Response to the comments from referee on "Process-based diagnostics using atmospheric budget analysis and nudging technique to identify sources of model systematic errors" by C. Matsukawa et al.

We thank the referee #1 for their valuable comments and constructive suggestions for improving the paper. We will make extensive corrections to a revised manuscript. The comments by referee #1 are listed below in black italics, followed by our responses in blue.

Response to 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 based on your comments and our responses to them below. Your suggestions of some additional references are very informative. We will cite more related work in a revised manuscript.

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. Figures 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 16, and 17 will be replotted with given unique identifying letters in each panel.

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" and additional papers could be cited there as an example of previous studies.

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.

As the referee #1 mentioned, the residual term in our study is interpreted to be qualitatively equivalent to estimated unresolved forcing and contain 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. However, we would like to emphasize that the model tendencies diagnostics (i.e., Figs. 6, 7, 10, 11, 12(g-j), 13(g-j), 14(g-j)) and atmospheric budget diagnostics (i.e., Figs. 4, 5, 8, 9, 12(c-f), 13(c-f), 14(c-f) are completely independent in their calculations. Model tendencies are calculated in the model 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)) after the completion of the time integration. The atmospheric budgets calculation doesn't require the model tendencies calculated by dynamical core and physics parameterizations at all. That's why we think that the atmospheric budget diagnostics can be compared with the model tendencies to demonstrate that the resolved term and the unresolved term in the atmospheric budget equation are qualitatively equivalent to the model tendencies of dynamics and physics parameterizations respectively (as shown by a comparison between Fig. 4(a) and Fig. 6(a) and between Fig. 8(a) and Fig. 10(a)). As you mentioned, 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.
We agree to your comment and a single, early, concise explanation for this will be added to the manuscript. We will create Section 2.5 and 2.6 newly and move some explanations in Section 4 to explain how these two independent diagnostics link and the difference in budgets/tendencies between CNTL and GLN can be interpreted in a revised manuscript.

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 we mentioned above, the model tendencies diagnostics and atmospheric budget diagnostics are completely independent. Therefore, the similarity between Fig. 4 and Fig. 6 is not necessarily true by definition and is worth being demonstrated. We agree that the earlier explanation of independency between the model tendencies diagnostics and the atmospheric budget diagnostics will improve this paper. We will create Section 2.5 and 2.6 newly to explain how these two diagnostics link in a 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 stronger 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) and 8(b) exhibit the error in the atmospheric budgets of CNTL against GLN. On the other hand, the model tendencies shown in Fig. 6(b) and Fig. 10(b) are just the difference between experiments and doesn't show any error in the model dynamics or model physics parameterizations at all.

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.

To provide a little more detail , calculations of the model tendencies diagnostics (i.e., Figs. 6, 7, 10, 11, 12(g-j), 13(g-j), 14(g-j)) and atmospheric budget diagnostics (i.e., Figs. 4, 5, 8, 9, 12(c-f), 13(c-f), 14(c-f) are independent. The atmospheric budgets diagnostics don't require the model tendencies calculated by dynamical core and physics parameterizations during the model integration. That's why we think that the atmospheric budget diagnostics (Fig. 4) can be evaluated against the model tendencies (Fig. 6) to demonstrate that the resolved term and the unresolved term in the atmospheric budget equation (Fig. 4) are qualitatively equivalent

to the model tendencies of model dynamics and physics parameterizations (Fig. 6) respectively. Figs 4 and 8 (zonal momentum budget and thermal budget) are necessary to compare with Figs 6 and 10 (model zonal wind tendency and temperature tendency).

Figure 5 shows a breakdown of the resolved processes which consist of mean stationary flow component, stationary eddy component, transient eddy component, and Coriolis forcing as shown by Eq. (2). Therefore, the resolved processes shown in the middle panel of Fig. 4 is completely equal to the sum of the 4 components shown in Fig. 5 (as mentioned in the caption of Fig. 4). Similarly, Fig. 7 shows a breakdown of the physics parameterization tendency which consist of convection, boundary layer, and gravity wave drag. Therefore, physics parameterization tendency shown in the third column is completely equal to the sum of the 3 parameterization tendencies shown in Fig. 7.

Similarly, Figs 8-11 show the comparison between two independent diagnostics and their breakdown for temperature.

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:

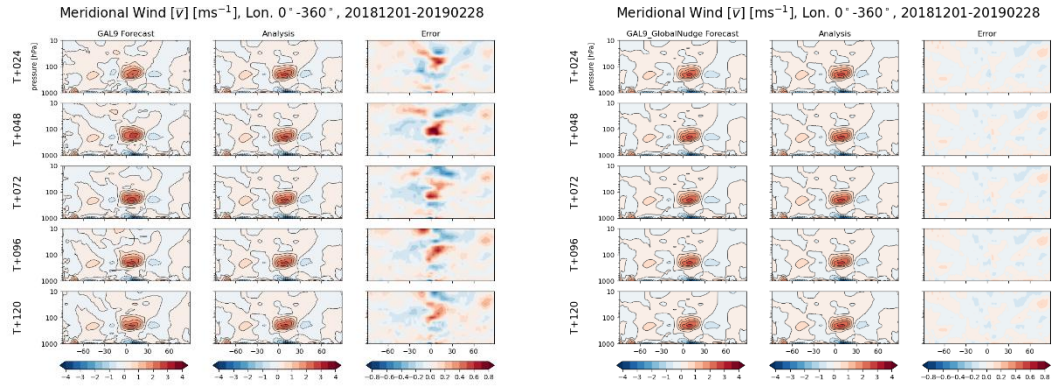

Figure A1 (Left) Zonal-mean meridional wind of CNTL and MetUM analysis, and error in meridional wind of CNTL against analysis from Day 1 to Day 5, and (Right) zonal-mean meridional wind of GLN and MetUM analysis, and error in meridional wind of GLN against analysis from Day 1 to Day 5. As can be seen 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 because 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 relatively small 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 $t_i+\Delta t$. This equation would be helpful and we will include it in Section 2.6 newly added.
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) with Fig. 6(a-3), 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 L330-331.

