We thank all referees for their expert comments and suggestions. To answer them, we have produced several changes to the manuscript and also enlarged our co-author team by Dr. Kuchař, who computed the residual tendencies for specific dynamics simulations, as an upper estimate for the nudging strength.

This document contains the original comments from all referees (in italics), along with our responses (in blue). For Ref. 2, we're also including the additional conversation from EGUsphere between Ref. 2 (in red) and the author team (in green).

Referee 1

The manuscript by Zajicek et al. presents Contrasting the gravity wave forcing between nudged and free-running models and reanalysis. I would like to thank the authors for this study. It represents a meaningful contribution toward understanding parameterization schemes and variability in Earth system models. This study compares gravity wave drag in high-resolution reanalysis with parameterized gravity wave drag in three climate model simulations, two nudged and one free-running. It finds that the models and reanalysis look quite similar in overall distribution and extremes, but their dynamical impact differs in the lower stratosphere. The key result is that in the models, parameterized gravity wave drag is strongly linked to Rossby wave forcing, while in ERA5 reanalysis that link is not seen. This suggests that climate models may represent the dynamical role of gravity waves differently from the real atmosphere when gravity waves are explicitly resolved. Overall, the manuscript is in good shape, and I recommend acceptance after minor revision. However, there are several minor issues and points of clarification that, if addressed, would improve the readability, consistency, and overall quality of the paper.

We would like to thank the referee for their helpful comments and suggestions that helped us to improve the quality of our manuscript.

Major Comments:

1. The motivation for using two nudged simulations and one free-running simulation is not sufficiently clear in the Introduction. This should be better explained and integrated into the overall flow of the section.

Thank you very much for the comment. We agree and expand the sentence originally centered around L155 to a full paragraph:

The goal of this study is to validate the dynamical interactions of parameterized GWD in the valve layer in global climate models using a quasi-observational dataset. For this, we analyse the dynamics associated with the resolved GW drag in ERA5 (Procházková et al., 2025b), as the best proxy for the real atmosphere GW forcing, and confront it with dynamical effects of parameterized GWs in nudged and transient climate model simulations....

->

The goal of this study is to compare the dynamical interactions of parameterized GWD in the valve layer in global climate models with a GW resolving dataset. In the previous literature, parameterized GWD effects in climate models were analyzed using either specified dynamics simulations (e.g. Kuchař et al., 2020; Šácha et al., 2021) ensuring realistic circulation backgrounds or free running model simulations (e.g. Eichinger et al., 2020; Hájková et al. 2024) enabling a cleaner attribution of dynamical effects. Here we aim to bridge the two approaches by analyzing an ensemble of opportunity consisting of two specified dynamics simulations of two different state-of-the-art climate models (with different parameterization schemes and nudging procedures), and a transient simulation of one of the models facilitating the assessment of the difference of responses between freely evolving and specified dynamics. The climate model results are then compared with the resolved GW drag in ERA5 (Procházková et al., 2025b), as the best proxy for the real atmosphere GW forcing, and the associated dynamical interactions are contrasted across the datasets.

The paper is organized as follows....

2. The manuscript would benefit from clearer presentation of the model descriptions. Currently, the information is somewhat scattered. Consider adding a dedicated subsection in the methodology specifically for model descriptions. Also, summarizing key model details in a table for clarity and readability.

Thank you for this suggestion. We also realize that the model description could be expanded and presented in a clearer manner. The model description section and summarizing table with model details will be included in the revised manuscript.

Minor Comments:

1. Abstract: Consider toning down the sentence starting with "However... non-existent," as it may be overstated. Include the names of the three Earth system models used in the study.

Thank you for the comment. We revise the text in the abstract accordingly:

However, any constraints on the parameterized gravity wave effects, especially on global spatial and climatological temporal scales, are practically non-existent.

->

However, constraining the parameterized gravity wave effects, especially on global spatial and climatological temporal scales, is extremely challenging.

Regarding naming the ESMs explicitly in the abstract, we decided not to do this due to the word limit on abstracts, but we have rewritten this particular sentence accordingly:

Here, we compare the recently published resolved gravity wave drag estimates and effects from the current generation high-resolution reanalysis with the climatology and dynamics in three different Earth system model simulations.

->

Here we compare the recently published resolved gravity wave (GW) drag estimates and effects from the current generation high-resolution reanalysis with the climatology and dynamics of the parameterized GW drag from two different ESMs and simulation types.

2. Introduction: Around line 30, the paragraph would benefit from additional citations to better support the statements made.

Thank you very much for spotting this. Indeed, the references to the existing GW parameterization schemes somehow slipped from the original text. This is now revised.

3. Terminology and consistency: Ensure consistent use of model naming (e.g., “CMAM-SD” vs. “CMAM-sd”). Clarify terminology related to seasons: “NH winter” and “boreal winter” appear to be used interchangeably; this should be stated clearly and used consistently throughout the manuscript. Specify explicitly which months are included in the analysis of NH winter.

Thank you for pointing out the inconsistency. The terminology has been standardized. “Boreal winter” has been specified in both the data section and the image captions.

4. Clarity and flow: Around line 100, the sudden mention of NH hotspots is somewhat abrupt. Consider better connecting this to Section 2.2 or providing additional context.

Thank you for spotting this. We move the original sentence from L102-103 to the Section 2.2. The paragraph around line 100 also changes to reflect comments from other reviewers.

->

.... Himalayas (HI, 70–102.5∘E and 20–40∘N), West America (WA, 235–257.5∘E and 27.5–52∘N) and East Asia (EA, 110–145∘E and 30–48∘N). These regions are among the main orographic hotspots of Northern Hemisphere, and it can be assumed that orographic GWs will dominate the resolved GW modes here in ERA5. The representative....

5. Figures: Figure captions would benefit from more detailed descriptions to help readers interpret the results without referring extensively to the main text.

For each relevant figures in the paper, we have expanded the description to include the following:

GWD for ERA5 is based on resolved GWs (see (Procházková et al., 2025a)). GWD for ESCiMo, ESCiMo-SD and CMAM-SD simulations is a sum of parameterized orographic and non-orographic GWD from the models.

6. Conclusions: These sections are relatively long. Consider streamlining and optimizing them to improve readability and focus. Alternatively, consider adding the section with the name "Discussion" and arranging some of the text there from the "Conclusion" section.

Thank you very much for the suggestion. As requested, we have modified and edited parts of the text in the section to improve its readability and focus. We decided not to add a new section, because we prefer to follow the recommended manuscript composition for WCD and also we try to follow the guidelines from a related Copernicus journal on the structure of the Conclusions. However, we moved part of the text regarding ERA5 to Methodology section making the Conclusions shorter.

https://www.weather-climate-dynamics.net/submission.html

https://www.atmospheric-chemistry-and-physics.net/policies/guidelines_for_authors.html

7. Citation practice: Around line 250, the use of "personal communication with Prof. Masatomo Fujiwara" should be reconsidered. If possible, replace it with a published reference or clarify its necessity.

We have followed your suggestion and based on a published dataset of monthly mean TEM tendencies (Serva, 2022) we have estimated the upper bound of unresolved GW contributions as half of the resolved GWD in the valve layer (between 100 and 50 hPa), which is documented in a new figure A2 in the Appendix. The sentence no longer refers to the *"personal communication with Prof. Masatomo Fujiwara"*.

