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 #1

Major issue 1 (Bounding Layer Mechanism: Insufficient Mechanistic Explanation):

> The manuscript's primary scientific contribution concerns the bounding layer strategy (Section 3.2, lines 221-227). The authors claim that applying ReLU activation functions to precipitation outputs "facilitates the learning of forecasting for sparse and intermittent variables" by enabling negative output space to serve as a proxy for no-rain classification. Figure 12 presents compelling evidence of structured spatial patterns in pre-activation space, showing strongly negative values over arid regions (Sahara Desert) with smooth gradients near precipitation events.
>
> This observation represents the paper's most interesting result, yet receives inadequate mechanistic analysis. During backpropagation, the ReLU derivative is zero for all negative inputs ($\frac{\partial \text{ReLU}}{\partial x} = 0$ for $x < 0$), suggesting gradient information should not flow to regions predicting negative values. How, then, does the model develop sophisticated spatial structure visible in Figure 12? What is the mathematical relationship between the MSE loss gradient and the learned negative space structure? Does this structure emerge from autoregressive training, where negative predictions at intermediate steps influence subsequent forecasts through rollout? Can the information content in the negative space be quantified? How does this compare to alternative formulations such as LeakyReLU (mentioned line 391 but not evaluated)?
>
> The authors should provide mechanistic explanation for the bounding layer's behavior. This need not require expensive additional experiments rather a clear theoretical model of gradient flow during rollout training would suffice. At minimum, the structured negative space in Figure 12 deserves quantitative analysis rather than qualitative description. Understanding this mechanism is critical for generalizing the approach to other sparse variables.

Response:

We thank the reviewer for identifying the effects of ReLU activation on precipitation outputs as one of the paper's most significant results. We agree that a mechanistic explanation is required to clarify how structured patterns

emerge in the negative pre-activation space despite the zero-gradient property of the ReLU function ($\partial \text{ReLU}/\partial x = 0$ for $x < 0$).

We clarify that this structure does not emerge from autoregressive training, as total precipitation (tp) is a diagnostic variable in AIFS and is not cycled back into the model state. Instead, the structure is an emergent feature arising from the shared representation of atmospheric states.

The model encodes input prognostic ($\mathbf{X}_t$) and forcing variables ($\mathbf{F}_t$) into a high-dimensional latent space (z_t) via an encoder:

$$z_t = \text{Encoder}(\mathbf{X}_t, \mathbf{F}_t) \tag{1}$$

This latent state is evolved to the next time-step through the processor (e.g., via attention-based computations):

$$z_{t+6} = \mathcal{F}(z_t) \tag{2}$$

and then decoded back into the physical space to obtain the forecast at t+6 of prognostic ($\mathbf{X}_{t+6}$) and diagnostic ($\mathbf{D}_{t+6}$) variables. It is worth mentioning here that z_{t+6} encodes the physical state of all the prognostic variables in a shared representation space and the diagnostic variables are decoded from it. The diagnostic precipitation output is thus produced by a specific decoder head:

$$\eta_{t+6} = \text{Decoder}_{\text{tp}}(z_{t+6}) \tag{3}$$

where η represents the pre-activation total precipitation. The final physical output is obtained via the bounding layer:

$$\text{tp}_{t+6} = \text{ReLU}(\eta_{t+6}) = \max(0, \eta_{t+6}) \tag{4}$$

Because $\text{Decoder}_{\text{tp}}$ maps from a latent space optimized for smooth gradients (z_{t+6}), η inherits this spatial structure. The precipitation decoder head learns a smooth mapping from the latent space encoding the moisture state of the system to physical precipitation in the positive regime ($\eta > 0$), where gradients are active. Because neural networks are continuous functions biased toward smoothness, this "moisture-to-precipitation" logic naturally extrapolates into the negative regime. As moisture variables decrease, the decoder continues to output decreasing values, pushing η into the negative space.

