The comments by referee #1 and referee #2 are listed below in black italics, followed by our point-by-point responses in blue and relevant author's changes in the manuscript in green. Note that a specific number of lines in green in this document is applicable to the revised manuscript with track changes.

Response to comments by Referee #1

Under the category "General Comments"

This paper addresses the interesting questions regarding the development of model errors in temperature and zonal wind in the Met Office UM, using a nudging relaxation and a tendency budget decomposition. These methods are well described, and used to come to some interesting results. In the most part the paper is well written, the results well presented and the conclusions good, however I would suggest a substantial re-write of Section 4 and additional reference to related work before the paper is accepted for publication.

Thank you for your suggestive comments. We agree that some descriptions of atmospheric budget diagnostics, particularly an interpretation of the residual term, are confusing and a substantial re-write is needed for improving the manuscript. Your suggestions of some additional references are very informative. As responded below in "Major comments" section in detail, new sections (Section 2.5 and 2.6) have been created in earlier part of the revised manuscript for describing the detail of a correspondence between model tendencies and budget diagnostics, and a validation of budget components against the global nudging experiment. Some of the descriptions in Section 4 and other parts of the manuscript have been moved to the new sections.

In several places the text would be easier to follow if each panel were given unique identifying letters. Figures 2 and 12 are good, figures 3-11 would benefit from additional labels.

We agree to your suggestive comments. As described below in "Additional specific comments" section in detail, Figures 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 16, and 17 have been replotted with given unique identifying letters in each panel. Additionally, based on the suggestive comments from editors, the color schemes in some of the figures (i.e., Figs 2, 12, 13, 14, 15) have been revised for readers with color vision deficiencies.

Under the category "Major comments"

L228-238, L430-432: As this is not the first article to identify temperature and wind biases in

models, it would be appropriate to provide in the introduction at least a brief review or acknowledgment of the previous work on model temperature and wind biases (particularly the lower stratosphere zonal wind, and tropical tropopause temperature) (in addition to the brief mention of two examples on L555 at the end) and some summary of the existing knowledge. By providing a sufficient overview of the related work that has already been published you will be able to make it much clearer to readers the original contribution that this article makes in the context of the current state of knowledge.

In the conclusion part of the manuscript, the two papers, Hardiman et al. (2015) and Bland et al. (2021), have been being cited to discuss further studies. However, as the referee #1 suggested, it would be better if these papers could be moved to Section 1 "Introduction".
L68-L73, L652-L656: The two papers which are dedicated to investigating temperature biases in the MetUM, Hardiman et al. (2015) and Bland et al. (2021), have been moved from the Conclusion part to the Introduction part and summarized.

As I understand it, the residual term for CNTL is the sum of the parameterised processes, any numerical integration errors and differences arising from the primitive budget equations being different from the actual equations of the UM. The CNTL-GLN difference in the residual is the difference of these in the CNTL and GLN plus the nudging, which is the very short timescale model error in the UM. However there is no single, early, concise explanation of this.
While various aspects of the above are written in various places in various ways, references to components of this and interpretations of the residual often seem confusing. For example lines 302-304 present the similarity as if it were not true by definition. The article could benefit from a concise re-write of lines 327-335 to clearly explain the interpretation of the residual, and moved earlier in the section. Then any subsequent discussion of the residual or any of its components can be re-considered in the context of the earlier explanation to make the section overall more cohesive and consistent.

As the referee #1 mentioned, the residual term in our study is interpreted to be qualitatively equivalent to estimated unresolved forcing and contains the sum of the parameterized processes, any numerical integration errors, any budget diagnostics calculation errors (e.g., conversion from the model levels to the pressure levels, second-order centered difference scheme for meridional and vertical derivatives calculation, etc.), and differences of the governing equations of the UM and the primitive budget equations. Model tendencies are calculated in the model online during the time integration at each model level using dynamical

core and physics parameterizations. On the other hand, atmospheric budgets used in our study are evaluated from forecast data (specifically U, V, T, and Omega fields) on pressure levels as shown by the budget equations (Eq. (2) and (3)) offline after the completion of the time integration. The discrepancy between the governing equations of the UM (non-hydrostatic, fully compressible deep-atmosphere equations of motion) and the primitive equations could account for a difference between the model tendencies and atmospheric budget diagnostics.

