Thank you so much for your insightful comments.
These comments are helpful to make the manuscript better and easier for readers.
Our responses to your comments are written below.

As I understand it, the residual term for CNTL…
… 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.
…
Thank you for clarifying the agreement on the points that the tendencies and budgets are linked due to the governing equations, and that it is inappropriate to consider them independent. This did not come across in the original text and is useful to have clarified.
The only further note I have on section 2.5 is for line 259, I might suggest changing "but is not necessarily true by definition" to "but they are not necessarily exactly equivalent".
I assume based on lines 260-266 that this is the intended meaning, but as written I do not think the sentence means that. As written is seems something of a contradiction, as is the derivation of the budget not also the definition of the components?

Thank you for your proposal for an alternative description. As you indicated, the additional explanation of the description "but is not necessarily true by definition" follows in lines 260-266. I agree with you that "but they are not necessarily exactly equivalent" is more appropriate to explain the potential discrepancy between the budget diagnostics and the model tendencies. We have changed "but is not necessarily true by definition" to "but they are not necessarily exactly equivalent" for line 259.

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…
It may be the case that I was very unclear in my first review, and phrased this badly.
I do not think it is adequately addressed as to why the CNTL-GLN nudging in figure 6(h) appears to have a similar character but with opposite sign to the dynamics in figure 6(f), and whether this is important to the conclusions.

The reason why the CNTL-GLN nudging in Fig. 6(h) appears to have a similar character but with opposite sign to the dynamics in Fig. 6(f) will be explained below.

The similarity is noted in line 402, but in lines 413-414 you write that nudging increments are comparable to forecast errors. The forecast error in figure 1(c, e) has a similar character to the total tendency error in 4(d) and 6(e), however these do not have similar character to the resolved processes error (4e), the Coriolis error (5h) or the dynamics error (6f), which all instead are comparable to the nudging error (6h). Therefore it does not seem that nudging increments are comparable to forecast erros.

Figure 1(c,e) exhibits the mean error of the snapshot at Day 5 and 10 respectively, which are the medium-range timescale. The total tendency error averaged from Day 0 to Day 5 (Fig. 4(d) against GLN) is almost proportional to the error of the snapshot at Day 5 as indicated by Eq. (4). The structures and/or signs of the error could change within the first 5 days as illustrated by Fig. 2(d). Therefore, the error structure/sign in very-short-range timescale (e.g., 6 hours) does not necessarily last up to the medium-range timescale (e.g., 5 days).
Lines 413-414 mention that the nudging increments are comparable to the forecast errors "at least those growing over the 6-hour timescale", which is the very short-range timescale. The rationale behind this description is the nudging forcing formulation, the second term in the RHS of Eq. (1). The nudging forcing relaxes the model variables towards the analysis and reduces the error at the very end of each model timestep. The nudging forcing is proportional to the instantaneous error, $X_M - X_A$, at that point (similar to an analysis increment in data assimilation). Therefore, as you mentioned, we acknowledge that nudging increments are not necessarily comparable to the forecast errors themselves at the medium-range timescale. The result that the forecast error in Fig. 1(c,e) does not have similar character to the quasi-error in the resolved processes (Fig. 4(e)) or the unresolved processes (Fig. 4(f)) implies that the forecast error results from a balance/compensation between the resolved processes error and the unresolved processes error.

It is acknowledged in line 450 that the thermal budget and the momentum budgets are different. Here it is noted that there is similarity between the residual (8f) and the total (8d). This similarity is used to conclude deficiencies in physics parametrizations from the thermal budget. However by using the phrase "unlike the momentum budget", this highlights my confusion that you also conclude for the momentum budget that it is a deficiency in the physics parametrizations in lines 413-417, with similar phrasing, despite the opposite result. I interpret line 415-417 as saying that the model physics should be applying the same forcing that the nudging is, because you are using a nudged forecast as the truth.

A first look at figure 6 had me confused, as it seemed that the majority of the forcing of the nudging to move the GLN field towards the analysis was being directly un-done and opposed by the tendencies of the model dynamics. The majority of the nudging tendency being directly counteracted seemed to be an odd result, but as I have not personally used nudging in the same way as you are here in this study, it would be very useful if you could please provide your interpretation of what the cancellation of terms in figure 6(f) and 6(h) means, in terms of what the nudging is doing, how the model is responding, and how the differences between CNTL and GLN should be interpreted in light of this.

I feel like I really ought to be able to understand this, and so it may be likely that someone reading this article in the future would get similarly confused.

