Response to Reviewer#2 for the paper: A NEURAL NETWORK-BASED OBSERVATION OPERATOR FOR WEATHER RADAR DATA ASSIMILATION

Marco Stefanelli[1], **Žiga Zaplotnik**[2,1], **Gregor Skok**[1]

[1]University of Ljubljana, Faculty of Mathematics and Physics,
Jadranska Cesta 19, 1000 Ljubljana, Slovenia
[2]European Centre for Medium-Range Weather Forecasts,
Robert-Schuman-Platz 3, 53175 Bonn, Germany

Corresponding author: Marco Stefanelli
marco.stefanelli@fmf.uni-lj.si

We want to thank the Reviewer for a thorough review and the valuable comments and suggestions to improve the manuscript's quality. As a result, notable improvements were made to the manuscript.

Below, we provide replies (in blue) to the specific points raised by the Reviewer (in black).

1. Summary evaluation

I was intrigued by the premise and results of the manuscript EGUSphere-2026-77v1 entitled "A NEURAL NETWORK-BASED OBSERVATION OPERATOR FOR WEATHER RADAR DATA ASSIMILATION" by Stefanelli et al.. The authors have used a very unusual approach to obtain expected radar observations given model fields, a necessary first step towards radar data assimilation.

Instead of trying to simulate the radar reflectivity from the model fields, they devised an approach that learned how to reproduce what an actual radar observes from model fields by training a machine-learning based machine to do so. They then demonstrated the use of the machine by showing that radar data could be used to find new model states that would reduce the mismatch between simulated and observed radar data.
That stated, I believe the authors focused too much of their attention on studying the more expected results of their work while not critically analyzing the much more interesting and novel ones.

We thank the Reviewer for their insightful comment and for recognizing the novelty of the proposed framework. We would like to clarify that our methodology consists of training a neural-network-based observation operator, denoted here as $\mathcal{H}$, to simulate the radar reflectivity, which is compared to the corresponding observed radar reflectivity in the process of data assimilation. The network is trained on paired samples of model states and radar observations and learns the conditional relationship

$$\hat{\mathbf{y}} = \mathcal{H}(\mathbf{x}),$$

where $\mathbf{x}$ collects the model predictors t, u, v, and r at 975, 925, 850, and 800 hPa, together with msl, $t2m$, and $r2m$ and $\hat{\mathbf{y}}$ is a model equivalent of observation vector $\mathbf{y}$. In this sense, the machine-learning model serves as an explicit (data-driven) observation operator instead of an explicit microphysics-to-reflectivity forward model, while still providing reflectivity in observation space consistent with what the radar observes. Once trained, $\mathcal{H}$ is used within our

data assimilation framework (3DVar) to evaluate the mismatch between measured and modeled reflectivity (the innovation) and to guide the search for the optimal model state (the analysis) that reduces this mismatch while respecting the assumed obseration error variances and background-error covariances. Importantly, the role of $\mathcal{H}$ is analogous to that of a conventional observation operator in data assimilation: it provides the mapping from model space to observation space required to compare with radar data and compute the corresponding innovations.

We appreciate the Reviewer's perspective and agree that a deeper analysis of the more novel outcomes is valuable. However, we would like to clarify the primary motivation and structure of the current manuscript. The presented study is intended as a proof-of-concept rather than a system ready for immediate implementation. Nevertheless, future work will focus on refining the methodology to bring it closer to operational testing and deployment.

Because the proposed methodology is entirely new, several results that might seem expected within the context of traditional variational data assimilation, such as a consistent reduction of the reflectivity misfit when using the learned operator within the 3DVar workflow, were not guaranteed in our framework. In particular, obtaining consistent reflectivity by applying the trained observation operator on analysis states is a non-trivial result. While the network is trained to emulate the radar observations from background model fields, there is no guarantee that it will remain stable and beneficial when applied to unseen model states. In other words, the analysis can populate regions of the model-state space that differ from those most frequently represented in the training data. In such regimes, the learned mapping could, in principle, produce spurious reflectivity responses. Demonstrating that the observation operator remains well-behaved under such inputs, and that the resulting innovations lead to a systematic reduction of the reflectivity misfit, is therefore an essential component of our proof-of-concept study. Consequently, the paper is intentionally framed to demonstrate that a neural-network-based observation operator can be successfully constructed and utilised to enable a functional end-to-end assimilation experiment.

Furthermore, following the Reviewer's suggestion, we have complemented the case-study analysis by including additional examples of $\mathcal{H}$ reproducing observed fields, alongside a long-term statistical evaluation (detailed later in this document). In the revised manuscript, we have more explicitly defined the study's objectives and ensured that the long-term statistics are better integrated.