The follow-up comment by referee #1 in terms of the tendency diagnostics is essentially correct, although budgets can be calculated using every model time step fields, since mean fields denoted by overbars, for instance $[\overline{u'v'}] = [\overline{uv} - \bar{u}\bar{v}]$ and $[\bar{u}^*\bar{v}^*] = [\bar{u}\bar{v}] - [\bar{u}][\bar{v}]$, are derived using every model time step fields.

In contrast to our study, previous studies (e.g., Milton and Wilson (1996)) present the budget residual using a different definition that directly refers to the model physics tendencies. Since the budget calculations in this study do not refer to model tendencies explicitly, we have mentioned that they are considered to be independent, to emphasize the difference in the definition compared to previous studies. However, they are implicitly linked through their governing equations, with different assumption and approximation as indicated by referee #1. We think it is inappropriate to consider the two diagnostics to be completely independent.

A single, early, concise explanation for this have been added to the manuscript. We have created two new sections, 2.5 and 2.6, to explain how these two diagnostics are linked and the difference in budgets/tendencies between CNTL and GLN. We have moved some explanations in Section 4 and other parts of the manuscript to bring together similar descriptions into one part, Section 2.5 and 2.6.

L246-L266: Section 2.5 has been added to explain interpretations of atmospheric budgets and model tendencies, and importance of the comparison to demonstrate their quantitative correspondence.

L267-L301: Section 2.6 has been added to describe how we validate the atmospheric budgets in CNTL against GLN to decompose an error in the total tendencies into individual budget components. For ease of understanding, we refer to a difference in budget components between CNTL and GLN as "quasi-error" in the revised manuscript.

L167-L168, L222-L223, L356-L358, L381-L382, L392-L400, L409-L413: The descriptions on the diagnostics that had sat in Section 4 and other parts have been moved to Section 2.5 and 2.6 for earlier explanations.

L358, L365, L401, L407, L414-L415, L417, L422-L423, L447, L451, L486, L487, L491, L512, L517, L528, L601-L602, L603, L604, L606-L607: We refer to a difference in budget components between CNTL and GLN as "quasi-error" in the revised manuscript.

A consequence of this is: taking the mean CNTL-GLN tendency from dynamics (4b, 6b, 8b, 10b), will this not be disproportionately dominated by the model dynamics bias in the first 6 hours being balanced by the nudging, therefore masking any information about differences between CNTL and GLN at timescales longer than this? If this is what you are trying to address in lines 320-324 it is not clear. With terms being so heavily dominated by the nudging, what information can be gotten from the dynamics/residual at timescales of longer than 24 hours in any quantities including data from GLN?

In the atmospheric budget analysis shown in Fig. 4 and Fig. 8, we use GLN with 6-hour relaxation timescale to make quasi-analysis data and use as a best estimate of truth of atmospheric budgets and verify the budgets in CNTL against GLN. In the nudging experiments, forced variables (i.e., U, V, and T) are relaxed towards the MetUM analysis data every model time step throughout the integration up to 15 days in this study. Therefore, the GLN has small error against MetUM analysis as shown in Fig. 2(d) even at longer forecast lead time than 24 hours. This is the reason we use the atmospheric budgets in GLN as a best estimate of truth (GLN is equivalent to a data assimilation/analysis as used in this study).
The choice of the relaxation timescale of nudging experiments is arbitrary. The shorter the relaxation timescale is, the more strongly the model prognostic variables are relaxed towards forcing data (i.e. MetUM analysis data) throughout the model integration. As described in the cited previous study (Telford et al. 2008), too long relaxation timescale is ineffective to use a nudging experiment as a best estimate of truth, but too short relaxation timescale nudging could make the model unstable.
Differences in the atmospheric budgets between CNTL and GLN shown in Figs 4~~(b)~~(d-f) and 8~~(b)~~(d-f) exhibit the error in the atmospheric budgets of CNTL against GLN. On the other hand, the model tendencies shown in Fig. 6~~(b)~~(e-h) and Fig. 10~~(b)~~(e-h) are just the difference between experiments and doesn't show any error in the model dynamics or model physics parameterizations at all.
L132-L136: In order to explain how the nudging works and why the 6-hour relaxation timescale was selected, we have added some explanations described above.
L168-L169: The reason why relatively short relaxation time scale is used in the nudging experiment in this study has been described.
L296-L301: The last paragraph in Section 2.6 explains interpretations of differences in model tendencies between CNTL and GLN, which is a difference in the response to the different grid-box mean fields rather than error in the model tendencies.
L392-L400, L403: Explanations of non-negligible error in GLN budgets against truth have