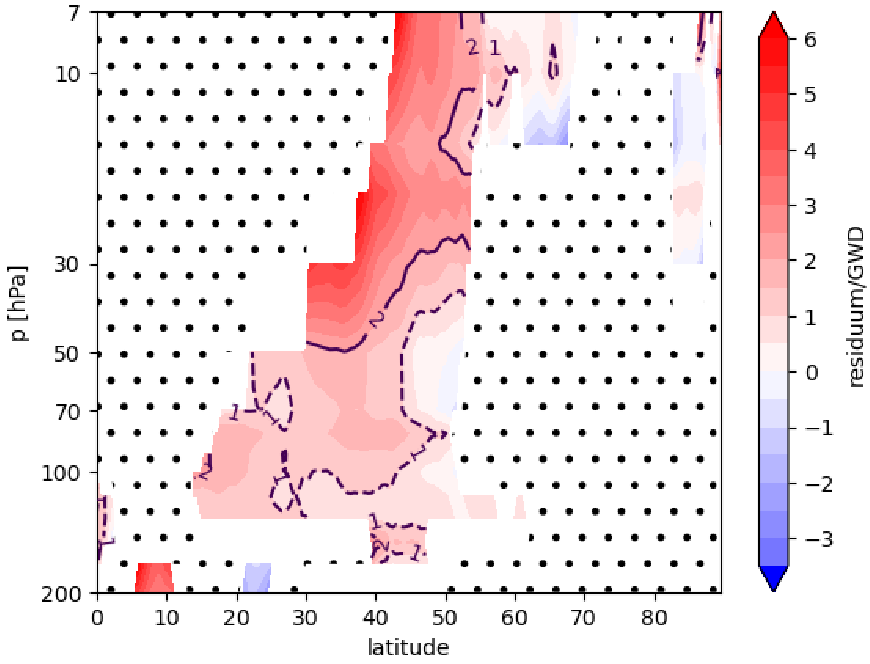


Fig. A2: Fraction between TEM residuum term (including resolved GWD, parametrized tendencies and analysis increment) and the resolved GWD. Only values with |GWD|>0.1 m/s/day are shown, regions out of the range are dotted. The dashed and solid lines represent the contours 1 (same value for residuum and resolved GWD) and 2 (resolved GWD forms half of the residual term), respectively.

Serva, F. (2022). Transformed Eulerian mean data from the ERA5 reanalysis (monthly means) (Version 0.1.1) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7081721

Referee 2

This manuscript compares the dynamical effects of gravity wave drag and Rossby wave drag on the zonal-mean circulation. The authors compare the resolved gravity wave drag diagnosed from the ERA5 reanalysis with the parameterized gravity wave drag from two nudged models and one free-running configuration. Their results indicate a broadly consistent GWD climatology between ERA5 and the models, alongside a weak correlation between GWD and Rossby wave drag within the models. The dynamical relationship between waves of different horizontal scales and their potential compensatory effects is a scientifically valuable question, and the multimodel framework adopted here is a legitimate approach. However, the manuscript has significant weaknesses in methodology, and I disagree with the interpretation of key results. For the reasons detailed below, I recommend rejection.

We would like to thank the referee for their time invested to the review of our manuscript and for clarifying their comments in a reply to our response.

Major Comments

1. Inconsistency in the ERA5 GWD dataset and its comparison with model parameterizations. The authors present the resolved GWD from ERA5 as a "quasi-observational dataset." While ERA5 is a high-quality product with a nominal horizontal resolution of 0.25◦ , its effective resolution is typically estimated to be five to ten times coarser, leaving a substantial fraction of the gravity wave spectrum in the sub-grid regime and therefore handled by the model's own parameterization scheme. As the authors themselves acknowledge in the Conclusion section, ERA5 does not resolve the total gravity wave drag. The comparison performed in the paper is therefore not symmetric: the resolved GWD from ERA5 (which excludes parameterized contributions) is compared to the parameterized GWD in the lower-resolution models, where the resolved contribution is presumably negligible.

In the paper, in the methodological part, we clearly discuss how the part of the spectrum constituting the resolved GW drag estimates overlaps with the part of the spectrum targeted by GW parameterization schemes of the models (see L103-105). Hence, this comment is very confusing for us.

The authors themselves acknowledge in the conclusion that one source of the GWD underestimation in ERA5 is the omission of the parameterized contribution, a point I fully agree with. My concern is precisely that the spectral overlap between ERA5 resolved GWD and the model parameterizations is only partial: the two datasets do not target the same range of horizontal scales. Lines 103–105 claim that the overlap is good, but if that were strictly the case, the remark in the conclusion would be unnecessary. The inconsistency between these two statements undermines the justification for the comparison as presented. This point is also mentionned by referee 3 in its first major comment.

Thank you for pointing out the apparent inconsistency. We now clarify in the revision that the overlap is good in a sense that the SSO parameterization schemes usually assume the dominant wave mode of a scale falling into the resolved spectrum in ERA5 and that we cannot account for the effects of the unresolved GWs in ERA5, because the parameterized GW drag product is not available for ERA5. We give more details on this aspect in our response to your next comment.

L103-106Notably, when comparing ERA5 to the climate model simulations with the horizontal resolutions described above, the GW spectrum contributing to resolved GWD in ERA5 aligns well with the spectrum targeted by GW parameterizations in the models. This spectrum in ERA5 spans roughly 100–2000 km, while the models typically target around 200 km (see Hájková and Šácha (2024)).

-> The analyzed GW spectrum in ERA5, which contributes to the resolved GWD estimates spans roughly 100–2000 km, which aligns well with the typical scales (around 200 km) of dominant modes assumed in the orographic GW parameterizations (Hájková and Šácha, 2024).

We have also significantly expanded the methodological section by including paragraph regarding the reliability of ERA5 GWD estimates:

In addition to the resolved component, ERA5 also uses GW parameterizations to supplement the effects of the GW spectrum it cannot resolve. However, this parametrized GWD is only available to users via a sum of all physics tendencies, which precludes a clean separation of the parameterized GW contribution and is not suitable for the goals of our paper. This is because the wind tendency from parameterized physics in the stratosphere also includes the contribution from the turbulence parameterization that is triggered by instabilities of the resolved GWs, which are captured in our resolved GWD estimates and would result in possible double-counting of the drag. Other than this, analysis increments also enter the momentum budget and they are not available for the analysis as well. Further, it has been shown that although ERA5 captures well the position and timing of directly observed GWs, it significantly underestimates their amplitudes and corresponding momentum fluxes \citep{guptaJAS,lear2024}, and hence also GWD. As a consequence, the resolved GWD derived from ERA5 should be regarded as a conservative lower bound on the true gravity wave forcing in the atmosphere, even though the resolved spectrum overlaps well with the range targeted by orographic GW parameterizations in models.

In conclusions we have added a new sentence in reaction to a comment by Ref#1:

L250-253: Preliminary results from closing the zonal mean momentum budget in ERA5 similarly to older reanalyses in the frame of the TEM equations (Fujiwara et al., 2024) indicate that the zonal mean resolved GWD presented here represents more than a half of the residual (including) drag in ERA5 in the valve layer (personal communication with Prof. Masatomo Fujiwara).

->

Preliminary results from closing the zonal mean momentum budget in ERA5 similarly to older reanalyses in the frame of the TEM equations (Fujiwara et al., 2024) indicate that the zonal-mean resolved GWD presented here represents more than half of the drag that can be attributable to GWs (both resolved, parameterized and missing) in ERA5 in the valve layer, because the ratio between the residual term (including resolved GWD, parameterized tendencies and analysis increment) and the resolved GWD ranges between 1 and 2 (see Fig. A2).