The norm in observation operators H() is to only use model variables x to devise H. You chose the unusual path to use a combination of observations y and model variables x to determine H.

Indeed, we used model state variables $\mathbf{x}$ at analysis time as input and radar observation $\mathbf{y}$ as target to learn the neural-network based mapping function $\mathcal{H}$. The rationale behind this approach was to use the most accurrate estimates of the true atmospheric conditions for both the input as well as the target output. To do this, our input describes the model state based on variables (t, q, u, v, $r2m$, $t2m$, $mslp$) from ALADIN analyses at 4.4 km resolution. The choice of analyses was obvious: it is the best estimate of the current atmospheric state, and no available observing system (or a combination of such systems) provides a comprehensive estimate of the instantaneous full atmospheric state. For the target output, we opted for observed radar reflectivities as the best estimate of the current precipitation field, rather than simulated radar reflectivities. We have added this description in lines 199–205 in the revised manuscript.

It is important to note that our goal was not merely to emulate existing observation operators that simulate radar reflectivity from both prognostic model variables (e.g. liquid water content, ice content) and diagnostic model variables (such as hydrometeors). The performance of such emulator would be inherently bounded by the quality of the traditional observation operator it mimics. Instead, our ultimate objective is to develop an even better operator capable of capturing processes that are not fully represented by traditional physical observation operators.

This study documents our attempt with a modest convolutional neural network and a limited number of vertical levels.

Furthermore, as mentioned in the previous response, we intentionally trained the operator on dynamic and thermodynamic prognostic variables. This allows the observed radar reflectivities to directly influence these model variables during assimilation, potentially improving the initial conditions and ensuring a more persistent impact on model predictability.
This strategy was specifically designed with the aim to infer basic model variables from complex observations (in our case radar reflectivity) in scenarios where the precise relationship between basic model variables and the quantity that is observed is either unknown exactly or cannot be analytically described in a comprehensive manner. By leveraging the ability of neural networks to approximate high-dimensional, non-linear mappings, we bridge the gap between the model's prognostic core and the observed physical phenomena.

This has a set of advantages (fewer assumptions on microphysics and scattering, natural ability to simulate many radar artifacts...) and pose additional challenges (decoupling between the real world shaping observations and the simulated world shaping model fields due to initial condition and model errors) compared to the traditional approach. And while a few of the advantages were mentioned in the introduction, none of the challenges were, and they were not reflected upon after the Introduction.

We agree that the manuscript did not sufficiently discuss the specific limitations of the proposed NN-based observation operator. Although observations are not used as input variables in $\mathcal{H}$, they are used as target values during training, and this has important implications that deserve to be made explicit.

First, because the network is trained to reproduce the observed signal, the learned operator may capture observational features that are not causally represented in the model state. In particular, the NN may absorb effects related to the observing system, unresolved processes, sampling issues, or radar artifacts that are present in $\mathbf{y}$ but not explicitly encoded in the state input variables $\mathbf{x}$. As a consequence, $\mathcal{H}(\mathbf{x})$ may reproduce observations well without necessarily representing the underlying physical relationship in a fully faithful way. In our methodology, this is partly alleviated by a quality-control and interpolation strategies described in the paper.

Second, the generalization capability of the learned operator may be fragile outside the training distribution. The NN may perform well for meteorological situations and model-error structures similar to those represented in the training dataset, but its behavior may degrade in different regimes. This is particularly relevant in data assimilation, where the background state may differ substantially from the situations encountered during training.

Third, the innovation $\mathbf{y} - \mathcal{H}(\mathbf{x}_b)$ becomes more difficult to interpret physically. In a traditional observation operator, the innovation can often be related more directly to deficiencies in the model state through a physically motivated forward relationship. In the present case, part of the mismatch may instead reflect limitations of the learned mapping itself, including statistical compensation introduced during training, and not only meteorological differences between the model state and reality.

Fourth, while the analysis is considered to represent the best estimate of the full atmospheric state, it is still susceptible to errors associated with DA method, observation errors and background errors. As a consequence, the model state might not be favourable for precipitation formation, while simultaneously, precipitation is observed by radar system. Despite that, we consider these model fields as the best available estimate of the truth and therefore take them as input data to train the neural network.

