

Response-CC1:

This study holds significant scientific importance and application value. By integrating large-scale information extracted from machine learning models to optimize the physics-driven model, it significantly improves its accuracy and generalization capability. The writing is of high quality, and the structure is well-organized and easy to follow. However, I believe the study still needs to further strengthen its emphasis on its key strengths.

We thank the reviewer for the positive assessment of our work and for recognizing its scientific importance and application value. We are also grateful for the constructive suggestion to further emphasize the key strengths of our study.

Major comments

Introduction: The authors present a relatively comprehensive discussion of physics-driven models and machine learning models. However, two points warrant attention:

1) The authors note that the development of physics-driven models is primarily constrained by limitations in the representation of physical processes. While this statement is technically accurate, it is worth noting that the key advantage of physics-driven models over machine learning models lies in their physical interpretability. Therefore, in the context of this study, this point may be reconsidered or omitted.

Thank you for your suggestion. The original statement was indeed inappropriate. "Understanding of physical processes" is both a bottleneck for physic models and their key advantage over ML models. We have revised the first sentence of the abstract accordingly.

Original: Traditional numerical weather prediction (NWP) models are constrained by limitations in the representation of physical processes and computational resources, resulting in lengthy development cycles and relatively slow improvements in forecast skill.

Revised: The development of traditional numerical weather prediction (NWP) relies on continuous advances in observation technology, data assimilation methods, numerical and parameterization algorithms, and the steady growth of computational resources, resulting in lengthy development cycles and relatively slow improvements in forecast skill.

Lines: 12-14

2) As noted by the authors, several studies have already improved global forecasts by extracting large-scale circulations from machine learning models (Lines 141-152). However, in the following paragraph, the authors directly introduce the online correction system based on the

nudging method, which I find somewhat confusing. In my view, it would be beneficial for the authors to first summarize the current state of research and clearly articulate the existing problems—namely, the motivation for conducting this study. This would help better highlight the importance of the research.

Thank you for your thoughtful and constructive suggestion. We agree that the transition from the literature review to the introduction of our online correction system is currently too abrupt, and that a clearer articulation of the research gap and motivation would significantly improve the readability and impact of the introduction.

From a scientific perspective, we objectively recognize that there are no major differences between the work of ECCC (Husain et al., 2025) and our own. Both independently arrived at the same conceptual approach. We started this research in early 2024, while the paper by Husain et al. (2025) was published in March 2025. During our work, we did not refer to their method. Our scheme is not an improvement or extension of Husain's work. We fully acknowledge that their independent work is excellent and has advanced the field.

The process of our research is as follows. Our team have background in dynamical cores and variational data assimilation. Previously, we conducted some work on reference profiles and implemented 3-D and 4-D reference profiles based on the CMA-GFS model (Su et al., 2025, doi: 10.1007/s13351-025-4114-5). Our initial idea was to introduce forecasts from the FuXi model as a time-varying 4-D reference profile into the dynamical solver, so that the reference state would stay close to the real atmosphere during integration and thereby improve the spatial discretization accuracy of the dynamical core. However, after implementing this method, we did not obtain significant improvements. Naturally, we then thought that direct nudging would certainly yield better effects. However, FuXi exhibits overly smoothed small-scale features and a rapidly decaying kinetic energy spectrum (KES). This led us to the idea of using spectral methods to separate the large-scale components before applying nudging. Since the 4D-Var module in CMA-GFS already contains spectral-grid transformation routines, the implementation was straightforward.

From a technical perspective: The inference module for FuXi and the preprocessing module connected to CMA-GFS were already completed during the development of 4DRef; For the vertical nudging coefficients, since FuXi output only contain 13 pressure levels, which are sparse near the surface and at the upper levels, applying a vertical profile and nudging only the middle levels became a necessary choice; The truncation wavenumber was determined through our own tests based on KES and real forecasts.

Therefore, objectively speaking, both the work of ECCC (Husain et al., 2025) and ECMWF (Polichtchouk et al., 2024, 2026), as well as our own work, have developed similar forecast systems based on the Spectral Nudging (SN) method using their own physical and ML model. ECCC is the first center to implement this approach. ECMWF, by contrast, trained AIFS on model levels, thereby addressing the issue of sparse vertical levels in the ML model, and established an ensemble forecasting system using the SN method. Our work indeed does not represent a novel scientific breakthrough. We did not summarize the limitations of the ECMWF

and ECCC methods in the introduction, as doing so would imply an intent to solve these problems, which was not my objective.

After the methodology section, We have added a table comparing the key differences between the ECCC, ECMWF, and our own work across various aspects to facilitate readers' comparison, as following:

TABLE 2. the differences in technical details of the SN method.

	ECCC	ECMWF	CMA
Physics model	GEM	IFS/IFS-ENS	GRAPES-GFS
ML model	GraphCast	AIFS/AIFS-ENS	FuXi
Spectral Transform	Discrete cosine	Spherical Harmonics	Spherical Harmonics
Method	transform (DCT)	expansion in IFS	expansion in GRAPES-4Dvar
Truncation	Soft cutoff between	T21	T21
wavenumber or	2750-2250km		
wavelength			
Nudging variable	u,v,virtual temperature	vorticity, specific humidity, virtual temperature	u,v,exner pressure,potential temperature
Nudging vertical	850hPa-250hPa	Model levels 50-137	600hPa-200hPa
profiles		(approximately surface to 56hPa)	
Nudging relaxation	12hour	12hour	6hour
time			

Lines: 143-146

Line: 305

3) The authors state that the FuXi model is driven by ERA5 reanalysis data to produce forecast fields, which then supply large-scale circulations to the physical model. I have a concern: given that reanalysis data are generally not accessible in real time for operational applications, how feasible is this method in practice?