Furthermore, the nudging term in the thermal budget is not mentioned at all in the text of section 4.2, and as with the momentum is also of similar character and opposite sign to the resolved tendency, particularly at the lowest pressures.

According to Eq. (2) and (3), the residual terms are interpreted as an estimation of Friction/diabatic heating or parameterized forcing terms (i.e., model physics). Figure 4(c) looks like Fig. 6(c), which suggests that estimated offline unresolved tendency shown in Fig. 4(c) could approximate online physics parameterization forcing in CNTL. However, with respect to their difference between CNTL and GLN, Fig. 6(g) shows little difference, and Fig. 4(f) looks like Fig. 6(h), which exhibits nudging tendency in GLN with the reversed sign.

We acknowledge this, and this can be accounted for by decomposing the error in physics parametrization into atmospheric fields and scheme. In a sense, a particular physics parametrization to calculate tendencies is considered as a multivariable function as follows:

$$\left(\frac{\partial u}{\partial t}\right)_{\mathrm{Phy,CNTL}} = F_{u,\mathrm{Phy,Mdl}}(\phi_{1,\mathrm{CNTL}}, \cdots, \phi_{N,\mathrm{CNTL}})$$

where $F_{\mathrm{Phy,Mdl}}$ denotes a given parameterization scheme employed by the model (in other words, one subroutine in Fortran code). As described in L296-L301 in Section 2.6, the scheme $F_{\mathrm{Phy,Mdl}}$ used in CNTL and GLN is identical, while the input of the scheme such as background atmospheric fields is different between $\phi_{n,\mathrm{CNTL}}$ in CNTL and $\phi_{n,\mathrm{GLN}}$ in GLN. However, the scheme $F_{\mathrm{Phy,Mdl}}$ itself must have deficiencies against unknown truth of physics parameterization scheme $F_{\mathrm{Phy,Truth}}$. An actual error in tendencies calculated by physics parameterizations we would like to know can be described as follows:

$$\begin{aligned}\left(\frac{\partial u}{\partial t}\right)_{\mathrm{Phy,Error}} &= F_{u,\mathrm{Phy,Mdl}}(\phi_{1,\mathrm{CNTL}}, \cdots, \phi_{N,\mathrm{CNTL}}) - F_{u,\mathrm{Phy,Truth}}(\phi_{1,\mathrm{Truth}}, \cdots, \phi_{N,\mathrm{Truth}}) \\ &\neq F_{u,\mathrm{Phy,Mdl}}(\phi_{1,\mathrm{CNTL}}, \cdots, \phi_{N,\mathrm{CNTL}}) - F_{u,\mathrm{Phy,Mdl}}(\phi_{1,\mathrm{GLN}}, \cdots, \phi_{N,\mathrm{GLN}})\end{aligned}$$

The second line of this equation, $F_{\mathrm{Phy,Mdl}}(\phi_{1,\mathrm{CNTL}}, \cdots, \phi_{N,\mathrm{CNTL}}) - F_{\mathrm{Phy,Mdl}}(\phi_{1,\mathrm{GLN}}, \cdots, \phi_{N,\mathrm{GLN}})$, indicates what is shown in Fig. 6(g). Little difference in physics parameterization tendency between CNTL and GLN, shown in Fig. 6(g), implies that $F_{\mathrm{Phy,Mdl}}$ could be insensitive to differences in the inputs of the parameterization ϕ_n. Furthermore, this implies that $F_{\mathrm{Phy,Mdl}}$ itself has deficiencies against unknown $F_{\mathrm{Phy,Truth}}$. For this reason, Fig. 6(g) is not a good estimate of physics errors, and thus we diagnose the budget to estimate actual physics tendency using GLN experiment and estimate physics errors $\left(\frac{\overline{\partial u}}{\partial t}\right)_{\mathrm{Phy,Error}}$ in Fig. 4(f). In other words, provided that the physics scheme is very poor (it is expected to be the case), the physics tendency calculated using the scheme should have a large error even if accurate atmospheric fields are fed into the scheme.

In terms of L450 versus L413-417, physics parameterization schemes $F_{\mathrm{Phy,Mdl}}$ have deficiencies against $F_{\mathrm{Phy,Truth}}$ and cause substantial errors in the short-range timescale (e.g., earlier than Day 1) for both zonal wind and temperature. Meanwhile, the similarity in temperature suggests that the error caused by the physics is significant even at Day 5 whereas the non-similarity in zonal wind suggests that the error at Day 5 results from a balance/compensation between the physics error and the dynamical response to the errors.

