

Response to report 1:

This study has clear methodological value in the integration of ML and physical models, with a relatively solid technical implementation. However, there is room for improvement in the positioning of novelty, the physical basis for key parameter selection, the completeness of experimental design, and the depth of result interpretation. The main comments are as follows:

1. The abstract and introduction position this work as a "proof-of-concept study," yet the main text repeatedly implies that this is an operational application-oriented system development. It is recommended to clearly distinguish the boundary between proof-of-concept and operational system prototyping.

Thank you for your valuable comments. This is an essential point that requires further clarification.

We have added relevant notes in the abstract to specify that the experimental results in this paper are based on ERA5 data rather than operational configurations, aiming to provide valid evidence for potential future operational applications (Lines 21-22, 27-30). Similar statements have also been incorporated in the conclusion section (Lines 492-495, 508-509).

Furthermore, as staff from the operational center, our ultimate interest in this technique lies in its operational implementation to improve the performance of the operational system. In practice, we have so far developed a fully-independent, hybrid system. Using CMA-GFS assimilated analyses as the initial conditions, systematic biases are adjusted via correction relationships trained by a Transformer model before driving the FuXi model. This hybrid system extends the forecast leading time of CMA-GFS from 8.1 days to 9.3 days, and effectively improves skill for high-impact weather in the medium-to-late forecast range (mainly beyond day 3)..

This study adopts ERA5 as initial conditions for proof-of-concept testing. For operational deployment, the initial conditions need to be switched to CMA-GFS assimilated analyses. Proof-of-concept validation serves as the first step, while operational implementation is our primary objective. The dividing line between them lies in whether CMA-GFS assimilated

analyses are used as initial conditions. At present, we are optimizing and refining the FuXi model driven by CMA-GFS assimilated analyses, with the aim of putting this system into operational use as soon as possible.

2. The selection process for the truncation wavenumber on pages 10–11 contains a logical leap. Figure 2 shows that at 850 hPa and 500 hPa, the kinetic energy spectra of FuXi and CMA-GFS are essentially consistent up to wavenumber 63, yet the authors ultimately select T21 (~2000 km) rather than T63 (~700 km). Although Figure 3 shows that T21 exhibits less attenuation at 10-day forecast lead times, the authors do not explain why a larger truncation wavenumber was not chosen in the middle troposphere to retain more useful information.

Thank you for your comments, which help us further clarify this key issue.

We aim for the hybrid system to incorporate large-scale circulations from the FuXi model without degrading the physical model's capability for meso- and small-scale systems; in other words, meso- and small-scale features should not become overly smoothed. The selection of the truncation wave number T21 is justified by the following two sets of experimental results:

1) We conducted multiple case study of high-impact weather forecasts using T21 and T42, with one typical case displayed in Figure 4. We also carried out batch winter and summer experiments to systematically evaluate and compare the two schemes in terms of large-scale circulation (Figure 5), forecast skill stability (Figure 6), vertical bias profiles and cross-sections (Figures 7-9), and precipitation ETS score (Figure 10). All results show that T21 and T42 achieve nearly identical forecast performance, and T42 provides no noticeable improvement. Given the negligible visual differences between the two set, only the T21 results are presented in Figures 6-10, as previously stated in the manuscript (Lines 385-386).

2) Kinetic energy spectrum comparisons reveal that by day-10 of integration, T21, T42 and T63 all exhibit kinetic energy decay of varying magnitudes. Larger spectral truncations introduce more FuXi-model-derived circulations during integration, leading to more pronounced decay within the coupled system (Figure 3).

Given that T42 does not deliver superior forecast performance relative to T21, we adopted the T21 to prevent mesoscale circulations in the hybrid system from becoming excessively smoothed by AI-model influences at later forecast lead times (around day 10). For the same

reason, T63 was also not considered.

Corresponding explanatory text has been added to the manuscript (Lines 307-309) to help readers follow our rationale for this choice.

3.The vertical nudging coefficient distribution in Figure 1 is described as "determined via multiple sets of seasonal sensitivity experiments," but the main text provides no details regarding the design, evaluation metrics, or results of these sensitivity experiments.

Thank you for your suggestion. The determination of vertical nudging coefficients does require further clarification.