Thank you for your suggestions, these are indeed key issues to address in our future work.

Our current work focuses on conceptual verification to confirm the feasibility of the SN method and the correctness of the system configuration. The ERA5 dataset is adopted here to initialize the FuXi model, ensuring optimal simulation performance.

To operationalize this system at CMA, we will run the FuXi model initialized with analysis fields from the CMA-GFS data assimilation cycle. Our tests show that initializing FuXi with CMA-GFS analysis instead of ERA5 reduces the model's predictable lead time by approximately 1–2 days across seasons and regions, a common issue in other ML models.

We plan to address this through two research directions: 1) Using Transformer-based neural networks to adjust CMA-GFS analysis fields to better align with ERA5 reanalysis data before applying them to FuXi. 2) Fine-tuning or retraining FuXi with CMA-GFS reanalysis data (derived from the CMA-GFS system) and CMA-GFS analysis fields to enhance its adaptability. Preliminary results from the first approach indicate that the predictable lead time can be extended by approximately one day, particularly over the Southern Hemisphere.

Relevant discussions have also been included in the first part of the future work plan.

Lines: 472-478

Minor comments

1) Line 148: Why is “truncation wavenumber” particularly noted here, unless it has special significance?

Thank you for your suggestions.

The truncation wavenumber is a key parameter in the SN method and does not need to be mentioned here; instead, it can be presented later in the implementation section. As you suggested, We have removed the description of truncation wavenumber at this point .

Lines: 137-139

2) Lines 262-278: the truncation wavenumber is determined mainly based on the KES differences between the CMA-GFS and FuXi models, illustrated using forecasts from four initialization dates. Given that this selection (42 instead of 21) is derived from a limited set of cases, I am concerned about its representativeness and robustness.

We thank the reviewer for raising this valid concern. We agree that determining the truncation wavenumber based on four initialization dates may raise questions about representativeness and robustness. We would like to clarify and address this issue as follows.

Based on previous experience, KES may vary with forecast lead time, model resolution, and diffusion scheme, but differs little on different dates. Here we selected one day from each of the four seasons, and their KES profiles are broadly consistent when examined individually.

Furthermore, the choice between truncation wavenumber T42 and T21 is not determined by KES alone. In Section 3 of this paper, both the case verification in Section 3.1 and the batch experiments January and July in Section 3.2 provide detailed comparisons between T42 and T21. Since no significant difference is observed, we adopted T21 as a conservative choice. This decision helps preserve more characteristics of the physics model and prevents excessive smoothing of forecast fields.

In addition, relevant studies from ECMWF (Polichtchouk et al., 2024, 2026) also adopted T21 as the truncation wavenumber, following comprehensive comparisons and validation. The selection of the truncation wavenumber is explained in Polichtchouk et al. (2026) (<https://arxiv.org/html/2603.05570v1>) as follows: “In deterministic hybrid systems, nudging beyond wavenumber 21 risks introducing excessive smoothing, since deterministic machine-learned models tend to suppress mesoscale variability. Probabilistic models such as AIFS-ENS do not exhibit this behaviour (see Figure 1) and could, in principle, support nudging at higher wavenumbers. We tested cut-off wavenumbers T42 and T85 in addition to T21. Nudging to T42 yielded only marginal further improvements (typically 1–2% for upper-air variables), while T85 provided no additional benefit. We therefore adopt T21 for this study as a conservative and robust choice that limits the degree of machine-learned intervention on the physics-based model.”

3) Line 353: For the case study, are there any quantitative comparative results available?

Thank you for your suggestion.

The case study is intended to intuitively demonstrate that the SN system can combine the large-scale circulation forecasts from the ML model with the intensity forecasts from the physics model. Since it only involves a comparison at a single forecast time, quantitative verification is not provided.

In the subsequent batch experiments for January and July, we present quantitative verifications including ACC, RMSE, and ETS. In particular, for the verification of tropical cyclones over the western North Pacific in 2024, we systematically provide quantitative results for typhoon track error, central pressure error, and maximum wind speed error.

Response-RC1:

The paper describes research tests of a system to nudge the CMA-GFS physical model with the FuXi machine learning model to improve the large-scale evolution of the physical model whilst retaining its benefits for small scale detail.

Fundamentally, the methodology is very similar to the referenced Husain et al. (2025) paper, but using different physical and ML models. It shares the same limitations in terms of the coarse vertical resolution of the output from the ML model and inconsistent analyses between the physical and ML models. As the authors note, these limitations were addressed by Polichtchouk et al. (2024) who used an ML model with much higher vertical resolution to gain considerably improved results from nudging, especially in the lower troposphere. The authors of the present paper do outline their plans to address these limitations in their discussion section.

Hence, this paper doesn't necessarily advance the science, however it does document the repeated test of a published method (an important aspect of science) and some common similarities in results are obtained using different models, which is useful for other centers considering using the nudging approach. The paper is clear and well written and, therefore, I believe this paper is a useful addition to the literature and should be published in EGU sphere.