On the other hand, a dynamical core itself generally has smaller deficiencies than a physics parametrization scheme because the governing equations are basically well established. In addition, model dynamics could be more sensitive to the atmospheric fields than model physics as the governing equation itself suggests - for example, wind fields strongly affect the advective term or the Coriolis term.

What is exhibited in Fig. 6(e-h) is written as follows:

$$\left(\frac{\overline{\partial u}}{\partial t}\right)_{\mathrm{Total,CNTL}} - \left(\frac{\overline{\partial u}}{\partial t}\right)_{\mathrm{Total,GLN}}$$

$$= \left(F_{u,\mathrm{Dyn,Mdl}}\left(\phi_{1,\mathrm{CNTL}}, \dots, \phi_{N,\mathrm{CNTL}}\right) - F_{u,\mathrm{Dyn,Mdl}}\left(\phi_{1,\mathrm{GLN}}, \dots, \phi_{N,\mathrm{GLN}}\right)\right)$$

$$+ \left(F_{u,\mathrm{Phy,Mdl}}\left(\phi_{1,\mathrm{CNTL}}, \dots, \phi_{N,\mathrm{CNTL}}\right) - F_{u,\mathrm{Phy,Mdl}}\left(\phi_{1,\mathrm{GLN}}, \dots, \phi_{N,\mathrm{GLN}}\right)\right)$$

$$+ \left(-F_{u,\mathrm{Ndg,Mdl}}(u_{\mathrm{GLN}})\right)$$

Figure 6(e,f,g,h) indicates the LHS, the first term, the second term, and the third term in the RHS, respectively. As mentioned above, the second term in the RHS is relatively small because of the physics scheme's insensitivity to the atmospheric fields. The LHS depends on forecast lead time until the error saturates. When the LHS and the second term in the RHS are small, the first term in the RHS should have similar structures with the opposite sign to the nudging forcing to close the tendency budget.

Although further investigations are needed for justification, a scientific rationale of similar structures with the opposite sign between model dynamics response (Fig. 6(f)) and the nudging forcing (Fig. 6(h)) attempts to be explained here with an extremely simplified example. Since Fig. 5(h) indicates that the dynamics tendency difference (Fig. 6(f)) stems from Coriolis forcing difference between CNTL and GLN, the Coriolis response to the nudging forcing in GLN is addressed here. If positive zonal wind forcing were exerted by the nudging, negative zonal wind error in CNTL (imagine too strong parameterized drag, for example) would be removed in GLN. This change in zonal wind velocity leads to larger southward (northerly) Coriolis forcing in meridional momentum than CNTL, and then meridional wind velocity should decrease in GLN relative to CNTL. The decreased meridional wind is expected to cause weaker eastward (westerly) or stronger westward (easterly) Coriolis forcing. This response of Coriolis forcing to the positive nudging forcing is interpreted as a counteraction, which accounts for the opposite sign of the model dynamics against the nudging forcing.

Another simplified example for temperature may also be able to account for the dynamical response (Fig. 10(f)) to the nudging forcing (Fig. 10(h)). Since Fig. 9(d) indicates that the dynamics tendency difference (Fig. 10(f)) is dominated by the mean stationary flow component, mean flow response to the nudging forcing is focused on. If positive temperature forcing would be exerted by nudging (i.e., artificial heating), negative temperature error in CNTL should be removed in GLN. This change in temperature causes vertical motions relative to CNTL and then the artificial heating is expected to be compensated by diabatic cooling with ascents or mean flow advection. This is consistent with the opposite sign of the model dynamics against the nudging forcing in temperature.

This explanation may be too simplified, and it might be not necessarily the case in the model. Please keep in mind that there must be more complex interactions in GLN among model dynamics, model physics, and artificial nudging forcing.

To summarize what happens in GLN, nudging forcing reduces errors caused mainly by deficient model physics at very short-range timescale, and then dynamical core responds to constrained atmospheric fields by the artificial forcing without physical consistency to counteract nudging forcing whereas model physics doesn't respond to the forcing strongly.

In the manuscript, how the errors are generated in CNTL against GLN is addressed for better understanding of error sources. In L413-418 for zonal wind and L450-455 for temperature, we intend to explain our interpretations from CNTL's error perspective (i.e., a dynamical response in CNTL to the error) rather than what happens in GLN (i.e., a dynamical response in GLN to the nudging forcing which looks like the error with the reversed sign), while we think they indicate one thing seen from different perspectives (either 'CNTL minus GLN' shown in this paper or 'GLN minus CNTL').