been moved to Section 2.6.

It is not clear that figure 4 provides any useful information not contained within figures 5 and 6 (and similarly that figure 8 does not show anything not in figures 9-11) other than the contour overlay which could be moved. It may be possible to make this section more concise by removing figures 4 & 8.

We feel that Fig. 4 does provide key information on overall model systematic bias in zonal wind whereas Fig. 5 provides the breakdown of that error into budget components coming from different aspects of the general circulation in the model (CNTL) and analysis (GLN). Figure 6 shows the equivalent breakdown from the tendency diagnostics (on model levels) which brings in the role of parametrized sub-grid scale physical forcing (gravity wave drag etc.) alongside the resolved circulation or Dynamics (made up of the mean flow, stationary & transient eddies, and Coriolis terms shown in Fig. 5.). Similar arguments apply for Fig. 8 with respect to Figs. 9 and 11. We think Figs. 4 and 8 are essential for conclusion of this study in order to demonstrate a correspondence between budget diagnostics shown in Figs. 4(a-c) and 8(a-c) and model tendencies shown in Figs. 6(a-c) and 10(a-c) and exhibit a quasi-error in resolved processes and unresolved processes against GLN shown in Figs. 4(d-f) and 8(d-f).
Figures 4-11: Their captions have been revised for clarification.

L323: This information not shown would perhaps benefit from being shown. With such an emphasis on the Coriolis term in the zonal wind budget, the article would benefit from the inclusion of further discussion of any biases in the meridional wind field.

Thank you so much for your informative comments.
An example of meridional wind error of CNTL and GLN against analysis is shown below:

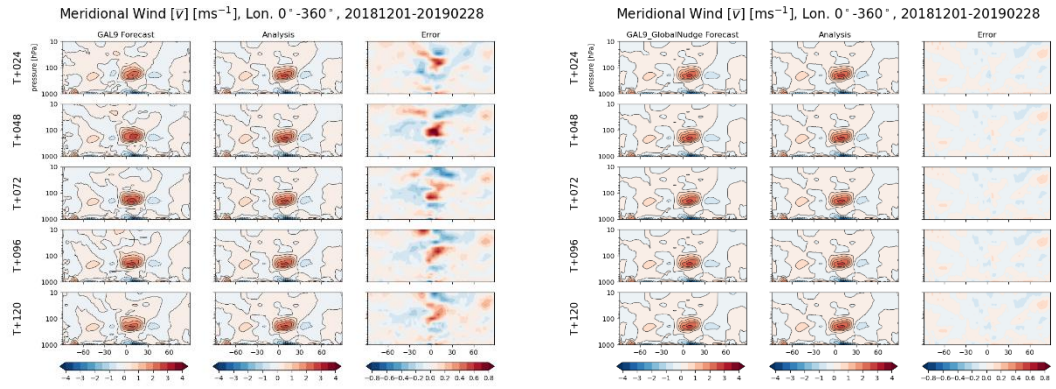

Figure A1 (Left) Zonal-mean meridional wind of CNTL and MetUM analysis, and error in

meridional wind of CNTL against analysis from Day 1 to Day 5, and (Right) zonal-mean meridional wind of GLN and MetUM analysis, and error in meridional wind of GLN against analysis from Day 1 to Day 5. As can be seen from Fig. A1, the GLN has a much smaller error in meridional wind than CNTL.