While the precipitation head receives no direct gradients when $\eta < 0$, the latent variables that serve as its input are not static. These latent features are shared with prognostic variables (e.g., specific humidity q, total water content tcw, etc) and receive continuous gradient information from their respective loss functions. Consequently, the negative space of the tp field is "indirectly learned"; it is a projection of a latent space that is being rigorously optimized.

Ultimately, this reveals that the optimization of the shared latent space is driven by the collective constraints of all output variables. In this framework, the negative pre-activation space for precipitation serves as a "saturation deficit" proxy that is kept physically consistent by the gradients flowing from prognostic moisture fields. The shared representation of the atmosphere in the latent

space allows the model to maintain a sophisticated, structured representation of dryness even in the absence of direct precipitation gradients.

To provide empirical weight to this mechanistic theory, we investigate the information content within the pre-activation space η by partitioning the model output into three distinct physical regimes: the negative (non-precipitating) space, the light precipitation regime (0–0.5 mm/6h), and the moderate precipitation regime (0.5–10 mm/6h).

We hypothesize that the pre-activation space η undergoes a fundamental physical decoupling as it transitions from dry to wet conditions. In the negative (non-precipitating) regime, the absence of precipitation is a deterministic function of low humidity; thus, the decoder should preserve a strong linear mapping from the prognostic moisture fields.

Conversely, we expect this linear correlation to weaken in the light precipitation regime ($0 < \eta \leq 0.5$ mm). While moisture remains a necessary condition for rain, the exact accumulation at these low intensities becomes increasingly stochastic, influenced by non-linear factors such as sub-grid scale turbulence, cloud-base evaporation, and microphysical uncertainties. These processes act as "interference," decoupling the surface precipitation from the column moisture signal.

We performed a global correlation analysis on a single forecast issued at 01/06/2023 00:00 UTC. For this experiment, we utilize the AIFS revised model without the final bounding layer on tp during inference, but activated during training. We focus our analysis on the first 120 hours (5 days) of the forecast.

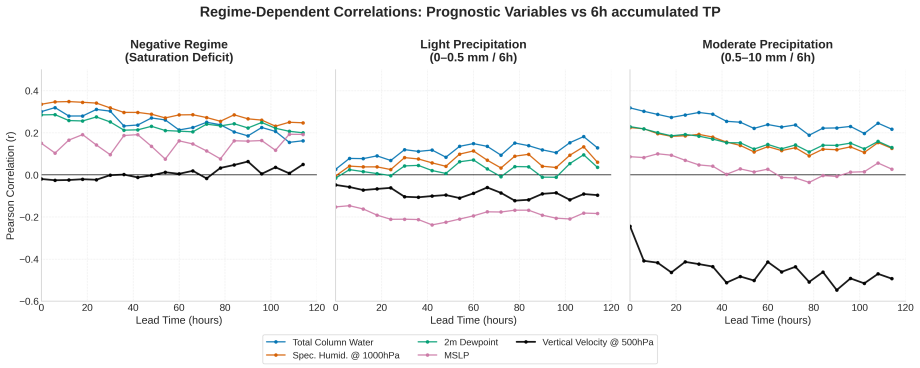

Figure 1: Regime-dependent correlations of pre-activation η (AIFS Revised, for a forecast issued the June 1, 2023 at 00 UTC. Pearson r between η and physical drivers across three regimes: (Left) Negative space ($\eta < 0$): high correlation with moisture variables (q_{1000}, TCW) identifies η as a structured saturation deficit proxy. (Center) Light rain ($0 < \eta \leq 0.5$ mm/6h): systematic weakening of correlation, likely associated with enhanced stochasticity in this regime. (Right) Moderate rain ($1 < \eta \leq 10$ mm/6h): transition to dynamic control, with vertical velocity(w_{500}) as the dominant predictor ($r \approx -0.5$). Analysis covers a 120-hour global forecast.