Thank you for your valuable comments and suggestions. We have carefully revised the manuscript in accordance with your advice.

As you correctly note, the method used in this paper is indeed very similar to that of Husain et al. (2025). We would like to clarify that we only became aware of their paper and research findings at a later stage of our work. In fact, we independently proposed the same scheme at around the same time (we initiated our research in early 2024, while their paper was published in March 2025). Our work progressed more slowly, but the conceptual convergence—using spectral nudging to couple an ML model with a physical NWP model—emerged independently from our own technical reasoning, which we have detailed in our response to the other reviewer.

We employed CMA's physical model (CMA-GFS) and the FuXi ML model to validate the nudging method. Beyond confirming the findings of Husain et al., we further demonstrate that using the ML model with spectral nudging not only significantly improves the prediction of large-scale circulation but also markedly enhances medium-range precipitation forecasting skill—the variable of greatest concern to operational forecasters. We believe this result strengthens the case for transitioning such hybrid systems into operational use, as it shows benefits for a key forecast variable that has been challenging for pure ML models.

We also acknowledge the two limitations of applying ML-based nudging within a physical model, as noted by the reviewer and also discussed in related work: first, the low vertical resolution of the ML model (FuXi provides only 13 pressure levels); second, the inconsistency

between the initial fields of the physical model and those used to train the ML model. Regarding the first limitation, one effective solution—as demonstrated by Polichtchouk et al. (2024)—is to retrain a new ML model using reanalysis data with more vertical levels. We are actively exploring this direction. Regarding the second limitation, we are currently conducting further work to address it. Our goal is to have the physical model and the ML model use the same analysis fields. The approach we plan to adopt is to develop a corresponding 4D-Var assimilation system that incorporates the hybrid physical-ML model (with ML nudging embedded) as a constraint, and then train a new ML model based on this integrated assimilation and hybrid modeling system.

We thank the reviewer again for the constructive feedback and for recognizing the value of our work as a useful replication and extension of an emerging methodology. We believe the revised manuscript, together with the clarifications provided above, adequately addresses the reviewer's concerns.

Minor comments:

- 1) Line 18 – suggest replacing 'foundational' with 'underpinning' to avoid any confusion with foundational AI models.

Thank you for your suggestion, The 'foundational' is replaced by 'underpinning' in line 18.

Line: 19

- 2) The authors cite a weakness of ML models as being “Progressive smoothing in long-range forecasts”. Whilst this can be the case if the target is RMSE, for which a smoother field can lead to a better score to avoid a 'double penalty' from positional errors, ML weather models do now exist which avoid this smoothing by not (solely) minimising on RMSE, such as the AIFS-CRPS. This should be acknowledged in the paper.

Thank you for your suggestions. We apologize for not covering the research progress of AIFS-CRPS in our prior work, and have revised this paragraph as advised. The underlined parts are the newly added content.

Progressive smoothing in long-range forecasts. ML models trained with the Mean Squared Error (MSE) loss function commonly exhibit a pronounced smoothing tendency as forecast lead time increases, i.e., forecast fields become increasingly smooth. This is also evident in a kinetic energy spectrum (KES), which shows evident dissipation of kinetic energy in meso- and small-scale systems (Kochkov et al. 2024; Husain et al. 2025). In contrast, Lang et al. (2025) adopted the almost fair Continuous Ranked Probability Score (afCRPS) as the loss function for the ensemble variant of the AIFS, which enables the model to generate stochastic

forecasts that preserve realistic atmospheric variability and maintain a physically consistent KES.

Lines: 83-89

3) Line 149 and subsequent use of 'typhoon'. Where the reference is not specifically to the Pacific basin, the more generic term of 'tropical cyclone' should be used.

Thank you for your suggestion, the 'typhoon' is replaced by 'tropical cyclone'.

Lines: 139-140

4) Line 296 – could the issues with nudging at smaller scales than T21 also be due to the poor vertical resolution of FuXi output and lack of nudging in the lower troposphere? Where centres have tried nudging to the model level AIFS, improved performance has been found to scales of T63 without the issues documented here.

Thank you for your question. We cannot confirm this for certain, but the following information is provided for your reference.

(1) A paper on ECMWF's hybrid ensemble prediction system constructed using the Spectral Nudging method has just been published on arXiv (Polichtchouk et al., 2026, <https://arxiv.org/html/2603.05570v1>). The AIFS-ENS is trained on model levels from 137 (at the surface) to 50 (at approximately 56 hPa). and the truncation wavenumber is also set to 21. The selection of the truncation wavenumber is explained in the paper as follows:

"In deterministic hybrid systems, nudging beyond wavenumber 21 risks introducing excessive smoothing, since deterministic machine-learned models tend to suppress mesoscale variability. Probabilistic models such as AIFS-ENS do not exhibit this behaviour (see Figure 1) and could, in principle, support nudging at higher wavenumbers. We tested cut-off wavenumbers T42 and T85 in addition to T21. Nudging to T42 yielded only marginal further improvements (typically 1–2% for upper-air variables), while T85 provided no additional benefit. We therefore adopt T21 for this study as a conservative and robust choice that limits the degree of machine-learned intervention on the physics-based model."

We have also reviewed the materials on ECMWF's deterministic forecasts with Spectral Nudging, and I found no relevant explanations regarding the T63 truncation wavenumber. There is no evidence to support an association between sparse vertical resolution and the truncation wavenumber.