We have evaluated the meridional momentum budget equation (not shown in this paper) that can be derived in the same manner as the zonal momentum budget equation (Eq. (2)). These additional budget diagnostics are expected to provide useful information on meridional wind fields. The reason why the meridional momentum budget diagnostics have been omitted from our paper is that they are less important than zonal momentum and thermal budgets in the context of this study. In the meridional momentum budget, northward geopotential gradient term and southward Coriolis forcing term are almost balanced (i.e. geostrophic balance) and other terms including temporal derivative term, momentum flux convergence term, and residual term are less dominant in the meridional momentum budget.

L330-331: Is this conclusion necessarily true? Unsure if the results provided support this.

Error of a given variable against analysis can be written below:

$$
\begin{aligned}
& x_e(t_i;\Delta t) \\
& \equiv x_f(t_i;\Delta t) - x_a(t_i+\Delta t) \\
& = \{x_f(t_i;\Delta t) - x_f(t_i;0)\} - \{x_a(t_i+\Delta t) - x_a(t_i)\} \\
& = \left\{\overline{\left(\frac{\partial x}{\partial t}\right)_{f,t_i}} - \overline{\left(\frac{\partial x}{\partial t}\right)_{a,t_i}}\right\}\Delta t
\end{aligned}
$$

where $x_e(t_i;\Delta t)$ is an error at a forecast lead time of Δt initialized at t_i, $x_f(t_i;\Delta t)$ is a forecast at a lead time of Δt initialized at t_i, $x_a(t_i+\Delta t)$ is an analysis at $t_i+\Delta t$.

Similarity between residual term error and total tendency error, which is proportional to the error of the corresponding variable itself as shown in the equation described above, suggests that the error of the corresponding variable itself stems from the residual term error. As shown by comparison of Fig. 4~~(a-3)~~(c) with Fig. 6~~(a-3)~~(c), the residual term in the budget equation is qualitatively equivalent to the physics parameterization representing unresolved processes in the model. These results suggest the conclusion described in L~~330-331~~412-413.

L268-L278: We have included the equation described above in Section 2.6 newly added for ease of understanding of relation between model errors and model tendency errors.

L409-L413: The descriptions on the interpretation of the residual term have been moved to Section 2.5 and 2.6 and written for earlier concise explanations.

L367-370: The sign of the CNTL-GLN difference in physics is the opposite of the sign of the

error in 10-b-3 and 10-b-1, which seems to suggest the opposite to what this sentence is saying.

We understand that you have questioned the opposite sign of the difference between two experiments shown in Figs. ~~10-b-3~~10(g) and ~~10-b-1~~10(e) particularly in the upper stratosphere above 35km altitude. Figures ~~4(b)~~4(d-f), 5~~(b)~~(e-h), 8~~(b)~~(d-f), and 9~~(b)~~(d-f) show differences in individual components of the atmospheric budget equations between CNTL and GLN, which can be interpreted as an error of the individual budget components against GLN, the best estimate of truth in this study. On the other hand, Figures 6~~(b)~~(e-h) and 10~~(b)~~(e-h) show differences in model tendencies between CNTL and GLN. For instance, a difference in physics parameterization tendencies is a difference in physics parameterization responses to CNTL and GLN grid-box mean fields given to the parameterization schemes. Physics parameterization scheme itself has deficiencies and that used in CNTL and GLN is identical. Figure 10~~(b)3~~(g) shows a difference in the response of physics parameterizations rather than error in physics parameterization tendencies. ~~L367-370~~L452-455 is not intended to describe this difference in physics parameterization tendencies.
L296-L301: It has been explained that the difference in physics parameterization tendencies between CNTL and GLN is a difference in physics parameterization responses to CNTL and GLN grid-box mean fields given to the same parameterization schemes.

L323-324: By "the error in Coriolis term against truth" do you mean the error in the Coriolis term in CNTL against truth? If so, state this.

As indicated by the referee #1, this sentence mentions "the error in the Coriolis term in CNTL against truth".
L406: We have revised this sentence.

Under the category "Additional specific comments"

All figures: It would benefit the reader if each individual panel were assigned a letter identifier (a), (b), etc.