We computed the Pearson correlation coefficient (r) between the pre-activation field η and five key physical drivers: Total Column Water (TCW), Specific Humidity (q_{1000}), 2m Dewpoint ($2d$), Mean Sea Level Pressure (MSLP), and mid-tropospheric Vertical Velocity (w_{500}). As shown in Figure 1, the results reveal a clear regime-dependent physical logic:

- **Negative Regime ($\eta < 0$):** We observe stable correlations ($r \approx 0.3$) with moisture variables (q_{1000}, TCW, and $2d$). This confirms that the negative space encodes a structured representation of the *saturation deficit*, kept physically consistent by gradients flowing from the prognostic moisture fields.
- **Light Precipitation ($0 < \eta \leq 0.5$ mm):** Correlation with specific humidity, 2m dewpoint and total column water is substantially reduced in this regime. The weaker relationships are consistent with a lower signal-to-noise ratio and increased sensitivity to small-scale or non-linear processes.
- **Moderate Precipitation ($1 < \eta \leq 10$ mm):** The model transitions to dynamic control. While moisture correlations remain moderate, Vertical Velocity (w_{500}) emerges as the primary physical driver ($r \approx -0.5$), illustrating the model's reliance on large-scale ascent to produce deterministic rainfall.

While presented as a targeted demonstration of internal model behaviour, the consistency of these signals across lead times suggests that this regime-specific transition is a fundamental structural property of the AIFS architecture. These results demonstrate that the negative pre-activation field encodes valuable information regarding a proxy for saturation deficit. We acknowledge that these correlations are computed from a single 5-day forecast, which limits the temporal sampling. However, the analysis is performed on a Gaussian reduced N320 grid, such that each 6-hourly forecast field contains more than 500,000 spatial evaluation points. Although based on one forecast initialization, the large number of grid-point samples per lead time provides a substantial statistical basis for examining the internal behaviour of the model.

Having established that the negative pre-activation space encodes physically meaningful information, we now turn to understanding why constraining it during training improves forecast skill for light precipitation. The mechanism can be understood by examining how the Mean Squared Error (MSE) interacts with model outputs in the vicinity of the zero-precipitation boundary for a non-bounded model:

1. **Scenario A (Non-physical negative dry prediction):** The model predicts a non-physical negative value ($tp = -0.2$ mm) for a dry observation ($tp_{obs} = 0$ mm). The gradient of the Mean Squared Error (MSE) is:

$$\frac{\partial \mathcal{L}}{\partial tp} = 2(tp - tp_{obs}) = 2(-0.2 - 0) = -0.4 \quad \text{(Push Up)} \tag{5}$$

2. **Scenario B (Underprediction):** The truth is light rain ($tp_{obs} = 0.45$ mm), but the model under-predicts the intensity ($tp = 0.25$ mm). The gradient is:
$$\frac{\partial \mathcal{L}}{\partial tp} = 2(0.25 - 0.45) = -0.4 \quad \text{(Push Up)} \tag{6}$$

3. **Scenario C (Overprediction):** The truth is dry or very light rain ($tp_{obs} = 0.05$ mm), but the model over-predicts the intensity ($tp = 0.25$ mm). The gradient is:
$$\frac{\partial \mathcal{L}}{\partial tp} = 2(0.25 - 0.05) = +0.4 \quad \text{(Push Down)} \tag{7}$$

Because non-physical negative dry predictions (Scenario A) and genuine drizzle underpredictions (Scenario B) produce identical upward gradients, the optimizer receives an ambiguous training signal in the vicinity of zero. The loss provides no information about why the correction is required — whether it reflects a physical regime transition (dry $\rightarrow$ drizzle) or merely a violation of the non-negativity constraint. One might expect the model to self-organize by learning to place dry predictions in a compact negative range, say, around -0.1 mm, thereby avoiding interference with the light-rain regime. However, this equilibrium is dynamically unstable under MSE. A dry prediction at -0.1 mm receives the same upward gradient as a genuine drizzle underprediction, so stochastic gradient updates continually push dry samples toward and across zero. As a result, no stable attractor can form in the negative space.