CMA-GFS employs a hybrid height-based terrain-following coordinate with 87 vertical levels, and the model top is at 63 km. The vertical nudging coefficients are calculated via Equation (1) according to the model level index k . The coefficients k_1 , k_2 , k_3 and k_4 are derived from sensitivity tests on typical cases across different seasons. We mainly examined the zonal mean vertical profiles of geopotential height and temperature biases on day-5 under different vertical nudging profiles. The adopted values for k_1 , k_2 , k_3 and k_4 are 25, 29, 38 and 46, which roughly correspond to 600, 500, 300 and 200 hPa above sea level.

$$\lambda_{\text{vertical}} = \begin{cases} 0.0, & \text{while } 1 \leq k \leq k_1 \\ \sin(\frac{\pi}{2} \cdot \frac{k-k_1}{k_2-k_1}), & \text{while } k_1 < k < k_2 \\ 1.0, & \text{while } k_2 \leq k \leq k_3, \\ \cos(\frac{\pi}{2} \cdot \frac{k-k_3}{k_4-k_3}), & \text{while } k_3 < k < k_4 \\ 0.0, & \text{while } k_4 \leq k \leq 87 \end{cases} \quad (1)$$

The coefficients were chosen after comprehensive comparison of multiple experimental runs. Owing to page-length limitations, results from the sensitivity tests are not included in the manuscript, and we provide some intermediate results as supplementary material in this response. In addition, the final coefficients are determined via approximate trade-offs, instead of being quantitatively derived directly from experiments.

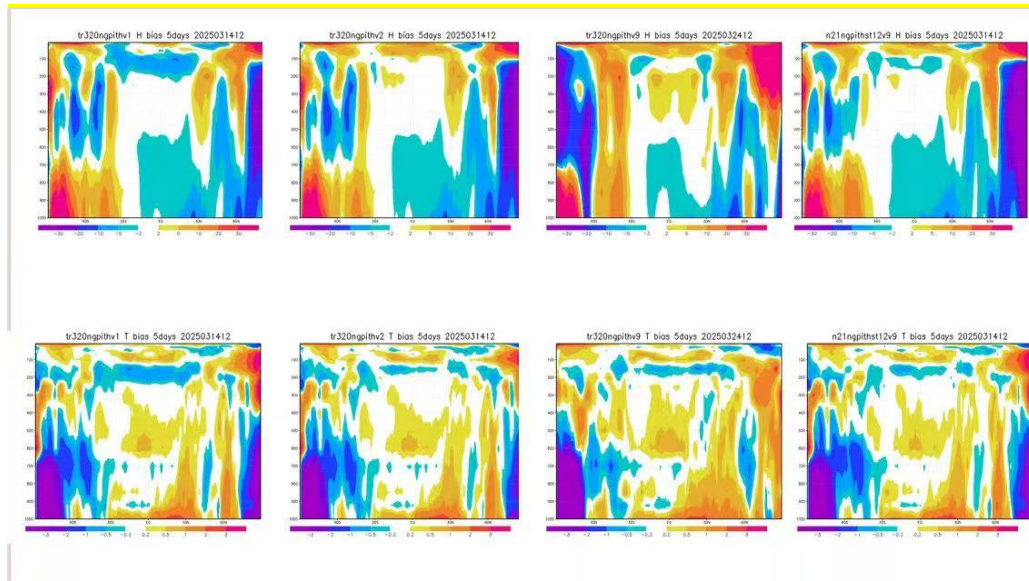


Figure 1. Zonal mean vertical profiles of forecast biases in geopotential height (upper) and temperature (lower) with different vertical nudging profiles on day-5 for a typical case.

Relevant descriptions have been added to the revised manuscript (lines 241-247).

4. Figure 11 shows that FuXi has a systematic underestimation in Vmax forecasts (MAE 15–20 m s^{-1}), while CMA-SN21 is close to CMA-GFS. However, the text does not explain why large-scale circulation correction can improve intensity forecasts—theoretically, intensity depends more on inner-core scale processes.

Thank you for your question. This point indeed requires further explanation.

In our view, the improved large-scale circulation yields better consistency among wind, pressure, temperature and moisture fields. Driven by physical processes within the numerical model, this contributes to improved intensity forecasts. Relevant explanations have been added to the manuscript (Lines 485-487).