(2) Husain et al. (2025) also discuss the influence of the vertical resolution of machine learning models on the Spectral Nudging system, as stated in the original text:

"This study employs the 13-pressure-level version of GraphCast with pretrained weights (learned features of the GNNs) that are available from Google DeepMind. Although a 37-level version is available, only the 13-level variant has been subjected to additional fine-tuning with ECMWF's operational analyses (2016-21), making it more skillful than the 37-level version."

Therefore, We are also curious whether the large scale circulation forecasting capability of ECMWF's 137 model level AIFS can maintain the performance of the version with 13 pressure level, We have not found any relevant comparison in the literature.

After discussing with the developers of the ML models, they generally agree that, owing to the contribution of the 500hPa MSE within the overall loss function, the version with 13 pressure levels achieves higher ACC and better RMSE scores.

This is out preliminary understanding, which may not be fully accurate: if a higher ACC for the 500hPa geopotential height is required, the 13-pressure-level version should be adopted. If comprehensive improvements for the middle and lower troposphere are prioritized, the model level version is preferable, though this will compromise some of the 500hPa scores.

Since the FuXi model we currently use only provides the 13-pressure-level version, We are unable to conduct relevant tests. If ML model products with more vertical levels become available in the future, we would very much like to carry out further experiments to investigate whether the forecasting performance of the physical model can be improved in a more comprehensive manner.

Lines: 241-243

5) Figure 4 – suggest making it clear that the “gridded merged precipitation product of the CMA” is observationally based and also add what sources are merged (gauge?, satellite(?), radar(?)).

Thank you for your suggestion. We apologize for the confusion. After re-checking the plotting script, we confirm that the data used in Figure 4 are gauge precipitation data, not the CMA's gridded merged product. The figure and caption have been revised accordingly.

For the reviewer's reference, the CMA does produce a multi-source merged gridded precipitation dataset at 0.01° resolution, which integrates ground-based gauge measurements, radar reflectivity-derived precipitation estimates, and satellite-retrieved precipitation estimates. However, this product was not used in Figure 4. We thank the reviewer for pointing out the need for clarity, and we have now ensured that the figure and its caption accurately state the use of station observations.

Lines: 330-335

6) Figures 8&9 – what is the bias with respect to. Is it own analysis, all compared with ERA5 or something else?

Thank you for your question. We did not explain this clearly before.

The biases in Figures 8 and 9 are calculated with respect to the model's own 0-hour forecast field. Since the model is cold-started with ERA5 data, the 0-hour field is also the ERA5 data. Relevant explanations have been added to the figure captions.

Lines: 389-390

7) I assume only deterministic models are used here. Throughout the paper, the framing is in terms of number of days of skilful prediction. The use of 0.6 on ACC is widely used, but fairly arbitrary. Ensembles provide the most useful forecast information, even when the skill of deterministic model is relatively high. Do the authors have plans to incorporate the spectral nudging into CMA's ensemble prediction system? Understanding what barriers would need to be overcome to achieve this would be a valuable addition to the discussion section

Thank you for your valuable suggestions. The current work is only based on deterministic forecasts. Relevant research on ensemble forecasts may be carried out in the future, mainly for the following two considerations:

(1) From the perspective of operational implementation, all current operational systems of the CMA (global, regional, and ensemble) are based on the GRAPES model (SISL, lat-lon). These will be gradually replaced starting in 2027 with the next generation dynamic core based on MCV (finite-volume, cubed-sphere). The replacement sequence is global first, then regional, and finally ensemble. The current SN system is developed based on the GRAPES global model. The next step is to migrate to the MCV global model and subsequently develop the MCV-SN ensemble forecast system.

(2) From a technological perspective, ECMWF provides a valuable benchmark (Polichtchouk et al., 2026) for establishing the SN ensemble forecast system. Since we have already established the workflow for the deterministic SN system, the key to extending this workflow to ensemble forecasting is to maintain reasonable spread among ensemble members. For example, we need to perform SN between the 16 corresponding members of GRAPES-GEPS and FuXi-ENS, rather than nudging all GRAPES-GEPS members toward a single FuXi deterministic forecast.

The FuXi-ENS system (Zhong et al., 2025, DOI: 10.1126/sciadv.adu2854) provides reasonable ensemble spread and good forecast skill, although its spread is slightly smaller than that of the GRAPES-GEPS system. GRAPES-GEPS includes initial perturbations and model perturbations; the latter are divided into large-scale, meso-scale, and small-scale perturbations. We propose replacing the large-scale component of model perturbations in GRAPES-GEPS with

the truncated large-scale component from FuXi-ENS forecasts, while retaining or enhancing the original meso-scale and small-scale components. This approach aims to ensure that the final SN-ENS system maintains adequate ensemble spread.

We have incorporated the corresponding technical content into the paper, which serves as the last point in the discussion section.

Lines: 487-494

Response-RC2:

This paper proposes an online correction framework that integrates a machine learning model (FuXi) with a physical numerical weather prediction model (CMA-GFS) through spectral nudging.

Overall, I find this work valuable as a careful and useful validation of an existing methodology. In particular, it demonstrates that the hybrid nudging framework can be successfully implemented within a different model system, which may be helpful for operational centers considering similar approaches. The manuscript is also clearly written and provides a detailed description of the workflow, which makes it easy to follow.