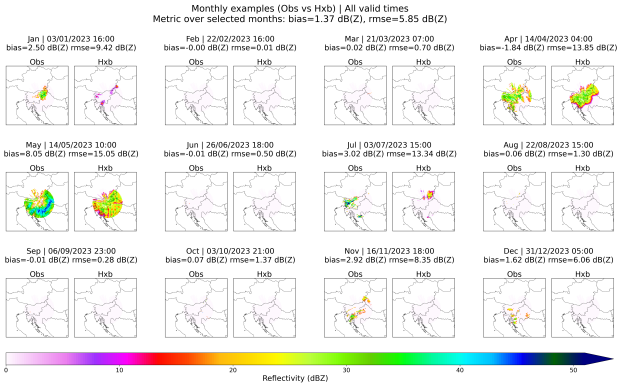
Finally, even if the resulting analysis increments improve the fit to observations, their physical interpretation may remain less transparent. In other words, good assimilation performance does not necessarily imply that the operator is extracting the correct physical information or transferring that information to the model variables in a fully reliable way.

We revised the manuscript in the Discussion and Conclusion section (lines 524–562) to discuss these limitations more explicitly and to clarify that, while the proposed NN-based operator offers clear practical advantages, it also introduces specific challenges in terms of physical interpretability, robustness, and model–observation consistency.

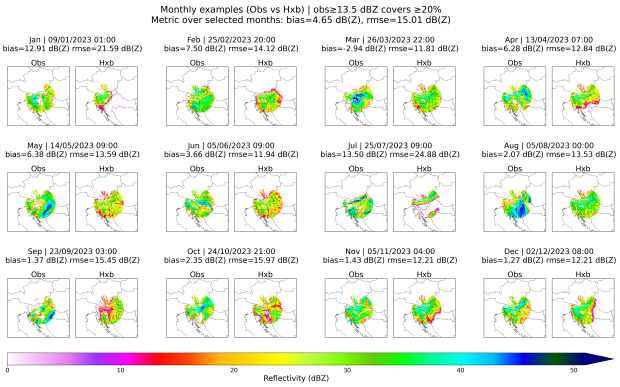
And while the machine was trained with 4 years of data with 2023 being used for testing, I only found results for four radar maps.

In response to the Reviewer's request, we present additional generated-reflectivity results for randomly selected cases from the 2023 testing period. Specifically, we show results for different reflectivity thresholds:

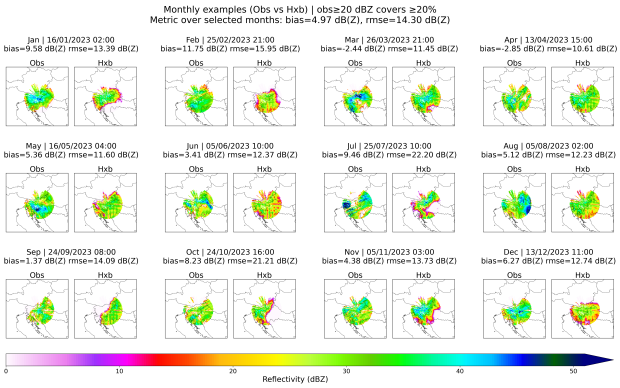
- All valid times. In this way, we also include clear-sky conditions.
- $\geq$ 13.5 dBZ. Above this threshold, we find all precipitation events. To select one timestep, we require at least 20% coverage of the radar-disc domain.
- $\geq$ 20 dBZ. Above this threshold, we find all events from light to heavy precipitation, and the probability of hail is above zero. To select one timestep, we require at least 20% coverage of the radar-disc domain.
- $\geq$ 32 dBZ. Above this threshold, we find all events from medium to heavy precipitation, and the probability of hail is above zero. To select one timestep, we require at least 20% coverage of the radar-disc domain.
- $\geq$ 44 dBZ. Above this threshold, we find all events of heavy precipitation, and the probability of hail is above zero. To select one timestep, we require at least 1% coverage of the radar-disc domain.
- $\geq$ 50 dBZ. Above this threshold, we find all events of heavy precipitation, and the probability of hail is above zero. To select one timestep, we require at least 1% coverage of the radar-disc domain.