5. Figure 2 uses an average of four dates for "24-hour forecasts" (October 1, 2024; January 1, April 1, and July 1, 2025), whereas Figure 3 uses a single case from "July 1, 2024." This shift from averaging to a single case weakens comparability.

Thank you for your question. This logic indeed requires further explanation.

Figure 2 uses multi-sample averaged KES to characterize the spectral properties of the

system, which follows standard practice.

When preparing Figure 3, we compared several different dates, including the four days in Figure 2. Across these cases, the 500 hPa KES show no decay on day-5, while varying magnitudes of KES decay emerge by day-10. To better illustrate this feature, we chose the case exhibiting the strongest decay, instead of showing multi-sample averaged KES as in Figure 2.

We have added corresponding notes in the revised manuscript to help readers better understand this reasoning (lines 301-304).

6. It is recommended to supplement the spatial distribution characteristics of precipitation improvement: does large-scale circulation correction primarily improve the systematic bias in rainband position, or does it also adjust the probability distribution of precipitation intensity? Furthermore, FuXi's precipitation in Figure 4d is shown as "<50 mm" with an overly smooth distribution. How CMA-SN21 produces extremes of 100–150 mm on this basis lacks in-depth analysis of this scale interaction mechanism.

Thank you for your question.

First, my sincere apologies: We have noticed that Figures 4c and 4e in the manuscript were swapped, and this error has been corrected (Line 348). As can be seen from the revised Figures 4e and 4f, the two nudging systems exhibit broadly consistent circulation patterns and rainband positions, which confirms the earlier mis-plotting.

Improvements in large-scale circulation do ameliorate rainband positioning, as illustrated in Figure 4. The original CMA-GFS predicts a southward shifted rainband with its precipitation center over Beijing (Figure 4c). After correcting the position of the Western Pacific Subtropical High, the two nudging systems (Figures 4e and 4f) predict precipitation centers that agree much better with observations.

The physical mechanism governing intensity forecasts is similar to that discussed in the previous question. With the improved circulation patterns, the wind, pressure, temperature and humidity fields achieve better mutual consistency. Driven by the model's internal physical parameterizations, the simulated precipitation intensity is more realistic and agrees better with observations.

Relevant descriptions have been added to the manuscript (lines 364-366).

7. Table 2 compares the technical details of the SN method among ECCC, ECMWF, and CMA, but does not include quantitative comparisons of forecast performance.

Thank you for your suggestions. We have considered them, yet it is challenging to conduct an objective and fair comparison of the forecast performance among the three systems.

We cannot run all these systems simultaneously, and can only provide approximate forecast leading time based on descriptions in the literature. However, both ECCC and ECMWF adopt operational configurations, where their AI models are initialized with their own initial conditions. By contrast, in this study the AI model is initialized with ERA5 data. A direct comparison would be unfair and misleading to readers.

As mentioned before, we have so far developed a fully-independent hybrid system. Using initial conditions from the CMA-GFS cycling system, systematic biases are adjusted via correction relationships trained by a Transformer model prior to driving the FuXi model. This hybrid system extends the forecast leading time of CMA-GFS from 8.1 days to 9.3 days.

Accordingly, approximate forecast leading time comparisons for the three hybrid systems under operational configurations are presented below:

ECCC: Boreal winter 2022, 10 days for NH, 8.9 days for SH; boreal summer 2022, 8.6 days for NH, 8.7 days for SH. Roughly estimated averages: 9.3 days for NH and 8.8 days for SH.

ECMWF: Based on the statements "gains of up to approximately 1.5 days in the tropics" and "up to 12-18 hours in the extra-tropics" from the literature, together with Table 1 in this manuscript, the estimated leading times are 9.5-9.8 days for both NH and SH.

CMA: Averaged over January, April, July and October 2025, 9.3 days for both NH and SH.

These rough estimates are provided merely for general reference and are not suitable for inclusion in the formal manuscript.

8. There are multiple instances of Chinglish expressions. For example, "the WPSH's western ridge at $\sim 115^{\circ}\text{E}$, northern edge $\sim 2^{\circ}$ south of Beijing—weaker, eastward and southward than observed" (page 15) has a fragmented sentence structure; "In addition" (page 7) should be

"In addition." It is recommended that the entire manuscript undergo professional language editing.

Thank you very much for your suggestions. We apologize for the numerous expression-related issues in the manuscript. The points you raised have been addressed in the revised manuscript (Lines 188, 357-360).