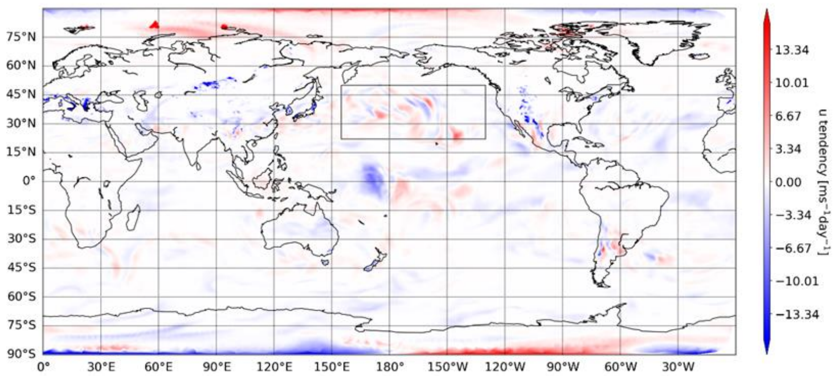
The authors justify excluding the ERA5 parameterized GWD by arguing that the associated diffusion term would be double-counted with the resolved drag. I do not follow this argument: at what stage of the budget decomposition would the parameterized drag appear in the resolved component $\nabla \cdot F$? Parameterized and resolved tendencies are computed separately in the IFS, so the risk of double counting seems unfounded. I strongly recommend that the authors add the ERA5 parameterized GWD to their ERA5 resolved GWD, and present the resulting total ERA5 GWD for comparison. If the authors are concerned about the turbulent diffusion contribution, they can mention it as a limitation of the comparison rather than using it as a justification for discarding the parameterized term altogether. That said, even the total ERA5 GWD should be interpreted with caution. While confidence in the large-scale ERA5 wind field is generally high, the GWD is a tendency that depends directly 1 on the IFS model physics and is also corrected by the analysis increment. ERA5 GWD is not a ground truth and carries its own biases, a point the authors should acknowledge explicitly when using it as a reference.

Let us emphasize that the ERA5 parameterized GWD product, which the reviewer calls for, is non-existent. As discussed around L248, ERA5 provides the wind tendency from parameterized physics, which, however, includes also the contribution from turbulent diffusion. In the regions of resolved GW instability, the turbulence parameterization smooths the resolved fields and the tendencies are sampled in the parameterized physics product. However, the attenuated resolved momentum fluxes are also captured by the EPF divergence mentioned by the reviewer and by our methodology for resolved drag computation (see Procházková et al., 2025, WCD for details), and hence inclusion of this product would inevitably result in double counting of the drag. We think that this is clear and does not need further explanation. The only possibility how to derive an estimate of the parameterized GWD in ERA5, would be to run the GW paramaterizations in IFS offline (one would need to have a precise information of all the tunning employed) and postprocess the whole ERA5, which is clearly impossible for this study.

I am aware that only the parameterized physics tendency product is available in ERA5, and this is precisely why I mention the turbulent diffusion contribution; we are referring to the same product, sorry about the naming. For this reason, I maintain that it is preferable to include this product, even if it contains a turbulent contribution, rather than omitting it altogether. The double-counting argument overlooks the role of spatial averaging. In the TEM framework, the relevant quantity is the zonal mean Xturb, which represents the effect of turbulence on the zonal-mean flow and enters the momentum budget as a distinct term from $\nabla \cdot F$. The effect of turbulence on the resolved waves projects onto X' turb, which does not contribute to Xturb but does contribute to $\nabla \cdot F$. These two contributions are orthogonal by definition. The same reasoning applies to the *area-averaging framework of Procházková et al. (2025), where ⟨Xturb⟩ and the wave-induced flux divergence at the domain boundaries are similarly orthogonal.*

Thank you very much for explaining your point. However, we cannot fully appreciate the arguments made about Xturb and how it enters the TEM equations. It seems to us that the reviewer makes an unsupported assumption by considering that "the effect of turbulence on the resolved waves projects onto X' turb" and vanishes in a mean.

Note that in our case we are concerned with a 4D quantity (lets say Xturb(lev,lat,lon,time)), which is a product of a turbulence parameterization in the underlying model triggered by a resolved GW instability. Hence, the magnitude of Xturb is unequally distributed over the wavelength and does not vanish in a mean over the wave, but sums up to GWD in the area-averaging framework of Procházková et al. (2025). This is documented by an example of the physics tendency output for one timestep at the level 70 hPa. Apart from the clear forcing from orographic GW parametrisation, the plot also shows a wave induced Xturb nicely copying the phases of the resolved GWs in the northern Pacific. The average of these perturbations taken over the black box is nonzero, as shown in the next plot, and would contribute to the mean tendency over this region. Clearly, this demonstrates that the turbulence contribution does not average out (except maybe some rare cases) over domains bigger than the wavelength of the GWs that induce it.



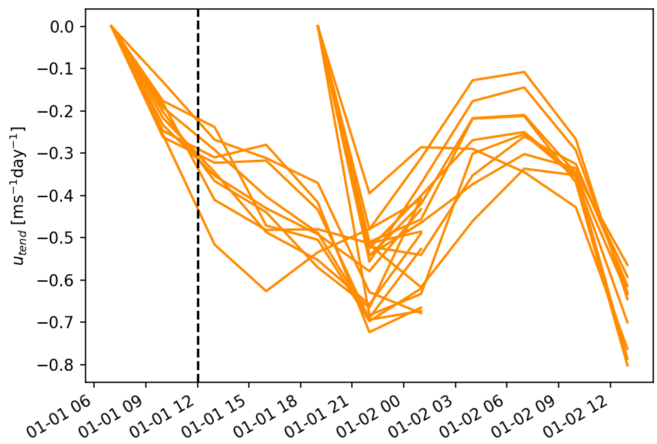


Figure R1 showing the U tendency from parametrized physics for 1. 1. 1979, 12:00 at 70 hPa (upper plot). Apart from the forcing from orographic GW parameterization about significant orography, the map also shows the contribution from the turbulence, as highlighted by the black box over the north Pacific. The average of these perturbations is nonzero, as shown in the bottom plot, which shows the U tendency averaged over the north Pacific box for individual forecast simulations.

Additionally, there are other technical issues that make the parametrized tendency not suitable for utilization as a proxy for parameterized GW drag in ERA5: The product of parametrized physics tendency does not describe the tendency of zonal and meridional wind but rather the tendency of wind rescaled by cosine of latitude, which complicates the interpretation (rescaling back, as in the figures above, does not give precise result due to the advection). Moreover, the parametrized tendency is given for the forecast simulations used to create ERA5 dataset only, with a non-negligible spread between the forecast values, as visible in the figure above.

No action is taken in the revised manuscript concerning this point.

2. The motivation for combining nudged and free-running models is insufficiently justified. The study uses two nudged models and one free-running configuration, but the rationale for this choice is never made clear. More importantly, the analysis treats the three configurations essentially interchangeably: no conclusions are drawn about what the nudging itself contributes to, or how it affects, the diagnosed GWD or its compensation with EPFD. The reader is therefore left wondering why the free-running model is included at all, and conversely, why results from the nudged configurations are not interpreted in light of the constraint imposed by the nudging. The authors should either provide a clear scientific justification for combining nudged and free-running configurations and discuss the implications of nudging for each result presented, or *separate the analysis into two clearly motivated threads, one focused on the role of the nudging itself, and one on the intrinsic model differences.*

We suspect that the reviewer did not read the manuscript in detail. The rationale is given in the paragraph around L95. The question of nudging is answered under the following comment.

