Response to Reviewers

We thank the reviewers for their careful reading of our manuscript and for the insightful comments. We address each point below and will revise the manuscript accordingly.

Reviewer #3

Major Comments

Major Comment 1:

> This should be very easy to address, but the paper needs to better discuss the relation between AIFS-Single and AIFS-ENS. Is the only difference between AIFS-Single and AIFS-CRPS the use layer-norm noise and a CRPS dominated training loss? In particular are these additional variables and bounding-layer strategies now deployed in the AIFS-ENS? If these improvements are not clearly planned for incorporation into the AIFS-CRPS, why not?

We thank the reviewer for this comment. The relationship between AIFS-Single and AIFS-CRPS, including their architectural similarities and differences, is documented in Lang et al. 2024b. The present paper focuses exclusively on the deterministic AIFS-Single system; a detailed comparison with the ensemble configuration is outside its scope. We note that work towards a unified framework for both systems is ongoing and is planned for a future release.

Major Comment 2:

> How much of the improvement in the update is simply due to including 2021 and 2022 in the training data. I imagine not that much, but there are potentially lots of other differences in the training schedule that could be responsible for the improvement in ACC for Z500 and T850 (Fig. 6). How does RMSE compare for these fields? The authors should try to isolate the source of this improvement, particularly if it turns out to be a better treatment of the loss weighting in the stratosphere.

Response: We thank the reviewer for this important question. To directly address it, we have added a controlled comparison to the revised manuscript (Figure 5). We trained an additional configuration, AIFS revised (2020 data), which uses the same revised architecture, loss weights, and training methodology as the full revised system, but with the training data restricted to the same period used for AIFS previous (ERA5 up to 2020, operational analysis fine-tuning for 2019–2020 only). Comparing the three configurations, AIFS revised

full, AIFS revised (2020 data), and AIFS previous, for Z500 ACC and T850 ACC isolates the contribution of the data expansion from the other changes.

The results show that the data expansion (ERA5 extended to 2022, operational fine-tuning expanded to 2016–2022) accounts for the largest share of the improvement in ACC. The remaining gain, attributable to the revised architecture and training methodology, is smaller but nonetheless present, and is more clearly visible in spectral characteristics than in ACC: as shown in Figure 6, AIFS revised (2020 data) already exhibits reduced mesoscale smoothing relative to AIFS previous despite comparable ACC scores, demonstrating that ACC alone is not sufficient to capture all aspects of forecast quality improvement.

Regarding RMSE: RMSE for Z500 and T850 is included in the scorecard (Figure 7), which shows consistent improvements across all variables, lead times, and regions.

Regarding stratospheric loss weights: the stratospheric improvement is attributed to the imposition of a minimum loss weight (0.2) for upper-level pressure levels, as described in Section 2. This is confirmed by Figure 9, which shows substantial improvement at 100 and 50 hPa in the revised system. This change is present in both the full and 2020-data configurations and therefore contributes to the non-data component of the improvement seen in Figure 5.

Major Comment 3:

> Further details about expanding the comparison with the other baselines
>
> Figs. 8, & 9a,b: why which switch from ACC (in Figs. 6 & 7) to RMSE for these variables. For a more thorough analysis, please plot both both RMSE and ACC for all of these cases?
>
> Fig. 7 caption: The evaluation for " the whole of 2023" - what does that mean, Presumably from caption to Fig. 5, twice daily (00 and 12 UTC forecasts for every day of the year? Maybe the authors can clearly establish this in the text and then only note any exceptions. (Sorry if I missed such a sentence.)
>
> Fig. 9: Why the different time ranges: (a) ssd is MAM, (b) is full year, (c) total cloud cover is JJA. Without further discussions this seems like it could be cherry picking.

Response:

ACC vs RMSE: ACC for Z500 and T850, RMSE for 2-metre temperature and 10-metre winds are standard headline metrics in NWP verification. ACC is the conventional choice for upper-air fields where anomalies from climatology matter most, while RMSE is typical for surface variables where absolute errors are more operationally relevant. Both metrics appear in the scorecard (Figure 7), which provides a comprehensive overview across all variables.

Forecast frequency: We apologise for the lack of clarity. Unless otherwise stated, all verification is based on twice-daily forecasts initialised at 00 and

12 UTC for the full year 2023. This is mentioned in the Figure 6 caption, but we have now added an explicit statement at the beginning of Section 4 to establish the default verification protocol.

Different time ranges: The different evaluation periods in Figure 11 reflect data availability, not selective presentation. Surface solar radiation is verified against CM SAF satellite observations, which were available only for MAM 2023 at the time of evaluation. Total cloud cover uses JJA 2023 also for data availability at the time of evaluation. The 100-metre wind verification uses the full year. We have clarified this in the revised caption. Three months of global, twice-daily forecasts already constitutes a statistically robust sample.

Minor comments

> l. 137: Do the authors mean 64 A100 80 GB GPUs (or maybe 64 of the 40 GB A100s)?

A100 40GB GPUs. This is now explicitly stated in the text.

> l. 190-192: Consider referencing Subramaniam et al., 2025 (arXiv:2506.08285) who add a loss penalty to effectively obtain a model that respects hydrostatic balance

This work is now cited.

> Fig. 5: Framed rectangles are difficult to see. I suggest using a thick underline for statistically significant results, though the authors may have a better idea. Perhaps even better, they could plot only those values that are significant at the 95% level.

We believe the current convention is already readable. For instance, in the Northern Hemisphere nearly all RMSE and ACC scores are framed (significant at 95%), with the exception of wind speed at 850 hPa at day 10 against observations. In the Southern Hemisphere, Z500 ACC scores against analysis are significant up to day 5 but not beyond, which is immediately apparent from the unframed boxes. This framed/unframed convention is standard in ECMWF scorecards and familiar to the NWP community.

> l. 258-259 and/or Fig. 6 caption: please be more specific about "medium range" — which seems like 1-day skill improvement at about 1-week lead time

Generally, the medium range term is used for forecast from 3 to 10 days in advance. This is now explicitly stated in the text.

> l. 275, Fig. 10: please do plot "the three main global regions" as suggested in the text and caption instead of results for just the northern and southern hemispheres.

We thank the reviewer for pointing this out. The text and caption have been corrected to accurately reflect that Figure 12 shows results for the Northern and Southern Hemispheres. The tropics verification is already included in the scorecard (Figure 7).

> l. 391: The Leaky ReLU would indeed allow calculations of gradients in the negative space, but it seems like it would open the door to negative precipitation amounts as well. There are a variety of possible ways to construct loss functions that handle "no rain" separately. A more thorough discussion of this would be useful.

We thank the reviewer for this valid point. With a small leak factor (e.g., $\alpha = 0.01$), the loss contribution from negative predictions is attenuated by α^2, making it negligible. The model would therefore still be strongly incentivised to push dry predictions deep into the negative space, and we expect similar regime separation to emerge. The main practical difference is that LeakyReLU produces non-physical slightly negative output values at inference, requiring post-processing clipping. We have expanded the discussion in Section 5 to clarify this and to discuss alternative formulations such as asymmetric loss functions or dedicated classification heads for the no-rain state, as promising directions for future work.