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 #2

General comment 1:

> I think the main benefit of ReLU training is that it reduce the small positive drizzle, which is nice and likely robust to my concerns below. It's unclear how this impacts the skill scores though or if it could be achieved by other means e.g. threshold all precip $< 0.1mm/day$ to 0 in the old AIFS.

Response: We agree that drizzle reduction is the most direct benefit of ReLU bounding. However, post-processing thresholding and training-time bounding are fundamentally different interventions. Without bounding during training, the MSE gradient ambiguity at zero causes dry events and drizzle to collapse into an indistinguishable region around zero (see our response to Reviewer 1, Major Issue 1, and the expanded Section 4.1). Thresholding can remove false positives from this contaminated distribution but cannot undo the representational damage. ReLU bounding resolves the ambiguity at source, enabling the model to cleanly separate dry and wet regimes during training. The impact on skill scores is documented in Section 4.1 through the controlled comparison between the bounded and non-bounded configurations.

General comment 2:

> I'm surprised in 2026 to be reviewing a paper on AI weather forecasting without probabilistic skill assessments. The introduction acknowledges this but then goes on to employ known blurring techniques like multistep training with MSE loss. Figure 15 really highlights the blurriness issue. This confound all their skill scores, and makes it impossible to conclude that their proposed modifications are genuinely helpful or if the model is just blurrier. The manuscript shows no results that would contradict this like power spectra or precipitation pdfs. Many of their skill improvements could alternatively be achieved by ensemble averaging. I have some hope that their findings are robust since the skill is also better for short lead times where the blurring impact is less important, but still this needs to be assessed better. One possible idea short of a full-blown probabilistic assessment is to compare probabilistic scores on lagged ensembles built from their existing deterministic hind-casts [1].

Response: We thank the reviewer for their comment. We agree that MSE-trained deterministic models produce spatially smoother fields at longer lead times. However, we believe the evidence presented in the revised manuscript is sufficient to demonstrate that the reported improvements are not attributable to increased blurring.

First, the revised manuscript now includes Z500 power spectra (Figure 6), added in response to Reviewer 1. These spectra compare the revised AIFS, the previous AIFS, and the IFS analysis across lead times and show that the revised model exhibits comparable spectral characteristics to the previous version, with no additional loss of spatial variability. This directly addresses the concern that "the manuscript shows no results that would contradict this."

Second, the relevant comparison throughout the paper is between the revised AIFS and the previous AIFS version, both of which use MSE loss with multistep fine-tuning. The blurring associated with this training paradigm is therefore a shared baseline, not a confound. For the skill improvements to be an artifact of smoothing, the revised model would need to be more blurry than its predecessor, which the power spectra show is not the case.

Third, the bounding layer evaluation in Section 4.1 provides a fully controlled comparison: the bounded and non-bounded configurations differ only in the presence of the ReLU constraint, with identical training data, loss function, rollout length, and architecture. Any blurring confound is therefore perfectly controlled in this experiment. Moreover, the precipitation improvements are assessed through categorical scores (FBI, PSS) that measure the frequency distribution of precipitation events at fixed thresholds, not through RMSE.

Regarding probabilistic assessment: this manuscript specifically documents the deterministic AIFS configuration (AIFS Single). A probabilistic ensemble version, AIFS-CRPS, has been developed and is evaluated with proper probabilistic skill metrics in separate work Lang et al. 2024b.

General comment 3:

> I also have some concerns about the novelty of the proposed methods. Adding a few input/output variables (some of which other models already) and enforcing precip ¿ 0 using a relu are valuable model developments, but seem incremental. I would be surprised if others aren't dong that, though I will admit to not finding a specific reference off the top of my head.

Response: We thank the reviewer for their honest opinion. The contribution of this manuscript operates on two distinct levels.

First, as a Geoscientific Model Development paper, it documents the development and evaluation of an operational forecasting system, AIFS v1.1.0, that is deployed at ECMWF and used by national meteorological services worldwide. The systematic documentation of architectural choices, training procedures, and verification results for operational AI weather prediction systems is an important contribution to the community, consistent with GMD's scope and with how conventional NWP system upgrades are documented in the literature.

Second, the scientific contribution extends well beyond "adding a ReLU." The bounding framework includes variable-specific strategies (ReLU for non-negative variables, HardTanh for bounded variables such as total cloud cover, and FractionBounding to enforce inter-variable consistency between convective and total precipitation), but the key novelty lies in understanding why these constraints improve forecast skill. The mechanistic analysis presented in the revised Section 4.1 including the MSE gradient ambiguity theory at the zero-precipitation boundary, the regime-dependent correlation analysis of the pre-activation space, and the quantitative comparison of output distributions between bounded and non-bounded configurations provides, to our knowledge, the first theoretical explanation for how non-negativity enforcement reshapes the loss landscape and resolves drizzle contamination in the AIFS. We note that the reviewer was unable to identify a reference for this analysis, which we believe reflects its novelty rather than its incrementality.

We also note that concurrent work by Sha et al. (2025) has independently demonstrated the importance of non-negativity constraints for drizzle reduction, but within a combined conservation framework that does not isolate the specific mechanism. Our analysis complements theirs by providing an explanation for why non-negativity alone can play a very important role.

Specific comment (Figure 2 presentation):

> The use of scatter plots for plotting precipitation or any other map increases the pdf size greatly and more importantly adds visual noise that hinders assessment of fine-scale structures. I understand it is not entirely trivial to plot data on the octahedral grid, but surely there is some way to make quad-mesh or contour plots.

Response: Figures 2, 4, 14 and 15 are now produced using contour plots instead of scatter in the revised version of the manuscript.