The paragraph around L95 provides only a partial justification for the choice of nudged versus free-running configurations. It does not clearly explain what specific scientific insight each configuration is expected to contribute in the context of this study. More importantly, the authors do not address the central point of my comment: throughout the manuscript, results are never interpreted in light of whether the model is nudged or free-running. This omission remains unaddressed in their response.

Thank you very much for clarifying the original comment. We address the issue of justification for our choice of simulations under the major comment 1 of Ref#1, which results in the changes in the revision in a form of expanding the sentence originally centered around L155 to a full paragraph:

The goal of this study is to validate the dynamical interactions of parameterized GWD in the valve layer in global climate models using a quasi-observational dataset. For this, we analyse the dynamics associated with the resolved GW drag in ERA5 (Procházková et al., 2025b), as the best proxy for the real atmosphere GW forcing, and confront it with dynamical effects of parameterized GWs in nudged and transient climate model simulations....

->

The goal of this study is to compare the dynamical interactions of parameterized GWD in the valve layer in global climate models with a GW resolving dataset. In the previous literature, parameterized GWD effects in climate models were analyzed using either specified dynamics simulations (e.g. Kuchař et al., 2020; Šácha et al., 2021) ensuring realistic circulation backgrounds or free running model simulations (e.g. Eichinger et al., 2020; Hájková et al. 2024) enabling a cleaner attribution of dynamical effects. Here we aim to bridge the two approaches by analyzing an ensemble of opportunity consisting of two specified dynamics simulations of two different state-of-the-art climate models (with different parameterization schemes and nudging procedures), and a transient simulation of one of the models facilitating the assessment of the difference of responses between freely evolving and specified dynamics. The climate model results are then compared with the resolved GW drag in ERA5 (Procházková et al., 2025b), as the best proxy for the real atmosphere GW forcing, and the associated dynamical interactions are contrasted across the datasets.

The part of the comment concerning nudging is addressed under your next point.

3. The Rossby wave forcing in nudged models should not be interpreted as a model performance metric. Figure 2 shows the EPFD in both models and ERA5. The authors compare these fields as if differences or similarities reflect model skill. However, nudged models constrains its large-scale dynamics toward ERA-Interim. It is therefore expected by construction that the resolved EPFD in the nudged model resembles that of ERA5 (or ERA-Interim), and this agreement cannot be attributed to any intrinsic model ability to reproduce the observed climatology. This point should be clearly stated in the analysis of Figure 2. Failing to do so risks misleading the reader into thinking that the nudged models are validated against ERA5 in an independent sense.

We again suspect that the reviewer did not read the paper in detail, because the reviewer poses their opinions as facts, and formulates arguments probably based on personal beliefs, even though they are contradicting the factual results presented. Traditionally, what is discussed and criticized during the review process are the actual results of a manuscript. In this particular case, Fig. 2 clearly shows striking differences in the Rossby wave field between the two nudged simulations (CMAM-SD and ESCiMo-SD) in the lower stratosphere, in a strong contrast to the Ref#2 comment. Clearly, in the lower stratosphere it is the internal model dynamics that governs the planetary wave field. This is in accordance with Fig. S1 of Šácha et al. (2021, GRL), which shows a negligible dynamical influence of the nudging tendency in the lower stratosphere in CMAM-SD. Fig. 2 also nicely justifies the motivation and utility of having 2 different specified dynamics simulations and a free running simulation (demonstrating again the role of internal model dynamics, because of the similarity of the two EMAC simulations) in the comparison.

I will not comment further on the suggestion that I did not read the manuscript carefully; this is not a constructive basis for scientific exchange. Reviews are necessarily written from the reviewer's perspective and knowledge of the literature, and it is standard practice to raise concerns about methodological choices that may affect the interpretation of results. Regarding the scientific content: I did not claim that the Rossby wave fields must be identical between the nudged models and ERA5. My point was simply that the role of nudging should be discussed in the analysis of Figure 2, even if only to clarify that it has a negligible influence at the levels considered. If the authors can demonstrate that the nudging tendency is indeed negligible in the lower stratosphere for both nudged configurations, this should be stated explicitly in the manuscript. As it stands, the figure is presented without any mention of the nudging constraint, which risks misleading the reader regardless of whether the effect is large or small.

Thank you again for clarifying the original comment. We agree that incorporating an estimate on the strength of nudging and its relationship with other dynamical quantities will make the paper clearer. In the revised version we add two figures showing the estimated climatological strength of nudging (note that estimating it as a residual from TEM provides an upper bound for its strength, including other possible effects and the error of the method) and its correlation with the EPFD and GWD. Further, we add a figure to the Appendix showing the scatter plots when nudging is included. Overall, the new results confirm that nudging plays a minor role in the valve layer. But, they also provide an interesting additional information for how much drag is probably needed in the valve layer to match observations.

4. Differences in the zonal wind field are not interpreted in terms of the nudging protocol. At several places in the manuscript [L180–185, 255], the authors note similarities or differences between the model zonal winds and those from ERA5, without ever relating these differences to the nudging. This omission is problematic because:

- *For the nudged models, resemblance to ERA5 is expected and should be qualified by the fact that the nudging is performed toward ERA-Interim, residual differences between the two reanalyses will propagate into the model state.*

- *For the free-running model, differences in the zonal wind are meaningful and should be discussed. I recommend that the authors mention the nudging whenever they compare zonal 2 wind fields between models and ERA5, so as not to mislead the reader into interpreting constrained agreement as a model performance.*

The zonal wind field analysis has a minor role in the manuscript, and the results are presented only to give a context for the GWD fields above the selected regions. Regarding nudging, we urge the reviewer again to take a closer look at the respective part of Fig. 3, where there is no hint of a systematic nudging influence for the differences between the different simulations and ERA5 (in the LS, CMAM-SD (driven by ERA-I) agrees well with ERA5, while free running ESCiMo simulations follows ERA5 more closely than the -SD simulation (driven by ERA-I)).

My comment was not that nudging explains all differences, but simply that it should be mentioned when comparing zonal wind fields, so that the reader can interpret the results accordingly. Regarding the specific result that ESCiMo agrees more closely with ERA5 than ESCiMo-sd, this is not self-evident and deserves discussion rather than being used as evidence against a nudging influence. I agree that the zonal wind field analysis has a minor role in the manuscript, and this comment should indeed be considered together with the previous one.

Thank you again for clarifying the original comment. We follow your suggestion and expand the text by discussing the results also in the context of nudging (to ERA Interim), where appropriate (around former L180 in particular).

->

For the two specified dynamics simulations, additional forcing influences the resolved fields in the form of nudging (but towards ERA-Interim). Fig. 3 provides an upper bound for the nudging strength estimated as a residual from the TEM zonal momentum budget. In the valve layer, the residual contributes to the negative drag in CMAM-SD with a similar magnitude as the GWD, but reaches also further down inside the UTLS jet in the upper troposphere. For ESCIMO-SD, which has a stronger GWD in the valve layer, the residual in the valve layer is small but positive, probably compensating for the excessive GWD, but there are stronger residual tendencies of opposite signs at both meridional flanks of the UTLS jet in the upper troposphere. Despite this, the internal model dynamics seems to govern the zonal wind climatology in UTLS, because both ESCIMo simulations have a very similar structure and magnitude of the UTLS jet compared to CMAM-SD. On the other hand, the polar night jet is stronger in the free running EMAC simulation than in its specific dynamics run and also than in CMAM-SD (see the contours in Fig.1). In the upper stratosphere, we see strong residual tendencies in ESCIMo-SD, but note that the forcing from resolved waves quantified by EPFD is about 5 times stronger than the strength of nudging and GWD in this region, which has a comparable magnitude only in the confined region above the center of the subtropical jet (Šácha et al.,2019).

