels give a worse ZHD estimate than using the Saastamoinen approach, that includes no

vertical information at all, only surface information. This should be checked and clearly explained.

Response: Thank you for raising this point. This result is indeed somewhat surprising at first glance and deserves a clear explanation. We have checked this and revised it accordingly.

The key point is that 13-level data are not directly used to integrate ZHD from T and RH profiles. Given its coarse vertical resolution that prevents reliable direct integration, the 13-level setup first interpolates and extrapolates the coarse profile to a finer vertical grid prior to integration. Hence, the comparison is essentially between interpolation approaches, not a comparison of "with vertical information" versus "without vertical information."

ZHD is dominated by surface pressure (roughly 2.3 mm per hPa). In the 13-level set, the lowest layers are 1000 and 925 hPa with no intermediate level, so the finer vertical pressure must be interpolated or extrapolated across a large gap. A few hPa of error in this step translates directly into centimeter-level ZHD errors. In contrast, the Saastamoinen approach was itself derived from radiosonde vertical profiles and requires only an accurate surface pressure, which places it on the correct starting point. Therefore, the 13-level ZHD is worse not because vertical T/RH information is useless, but because a coarse profile cannot reliably constrain the surface/near-surface pressure that ZHD is most sensitive to.

We have added this explanation to the revised manuscript.

"*This unexpected finding arises because the 13-level data are not directly integrated. The coarse profile is first interpolated/extrapolated to a finer vertical grid before integration. Since ZHD is dominated by near-surface pressure, interpolation across the large 1000–925 hPa gap introduces large errors. The Saastamoinen approach, by contrast, requires only an accurate surface pressure and is therefore not penalized by coarse vertical resolution.*"

line 254: the aim of the paper is real-time provision of PWV nowcast/forecast. So, how relevant are the 5 to 15 day forecast results with this goal? Please comment.

Response: Thanks for the comment. Yes, our main target is real-time PWV nowcast/ forecast. The 5–15 day results are not for regular real-time operation. They help examine the error growth and performance limits of different AI models, and also evaluate their fault tolerance: unexpected delays in initial fields in real operations may force us to adopt longer forecast lead times. We have added this clarification in the text.

The added text: "*These over 5-day forecasts are not designed for routine real-time PWV applications. They are used to characterize error evolution with lead time and also assess model fault tolerance, as operational delays in initial fields may require longer forecast horizons in practice.*".

line 263: as you don't use the vertical profile information from those 489 radiosonde stations, the sample of ground-based sites could be extended with surface meteorological stations (e.g. automatic weather stations), so that a much larger sample could be obtained.

Response: Thanks for the suggestion. Radiosonde stations are ground-based sites, but we use them because their surface observations are inherently time-matched with the profile-derived reference PWV. Automatic weather stations provide surface data but generally lack the collocated vertical profile reference PWV needed for validation. Thus, we keep radiosonde

stations for this comparison.

lines 265-268: you still haven't described the GPT3 model in great detail at this point. Should have been done long before. Mention here for which parameters (ZHD, Tm, other?) you have used the GPT3 empirical model here. Does your analysis imply that the GPT3 model for calculating Tm is worse than the Bevis Tm-Ts relationship?
Response: Thank you for the comment. We would like to clarify that GPT3 and the Bevis model are not directly comparable, as they belong to different modelling frameworks. GPT3 performs temporal climatological modelling on pre-integrated Tm and provides gridded parameters as a function of DOY, latitude and longitude, Tm = GPT3 (DOY, Lat, Lon). By contrast, the Bevis model is a vertically empirical parameterization that describes the statistical relation between surface temperature Ts and Tm, Tm = Bevis (Ts). GPT3 does not build a Tm–Ts relationship, and Bevis cannot output Tm without surface temperature input. We have clarified which parameters (ZHD and Tm) are obtained from GPT3 and added this distinction in the manuscript to avoid an inappropriate direct comparison between the two models.

