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.

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 Method	Discrete cosine transform (DCT)	Spherical Harmonics expansion in IFS	Spherical Harmonics expansion in GRAPES-4Dvar
Truncation wavenumber or wavelength	Soft cutoff between 2750-2250km	T21	T21
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

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.

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 .

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.