Importantly, the instability is locally asymmetric around $tp = 0$. For small $tp = \epsilon$ with $|\epsilon| \ll 1$,
$$\frac{\partial \mathcal{L}}{\partial tp} = 2(\epsilon - tp_{\text{obs}}).$$
In the neighbourhood of zero, the target distribution is one-sided: $tp_{\text{obs}} \geq 0$, with strictly positive drizzle values arbitrarily close to zero but no negative observations. Let
$$\mu = \mathbb{E}[tp_{\text{obs}} \mid tp_{\text{obs}} \approx 0], \quad \text{with } \mu > 0.$$
Then
$$\mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial tp}\right] = 2(\epsilon - \mu).$$
Hence the expected gradient is negative for all $\epsilon < \mu$, including the negative space. The only stationary point of the expected dynamics is $\epsilon = \mu > 0$, which lies strictly on the positive side. Zero is therefore not a locally stable fixed point under MSE; stochastic gradient updates induce a systematic drift that transports dry predictions across the boundary into weakly positive values.

As a consequence, dry predictions do not concentrate at a stable negative value but instead occupy a diffuse region centered on zero, extending into both the negative and weakly positive ranges. The interval just above zero therefore

contains a superposition of displaced dry cases and genuine drizzle events. This overlap reduces representational separability and compresses the effective dynamic range available to encode variability within the light-precipitation regime.

By enforcing non-negativity through a ReLU constraint during training, negative pre-activations are projected to zero before loss evaluation. As a result, dry samples no longer generate corrective gradients within the negative space. Zero becomes a hard boundary rather than a distributional equilibrium, and the dry regime collapses deterministically onto this boundary point. The positive axis is therefore freed to encode light-rain variability without contamination from non-physical corrective gradients.

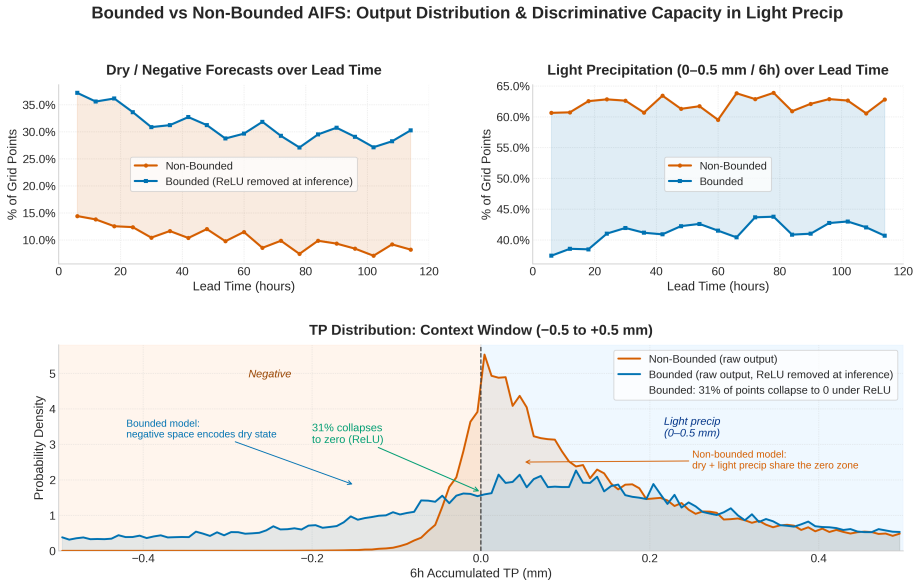

Figure 2: Output distribution and discriminative capacity in the light-precipitation regime for bounded and non-bounded AIFS. The bounded model's ReLU is removed at inference to expose raw pre-activations. **(Top left)** The non-bounded model produces dry or negative outputs at only ∼10 % of grid points versus ∼30 % for the bounded model. A persistent 20-percentage-point gap across all lead times. **(Top right)** The non-bounded model assigns ∼60 % of grid points to the light-precipitation bin (0-0.5 mm / 6h) versus ∼40 % for the bounded model, an excess whose magnitude mirrors the dry-detection deficit almost exactly. **(Bottom)** Pre-activation density near zero. The non-bounded model concentrates dry and drizzle cases in an indistinguishable spike around zero; the bounded model distributes dry-state density broadly across the negative space, with 31 % of pre-activations collapsing cleanly to zero under ReLU at inference.

