

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:

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

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

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

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:

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

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

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

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 < η ≤ 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.

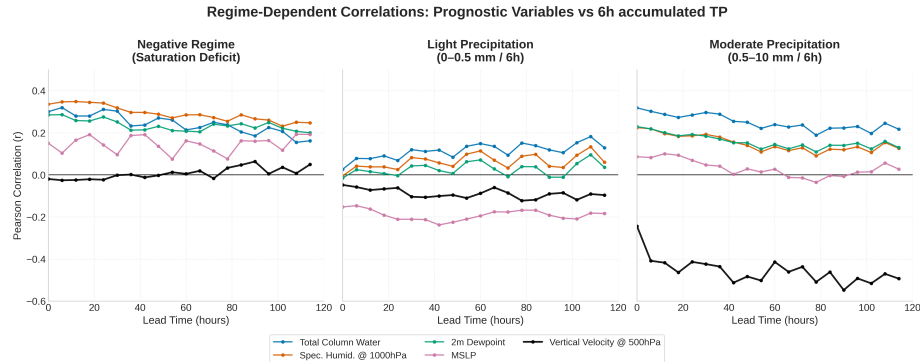


Figure 1: Regime-dependent correlations of pre-activation η (AIFS Revised, for a forecast issued the June 1, 2023 at 00 UTC). Pearson r between η and physical drivers across three regimes: (Left) Negative space ($\eta < 0$): high correlation with moisture variables (q_{1000} , TCW) identifies η as a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

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

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:

$$\frac{\partial \mathcal{L}}{\partial tp} = 2(0.25 - 0.45) = -0.4 \quad (\text{Push Up}) \quad (6)$$

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:

$$\frac{\partial \mathcal{L}}{\partial tp} = 2(0.25 - 0.05) = +0.4 \quad (\text{Push Down}) \quad (7)$$

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 $|\epsilon| \ll 1$,

$$\frac{\partial \mathcal{L}}{\partial tp} = 2(\epsilon - tp_{obs}).$$

In the neighbourhood of zero, the target distribution is one-sided: $tp_{obs} \geq 0$, with strictly positive drizzle values arbitrarily close to zero but no negative observations. Let

$$\mu = \mathbb{E}[tp_{obs} \mid tp_{obs} \approx 0], \quad \text{with } \mu > 0.$$

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.

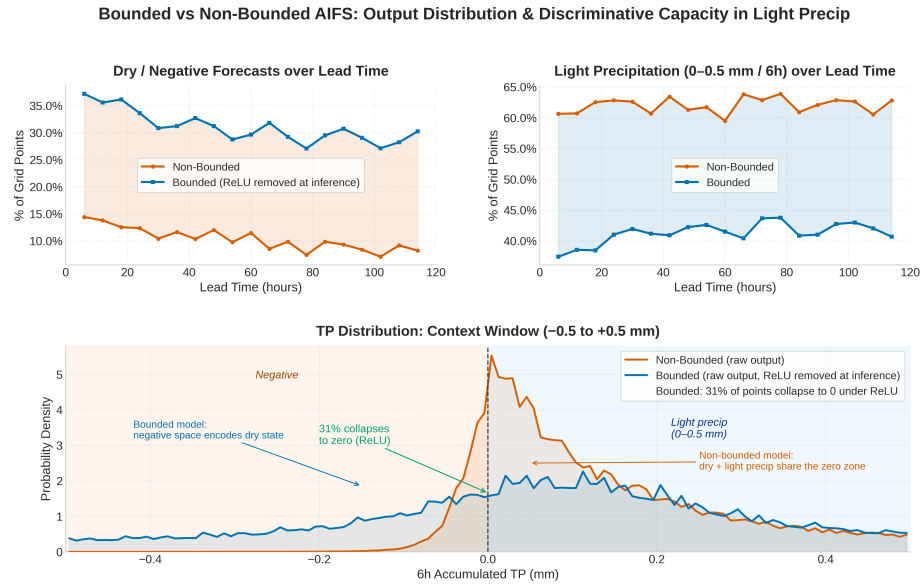


Figure 2: Output distribution and discriminative capacity in the light-precipitation regime for bounded and non-bounded AIFS. The bounded model’s ReLU is removed at inference to expose raw pre-activations. **(Top left)** The non-bounded model produces dry or negative outputs at only $\sim 10\%$ of grid points versus $\sim 30\%$ for the bounded model. A persistent 20-percentage-point gap across all lead times. **(Top right)** The non-bounded model assigns $\sim 60\%$ of grid points to the light-precipitation bin (0-0.5 mm / 6h) versus $\sim 40\%$ 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 $\sim 10\%$ of grid points, compared to $\sim 30\%$ 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/10.1029/2025GL118478>) 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.

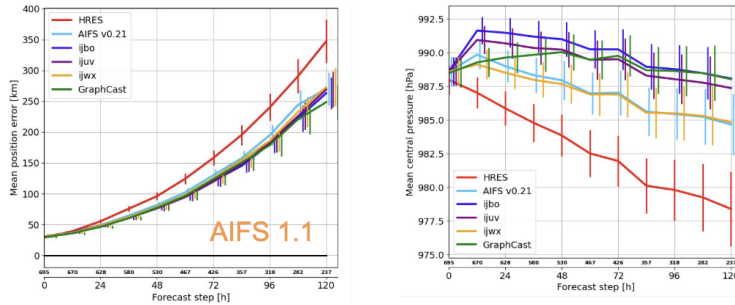


Figure 3: Tropical cyclone verification for the 2023 season: track error (left) and intensity error (right) as a function of lead time for AIFS revised (orange) and AIFS previous (light blue). Results are provided to the reviewer for reference and are not included in the revised manuscript.

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.