With respect to thermal budgets and tendencies (Figs. 8-11), some of the physics parameterizations (especially longwave radiation scheme) could be more sensitive to the atmospheric fields than those for zonal momentum. In terms of the opposite impacts of the parameterization (Fig. 10(g)) relative to the total tendency (Fig. 10(e)) in the upper troposphere, a decomposition of the physics parameterization difference into the individual processes (i.e., convection, radiation) suggests that the longwave radiation scheme seems to generate more cooling with constrained higher temperature fields and less cooling with constrained lower temperature fields (see Fig. B1 below). We suppose that this could be accounted for by Stefan–Boltzmann law as first approximation, although what actually happens in the longwave radiation scheme is much more complex. This implies that the longwave radiation scheme consisting of physics parameterizations works to counteract a part of the temperature change.

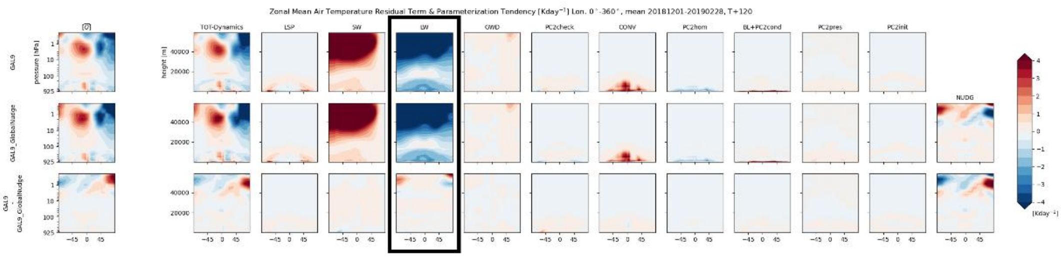

Fig. B1 Latitude-height cross section of differences in the residual term of the zonal-mean thermal budget equation (the leftmost column) and zonal-mean temperature tendencies of individual processes in (Top) CNTL, (Middle) GLN, (Bottom) CNTL-GLN. The panel with black bold box shows longwave radiation tendencies. Note that the colour scale and the plotted uppermost altitude are different from those in Fig. 10 in the manuscript.

Hopefully the explanation above will clear up reviewer 1's questions.
If we added some explanations about what happens in GLN, descriptions from two perspectives, CNTL and GLN, coexist, and thereby they may bring some confusion. To avoid additional confusion, we would rather not include detailed explanations about what happens in GLN and how dynamical core and physics parameterizations response to the nudging forcing in the manuscript.

L417-420: Less usefulness of the tendency diagnostics to estimate correct unresolved physics tendency is described.
L453-461: The description on temperature error is improved, and the nudging forcing on temperature is mentioned.

L330-331: Is this conclusion necessarily true? ⋯ (now L413)
... 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 ...
I can agree that this would be true, however, as before, I do not see similarity between residual term error and total tendency error in figure 4. And, you also seem to acknowledge that this similarity is not there, in line 450.

In general, even if some physics parameterization have deficiencies ($F_{\mathrm{Phy,Mdl}}$ versus $F_{\mathrm{Phy,Truth}}$) and the residual term causes substantial errors, the errors caused by model dynamics (including a correct dynamical response to incorrect unresolved forcing) may hide the errors caused by model physics parameterizations in longer forecast lead time. It is quite important to disentangle errors and decompose their sources into processes.
As you mentioned, we acknowledge that we do not see similarity between the residual term error shown in Fig. 4(f) and the total tendency error shown in Fig. 4(d). It is the forecast errors "at least those growing over the 6-hour timescale" that the nudging increments are comparable to in this study, as suggested by Eq. (1) or the nudging formulation. The error structure/sign in very short-range timescale (e.g., 6 hours) does not necessarily last up to medium-range timescale (e.g., 5 days). We have intended to demonstrate that the error structure is similar in temperature field between these two different timescales but not in zonal wind field. As you seem to acknowledge, the residual term error is similar to the total tendency error relatively in temperature (Fig. 8(f) versus Fig. 8(d)), but not similar in zonal wind velocity (Fig. 4(f) versus Fig. 4(d)). This is the reason why we conclude that temperature error in the medium-range timescale (Fig. 1(d)) is caused mainly by deficient physics parameterizations and zonal wind error in the timescale (Fig. 1(c)) results from the error compensation between deficient physics parameterizations and incorrect dynamical responses.