Figure 2 allows the gradient-ambiguity argument to be verified quantitatively. The three panels form a closed chain of evidence. The non-bounded

model produces dry or negative outputs at only ∼10 % of grid points, compared to ∼30 % for the bounded model. The top-right panel shows that the non-bounded model's light-precipitation frequency is inflated by almost exactly the same ∼20 percentage points. The non-bounded model is not detecting more drizzle; it is misclassifying displaced dry events as light rain. The bottom panel reveals the mechanism predicted by the expected-gradient analysis. The non-bounded model produces a narrow spike of density straddling zero, within which the dry and drizzle regimes are superimposed and statistically indistinguishable. The distribution is tightly concentrated near zero but exhibits a slight positive skew, consistent with the theoretical result that the local stationary point of the expected MSE gradient lies at a strictly positive value. In other words, the model attempts to encode dry states in the neighbourhood of zero, yet the systematic upward drift induced by $\mathbb{E}[\partial\mathcal{L}/\partial tp] < 0$ for $tp < \mu$ prevents zero from acting as a stable attractor. The consequence is a persistent displacement of dry samples into weakly positive values, producing the observed excess of light precipitation.

Although Figure 2 illustrates a single 5-day forecast, the behavior is systematic rather than case-specific. This interpretation is reinforced by the Frequency Bias Index (FBI) and Peirce Skill Score (PSS) shown in Figure 3 of the main article. The non-bounded configuration exhibits a pronounced positive frequency bias in the light-precipitation category, together with degraded discrimination skill, consistent with systematic misclassification of dry grid points as drizzle.

The mechanism described here provides a refined interpretation of recent findings in AI-driven precipitation forecasting. Sha et al. (2025) reported that drizzle bias is substantially reduced when physical constraints are applied, whereas terrain-following coordinates alone do not mitigate drizzle bias but instead improve extreme precipitation forecasts. Notably, their constraint framework combines global conservation principles with an explicit non-negativity correction.

The present analysis isolates the role of non-negativity enforcement and demonstrates that it addresses a fundamental gradient asymmetry at the zero-precipitation boundary. This mechanism operates at the level of local optimization dynamics and provides a distinct, mechanistically interpretable pathway for drizzle reduction. While Sha et al. (2025) demonstrate effectiveness of combining non-negativity with global conservation constraints, our analysis suggests that non-negativity merits investigation as an independent design element. The relative contributions of boundary enforcement versus conservation-based regularization, and their potential architecture dependence, remain important questions for future work.

Major issue 2 (Performance Attribution: Confounded Experimental Design):

> Lines 88-91 state that the revised training schedule (direct fine-tuning on operational analysis rather than sequential ERA5→ERA5→operational) "results in better forecast performance." This causal claim cannot be substantiated from the presented evidence. The revised model simul-

taneously modifies training schedule, adds new prognostic variables, implements bounding layers, and adjusts learning rate schedules. Section 4.1 attempts to isolate bounding layer effects through Figure 11, but this comparison still includes differences in training data extent (1979-2022 vs shorter periods) and other modifications.

The authors should either remove unsupported causal claims or clearly state that performance improvements result from combined system modifications. Revising "results in better forecast performance" to "is associated with improved forecast performance" would be appropriate. I recognize comprehensive ablation studies are computationally expensive, but making causal claims without supporting evidence is scientifically inappropriate. At minimum, explicitly acknowledging the confounded nature of these comparisons would improve scientific rigor.

Response: We thank the reviewer for raising this point. We agree that the simultaneous modification of multiple system components precludes a clean causal attribution of the total performance gain to any single factor, and we have updated the manuscript language accordingly, removing the direct causal claim that the revised training schedule "results in" better performance and replacing it with the reviewer's suggested formulation that these changes are "associated with" improved forecast skill.

