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 EGUsphere.

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.

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. See lines 82 to 88 of the revised manuscript.

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.

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’ in line 138-139 in revised manuscript.

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.

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.

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.

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.