Specific comment (L37–41):

> "Although such MSE-trained forecast models have been shown to smooth forecast fields at longer lead times to avoid the double-penalty of incorrectly positioned weather phenomena (Lam et al., 2023; Ben Bouallègue et al., 2024; Lang et al., 2024a; Bonavita, 2024)..."
>
> It would be appropriate to also cite the Brenowitz et al.2025 work on lagged ensembles, which explicitly links field blurring to multistep fine-tuning.

Response:
The work of Brenowitz et al. is also cited now.

Specific comment (L113–116):

"We have increased the characterization of the land surface in the model by including new prognostic variables"

Do these non-prognostic variables materially influence forecast skill? A targeted sensitivity experiment would be informative.

Response: We thank the reviewer for their comment. Their inclusion is motivated primarily by their importance for downstream applications: land surface variables such as soil moisture, runoff etc..., are operationally relevant outputs in their own right, used directly by hydrological models, agricultural forecasting systems, and land-surface monitoring services. Expanding the set of predicted land-surface variables therefore broadens the range of end-user applications that AIFS can support, independently of any impact on upper-air skill scores. We did not run a dedicated ablation for these variables in isolation. As the land-surface representation is extended in future AIFS versions, targeted ablation studies to quantify the impact of individual land variables on near-surface forecast skill (e.g., 2-metre temperature) are planned. We have added a brief statement to this effect in Section 5.

Specific comment (L218–221):

"Clipping the precipitation output in inference is a possibility and a common practice... However, we show that the introduction of bounding in the output during training has benefits beyond simply avoiding slightly negative or unphysical values..."

Clipping is likely to introduce bias. Since MSE-trained models predict the conditional mean, truncating negative values alone will generally induce a positive mean bias. It would be worth showing how the model climatology (averaged over initial times) differs spatially from IMERG..

Response: We think the concern is actually inverted. The gradient analysis in the revised Section 4.1 shows that the non-bounded model already carries a positive bias in light precipitation: the MSE dynamics systematically drift dry predictions across zero into weakly positive values. Training-time bounding removes this contamination rather than introducing new bias, dry states collapse cleanly onto zero, freeing the positive axis for genuine precipitation. This is distinct from inference-time clipping, which cannot undo the representational damage incurred during training. Regarding IMERG climatology comparison: this would be interesting but tangential, as the claim is about resolving an optimisation instability at the zero boundary, not about matching a specific observational climatology.

Specific comment (L262):

"The Frequency Bias Index (FBI) and Peirce Skill Score (PSS) are shown for the Northern Hemisphere for different thresholds."

> I'm not familiar with these metrics. Can you add citations and definitions? Perhaps histogram would be more familiar to the broader audience make the point that the model predicts too much light drizzle.

Response: We have added brief definitions and citations for both metrics in Section 4 of the revised manuscript. The Frequency Bias Index (FBI, Wilks 2019) is defined as the ratio of predicted to observed event frequency at a given threshold, FBI $= (H + FA)/(H + M)$, where H are hits, FA false alarms, and M misses. The Peirce Skill Score (PSS, also known as the Hanssen–Kuipers discriminant; Jolliffe 2011) is defined as PSS $= H/(H + M) - FA/(FA + CN)$, where CN are correct negatives.

Regarding the histogram suggestion: Section 4.1 of the revised manuscript now includes a panel showing the precipitation frequency distributions for the bounded and non-bounded models near the zero-precipitation boundary (Figure 17, bottom panel). This directly illustrates that the non-bounded model accumulates excess density in the light-precipitation range, consistent with the systematic misclassification described in the text.

Specific comment (L276–278):

> "The revised AIFS demonstrates approximately a one-day gain in forecast skill... The forecast fields also exhibit noticeable improvements, as illustrated in Figure 2..."
>
> These improvements may simply reflect increased spatial smoothing. As presented, the evidence is inconclusive.

Response: The revised manuscript now includes Z500 power spectra (Figure 6) showing that the revised AIFS has comparable spatial variability to its predecessor, the skill gains are not accompanied by additional smoothing. We also note that both models use the same MSE loss and rollout configuration, so any smoothing baseline is shared. See also our response to General Comment 2.

Specific comment (L311–314):

> "The results show that the improvement observed in total precipitation forecast skill in the revised AIFS version can mainly be attributed to constraining the output..."
>
> The improvement appears largely constant with lead time rather than growing, suggesting that it may not feed back on error growth and could potentially be achieved through post-processing. Figure 11 may be relevant in this context.

Response: The constant improvement with lead time is exactly what we would expect. Total precipitation is diagnostic in AIFS, it is not cycled back

into the model state. The bounding therefore corrects the same per-step zero-boundary contamination at each forecast step independently, rather than compounding through rollout. This is consistent with the mechanism. The key point remains that post-processing clips a contaminated distribution, whereas training-time bounding produces a fundamentally different distribution: the bounded model cleanly separates dry and drizzle regimes during training, and this cannot be recovered by truncating the output of a non-bounded model after the fact.

Specific comment (L329–331):

> "Unlike the previous AIFS version (Figure 4), the convective precipitation forecast is now consistent with the predicted total precipitation accumulation..." This figure does not demonstrate that training with this constraint is essential. It would be useful to assess the impact of applying similar consistency corrections as a post-processing step to the "AIFS revised – no bounding" configuration.

Response: Post-processing can enforce cp $\leq$ tp by clipping, but this only masks incoherent outputs without ensuring that the model learns a physically meaningful relationship between the two fields. With FractionBounding, cp is predicted as a learned fraction of tp, so the constraint is built into the model's representation, the network must learn to partition total precipitation into convective and large-scale components. We argue that this is preferable to training a model that produces inconsistent fields and correcting them afterwards. More generally, if a physical constraint can be enforced during training at no additional cost, there is little reason to defer it to post-processing.