From a scientific perspective, however, the main contribution appears to be a system-specific implementation of an already established paradigm rather than a fundamentally new methodological development. The approach is largely consistent with prior work such as Husain et al. (2025), including scale-selective spectral nudging, handling of coarse vertical ML outputs, and the use of vertical weighting to mitigate inconsistencies. While this is not a limitation in itself, it may be helpful for the authors to more clearly position their work relative to these studies and clarify whether there are specific aspects in which their implementation provides advantages.

We thank the reviewer for the insightful comments, which have helped us improve the clarity and impact of our manuscript.

We would like to clarify the origin of the methodology. We independently proposed and developed this online correction framework that integrates a machine learning model (FuXi) with a physical numerical weather prediction model (CMA-GFS) via Spectral Nudging. Late in our research, we came across the work of Husain et al. (2025), which is largely consistent with ours in both concept and methodology. Their work, of course, was initiated earlier than ours. We fully acknowledge that Husain et al. (2025) have done an excellent job in formulating and validating their version of the method. Their work is rigorous and valuable, and we do not claim any priority over them. Instead, we view this independent convergence as strong evidence that the hybrid Spectral-Nudging framework is a timely and promising direction for integrating machine learning into operational NWP.

The process of my research is as follows: Our team have background in dynamical cores and variational data assimilation. Previously, We conducted some work on reference profiles and implemented 3-D and 4-D reference profile based on the CMA-GFS model (Su et al., 2025, doi: 10.1007/s13351-025-4114-5). Our initial idea was to introduce forecasts from the FuXi model as a time-varying 4-D reference profile into the dynamical solver, so that the reference state would stay close to the real atmosphere during integration and thereby improve the spatial discretization accuracy of the dynamical core. However, after implementing this method, it did not get significant improvements. Naturally, We then thought that direct nudging would certainly yield better effects. However, FuXi exhibits overly smoothed small-scale features and

a rapidly decaying KES. This led me to the idea of using spectral methods to separate the large-scale components before applying nudging. Since the 4-Dvar module in CMA-GFS already contains spectral-grid transformation routines, the implementation was straightforward.

From a technical perspective: The inference module for FuXi and the preprocessing module connected to CMA-GFS were already completed during the development of 4DRef. For the vertical nudging coefficients, since FuXi output only contain 13 pressure levels, which are sparse near the surface and at the upper levels, applying a vertical profile and nudging only the middle levels became a necessary choice; The truncation wavenumber was determined through my own tests based on KES and real forecasts.

Therefore, objectively speaking, both the work of ECCC (Husain et al., 2025) and ECMWF (Polichtchouk et al., 2024, 2026), as well as my own work, have developed similar forecast systems based on the Spectral Nudging (SN) method using their own physical and ML model. ECCC is the first center to implement this approach. ECMWF, by contrast, trained AIFS on model levels, thereby addressing the issue of sparse vertical levels in the ML model, and established an ensemble forecasting system using the SN method. My work indeed does not represent a novel scientific breakthrough. We did not summarize the limitations of the ECMWF and ECCC methods in the introduction, as doing so would imply an intent to solve these problems, which was not my objective.

After the methodology section, I have added a table comparing the key differences between the ECCC, ECMWF, and my own work across various aspects to facilitate readers' comparison, as following:

TABLE 2. the differences in technical details of the SN method.

	ECCC	ECMWF	CMA
Physics model	GEM	IFS/IFS-ENS	GRAPES-GFS
ML model	GraphCast	AIFS/AIFS-ENS	FuXi
Spectral Transform	Discrete cosine	Spherical Harmonics	Spherical Harmonics
Method	transform (DCT)	expansion in IFS	expansion in GRAPES-4Dvar
Truncation	Soft cutoff between	T21	T21
wavenumber or	2750-2250km		
wavelength			
Nudging variable	u,v,virtual temperature	vorticity, specific humidity, virtual temperature	u,v,exner pressure,potential temperature

Nudging vertical profiles	850hPa-250hPa	Model levels 50-137 (approximately surface to 56hPa)	600hPa-200hPa
Nudging relaxation time	12hour	12hour	6hour

Lines: 143-146

Line: 305

One aspect that could benefit from further clarification is the preprocessing step used to map FuXi outputs from 13 pressure levels to the 87 model levels of CMA-GFS. This step is only briefly mentioned and not specified in detail. Since vertical interpolation is a critical component that can significantly influence the representation of atmospheric structure (especially gradients, stability, and boundary-layer processes), the lack of description raises concerns about reproducibility and scientific validity. It is unclear what interpolation scheme is used, how physical consistency is preserved, and to what extent this preprocessing step may introduce biases or damp important features before nudging is even applied.

We thank the reviewer for raising this important and technically critical point. We agree that the vertical interpolation from FuXi's 13 pressure levels to CMA-GFS's 87 model levels is a key preprocessing step, and our original manuscript did not provide sufficient detail. We substantially expanded the preprocessing procedure in the revised manuscript in lines 210-215 as follows:

The detail of preprocessing procedure is as follows: First, in the horizontal direction, the geopotential height, temperature, zonal and meridional winds (h , t , u , v) forecast by FuXi on 13 pressure levels are bilinearly interpolated from 0.25° resolution to the model resolution of 0.125° . Then, in the vertical direction, h , t , u , v on pressure levels are interpolated to p , t , u , v on the 87 model levels using cubic spline interpolation, based on the height coordinates of pressure levels and model levels. Finally, the Π and θ on the 87 model levels are

computed by $\Pi = \left(\frac{p}{p_0}\right)^{\frac{R}{C_p}}$ and $\theta = T \left(\frac{p}{p_0}\right)^{\frac{R}{C_p}}$, where p is pressure, p_0 is standard sea-level pressure. R is gas constant, C_p is specific heat capacity at constant pressure.