5. The compensation between GWD and EPFD is statistically weak and physically ambiguous. The authors propose a compensatory relationship between GWD and EPFD, motivated by the idealized framework of Cohen 2013. Several issues arise with this interpretation.

Cohen et al. (2013, JAS) indeed formulated the physical mechanisms leading to compensation between resolved and unresolved wave driving of BDC based on an idealized framework, but the robust emergence of this kind of dynamical interaction in comprehensive models has been demonstrated in numerous papers since then (e.g. White et al., 2017, JClim; White et al., 2018, GRL; Šácha et al., 2019, ACP; Eichinger et al., 2020, CliDyn; White et al., 2021, JAS; Šácha et al., 2021, GRL; Sigmond et al., 2023, GMD; Hájková and Šácha, 2024, CliDyn), which makes us wonder, how the reviewer can call it "an interpretation".

I do not question the existence of the compensation mechanism between resolved and parameterized wave driving, which is well established in the literature cited. My concern is specifically that the results presented in this manuscript do not provide evidence to support this.

Thank you for clarifying this, however, we do not agree. We think that our results clearly document the existence of a compensation mechanism between unresolved GW drag and resolved wave drag in the valve layer.

• The correlations presented are low: a coefficient r 2 ≈ 0.04 explains only a small part of the variance. At this level, confounding factors, foremost among them the state of the background zonal wind, can easily produce the observed scatter and slope. Furthermore, the slope varies from approximately -1 in the free-running model (which might suggest compensation) to nearly -3 in the nudged ESCiMO-SD simulation, undermining any universal interpretation.

In Fig. 6 you can see that the correlations are statistically significant and locally they reach higher values (in the models). By no means this result should be interpreted in the sense that GWD determines EPFD, because not only background zonal wind, but importantly the momentum fluxes from below determine the variability here. Given the fact that we use daily data without any threshold for GWD (sampling also negligible GWD episodes, which cannot produce any signal of dynamical interaction), we find the occurrence of significant anticorrelations remarkable, documenting the robustness of the compensating type of interaction between the unresolved and resolved drag in the valve layer. Regarding the Pearson correlation coefficient in Fig. 7, the reviewer should note that the coefficient provides only an estimate for the strength of the correlation. Since we are using Deming regression to consider changes in both GWD and EPFD, this coefficient does not describe the quality of the fit as in the case of least-square method. It should also not be mistaken with the degree of compensation measure defined by Cohen et al., (2013, JAS, eq. 1) or Eichinger et al. (2020, CliDyn, eq. 4) that are by definition constructed to give values in the interval <0,1>.

Regarding Figure 6, the anticorrelations are statistically significant, but a correlation coefficient of −0.2 remains weak. My concern is not whether the anticorrelation exists but whether it can be unambiguously attributed to a compensation response. Describing this as "clearly visible" overstates what the data show. Regarding Figure 7, I am aware that the Pearson coefficient does not describe the quality of a Deming regression fit as in the case of least-square method, but it gives an indication. My point was simply that the large spread in regression slopes across configurations and the variability in anticorrelations between models suggest that the signal is not as clear as claimed by the authors.

Thank you for continuing the discussion. We cannot write more than already written in our initial response (in green). The compensation mechanism and the physical processes behind it are well established and we find it remarkable that just by correlating the two dynamical fields in a comprehensive model without restricting the analysis to strong drag events, we derive a statistically significant signal of compensation in the region, where it is anticipated.

We have revised the section on the EPFD-GWD interaction and the phrase “clearly visible is no longer appears in the revised paper.

• The deceleration of the zonal-mean flow by Rossby wave breaking corresponds EPFD < 0. Points in the scatter plots where EPFD > 0 correspond to EP flux divergence, which physically indicates wave activity propagating away from the region, or local wave generation, not wave breaking drag. Including these points in a compensation analysis with GWD conflates physically distinct processes and should at minimum be discussed.

This is again a completely wrong statement, because the role of the GWD induced instabilities for in-situ sourcing of Rossby waves (EPFD > 0) is one of the possible interaction types between the unresolved and resolved drag, well documented in numerous studies (e.g. McLandress & McFarlane, 1993, JAS; Sato et al., 2018, JAS; Smith, 2003, JAS; Šácha et al., 2016, ACP; Yoo et al., 2024, JAS).

The authors' response actually confirms my point rather than contradicting it. In-situ sourcing of Rossby waves by GWD-induced instabilities (EPFD > 0) is indeed a documented interaction type, as the cited references show. However, this is a physically distinct mechanism from the compensation between resolved and parameterized wave driving in the sense of Cohen et al. (2013). The clarification provided in the authors' response is exactly what I was asking to see in the manuscript itself.

Thank you for clarifying this. We included similar clarification also in the revision.

6. Pervasive use of unsupported qualitative language. Throughout the manuscript, the authors employ numerous evaluative adjectives to characterise their results without providing any quantitative benchmark or reference to prior work. This is problematic on two grounds: without an objective reference, the reader cannot independently assess whether the claimed quality of agreement is meaningful; qualitative descriptors are inherently subjective, and a reader may reasonably disagree with their usage, which undermines confidence in the results as presented. The following lines are representative instances of this issue: [L131, 149, 152, 182, 185, 188, 191, 203, 220, 227, 230, 232]. For each occurrence, the authors should either replace the qualitative descriptor with a quantitative metric (e.g. bias magnitude, percentage difference, or a skill score) supported by the data shown, or anchor the claim explicitly to a published result, or refrain from commenting on the result quantitatively.

We have serious doubts about the relevance of this comment, because in the lines highlighted, we do not make any strong claims in the manuscript and mostly report on areas of agreement between the datasets. Note that we cannot quantitatively compare every aspect and pixel of the meridional distribution of any quantity. For instance, at L131 we write: The zonal mean GWD distribution is very similar for all the datasets…and then the paragraph continues by listing features that give basis for the initial sentence.

The situation is similar also for other instances, documenting how surreal this comment is.

I understand that a pixel-by-pixel quantitative comparison is neither feasible nor expected. My concern is different. Qualitative descriptors such as "very similar" placed at the beginning of a paragraph create an anchoring bias. A factor of two difference between ERA5 and the models could reasonably be described as "very similar" by some readers and "substantially different" by others, the descriptor carries no objective information. Moreover, for several instances the data themselves provide natural quantitative anchors that are not used. For example, at L200 the authors state that strong GWD events have a similar duration and vertical structure across datasets. Duration and vertical extent are measurable quantities that could be reported directly, as they are already implicit in the figures. This is the type of change I am requesting, not a comprehensive pixel-by-pixel comparison. I understand that a formal validation exercise is not the stated objective of the study. Nevertheless, providing objective elements of comparison does not require such an exercise, and would strengthen the manuscript given that comparison between datasets is one of its goals. If the editors and other reviewers find the current level of quantitative support sufficient, I will defer to their judgment. I maintain however that replacing subjective descriptors with measurable quantities where they are naturally available in the data would improve the clarity and reproducibility of the conclusions.