To go further and directly address the attribution question, we have added a new controlled experiment to Section 4 (Figure 5 in the revised manuscript). This three-way comparison evaluates: (1) the full AIFS revised system, (2) an AIFS revised model trained with limited data matching the previous version's extent (ERA5 up to 2020, rollout fine-tuning on 2019–2020 only), and (3) the previous AIFS version. The close agreement in ACC between configurations (2) and (3) demonstrates that the dominant portion of the overall performance gain is attributable to the expanded training dataset (ERA5 extended to 2022, rollout fine-tuning expanded from 2019–2020 to 2016–2022), with a smaller residual contribution from the other system modifications.

We note, however, that the close ACC agreement between configurations (2) and (3) should not be interpreted as equivalent forecast quality. The Z500 power spectra (Figure 6 in the revised manuscript) show that AIFS revised (2020) exhibits less mesoscale smoothing than the previous AIFS despite comparable ACC, indicating that the changes do contribute positively to forecast quality in ways not fully captured by ACC alone.

Finally, the comparison between "AIFS revised" and "AIFS revised (no bounding)" in Section 4.1 remains a fully controlled isolate of the bounding layer's effect, as these two configurations differ solely in the presence of the ReLU constraint on tp, with identical training data, architecture, and all other settings.

Major issue 3 (Subjective Design Tradeoffs: Insufficient Quantification):

> Lines 254-256 and 400-410 acknowledge a "subjective compromise between forecast realism and forecast skill measured by RMSE," where more aggressive rollout fine-tuning could improve headline scores at the cost of spatial field characteristics. The authors base this decision on spectral analysis not presented in the manuscript. What specific spectral characteristics were prioritized? What magnitude of RMSE degradation was accepted to achieve desired spectral properties? Without quantification, readers cannot assess the appropriateness of this tradeoff or replicate the training procedure.
>
> The spectral analysis underlying this design choice should be presented. Representative power spectra comparing aggressive fine-tuning versus the chosen configuration would clarify the tradeoff. Quantifying the approximate magnitude of this compromise requires no additional training runs—only analysis of existing model outputs. This documentation is essential for reproducibility.

Response:

We thank the reviewer for this comment. We agree that the original wording in the manuscript suggested a stronger and more subjective trade-off than was intended. We have revised the discussion accordingly to remove the term "subjective compromise" and clarify the rationale behind the training configuration.

First, no explicit optimisation of spectral characteristics was performed during model development. The training configuration, including the maximum rollout length of 12, was retained from the previous AIFS version to ensure methodological consistency. This parameter is known to influence the degree of smoothing, but it was not tuned in this work to prioritise specific spectral properties.

To address the reviewer's request for documentation, we have now added Z500 power spectra to the manuscript (see Figure 6 in the revised manuscript). These spectra compare the revised AIFS, the previous AIFS version, and the IFS analysis for JJA 2023 across multiple lead times. The results show that the revised model exhibits spectral characteristics that are broadly comparable to the previous AIFS across scales, including the 500 km range (zonal wavenumbers 70–90), with slightly improved agreement with the analysis at longer lead times.

Importantly, these comparable spectral characteristics are obtained alongside overall improvements in RMSE-based skill, as shown in the scorecard (Figure 7 in the revised manuscript). Thus, the skill gains are not associated with degraded spatial variability. We have revised the discussion in Sections 4 and 5 to clarify this aspect.

Major issue 4 (Physical Conservation Properties: Missing Diagnostics):

> Lines 188-190 acknowledge that the bounding strategy does not enforce mass or energy conservation, dismissing this with "we did not

> consider other constraints such as energy or mass conservation." For a production forecasting system deployed operationally, this warrants more thorough treatment. Do 10-day integrations accumulate substantial mass or energy errors? How do these compare to the physics-based IFS? Does violation magnitude correlate with forecast error? These questions can be addressed through straightforward post-hoc analysis requiring no model retraining. Brief discussion of whether conservation violations matter for medium-range prediction timescales would strengthen the manuscript.