Lines: 210-215

In the current work, the FuXi outputs are interpolated to the CMA-GFS vertical grid through this unspecified preprocessing step, and the resulting inconsistency is mitigated by applying a vertically varying nudging coefficient that limits the correction primarily to the mid–upper troposphere. However, this approach closely follows that of Husain et al. (2025), who employed a similar vertical weighting strategy to address the same issue. As such, it is unclear what methodological innovation is introduced here beyond adopting an existing workaround.

Thank you for your advice.

We thank the reviewer for this observation. We agree that our vertically varying nudging coefficient, which limits corrections primarily to the mid-upper troposphere, is technically similar to the strategy employed by Husain et al. (2025). However, we would like to clarify the following points.

As explained earlier, our work is not a replication or extension of Hussain's. We initially developed a 4D reference profile based on FuXi forecasts, and later shifted to building an SN-based hybrid system.

The vertical nudging profile was not introduced at the beginning. However, after applying SN, the deviations in the lower and upper levels increased. Since we are most concerned with the forecast leading time at 500 hPa, introducing a vertical coefficient was a natural and straightforward choice.

The vertical profile is only a temporary solution for constructing the hybrid system at the current stage and indeed lacks innovation. In the future, we will attempt to increase the vertical levels of the ML model or directly train FuXi on model levels.

Lines: 241-243

At the same time, the manuscript acknowledges alternative approaches, such as Polichtchouk et al. (2024), who address this limitation more fundamentally by increasing the vertical resolution of the ML model (e.g., 137 levels), thereby reducing the need for ad hoc vertical weighting. Given this, it would be important for the authors to clarify why a similar strategy is not adopted in the present study. Is the choice driven by computational constraints, data availability, or compatibility with FuXi? Without such justification, the current approach appears as a pragmatic but potentially suboptimal solution rather than a deliberate methodological design.

Thank you very much for your advice. This is indeed one of the key issues in our study, and we would like to address it from the following two perspectives.

(1) For machine learning models, a denser vertical hierarchy does not necessarily lead to better forecast performance.

Husain et al. (2025) also discuss the influence of the vertical resolution of machine learning models on the Spectral Nudging system, as stated in the original text: "This study employs the 13-pressure-level version of GraphCast with pretrained weights (learned features of the GNNs) that are available from Google DeepMind. Although a 37-level version is available, only the 13-level variant has been subjected to additional fine-tuning with ECMWF's operational analyses (2016-21), making it more skillful than the 37-level version."

For ML developers, increasing the vertical levels poses no technical difficulties, and computing power is no longer an issue. Yet most ML models still only offer a 13-level version. After discussing with the developers of the ML models, they generally agree that, owing to the contribution of the 500hPa MSE within the overall loss function, the version with 13 pressure levels achieves higher ACC and better RMSE scores.

Therefore, We are also curious whether the large scale circulation forecasting capability (500hPa ACC and RMSE) of ECMWF's 137 model level AIFS can maintain the performance of the version with 13 pressure level, We have not found any relevant comparison in the literature.

(2) Regarding the copyright and model availability:

FuXi was developed by the Institute of Artificial Intelligence at Fudan University and does not belong to CMA. Currently, we are using the publicly released version of FuXi, which includes only the inference code but not the training code. Therefore, we are temporarily unable to retrain the FuXi model.

In the next phase, before the spectral nudging system is put into operational use at CMA, we will communicate with the FuXi team to obtain the training code. We will then use initial fields from CMA-GFS to fine-tune or retrain the model. At the same time, we plan to attempt increasing the number of isobaric levels, or even train the model directly on model levels, to see whether the physical model can be improved in a more comprehensive manner.

This is my preliminary understanding, which may not be fully accurate: If a higher ACC for the 500hPa geopotential height is required, the 13-pressure-level version should be adopted. If comprehensive improvements for the middle and lower troposphere are prioritized, the model level version is preferable, though this will compromise some of the 500hPa scores. Since it is uncertain whether this is correct and there has been no rigorous comparative validation, We have not included this viewpoint in the paper.

We thank the reviewer again for raising this important issue. The descriptions regarding copyright issues and the plan to train FuXi with more levels in future work have been incorporated into the manuscript (lines 241-243) .

Regarding presentation, the introduction provides a broad survey of ML-based weather models, but the connection to the core contribution is not clearly articulated. The cited models

appear more as a catalogue than as elements that directly motivate the proposed method. Given that the main idea can be summarized concisely — combining ML-derived large-scale circulation with physics-based small-scale consistency via spectral nudging — the introduction could be significantly streamlined. More generally, this issue extends beyond the introduction to the entire manuscript. While the detailed narrative of the research process is informative, the paper would benefit from a more concise and structured presentation. In particular, lengthy descriptive explanations of intermediate attempts or design choices could be reduced, and key ideas could instead be conveyed more effectively through tables, figures, or mathematical formulations.