We went through the individual occurrences mentioned by the reviewer and we will answer them separately:

L131: As the similarity among datasets is described in the next sentences, we changed dot for column to make clear the explanation will follow.

The zonal mean GWD distribution is very similar for all the datasets.

-> The zonal mean GWD distribution is very similar for all the datasets:

L149: “The boreal winter climatological distribution is again grossly similar between the models and the reanalysis, (…)” as this is described by the second part of the sentence (matching upward/downward propagation), we do not make any changes here

L152: “the CMAM-SD simulation captures the EPFD climatology from ERA5 remarkably well in terms of both distribution and magnitude of the drag”

We included pattern correlation coefficient between CMAM-SD and ERA5.

the CMAM-SD simulation captures the EPFD climatology from ERA5 remarkably well in terms of both distribution and magnitude of the drag

-> the CMAM-SD simulation captures the EPFD climatology from ERA5 remarkably well in terms of both distribution and magnitude of the drag (pattern correlation approx. 0.975)

L182: “Especially above the West America and East Asia hotspots the model winds agree very well with ERA5 up to around 50 hPa.” – We think the level of agreement is clearly visualized in Fig. 3 and can be assessed by the reader based on their precision needs.

L185: “Above Himalayas we see a pronounced underestimation of the climatological zonal wind speed for both ESCiMo simulations across the UTLS region, but CMAM-SD follows the ERA5 wind profile remarkably well.” – same as the comment at L182.

L188: “The distribution of parameterized GWD is very similar across all model simulations.”

->“Generally, the distribution of parameterized GWD is very similar across all model simulations.”

As the aim of this statement was to contrast the similarity of the drag distribution of the models with the very different distribution of reanalysis, which is clearly visible from Fig. 4, we do not see any benefit of listing numbers describing the distributions. The different means of the distributions are visualized in the figure by dashed vertical lines.

L191: “When we rescale the datasets to have the same mean GWD, we see that the intermittency of the parameterized GWD for all simulations agrees remarkably well with that of the resolved GWD from ERA5.”

- Again, we think that the level of agreement can be qualitatively assessed from Fig. 4. As the distributions differ among the subdomains, we cannot imagine possible usage of parameters of the distributions that could be supplemented to quantify the agreement.

L200: “In Fig. 5 we see the best agreement among the models and ERA5 for the East Asia hotspot, where the strong GWD events have a similar duration and vertical structure across all datasets.”

->

“In Fig. 5 we see the best agreement among the models and ERA5 for the East Asia hotspot, where the strong GWD events have a similar duration and vertical structure across all datasets, with the strong negative anomaly lasting approximately 5 days between 150 and 40 hPa.”

L203: “This is also partly true for the West America hotspot, but here the parameterizations in the models do a good job in capturing the nontrivial vertical structure of the event in ERA5, which suggests (…)”

->

“This is also partly true for the West America hotspot, but here the parameterizations in the models do a good job in capturing the nontrivial vertical structure of the event in ERA5, with negative drag anomaly in the upper pressure levels. This suggests (…)”

L220: “The compensation mechanism between parameterized GWD and EPFD is again very clearly pronounced for the models. The relationship is very similar for both specific dynamics simulations.”

- This section is supported by the described Fig. 7, where the reader can assess the level of similarity with the help of the regression lines and correlations coefficients.

L227, L230 and L232: “(…), we have shown that there is good agreement in climatological features of the parameterized drag from models in the NH winter with the quasi-observational estimates from the reanalysis, both in a zonal mean and regionally. We have identified only minor differences in the structure, however, also substantial differences in the magnitude of GWD in the lower stratosphere. When accounting for the differences in the mean drag value, we have further shown that the probability distribution of the GWD magnitude is almost exactly similar between the datasets.”

- As this paragraph summarizes the important results from our point of view, it is in some sense inherently subjective. In our opinion, what we wrote here follows from the previous section and was illustrated there in a way allowing the reader to check the validity of the statement.

Minor comments

C1. [L29–33] A citation is needed to support the claims made here.

Citations have been included in revised paper.

C2. [L33–35] A citation is needed to support the claims made here.

Citations have been included in revised paper.

C3. [L40–45] A citation is needed to support the claims made here.

Citations have been included in revised paper.

C4. [L53] The stated goal of the study is "to validate the dynamical interactions of parameterized GWD [...] in GCMs using a quasi-observational dataset." The word "validate" seems too strong here, as the paper does not perform a formal validation exercise. The verb "compare" would be more appropriate and more defensible.

The word validate has been replaced with the word compare as suggested.

C5. Model naming is inconsistent throughout the manuscript: "CMAM-sd" and "CMAM-SD", as do "ESCiMo" and "EMAC". The authors should adopt a single, consistent name for each model and apply it throughout.

Done.

C6. Using sub-panel labels (a), (b), (c), (d) systematically in figure would make the description easier to follow.

Thank you for the suggestion. However, we decided not to make any changes in this regard since we believe the titles of the individual subplots are sufficient for the reader's orientation.

C7. Figure 1 "Boreal winter" should be specified explicitly as DJF in the caption or at least one time in the paper.

Done.

C8. [L123] The 55% threshold used as a selection criterion should be justified: what is the sensitivity of the results to this specific value? A brief sensitivity test or a reference supporting this choice would strengthen the methodology.

The threshold was taken directly from the study Kuchař et al., 2020. We reformulated the text to make clear that it is a specific method for GWD time series:

“To identify these events, we apply a peak-detection algorithm following the methodology of Kuchař et al. (2020), as designed for time series of parametrized GWD.”

and added another reference to the last sentence describing the algorithm

“and well-isolated episodes of enhanced GWD (Kuchař et al, 2020)”

C9. [L125] The term "anomalies" is not defined. Does this refer to deviations from the daily climatology? The reference period and the nature of the baseline should be stated explicitly.

We have changed the sentence to explain how the anomalies are constructed:

The statistical significance of composite anomalies were constructed using a bootstrap method based on 10 000 samples and subsequently corrected for multiple testing using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995).

->

The statistical significance of the composite anomalies, obtained by subtracting daily values from the seasonal long-term climatology, was assessed using a bootstrap method with 10,000 samples and subsequently corrected for multiple testing using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995).

C10. Throughout the manuscript, the terminology implicitly places ERA5 in the role of the observational reference and all other datasets in the role of "models". However, ERA5 is itself a reanalysis product and is strongly dependent on the underlying forecast model (IFS Cycle 41r2). Despite data assimilation, ERA5 cannot be treated as independent from the model physics. It would be appropriate to remind the reader of this at the relevant points, particularly where the reanalysis is used as the reference for evaluating GCM behaviour.

Thank you for the comment. Within the manuscript, we always refer to the dataset as quasi-observational to avoid this type of confusion. The problem of dependence on model physics is unavoidable due to limited observations and data from high-resolution models as well as reanalyzes are commonly used as a reference for validating behavior of other models. We reformulated the Methodology section to include a clear description of the limitations of the ERA5 GWD estimates.

C11. Figure 3 The peak in GWD can plausibly be connected to the position of the zero-wind line in the models, consistent with critical-level absorption of gravity waves.

In general, the GWD distribution reflects the saturation criterion utilized in the parameterizations, the zero wind line is a special case here.

C12. [L177–178] The remark about oblique gravity wave propagation appears without context or justification. No mechanism is invoked, no reference is provided, and it is unclear what in the results motivates this claim. The authors should either ground this statement in an established theoretical or observational basis, or remove it.