We agree to your suggestive comments. Figures 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 16, and 17 have been replotted with given unique identifying letters in each panel. Additionally, based on the suggestive comments from editors, the color schemes in some of the figures (i.e., Figs 2, 12, 13, 14, 15) have been revised for readers with color vision deficiencies.
Figure 1 has subtitles with given unique identifying letters (a-f) for each panel with their explanations in the caption.

Figure 2 has been changed to a new color scheme using the Coblis, as editors provided suggestive comments.
Figure 3 has subtitles with given unique identifying letters (a-d) for each panel with their explanations in the caption.
Figure 4 has subtitles with given unique identifying letters (a-f) for each panel with their explanations in the caption.
Figure 5 has subtitles with given unique identifying letters (a-h) for each panel with their explanations in the caption.
Figure 6 has subtitles with given unique identifying letters (a-h) for each panel with their explanations in the caption.
Figure 7 has subtitles with given unique identifying letters (a-c) for each panel with their explanations in the caption.
Figure 8 has subtitles with given unique identifying letters (a-f) for each panel with their explanations in the caption.
Figure 9 has subtitles with given unique identifying letters (a-f) for each panel with their explanations in the caption.
Figure 10 has subtitles with given unique identifying letters (a-h) for each panel with their explanations in the caption.
Figure 11 has subtitles with given unique identifying letters (a-g) for each panel with their explanations in the caption.
Figure 12 has been changed to a new color scheme using the Coblis, as editors provided suggestive comments.
Figure 13 has been changed to a new color scheme using the Coblis, as editors provided suggestive comments.
Figure 14 has been changed to a new color scheme using the Coblis, as editors provided suggestive comments.
Figure 15 has been changed to a new color scheme using the Coblis, as editors provided suggestive comments.
Figure 16 has subtitles with given unique identifying letters (a1-d6) for each panel with their explanations in the caption.
Figure 17 has subtitles with given unique identifying letters (a1-d6) for each panel.
L321, L347, L350, L369, L371, L379, L383, L384, L385, L387, L402, L415, L423, L424, L432, L435, L436-L437, L441, L447, L448, L452, L458, L474, L508, L522, L567, L569, L570, L578, L587: Changes due to a reference to the panels.

L33-34: There have been a wide variety of diagnostic methods… It would be appropriate to

list some.

Thank you so much for your suggestive comments.
L32-L37: Some papers have been cited as an example of the diagnostic methods in the revised manuscript including single-column models experiments (e.g., Duynkerke et al. 2004; Lenderink et al. 2004; Svensson et al. 2011), model intercomparison projects (e.g., van Niekerk et al. 2020, Elvidge et al. 2019), WGNE conferences on model systematic errors (Frassoni et al. 2023), potential vorticity budget diagnostics (e.g., Chagnon et al. 2013; Saffin et al. 2016), semi-geostrophic balance tool (Sánchez et al. 2020), perturbed parameter ensemble technique (Sexton et al. 2019; Karmalkar et al. 2019; Williams et al. 2020).

L124-128: I can discern what you mean, but it would be beneficial to re-write these sentences to make them clearer.

Thank you so much for your comments. The sentences in L~~124~~138-L~~128~~142 have been rewritten in the revised manuscript.
L138-L142: An explanation of the nudging increments or nudging tendencies has been revised to make it clearer.
L413-L414: A similar description here has been revised as well.

L191: This is not a general property of all partial differential equations, so remove these opening 4 words (or re-phrase the sentence to make it more accurate).

In this paragraph, we would like to explain a difference in the budget equations between the partial differential equations themselves shown in Eq. (2, 3) and discretized version of the equations. Equations considered in the numerical model or the diagnosed forecast/analysis data are the discretized ones. As you mentioned, this is not a general property of all partial differential equations.
L208-L211: The sentences have been written for clarification.

L253-254: I do not see evidence for this statement. Provide evidence, or clarify that this is not shown.