Thank you for this constructive suggestion. We agree that the previous introduction was indeed somewhat broad and did not sufficiently focus on the core issues.

Following your advice, we have streamlined the overly general literature survey in the introduction and added or expanded content specifically related to the SN method. In addition, we have condensed the descriptions of the subsequent experimental procedures and scheme selections to improve clarity and conciseness.

We believe these revisions will make the manuscript more focused and easier to follow. We thank the reviewer again for helping us improve the presentation.

Lines: 101-107

Line: 224

Line: 275

Lines: 284-287

Lines: 334-347

Lines: 357-365

Lines: 370-376

Lines: 391-397

Lines: 405-412

Lines: 420-425

Lines: 436-450

In addition, the paper does not sufficiently justify why spectral nudging is appropriate in this global modeling context. The evaluation is also limited to internal comparisons within a single modeling framework, without benchmarking against other state-of-the-art ML or hybrid systems, including closely related work such as Husain et al. (2025). This makes it difficult to assess the broader competitiveness or generality of the approach.

We thank the reviewer for this important comment.

The main purpose of our work is to establish a hybrid system of CMA-GFS and FuXi based on the SN method. In the introduction, we discussed the relative strengths and weaknesses of ML models and physical models, and then pointed out that the SN method can effectively combine the advantages of both.

During the evaluation, comparisons were mainly conducted among three systems: CMA-GFS, FuXi, and CMA-SN, where FuXi represents the state-of-the-art ML model. In the introduction, the description of the comparison platform of CMA and ECMWF (Table 1) demonstrates that the current FuXi model achieves world-class performance in large-scale circulation patterns in both winter and summer. In addition, we sincerely apologize that I am currently unable to run the hybrid model as described in Husain's paper, so a comparative evaluation with it cannot be performed.

Response-RC3:

This manuscript develops a hybrid forecast system by combining the CMA-GFS physical model and the FuXi ML model using the spectral nudging (SN) method. It takes FuXi's strength in large-scale circulation forecasts and CMA-GFS's strength in small-scale details. The hybrid system performs very well and offers useful references for operational NWP centers. I recommend acceptance after minor revisions.

The vertical nudging profile you used is similar to the ECCC approach. Why didn't you follow the ECMWF AIFS method and train the ML model directly on model levels instead of pressure levels?

Thank you very much for your advice. This is indeed one of the key issues in our study, and we would like to address it from the following two perspectives.

(1) For machine learning models, a denser vertical hierarchy does not necessarily lead to better forecast performance.

Husain et al. (2025) also discuss the influence of the vertical resolution of machine learning models on the Spectral Nudging system, as stated in the original text: "This study employs the 13-pressure-level version of GraphCast with pretrained weights (learned features of the GNNs) that are available from Google DeepMind. Although a 37-level version is available, only the 13-level variant has been subjected to additional fine-tuning with ECMWF's operational analyses (2016-21), making it more skillful than the 37-level version."

For ML developers, increasing the vertical levels poses no technical difficulties, and computing power is no longer an issue. Yet most ML models still only offer a 13-level version. After discussing with the developers of the ML models, they generally agree that, owing to the contribution of the 500hPa MSE within the overall loss function, the version with 13 pressure levels achieves higher ACC and better RMSE scores.

Therefore, We are also curious whether the large scale circulation forecasting capability (500hPa ACC and RMSE) of ECMWF's 137 model level AIFS can maintain the performance of the version with 13 pressure level, We have not found any relevant comparison in the literature.

(2) Regarding the copyright and model availability

FuXi was developed by the Institute of Artificial Intelligence at Fudan University and does not belong to CMA. Currently, we are using the publicly released version of FuXi, which includes only the inference code but not the training code. Therefore, we are temporarily unable to retrain the FuXi model.

In the next phase, before the spectral nudging system is put into operational use at CMA, we will communicate with the FuXi team to obtain the training code. We will then use initial fields from CMA-GFS to fine-tune or retrain the model. At the same time, we plan to attempt increasing the number of isobaric levels, or even train the model directly on model levels, to see whether the physical model can be improved in a more comprehensive manner.

This is my preliminary understanding, which may not be fully accurate: If a higher ACC for the 500hPa geopotential height is required, the 13-pressure-level version should be adopted. If comprehensive improvements for the middle and lower troposphere are prioritized, the model level version is preferable, though this will compromise some of the 500hPa scores. Since it is uncertain whether this is correct and there has been no rigorous comparative validation, We have not included this viewpoint in the paper.

We thank the reviewer again for raising this important issue. The descriptions regarding copyright issues and the plan to train FuXi with more levels in future work have been incorporated into the manuscript.

Lines: 241-243

Please explain your work's difference from ECMWF and ECCO. Highlight your improvements so readers can easily see your novelty.

Thank you for your suggestion, which have helped us improve the clarity and impact of our manuscript.

We would like to clarify the origin of the methodology. We independently proposed and developed this online correction framework that integrates a machine learning model (FuXi) with a physical numerical weather prediction model (CMA-GFS) via Spectral Nudging. Late in our research, we came across the work of Husain et al. (2025), which is largely consistent with ours in both concept and methodology. Their work, of course, was initiated earlier than ours. We fully acknowledge that Husain et al. (2025) have done an excellent job in formulating and validating their version of the method. Their work is rigorous and valuable, and we do not claim any priority over them. Instead, we view this independent convergence as strong evidence that the hybrid Spectral-Nudging framework is a timely and promising direction for integrating machine learning into operational NWP.