Thank you for pointing out this claim is not justified. We added a reference to Procházková et al., 2025, Fig. 10, where the importance of oblique propagation at the hotspots is visible.

C13. [L190] The rescaling of the GWD mean is performed without apparent justification. Unsurprisingly, the rescaled fields show better agreement.

The purpose of the rescaling is to allow a straightforward comparison of the model GWD distributions with ERA5, which has lower GWD values due to the different nature of its data source, as discussed in previous comments. The rescaling has little effect on the shape of the model distributions because their means are similar, with the exception of CMAM-SD over Western America. We modify the relevant sentence:

When we rescale the datasets to have the same mean GWD (shown by the dashed line in the top row), we see that the intermittency of the parameterized GWD for all simulations agrees remarkably well with that of the resolved GWD from ERA5.

->

When we rescale the datasets to have the same mean GWD (shown by the dashed line in the top row), enabling comparison with ERA5, the intermittency of the parameterized GWD in all simulations agrees remarkably well with that of the resolved GWD from ERA5.

C14. [L192–194] The interpretation of the ERA5 departure after rescaling is questionable. Anomalies with respect to a rescaled mean depend on the rescaling procedure and may not carry a straightforward physical interpretation. Any conclusion drawn from this analysis must be supported by a clear argument.

We agree that distributions defined with respect to a rescaled mean should not be interpreted as an independent physical quantity in the same way as distribution from the original data. Our intention in this part of the analysis was not to assign direct physical meaning to the rescaled distribution itself, but to use the rescaling only as a normalization procedure that allows a like-for-like comparison of the ERA5-derived GWD with the model parameterized GWD.

C15. Figure 5 The dashed vertical green line appearing in the figure is not described in the caption. In addition, the GWD anomaly values themselves are never discussed in the text, only the region of statistical significance receives attention. The authors should either comment on the magnitude of the anomalies or, if the values are not the focus, simply overlay the region of statistical significance on the figure without displaying the anomaly field itself, or at least state explicitly that the anomaly magnitudes are not the object of the analysis.

The caption of Fig. 5 has been updated to include a description of green lines. The paragraph describing the Fig. 5 has been extended:

In Fig. 5, we focus on the temporal and vertical evolution during strong GWD events (detected at 70 hPa) through the composite analysis of averaged GWD anomalies for each hotspot on the 20 day timescale. In all three hotspots, strongest GWD anomalies, exceeding -4 m s^{-1} day^{-1} for models and -2 m s^{-1} day^{-1} for ERA5, are located around the composite level and during the peak day (lag=0). However, statistically significant GWD anomalies were also detected over a period of several days around the peak day along with the vertical structure. The best agreement among the models and ERA5…

C16. Figures 6 and 7 As raised in Major Comment 5, the correlation found here is low and the interpretation appears to reflect confirmation bias rather than emerging robustly from the data. The authors should either adopt a more cautious framing or ground the interpretation in prior literature using comparable diagnostic criteria. Additionally, a simple spatial or temporal shift between the GWD and EPFD fields might be more physically appropriate for detecting a compensatory response, given that such compensation may operate with a lag; this possibility should be explored or at least discussed.

See our response to your Major Comment 5. Besides this, thank you for pointing out the issue of a time-scale and the lag of the response. This issue was studied in Šácha et al. (2021, GRL) and it is indeed another aspect that may cause the lower value of anti-correlation that we see. This is now clarified in the text and the more we find it remarkable that a significant anti-correlation emerges using our simple methodology here.

C17. [L250] The motivation for including the analyse momentum term from ERA5 is unclear. If the goal is to compare GWD magnitudes across datasets, which appears to be the purpose of the study, then adding a budget residual term for ERA5 is not the relevant quantity. Moreover, if one were to close the budget symmetrically, the nudging tendency would need to be added on the model side. Closing the zonal momentum budget does not appear to be the stated objective of this paper. Separately, the free-running model provides the most direct test of the dynamical consequences of a mis-calibrated GWD, since its zonal-mean state evolves freely; this aspect deserves more prominence in the discussion.

Thank you for this comment. You are correct that our intention has not been to close the budget across the datasets, but here we wanted to estimate how much of drag attributable to GWs in ERA5 in the valve layer we are missing, because of the impossibility of including the parameterized GWD and analysis increments. We are now adding a figure for this in the Appendix and clarify our motivation better. Also, we clarify an important point that is needed to properly understand this result – directly computed zonal mean GWD in the valve layer is almost perfectly identical to EPFD from corresponding GW modes. This is analogical with the statement that linear GWs do not produce a Stokes drift.

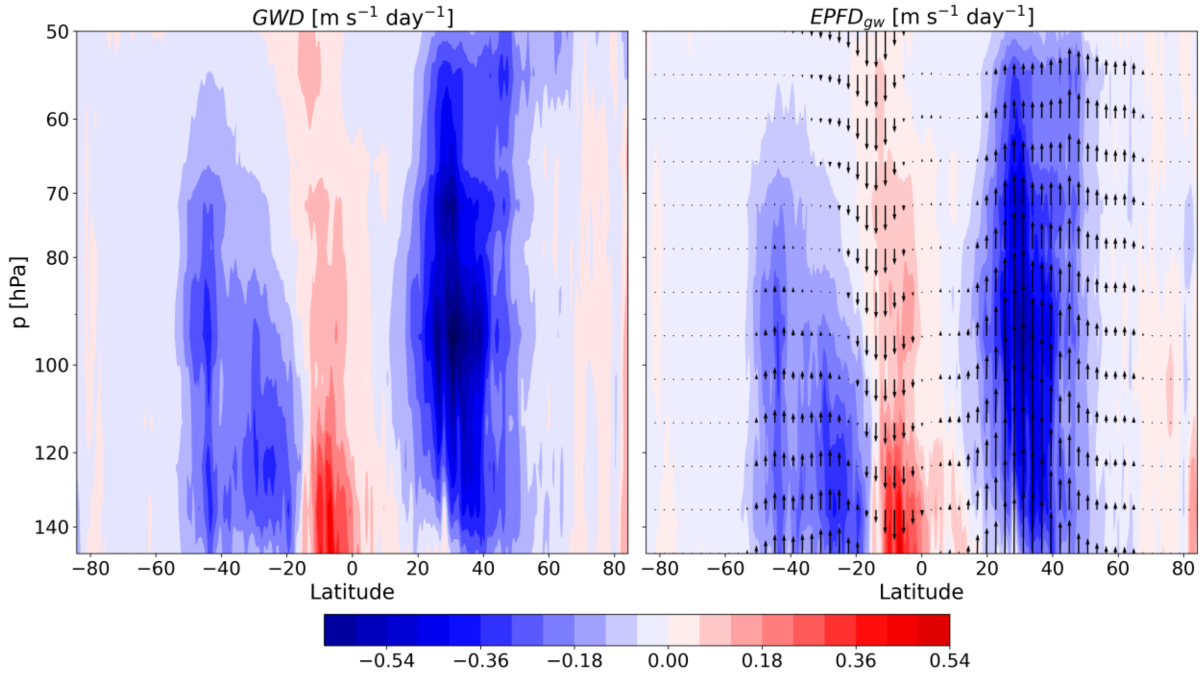


Fig. R2: Comparison of the GWD and the EPFD (with EPF displayed as arrows), when computed from the GW modes only. Figure shows monthly average from a period starting at 20 June 2020. Note smaller vertical extent compared to other figures in the paper. The difference between the average GWD and EPFD has a magnitude of approximately 15% of GWD, with the correlation between the time series for individual latitudes and pressure levels ranging from 0.7 in the mid- and high-latitudes to values close to 1 in the equatorial region (not shown).