Response:

For medium-range timescales (up to 10 days), we believe that conservation violations are unlikely to materially affect forecast skill, as the dominant error source at these lead times is chaotic error growth rather than systematic drift. However, we acknowledge that this is an assumption rather than a demonstrated result. Conservation enforcement is expected to become increasingly important as AIFS is extended toward longer integration timescales, where small systematic imbalances can accumulate and affect the model's climatological equilibrium.

We have added a brief discussion of this point in Section 5 of the revised manuscript, noting that conservation constraints represent an important direction for future work.

Major issue 5 (Missing Contextualization with Related Work):

> The manuscript does not adequately position this work within the broader context of physically-constrained machine learning weather models. Multiple operational systems now implement similar physical constraints. The CREDIT framework (Chen et al., 2025, Nature Communications Climate, https://www.nature.com/articles/s41612-025-01125-6; Chen et al., 2025, Geophysical Research Letters, https://agupubs.onlinelibrary.wiley.com/doi implements comparable bounding strategies for physical constraints in operational settings. Harder et al. (2024) provides theoretical framework for hard-constraint approaches. Kent et al. (2025) and Bonev et al. (2025) present alternative constraint methodologies.
>
> Particularly relevant is recent work by Sha et al. (2025, Geophysical Research Letters, https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2025GL118478) addressing the identical problem of blurry precipitation forecasts and drizzle bias in AIWP models. While AIFS uses ReLU activation functions for bounding, Sha et al. demonstrate that terrain-following coordinates combined with global mass and energy conservation constraints provide comparable benefits for reducing drizzle bias and improving extreme precipitation forecasts. This represents an alternative technical approach to the same fundamental challenge, suggesting multiple pathways exist for addressing sparse variable prediction in ML weather models.

> The manuscript would benefit from discussing how this implementation compares to related approaches in GraphCast, Pangu-Weather, FuXi, and CREDIT models. Do other systems exhibit similar light precipitation biases? How do different architectural choices and constraint methodologies affect this common challenge? Comparing the ReLU bounding approach to alternatives like terrain-following coordinates or other constraint formulations would strengthen the contribution by clarifying the relative advantages and positioning the work within the field.

Response: We agree with the reviewer that the original manuscript did not adequately position the bounding strategy within the broader context of physically constrained ML weather models. We have revised the manuscript to address this.

In Section 3, we have added a new contextualisation paragraph (before the description of our bounding approach) that discusses the CREDIT platform Schreck et al. (2025) and recent work by Sha et al. (2025), who implemented global mass and energy conservation constraints within FuXi and demonstrated a reduction of drizzle bias, as well as a companion study showing that terrain-following coordinates improve extreme precipitation forecasts. We also cite the theoretical framework of Harder et al. (2024) for hard-constraint approaches, and the alternative constraint methodologies of Kent et al. (2025) and Bonev et al. (2025).

In the Discussion (Section 5), we have expanded the text to explicitly note that positive frequency bias in the drizzle regime is a recurring feature across architecturally diverse systems, including GraphCast, Pangu-Weather, FuXi, and CREDIT, all of which are trained with symmetric regression losses on non-negative, intermittent variables. The drizzle problem therefore appears largely independent of backbone architecture and is instead tied to how precipitation is parameterised and constrained during training. This framing directly addresses the reviewer's question about whether other systems exhibit similar biases.

Regarding the comparison between our ReLU bounding approach and the conservation-based constraints of Sha et al. (2025): our mechanistic analysis (detailed in the response to Major Issue 1 and incorporated into the revised manuscript in Section 4.1) isolates the specific role of non-negativity enforcement in resolving gradient ambiguity at the zero-precipitation boundary. While Sha et al.'s constraint framework combines conservation principles with non-negativity corrections, our analysis suggests that non-negativity alone addresses a fundamental optimisation instability. This distinction clarifies the complementary nature of these approaches rather than positioning them as competing alternatives.

We note, however, that this manuscript documents the development and evaluation of an operational forecasting system rather than providing a systematic review or benchmark of all constraint methodologies. A comprehensive comparison of bounding strategies, conservation schemes, terrain-following coordinates, and alternative loss formulations across different model architectures

would be a valuable contribution but lies beyond the scope of this work. We believe the revised manuscript now appropriately situates our approach within the field while maintaining focus on the operational system description.