The process of my research is as follows: Our team have background in dynamical cores and variational data assimilation. Previously, We conducted some work on reference profiles and implemented 3-D and 4-D reference profile based on the CMA-GFS model (Su et al., 2025, doi: 10.1007/s13351-025-4114-5). Our initial idea was to introduce forecasts from the FuXi model as a time-varying 4-D reference profile into the dynamical solver, so that the reference state would stay close to the real atmosphere during integration and thereby improve the spatial discretization accuracy of the dynamical core. However, after implementing this method,

it did not get significant improvements. Naturally, We then thought that direct nudging would certainly yield better effects. However, FuXi exhibits overly smoothed small-scale features and a rapidly decaying KES. This led me to the idea of using spectral methods to separate the large-scale components before applying nudging. Since the 4-Dvar module in CMA-GFS already contains spectral-grid transformation routines, the implementation was straightforward.

From a technical perspective: The inference module for FuXi and the preprocessing module connected to CMA-GFS were already completed during the development of 4DRef. For the vertical nudging coefficients, since FuXi output only contain 13 pressure levels, which are sparse near the surface and at the upper levels, applying a vertical profile and nudging only the middle levels became a necessary choice; The truncation wavenumber was determined through my own tests based on KES and real forecasts.

Therefore, objectively speaking, both the work of ECCC (Husain et al., 2025) and ECMWF (Polichtchouk et al., 2024, 2026), as well as my own work, have developed similar forecast systems based on the Spectral Nudging (SN) method using their own physical and ML model. ECCC is the first center to implement this approach. ECMWF, by contrast, trained AIFS on model levels, thereby addressing the issue of sparse vertical levels in the ML model, and established an ensemble forecasting system using the SN method. My work indeed does not represent a novel scientific breakthrough. We did not summarize the limitations of the ECMWF and ECCC methods in the introduction, as doing so would imply an intent to solve these problems, which was not my objective.

After the methodology section, I have added a table comparing the key differences between the ECCC, ECMWF, and my own work across various aspects to facilitate readers' comparison, as following:

TABLE 2. the differences in technical details of the SN method.

	ECCC	ECMWF	CMA
Physics model	GEM	IFS/IFS-ENS	GRAPES-GFS
ML model	GraphCast	AIFS/AIFS-ENS	FuXi
Spectral	Discrete cosine	Spherical Harmonics	Spherical Harmonics
Transform	transform (DCT)	expansion in IFS	expansion in
Method			GRAPES-4Dvar
Truncation	Soft cutoff between	T21	T21
wavenumber or	2750-2250km		
wavelength			
Nudging variable	u,v,virtual temperature	vorticity, specific	u,v,exner

		humidity, virtual temperature	pressure,potential temperature
Nudging vertical profiles	850hPa-250hPa	Model levels 50-137 (approximately surface to 56hPa)	600hPa-200hPa
Nudging relaxation time	12hour	12hour	6hour

Line: 305

You initialized FuXi with ERA5 reanalysis, which works for research but not for real-time operations. How do you plan to fix this when moving to operational runs with real-time analysis data?

Thank you for your suggestions, these are indeed key issues to address in our future work.

Our current work focuses on conceptual verification to confirm the feasibility of the SN method and the correctness of the system configuration. The ERA5 dataset is adopted here to initialize the FuXi model, ensuring optimal simulation performance.

To operationalize this system at CMA, we will run the FuXi model initialized with analysis fields from the GRAPES data assimilation cycle. Our tests show that initializing FuXi with GRAPES analysis instead of ERA5 reduces the model's predictable lead time by approximately 1–2 days across seasons and regions, a common issue in other ML models.

We plan to address this through two research directions: 1) Using Transformer-based neural networks to adjust GRAPES analysis fields to better align with ERA5 reanalysis data before applying them to FuXi. 2) Fine-tuning or retraining FuXi with GRAPES reanalysis data (derived from the GRAPES system) and GRAPES analysis fields to enhance its adaptability. Preliminary results from the first approach indicate that the predictable lead time can be extended by approximately one day, particularly over the Southern Hemisphere.

Relevant discussions have also been included in the first part of the future work plan.

Lines: 472-478

Do you plan to extend this SN method to CMA regional models for better tropical cyclone simulation?

Thank you for your advice. They are highly meaningful for the development of regional models, particularly for the prediction of tropical cyclones.

Large-scale steering flow is crucial for typhoon track forecasting, and relevant work is currently underway. Specifically, we plan to extract large-scale circulations from FuXi outputs, and apply online correction to the 1 km-resolution CMA-MESO (national domain) system. In the future, we may also use outputs from the global SN system as initial and boundary conditions for regional model, to improve the performance of regional model in large-scale circulations.

Since a global SN system has already been established, constructing a regional SN system will be relatively straightforward—the overall workflow can be directly adapted from the global one. The main difference lies in: the global model employs a 4D-Var assimilation system, whose core routines (including transformations between lat-lon grids and Gaussian grids, spherical harmonic expansions, etc.) are already developed and do not need to be rebuilt. For regional models, however, corresponding modules must be newly developed. We plan to either reference relevant modules from the WRF model, or develop the code that directly performs expansion and truncation using the discrete cosine transform (DCT) on a regular lat-lon grid.