Prior to the initial submission, our manuscript had been polished by a professional editing service. However, we implemented substantial revisions to address comments from the open-discussion phase afterwards, without further review by the professional editing service. We have engaged a professional English language service provider to improve the sections where revisions were made.

Thank you again for your understanding.

Response to report 2:

This is a study on improving traditional physical model forecasts by using large-scale circulation information from machine learning models. The authors use CMA-GFS as the physical model and FuXi as the machine learning model, and introduce a spectral nudging term online into the model integration process, with the aim of improving medium-range large-scale circulation forecasts while retaining the advantages of the physical model in precipitation intensity, tropical cyclone intensity, and meso- and small-scale structures. The paper further evaluates the hybrid system through monthly batch experiments, heavy precipitation events, and tropical cyclone cases. The results have certain reference value for the integration of machine learning models and traditional numerical models and for their operational application.

The authors have responded to the previous round of review comments in a relatively serious and comprehensive manner, supplemented comparisons with related studies from ECCC and ECMWF, further clarified issues such as the spectral truncation wavenumber, the vertical nudging range, and ERA5 initialization, and objectively acknowledged that the main contribution of this study is the independent application of an existing method to the CMA-GFS–FuXi system rather than the proposal of a completely new spectral nudging theory. The revised manuscript has shown clear improvements in research positioning, methodological description, and discussion of the results, and the main conclusions are basically supported by the experimental results. However, a few issues related to methodological reproducibility, the description of verification data, and some textual expressions still need to be improved. Minor revision is recommended.

Major comments

1. The authors have already clearly stated in their response that the basic idea adopted in this study is relatively similar to those of Husain et al. (2025) and Polichtchouk et al. (2024, 2026),

and that the main value of this study lies in the independent implementation of the method using a different combination of physical and machine learning models, as well as in the further verification of its effects on precipitation and western North Pacific tropical cyclone forecasts. It is recommended that this relatively objective positioning be maintained in the abstract, at the end of the introduction, and in the conclusions. Expressions such as “providing a new pathway for operational NWP” and “robust forecasting capability” may be appropriately weakened to “demonstrating the potential of this approach” or “providing additional evidence for possible operational application”, so as to avoid giving readers the impression that this study proposes a completely new hybrid forecasting framework.

Thank you for your suggestions.

Indeed, it is necessary to more clearly state the positioning and contributions of this paper in key sections including the abstract, introduction, and conclusions. Revisions have been made accordingly as you recommended, specifically in the Abstract (Lines 27-30), Introduction (Lines 146-149), and Conclusions (Lines 492-495).

2. The manuscript provides the basic forms of the temporal nudging coefficient and the vertical nudging coefficient, but it is still recommended to clarify the following points: within each 6-hour nudging window, whether the reference field uses the FuXi forecast at the end of the window or the temporally interpolated fields between two adjacent FuXi output times; how the cumulative integration step number N is defined; the specific variation of the vertical nudging coefficient outside 600–200 hPa; and how below-ground pressure levels and upper-level extrapolation are handled when pressure, temperature, and wind fields on 13 pressure levels are interpolated to 87 model levels. The current manuscript already states that bilinear interpolation and cubic spline interpolation are used, but the above boundary treatments are still not sufficiently clear. These issues do not require additional experiments and only need to be clarified in the methodology section.

Thank you for your suggestions.

We indeed overlooked the description of the temporal coefficient in the original manuscript. Following your suggestion, we have added details regarding the evolution of N within each 6h time window (Lines 228-233). Physically, it is equivalent to interpolating the FuXi fields at the

current forecast time from two adjacent FuXi outputs.

As suggested, we have supplemented the description for the vertical nudging coefficient (Lines 241-247). It is set according to model level index and bears an approximate correspondence to pressure level. No nudging is applied outside the 600–200 hPa range.

We note that FuXi forecasts are only available for the 1000–50 hPa pressure range. Values at model levels below 1000 hPa or above 50 hPa are derived by extrapolation during transformation. However, owing to the vertical nudging coefficient shown in Figure 1, these extrapolated values are not fed into the model and have no impact on the hybrid system. Relevant descriptions have been added to the manuscript (Lines 215-219).