The sentence in L~~253~~328-~~254~~329 "*The rapid error growth and consistent drifts suggest that these errors are relevant to the model physics rather than changes in the atmospheric circulation.*" is based on the suggestion by previous work (e.g. Martin et al., 2010; Ma et al.,

2014; Martin et al., 2021) cited in the Introduction part.
L330: We have cited the previous studies (e.g. Martin et al., 2010; Ma et al., 2014; Martin et al., 2021) to support this suggestion.

L262-263 & Figures 2d, 3b, 12a, 12b: The independence of GLN state with forecast lead time seems strange. Is the GLN state at lead times >24 hours truly identical? Can you explain this or comment further on why the differences are zero or imperceptibly small?

The GLN at forecast lead times except for 0 hours should be truly identical.
In the GLN, forced variables (i.e., U, V, and T) are relaxed towards the MetUM analysis data every model time step throughout the integration up to 15 days. Mean error can be formulated as follows

$$\mathrm{ME} = \sum_{i=1}^{N} \phi_e(t_i; \Delta t) = \sum_{i=1}^{N} \{\phi_f(t_i; \Delta t) - \phi_a(t_i + \Delta t)\}$$

where $\phi_e(t_i; \Delta t)$ is an error at a forecast lead time of Δt initialized at t_i, $\phi_f(t_i; \Delta t)$ is a forecast at a lead time of Δt initialized at t_i, $\phi_a(t_i + \Delta t)$ is an analysis at $t_i + \Delta t$, and N is the number of the cases. For example, mean zonal wind error at 1-day forecast lead time is calculated from 90 cases initialized between at 30th Nov. 2018 and at 27th Feb. 2019, and mean zonal wind error in 15-day forecast lead time is calculated from 90 cases initialized between at 16th Nov. 2018 and 13th Feb. 2019. Regardless of forecast lead time, mean zonal wind error is calculated for the fixed validity period from 1st Dec. 2018 to 28th Feb. 2019. It is obvious that the MetUM analysis used as the forcing data doesn't depend on forecast lead time. Therefore, model prognostic variables forced in nudging experiments are relaxed towards the same analyzed fields regardless of forecast lead time, which results in the same forecast fields at specific validity time. This can account for the reason why the forced variables in the nudging experiments have an error independent of forecast lead time.
L340-L343: Additional concise explanations of consistent error in GLN regardless of forecast lead time have been added.

L292-293: a decrease in the tropics... State, a decrease in what in the tropics. Similarly for lines 293-295.

Thank you for your suggestive comment.
We have revised the manuscript as below:
L371: a decrease in *zonal-mean zonal wind in* the tropics with ...
L372-L373: a decrease in *zonal-mean zonal wind in* the mid-latitude with ...

L374: an increase in *zonal-mean zonal wind in* the subtropic with …

Figure 12-14: Are the differences between the dynamics and the sum of the resolved flow components, and the differences between the residual term and the sum of the nudging + physics, in panels f compared to panels j, larger than one would expect resulting from the model level v.s. pressure level comparison? They seem quite large by eye. Can you comment on this please?

In Figs 12-14, panels (f) show error in atmospheric budget components of CNTL against GLN, and panels (j) show differences in model tendencies between CNTL and GLN. As we mentioned above, a similarity between them is not necessarily true by definition even on the same vertical coordinate (i.e., given model level v.s. the model level, or given pressure level v.s. the pressure level). In addition to this, we compare the atmospheric budgets on given pressure levels (i.e., 50 hPa in Fig 12, 70 hPa in Fig 13, and 200 hPa in Fig. 14) in panels (c-f) and model tendencies on the specific model levels closest to the respective pressure levels. Thus, the larger difference between panels (f) and panels (j) than one would expect results from not only the difference of their vertical coordinate but also the two different diagnostics which are not necessarily equivalent by definition due to the different governing equations and the diagnostics errors.

L479-480: This result seems important but it is not shown in this article (nor do you state in the text that it is not shown). With such an emphasis on the Coriolis term in the zonal wind budget, the article would benefit from the inclusion of further discussion of any biases in the meridional wind field.