The revised text: "*Note that we also include results from the climatological model GPT3 (Landskron and Böhm, 2018) in the comparison, which acts as a static baseline to illustrate the accuracy degradation of AI foundation models with forecast lead time. GPT3 derives ZHD and Tm from DOY, latitude and longitude, and follows a different modelling paradigm from the Bevis Tm–Ts formulation; we do not aim to rank their performance.*"

line 284: the term "trend" is somewhat misleading in a meteorological context. Please use another word.
Response: We have replaced the term "trends" with "patterns" to avoid ambiguity in the meteorological context.

line 298: this has been reported already several times. Drop.
Response: It has been removed.

line 296: are the 8320 NGL GNSS sites used in Fig. 5 identical to the "more than 8000 GNSS stations" mentioned on line 257 in Fig. 4? Be consistent in the amount, please! What are the "requirements" mentioned in line 297 that defined the sample of 8320 stations?
Response: Yes, these two sets of stations are identical, we have revised the "more than 8000" to "8320". Only 8320 NGL GNSS sites satisfy this criterion. The "requirements" refer to the criterion that data completeness exceeded 95% in 2022. We have added this description in Section 2.3 where the GNSS data are introduced.

lines 306-307: this conclusion is hard to draw from the maps and colours in Fig. 5. Please consider another illustration to make this point clear.
Response: We appreciate this comment. We have revised the statement to be less absolute and use "generally slightly outperforms" instead of claiming consistent superiority for all cases, as FengWu's advantage over GraphCast is modest and the two models are close in some

height intervals.

The revised text: "*Among the three AI solutions, GraphCast and FengWu show substantially lower PWV RMSE than Pangu-Weather; overall, FengWu generally slightly outperforms GraphCast across the tested forecast durations and height bins.*".

lines 310-319: I think that the discussion of the spatial differences of the RMSE errors deserves much more in-depth analysis than is currently the case. Now, a lot of hypothesis are put forward, but this is just guessing and there are not very convincing (e.g. the reason why the GNSS PWV accuracies degrade at the edge of the continents). Also the finding that the RMSEs decrease at higher latitudes and higher altitudes should be evaluated by also considering relative errors as an accuracy metric, next to the absolute RMSE. Then, a more fair evaluation of the impact of the PWV value itself on the accuracy would be possible.
Response: Thank you this constructive comment. We agree that the original discussion was too speculative, and we have substantially rewritten lines 310–319.

First, we revised the unsupported hypotheses with softened causal wording: coastal/edge degradation is attributed mainly to representativeness error (GNSS point measurements over land vs. land–sea mixed grid cells), and high-latitude degradation to limited model-derived ZHD skill.

Second, the RMSE is aggregated over the full year per lead time. A relative error cannot be computed from these aggregates with all the original per-time values by mean value, and an instantaneous relative error would inject PWV diurnal/seasonal cycles and station clustering, confounding the real accuracy.

Third, since all atmospheric variables were normalized prior to training by these AI foundation models, the models' errors are structured in normalized space. This largely suppresses the spatial contrast that a relative-error metric would need to expose, limiting its usefulness for analyzing spatial characteristics.

We therefore did not add relative RMSE as a primary metric, but explicitly note that lower absolute RMSE at higher latitude/altitude is consistent with lower background PWV and does not imply higher proportional accuracy. This is now flagged as a limitation.

The revised text: "*Moreover, the GNSS PWV accuracies degrade faster at the edge of continents, especially in narrow peninsulas or archipelagic regions, such as the Antarctic Peninsula, the Aleutian Islands, and the Newfoundland Islands region of Canandaigua, among others. The former may be caused by the lower accuracy of ZHD in high latitude regions (see Figure 3), while the latter may be caused by the sharp fluctuation of water vapor and the decrease in the accuracy of Tm. For the dependence of PWV RMSE on station height, overall, the absolute PWV RMSE decreases with increasing height, which is consistent with the fact that background PWV decreases with height; however, this does not necessarily indicate higher proportional accuracy. This is likely related to representativeness error from land–sea contrast and grid averaging, rather than a definitive attribution. In addition, the difference between the results of Pangu-Weather and the other two solutions becomes smaller as the forecast length increases, as well as the height increases. In contrast, the difference between the results of GraphCast and FengWu solutions gets larger as the height increases.*".