3. The precipitation ETS and the errors in tropical cyclone track, minimum sea level pressure, and maximum wind speed are key results supporting the application value of the paper. It is recommended that the methodology section clearly provide the name of the precipitation observation dataset, its temporal and spatial resolutions, the matching method between the model and observations, the statistical domain of ETS, and the significance testing method. For the tropical cyclone experiments, the authors should also specify the best-track dataset used, the tropical cyclone center identification method in the models, the search radius for maximum wind speed, and the meaning of the sample numbers corresponding to different forecast lead times in the figure. The authors have emphasized in their response that medium-range precipitation forecasting is an important result distinguishing this study from previous work; therefore, the description of the corresponding verification methods should be more complete.

Thank you for your suggestion about precipitation verification. The precipitation observation dataset used for ETS scores is the China national-station 6-hour precipitation dataset (~2515 stations), which has a temporal resolution of 6-hour (24-h accumulated precipitation is also supported) and consists of discrete station observations. For model-observation matching, model grided forecasts are bi linearly interpolated to station geographic coordinates, forecast accumulation windows are temporally aligned with station observations, and point-to-point pairs are adopted to build contingency tables for ETS computation. The ETS statistical domain covers Chinese mainland within 73°E-136°E,

18°N-54°N, and the significance test is implemented via non-parametric bootstrap resampling to obtain the 95% confidence interval ($\alpha=0.05$) for ETS scores.

Relevant descriptions have been added to the manuscript (Lines 446-452).

Thank you for your question regarding tropical cyclone verification. The best-track dataset is sourced from the Shanghai Typhoon Institute (STI) of CMA (<https://tcdata.typhoon.org.cn/>). TC center positions are determined using a two-step algorithm. First, first-guess estimation uses the TC position from the prior time step as the starting point. The search radius is computed from the maximum plausible TC translation speed (~150 km/hr) multiplied by the model output time interval. The location of minimum 1.5km altitude pressure perturbation within this domain yields the first-guess center. Second, vorticity centroid refinement calculates the 1.5km altitude vorticity centroid by spatially weighting relative vorticity within a 300km radius of the first-guess position. Representing the TC's rotational dynamical core, the vorticity centroid refines center localization and reduces biases induced by small-scale perturbations. The x-axis denotes forecast lead time and the number of valid samples.

Relevant descriptions have been added to the manuscript (Lines 461-462, 468-475).

4. The authors have explained that ERA5 is used to initialize FuXi in this study in order to ensure optimal performance under proof-of-concept conditions; if CMA-GFS analysis fields are used instead, the predictable lead time of FuXi may be shortened by approximately 1–2 days. Therefore, the current experiments are more appropriately defined as proof-of-concept and cannot fully represent the performance of a real-time operational system. It is recommended that a clear statement be added to the abstract or conclusions and that the conclusions related to operational application be limited to “potential operational application”, so as to avoid conclusions that exceed what can be supported by the current experimental design.

Thank you for your suggestions. It is indeed necessary to clarify the scope of this work. The experimental results presented are based on ERA5 data and do not represent the actual operational performance of the system. Corresponding revisions and additions have been made in the abstract (lines 21-22, 29-30), and explicit statements have also been incorporated in the discussion section of the conclusion (lines 508-509).

In practice, we have so far developed a fully-independent, hybrid system. Using initial conditions from the CMA-GFS cycling system, systematic biases are adjusted via correction relationships trained by a Transformer model before driving the FuXi model. This hybrid system extends the forecast leading time of CMA-GFS from 8.1 days to 9.3 days, and effectively improves skill for high-impact weather in the medium-to-late forecast range (mainly beyond day 3).

This study adopts ERA5 as initial conditions for proof-of-concept testing. For operational deployment, the initial conditions need to be switched to CMA-GFS assimilated analyses. Proof-of-concept validation serves as the first step, while operational implementation is our primary objective. At present, we are optimizing and refining the FuXi model driven by CMA-GFS assimilated analyses, with the aim of putting this system into operational use as soon as possible.

Minor comments

1. The potential temperature formula given in Lines 210–215 appears to be

$\theta = T(p/p_0)R/C_p$, whereas it should usually be $\theta = T(p_0/p)R/C_p$.

Please verify whether this is a typographical error and ensure that the formula is consistent with the calculation method used in the actual program.

Thank you for the suggestions. This was a typographical error; the program itself is correct. Revisions have been made in the manuscript (Line 214).