The Coriolis term in the zonal momentum budget equation, fv, is proportional to meridional wind speed, v. Therefore, the change in zonal-mean Coriolis term between specific two experiments shown in Fig. 16, $f\bar{v}_{\mathrm{EXP}} - f\bar{v}_{\mathrm{CNTL}} = f(\bar{v}_{\mathrm{EXP}} - \bar{v}_{\mathrm{CNTL}})$, is proportional to the impact of nudging on the meridional wind fields, although a latitudinal dependency of Coriolis parameter $f = 2\Omega \sin\phi$ needs to be considered. As shown in Figure A1 in this response, CNTL has a southerly wind bias in NH subtropics in the lower stratosphere. As can be seen in Fig. 5~~(b)4~~(h) and Fig. 16~~(a)5~~(a5), the error in Coriolis term against GLN can be reduced by temperature nudging, which implies changes in meridional wind. We would rather put the focus on the zonal momentum budgets and implications of meridional wind from Coriolis term diagnostics in this paper.

L567: A word "(not shown)" has been put in to just after "which is clearly demonstrated by a

meridional momentum budget analysis".

L483-485: Returning to earlier comments on the interpretations of the residual term – this idea would benefit from further explanation and possibly inclusion of the "not shown" material.

We hope that the earlier explanation rewritten in Section 2.5 and 2.6 in the revised manuscript could be beneficial. In the NHTrpT and NHTrpTrpT, temperature is relaxed towards the analysis but zonal wind is not relaxed at all. Therefore, zonal wind tendency due to nudging in the NHTrpT and NHTrpTrpT is completely zero. The sentence "The difference in the residual term of NHTrpT and NHTrpTrpT cannot be fully accounted for by changes in the physics parametrization tendencies (not shown)" describes a discrepancy of the difference in the residual term and the difference in the parameterization tendencies.
L570-L574: This sentence has been rewritten for clearer explanations.

Under the category "Technical Corrections"

L29: within the first few days
L133-134: 10 degrees in the horizontal … two model levels in the vertical
L136: from December 2018 to February 2019 (or from the December of 2018 to the February of 2019)
L206: the word 'definitely' doesn't need to be here
L242: we focus on three…
L262: but do not depend…
L538: "specially" is the wrong word to use here. One way to rewrite this would be for example "one promising way for operational …"

These technical corrections are really helpful.
All of the technical corrections have been reflected to the revised manuscript as follows:
L29: within *the* first few days
L149: 10 degrees in *the* horizontal … two model levels in *the* vertical
L152: from ~~the~~ December 2018 to ~~the~~ February 2019
L225: ~~definitely~~
L317: we focus *on* three
L337: but ~~does~~*do* not depend
L633: One promising way ~~specially in~~*for* operational

The reviewer would like to thank the authors for their comprehensive study on model-based diagnostics to identify sources of model systematic errors, which could be useful for any NWP and climate models. The method successfully detects sources and origins of model forecast errors in the Met Office Unified Model, which could be common in the state-of-the-art operational and research atmospheric GCMs. I think The paper is comprehensive and well organized as a WCD paper, so that I would already recommend minor revision, though I have several comments below before acceptance.

Thank you so much for your reviewing and suggestive comments. We respond to your specific comments below.

Under the category "Specific comments"

1) The relaxation timescale of 6 hours could be practically useful to reduce/erase the initial errors that can grow up in subsequent forecasts. I am curious that how sensitive is "tau" (timescale) to your conclusion? For example, if you choose 12 or 24 hours as tau, you could obtain consistent results, that is, could detect the same origins causing the NH wind field errors?

The relaxation timescale might be similar with the evaluation time of the Forecast Sensitivity to Observations (Hotta et al. 2017, Prive et al. 2020). The shorter timescales might be more practical for forecasts.

(Hotta et al. 2017, MWR, DOI:10.1175/MWR-D-16-0290.1; Prive et al. 0220, QJRMS, DOI:10.1002/qj.3909)

We use the GLN experiment to make quasi-analysis data and use it as a best estimate of truth of atmospheric budgets. In the nudging experiments, forced variables (i.e., U, V, and T) are relaxed towards the MetUM analysis data every model time step throughout the integration up to 15 days in this study. The relaxation timescale "tau" determines how strongly the model prognostic variables are relaxed towards forcing data rather than how long the nudging is applied to. Therefore, the GLN has small error against MetUM analysis as shown in Fig. 2(d) even at longer forecast lead time than 6 hours.