Referee 3

The manuscript presents an investigation of data from ERA5 reanalyses (with a nominal spatial resolution of about 30km), from CMAM (300km) and EsCiMo (> 300km, nudged to ERA-Interim). Via spectral filtering, waves with wavelengths < 2000km are extracted from the ERA5 data, and these are interpreted as gravity waves. The momentum-flux convergence obtained from these waves is interpreted as resolved gravity-wave drag, and the latter is compared with the parameterized gravity-wave drag in the two more coarsely resolved models. With a focus on boreal winter, it is shown that the resolved gravity-wave drag in ERA5 is weaker than in the parameterizations in the two other models. However, the patterns of the distribution of the zonal-mean fluxes are similar in many regards, and the same also holds for the Eliassen-Palm-flux divergence of the larger-scale waves in the ERA5 data and of the resolved waves in the two other models. After a rescaling of the gravity-wave drags, the histograms of their distributions are very similar, with ERA5 exhibiting least intermittency. The lagged evolution of extreme events also is reasonably similar, but over the Himalayas and East Asia the parameterized drag is more persistent. It is argued that some of the differences might be due to oblique *gravity-wave propagation, that is neglected in the parameterizations. Finally also the correlation between the gravity-wave drag and the Rossby-wave Eliassen-Palm-flux divergence is investigated. It is found that in the case of the two coarse-resolution models the correlation of the parameterized drag with the Rossby-wave Eliassen-Palm-flux divergence is in line with the mutual compensation first observed by Cohen et al. (2013), but not so for the resolved gravity-wave drag in ERA5. Here as well it is argued that this might be due to oblique gravity-wave propagation.*

We would like to thank Prof. Achatz for a positive evaluation of our manuscript and for his helpfull comments and suggestions that helped us to improve the manuscript.

Major comment:

The overall description of the analyses is clean, and it supplements the mounting evidence that the assumptions which present-day operational gravity-wave parameterizations are based upon limit the realism and reliability of the latter. This manuscript should eventually be published after the following comments have been taken into account.

(1) The inclusion of the parameterized gravity-wave drag from the ERA5 data is missing in the analyses. The parameterized drag in the coarser-resolved models is meant to represent gravity waves on all scales, and the omission of the effect of waves not resolved explicitly by ERA5 makes the present analysis incomplete. The conclusions presently drawn by the authors would be on much firmer ground after also taking the smallest-scale waves into account.

Thank you for the comment. We agree that inclusion of the parametrized GW tendencies would strengthen the results. However, such a product is unfortunately unavailable for ERA5 data. The forecasts used to produce ERA5 only contain tendencies caused by different parametrizations together (GW drag parametrization, vertical diffusion parametrization, convective momentum transport). As shown in a figure in the response to Reviewer 2, including this term in the analysis would additionally add a significant amount of drag that is already present in our resolved GWD estimates. That said, we have added a new figure in the Appendix, which is discussed in the conclusion, where we estimate the upper bound of the GW related drag in ERA5 that we miss in our analysis.

(2) In the abstract, the discussions and the conclusions I am missing the consideration of recent work on the limitations and possible improvements of the parameterization on gravity waves. (Bölöni et al., 2016, 2021; Kim et al., 2021) show that the steady-state assumption of traditional gravity-wave parameterizations limits their reliability, especially with regard to gravity-wave intermittency, as well as does the neglect of

oblique propagation. The effect of the latter has explicitly been demonstrated to be of relevance by (Banerjee et al., 2026; Kim et al., 2024; Kühner et al., 2026; Voelker et al., 2024), for the spatial distribution of the gravity-wave fluxes, the forcing of the QBO, the residual-mean circulation, and the zonal-mean flow. Finally it has also been shown that the neglect of the thermal forcing by gravity waves can also lead to significant errors (Kühner et al., 2026; Wei et al., 2019) in the simulated zonal-mean flow and in solar tides. All of these studies demonstrate how these effects can be incorporated without explicitly resolving gravity waves. I think a corresponding discussion would be in place.

Thank you for the suggestion and for providing the references. We added the following paragraph into the introduction:

The simplifying assumptions underlying traditional GW parameterization schemes are known to affect model dynamics adversely. For example, the steady-state assumption (Bölöni et al., 2016, 2021; Kim et al., 2021) neglects the GW life cycle and finite propagation speeds. Similarly, the omission of the thermal forcing exerted by GWs (Kühner et al., 2026; Wei et al.,2022) and the restriction to purely vertical propagation have been shown to affect the spatial distribution of gravity-wave fluxes, QBO forcing, the residual-mean circulation, and the zonal-mean flow (Kim et al., 2024; Voelker et al., 2024; Banerjee et al., 2026; Kühner et al., 2026).

Minor comments:

l. 12: It might be not necessary to resolve the waves. More general parameterization approaches might work as well.

We have made the following change:

the dynamical effects of gravity waves that we know from Earth system models can be different if gravity waves are resolved

->

the dynamical effects of gravity waves that we know from Earth system models can be different if gravity waves are represented more realistically

l. 177: replace pressumably by presumably

Done.

l. 180 – 185: Especially in the seasonal mean I believe there is no direct local connection between gravity-wave drag and winds forced by it. In the corresponding balances an anomalous drag causes an anomalous circulation, the latter influences the thermodynamic fields, and that again modifies the winds. This is well established in *the zonal-mean picture (which could be constructed using downward-control analysis), and in the present case, with local hot-spot regions, the situation is even more complex. Hence it might be not so surprising that the winds are not so different. Moreover, the gravity-wave drag is weak compared with the Rossby-wave impact. Maybe the latter just dominates?*

Thank you for the comment. We agree that in the zonal mean perspective, GWD is connected to the zonal mean zonal wind through the non-local dynamical effects. Locally, GWD is related to the zonal wind by the definition of the drag, i.e., it describes the wind tendency due to the gravity waves. From this point of view, an increase in GWD magnitude would imply a decrease in the zonal wind magnitude. As you suggest, it is very likely that this local effect is small compared to large scale dynamical and thermal influences and can also induce non-local dynamical responses that are not yet fully understood. We made the following change in the manuscript:

Surprisingly, much stronger GWD in the models than in ERA5 above 100 hPa is not clearly reflected in differences in the regionally averaged zonal wind.

->

Much stronger GWD in the models than in ERA5 above 100 hPa is not clearly reflected in differences in the regionally averaged zonal wind, most likely due to indirect effects such as an alteration of the Rossby-wave fields or due to dominating large scale dynamical and thermal influences.

Conclusions: Why not also consider whether more general gravity-wave parameterization approaches are more in agreement with high-resolution simulations and observations? Also one should be aware that highly resolved models are not guaranteed to be realistic in their gravity-wave fluxes, when compared to observations.

Thank you for this invaluable comment and suggestion. We changed the final sentence:

..Hence, for future work we see it vital to constrain the GWD dynamical effects from ensembles of high or ultra high resolution free running simulations of sufficiently long duration.

->

Hence, for future work we see it vital to constrain the GWD dynamical effects from a combination of theory, observations and ensembles of high or ultra-high resolution free-running simulations of sufficiently long duration and assess the validity of GW effects simulated using more general GW parameterizations (Voelker et al., 2024).