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 and 10-b-1 particularly in the upper stratosphere above 35km altitude. Figures 4(b), 5(b), 8(b), and 9(b) 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) and 10(b) 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 calculations. Physics parameterization scheme itself has deficiencies and that used in CNTL and GLN is identical. Figure 10(b)3 shows a difference in the response of physics parameterizations rather than error in physics parameterization tendencies. L367-370 doesn't describe this difference in physics parameterization tendencies.

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". We will revise 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 will be replotted with given unique identifying letters in each panel.

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.
The diagnostic methods and their activities to evaluate model systematic errors to be listed include, for instance, single-column models experiments (e.g., Duynkerke et al. 2004; Lenderink et al. 2004; Svensson et al. 2011), model intercomparison projects (e.g., Zadra et al. 2013; van Niekerk et al. 2020, Elvidge et al. 2019), WGNE conferences on model systematic errors (Zadra et al. 2018; 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). We will cite some of these papers in a revised manuscript.

- Duynkerke et al. 2004: Observations and numerical simulations of the diurnal cycle of the EUROCS stratocumulus case, QJRMS, https://doi.org/10.1256/qj.03.139
- Lenderink et al. 2004: The diurnal cycle of shallow cumulus clouds over land: A single-column model intercomparison study, QJRMS, https://doi.org/10.1256/qj.03.122

- Svensson et al. 2011: Evaluation of the Diurnal Cycle in the Atmospheric Boundary Layer Over Land as Represented by a Variety of Single-Column Models: The Second GABLS Experiment, *Boundary-Layer Meteorol*., https://doi.org/10.1007/s10546-011-9611-7
- Zadra et al. 2013: WGNE drag project, https://collaboration.cmc.ec.gc.ca/science/rpn/drag_project/
- van Niekerk et al. 2020: COnstraining ORographic Drag Effects (COORDE): A Model Comparison of Resolved and Parametrized Orographic Drag, JAMES, https://doi.org/10.1029/2020MS002160
- Elvidge et al. 2019: Uncertainty in the Representation of Orography in Weather and Climate Models and Implications for Parameterized Drag, https://doi.org/10.1029/2019MS001661
- Zadra et al. 2018: Systematic Errors in Weather and Climate Models: Nature, Origins, and Ways Forward, BAMS, https://doi.org/10.1175/BAMS-D-17-0287.1
- Frassoni et al. 2023: Systematic Errors in Weather and Climate Models: Challenges and Opportunities in Complex Coupled Modeling Systems, https://doi.org/10.1175/BAMS-D-23-0102.1
- Chagnon et al. 2013: Diabatic processes modifying potential vorticity in a North Atlantic cyclone, QJRMS, https://doi.org/10.1002/qj.2037
- Saffin et al. 2016: The non-conservation of potential vorticity by a dynamical core compared with the effects of parametrized physical processes, QJRMS, https://doi.org/10.1002/qj.2729
- Sánchez et al. (2020): Linking rapid forecast error growth to diabatic processes, https://doi.org/10.1002/qj.3861
- Sexton et al. (2019): Finding plausible and diverse variants of a climate model. Part 1: establishing the relationship between errors at weather and climate time scales, https://doi.org/10.1007/s00382-019-04625-3
- Karmalkar et al. (2019): Finding plausible and diverse variants of a climate model. Part II: development and validation of methodology, https://doi.org/10.1007/s00382-019-04617-3
- Williams et al. (2020): Addressing the causes of large-scale circulation error in the Met Office Unified Model, QJRMS, https://doi.org/10.1002/qj.3807

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.

These sentences describing the nudging increments or nudging tendencies will be revised to make them clearer.

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 is the discretized one. As you mentioned, this is not a general property of all partial differential equations. We think that the phrases "*In the partial differential equations of Eqs. (2) and (3),* " and "*the residual term diagnosed using spatially discretized data*" are clearer and will revise the manuscript so.

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

The sentence in L253-254 "*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. We will cite these papers again in L253-254.

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 on forecast lead time.

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 will revise the manuscript as below:
L292: a decrease in zonal-mean zonal wind in the tropics with …
L293: a decrease in zonal-mean zonal wind in the mid-latitude with …
L294: 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 tern 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 2 independent diagnostics which are not necessarily equivalent by definition.

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.

A word "(not shown)" will be put in to just after "which is clearly demonstrated by a meridional momentum budget analysis".
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, $f\bar{v}_{\mathrm{EXP}} - f\bar{v}_{\mathrm{CNTL}} = f(\bar{v}_{\mathrm{EXP}} - \bar{v}_{\mathrm{CNTL}})$, 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 and Fig. 16(a)5, the error in Coriolis term against GLN can be reduced by temperature nudging, which implies changes in meridional wind.
We will add further discussion of biases in the meridional wind fields related to model biases of zonal wind and temperature addressed in this study.

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.

As we mentioned above, calculations of the model tendencies and atmospheric budgets are completely independent. We suppose that the earlier explanation rewritten in a 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.
We will add some additional explanations above to a revised manuscript and rewrite this sentence.

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, and all of them will be reflected to the revised manuscript.