The choice of the relaxation timescale of nudging experiments is arbitrary. The shorter the relaxation timescale is, the more strongly the model prognostic variables are relaxed towards forcing data (i.e. MetUM analysis data) throughout the model integration. As described in the cited previous study (Telford et al. 2008), too long relaxation timescale is ineffective to use a nudging experiment as a best estimate of truth, but too short relaxation timescale nudging could make the model unstable.

This paper doesn't focus on a difference in the relaxation timescale, but it sounds interesting and might provide an interesting result which could be useful for understanding different timescales of error contribution from each component in the budget equations.
L132-L136: In order to explain how the nudging works and why the 6-hour relaxation timescale was selected, we have added some explanations described above.
L168-L169: The reason why relatively short relaxation time scale is used in the nudging experiment in this study is described.

2) Could you comment in your conclusion part about two topics that could be useful as future studies?

Thank you so much for your suggestive comments.
These specific comments are really helpful to improve our manuscript and our response to the two comments kindly given to us are described below.

- Budget analysis based on conservation values, potential vorticity or angular momentum (TEM or MIM), could be useful for your further budget analysis. Currently I only refer to a paper by Kobayashi and Iwasaki (2016), though there may be other important studies to be cited.
(Kobayashi and Iwasaki 2016, JGR-A, DOI:10.1002/2015JD023476)

Thank you so much for your informative suggestions.
L32-L37, L623-L626: We have cited papers on potential vorticity budget analysis (Chagnon et al. 2013; Saffin et al. 2016) in the Introduction part and angular momentum equations (TEM: Andrews and Mcintyre, 1976; MIM: e.g., Iwasaki, 1989, 1992; Tanaka et al., 2004) in the conclusion part.

- As a comparison with the experiment for 2018 boreal winter, additional experiment for an SH winter could be helpful to improve your conclusion (e.g., Yamazaki et al. 2023).
(Yamazaki et al. 2023, WAF, DOI:10.1175/WAF-D-22-0159.1)

We agree to your suggestive comments. Some conclusions in our study depend on seasons and hemispheres. For example, gravity wave drag, which could be one of the sources of the wind error, is expected to be more active in NH because of the orographic features and also in winter because of the wind fields and atmospheric stabilities near land surface. Even if the temperature bias around tropical tropopause wouldn't depend on season, balanced situation

of the compensating errors would change and result in different consequences. We are wondering if a focus on specific season and hemisphere makes this paper detailed and less confusing.

L647-L650: We have implemented some comments described above into the Conclusion part in a revised manuscript. We hope this could help to improve our conclusion and discussion.

Changes to the manuscript not mentioned above

L126-L128: We have changed the notation for the time interval of the model integration because of the distinction from the forecast length denoted by Δt in L273-L275. The time interval of the model integration adopted in this study has been clarified in L128.

L179: Rewritten for clarification

L202-L203: Rewritten for clarification

L337-L340: Two sentences describing potential errors even in the forced variables in nudging experiments have been moved to Section 2.6 newly created.

L353: Technical correction

L354: Technical correction

L360: Rewritten for clarification

L377-L379: This sentence is not necessary for our conclusion, so it has been deleted for ease of reading

L380: Rewritten for improvements

L382-L383: Parameterizations affecting sub-grid momentum transport has been listed for clarification.

L389-L390: Rewritten for improvements

L405-L406: Rewritten for ease of reading, associated with the “quasi-error” of CNTL against GLN

L476-L479: Rewritten for clarification of explanations of each panel in Figure 12.

L491: Rewritten for clarification

L500-L504: Two sentences have been reordered for ease of reading.

L519: Technical correction

L553, L554: Rephrased for improvements

L610: For clarity, error compensation in *what* is described.

L619: Rewritten for ease of reading

L642-L643: A revision has been added to make the meaning of the word “physical” clearer.