Minor issue 1 (Loss Weight Justification):

> Loss scaling factors in Table 1 (lines 126-133) are described as "chosen empirically," which is reasonable. However, line 129 states that vertical velocity and soil moisture are "deliberately down-weighted," implying specific motivation. Brief rationale would improve clarity.

Response: Vertical velocity is down-weighted due to known accuracy limitations in ERA5, particularly in convective regions. Soil moisture receives reduced weight for similar reasons, and additionally because the transition from ERA5-based pretraining to operational IFS analysis during fine-tuning introduces distributional inconsistencies; down-weighting mitigates the influence of this mismatch on training.

Minor issue 2 (Unsupported Generalizations):

> Line 173-175 claims "well-known characteristic of machine learning-based forecasts: a tendency to produce overly smooth spatial fields." Either provide specific citations (e.g., Ben Bouallègue et al., 2024; Bonavita, 2024) or remove "well-known."

Response:
These references have been provided in the revised version of the manuscript.

Minor issue 3 (Cyclone Performance Statement):

> Line 253 states tropical cyclone performance "is similar to that of the previous version" without supporting data.

Response: We have conducted an internal analysis (see the figure below) of tropical cyclone track and intensity errors for both model versions, covering the 2023 season. The results, shown in the supplementary figures provided to the reviewer, confirm that the two configurations perform comparably across lead times, with a marginal advantage for AIFS revised in both track and intensity metrics. Given that tropical cyclone verification is not a primary focus of this manuscript and the result does not materially alter its conclusions, we have chosen not to include these figures in the paper for brevity.

Minor issue 4 (Statistical Significance):

> Statistical Significance: Several PSS comparisons in Figure 3 between revised AIFS and IFS at 144-hour forecasts show marginal differences. Claims of "improvement" should be verified as statistically significant.

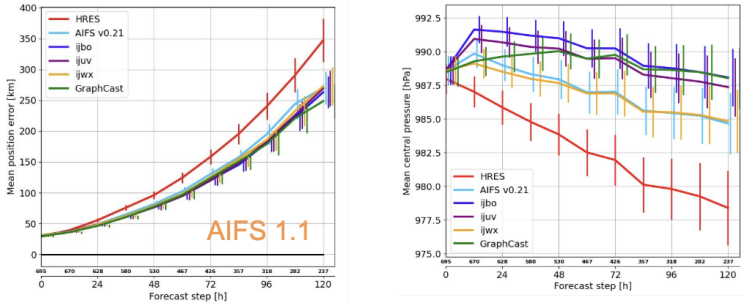

Figure 3: Tropical cyclone verification for the 2023 season: track error (left) and intensity error (right) as a function of lead time for AIFS revised (orange) and AIFS previous (light blue). Results are provided to the reviewer for reference and are not included in the revised manuscript.

Response: We have updated Figure 3 in the revised manuscript to include statistical significance information. For each threshold, we performed a paired Wilcoxon signed-rank test on the daily score values over the full 2023 verification period, testing whether the difference between each model and the previous AIFS version is statistically significant at the $p < 0.05$ level. Filled markers indicate significant differences; open markers indicate non-significant ones. This allows the reader to directly assess which improvements are robust and which fall within sampling variability. The revised figure caption describes this convention.

Minor issue 5 (Case Study Limitations):

> Section 4.2 case studies effectively demonstrate model capability but represent illustrative examples rather than systematic validation. How many extreme events were evaluated? What is the false alarm rate?

Response: The case studies presented are intended to be illustrative examples of model capability rather than a thorough or systematic evaluation. We acknowledge this limitation and encourage more extensive evaluation, including analysis of false alarm rates and a broader range of extreme events, as an important direction for future work.

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, $\mathrm{FBI} = (H + FA)/(H + M)$, 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 $\mathrm{PSS} = H/(H + M) - FA/(FA + CN)$, 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.

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.