2. The correspondence between wavenumber and wavelength in Lines 265–268 needs to be corrected. Wavenumbers smaller than 63 should correspond to wavelengths larger than approximately 700 km, rather than “wavelengths shorter than approximately 700 km”; wavenumbers smaller than 21 should correspond to wavelengths larger than approximately 2000 km. The current wording is inconsistent with the basic relationship between wavenumber and wavelength in spectral space.

Thank you for the suggestions. This was an oversight on our side and we have corrected

it in the manuscript (lines 280,282).

3. Table 2 provides a useful comparison of the technical differences among the ECCC, ECMWF, and present systems, but the model names should be unified. “GRAPES-GFS” is used in the table, whereas “CMA-GFS” is mainly used in the main text. The relationship between the two should be explained or the terminology should be unified.

Thank you for the suggestion. We have unified the model name in the manuscript. In addition, the relationship between CMA-GFS and GRAPES-GFS is explained in Table 3 (Line 327): the global medium-range NWP system independently developed by the CMA was originally named GRAPES-GFS. Later, to unify domestic model nomenclature, CMA renamed it CMA-GFS.

3. The units and formatting in Table 2 should be standardized. For example, “12hour” should be changed to “12 h”, “600hPa-200hPa” to “600–200 hPa”, and “2750-2250km” to “2750–2250 km”.

Thank you for your suggestion. We have revised the units and formatting in Table 2 accordingly (line 322).

4. Please clarify the exact meaning of “relative to the 0-hour forecast field” in Figures 8 and 9. If the comparison is made against the ERA5 analysis at the corresponding valid time, it should be written directly as “relative to ERA5 analysis”; if it is relative to the initial field of the same forecast, it should not be referred to as forecast error or forecast bias.

Thank you for your suggestion. Sorry for the unclear expression. As suggested, we have revised the sentence to “relative to ERA5 analysis” (line 410).

5. Please check the figure references in the main text. For example, when discussing tropical cyclone track, minimum pressure, and maximum wind speed results, the manuscript should not cite only “Figures 11a and 11d”, but should cite Figures 11a–f according to the actual content.

Thank you for your suggestions. The wording here was indeed incorrect and has been

revised as you recommended (line 476).

6. Figure 2 contains many curves for different pressure levels and models. It is recommended to appropriately improve the clarity of the legend and axis labels and to ensure consistent capitalization of “hPa”.

Thank you for your suggestions. We have adjusted the font size and weight of the legends and axis labels in Figure 2 (line 269) and Figure 3 (line 290). Besides, we are not entirely sure about the meaning of "consistent capitalization of hPa". Does it refer to the case of these three letters? In meteorological papers, the standard notation is generally “hPa”, rather than “HPA” or “hpa”.

7. There are still some spelling and grammatical errors in the manuscript, such as “In addiation”, “The detial”, “gas constan”, and “The publicly available version of the FuXi”. It is recommended that the entire manuscript undergo a systematic English-language proofreading.

Thank you very much for your suggestions. We apologize for the numerous spelling errors. The spelling mistakes you pointed out (Line 178, 210, 215) have been corrected. In addition, “FuXi” is the official capitalization for this model name.

Prior to the initial submission, our manuscript had been polished by a professional editing service. However, we implemented substantial revisions to address comments from the open-discussion phase afterwards, without further review by the professional editing service.

We have engaged a professional English language service provider to improve the sections where revisions were made. Thank you again for your understanding.

8. In the newly added description of AIFS-afCRPS, “Lang et al. (2025) adopted ..., it enables ...” contains a comma splice connecting two independent clauses. It may be revised to “..., which enables the model to ...”.

Thank you for your suggestions. The comma-splice error has been fixed. The pronoun “it” is replaced by “which” to form a non-restrictive relative clause (line 87).

9. It is recommended to standardize the use of the terms “machine learning model”, “ML-based weather prediction model”, and “ML model” throughout the manuscript, and to change “Machine learning” in the title to “Machine Learning”.

Thank you for your suggestion. “learning” in the title has been revised to “Learning”. For the abstract, “ML-based weather prediction model” is used at its first occurrence, and “ML model” is adopted consistently in subsequent mentions. All other instances of “machine learning” throughout the manuscript have also been unified as “ML”.