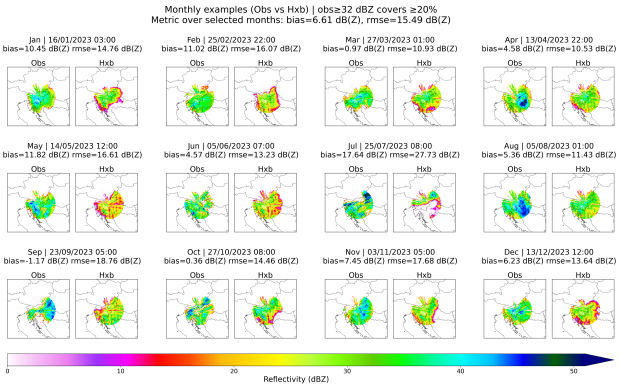
(a) All valid times



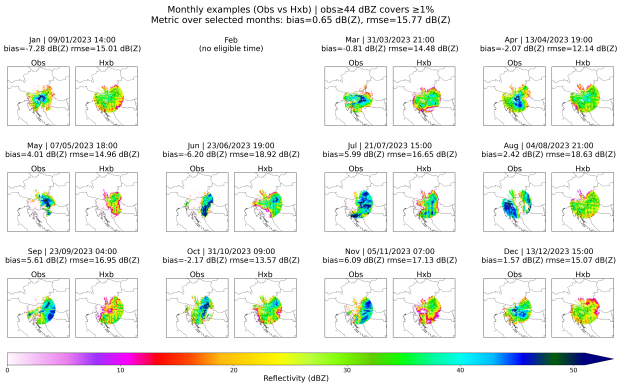
(b) Threshold $\geq$ 13.5 dBZ



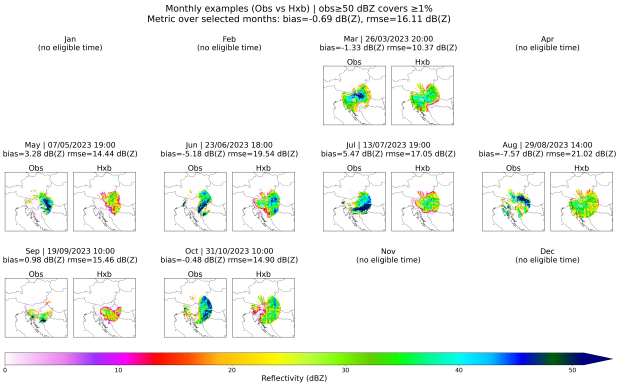
(c) Threshold $\geq$ 20 dBZ



(d) Threshold $\geq$ 32 dBZ



(e) Threshold $\geq$ 44 dBZ



(f) Threshold $\geq$ 50 dBZ

Figure 1: Additional generated-reflectivity results for randomly selected cases from the 2023 testing period, shown for different reflectivity thresholds. Panels show results for thresholds of (a) All valid times, (b) $\geq$ 13.5 dBZ, (c) $\geq$ 20 dBZ, (d) $\geq$ 32 dBZ, (e) $\geq$ 44 dBZ, and (f) $\geq$ 50 dBZ. (b)-(d) requires at least 20% coverage of the radar-disc domain. (e)-(f) requires at least 1% coverage of the radar-disc domain

Critically analyzing the performance of your indirect approach to devising H() should have been given considerably more emphasis, as the follow-up results, and your planned future work, are, in my opinion, expected: Once you have an H function, it is a trivial result that you can use 3-D Var to modify x and find a new one whose H(x) would be a better match to y than the original H(x). Given that the key novelty of your approach now and in the future rests on your ability to find a good H() using a combination of model and radar data, a better critical analysis of its performance would have been expected. Ideally, a comparison with a more traditional H() would be best, but I'd be happy with some longer-term statistics (echo coverage, biases, standard errors, etc.).

Because I am uncertain that these and other changes can be made in the time constraints associated with a conditional acceptance, I will recommend rejection of the current manuscript in its current form. But I encourage you to make the necessary modifications to your manuscript as your approach has the value of being much more original than many others.

We thank the Reviewer for this thoughtful and constructive comment. We agree that the main novelty of the study lies not only in showing that 3D-Var can improve the fit between $\mathcal{H}(\mathbf{x}_b)$ once an observation operator is available, but in assessing how well the proposed approach is able to construct a meaningful and useful radar observation operator. We acknowledge that this aspect was under-emphasized in the original submission.

In response, we have substantially expanded the evaluation of the derived observation operator and added a longer-term statistical analysis of its behavior in mapping model dynamic and thermodinamic variables into reflectivity observation space. Specifically, we now include figures and discussion showing:

1. Cumulative coverage of reflectivity (Fig. 2), and monthly mean echo coverage for observations and $\mathcal{H}(\mathbf{x}_b)$ at reflectivity thresholds of 13.5, 20, 32, 44, and 50 dBZ (Fig. 3) included to evaluate how well the derived observation operator reproduces the occurrence and areal extent of radar echo.
2. Distributions of innovation bias and RMSE for reflectivity thresholds of 13.5, 20, 32, 44, and 50 dBZ, included to provide a longer-term assessment of the variability and robustness of the observation-operator performance (Fig. 4). The figure shows histograms of the time series of domain-wide spatial bias and RMSE, computed on the radar-disc for each threshold-based sample selection. By displaying the full distributions rather than only temporal means, these panels characterize not only the typical error magnitude but also the spread, skewness, and threshold dependence of the innovation statistics.
3. Monthly mean innovation bias and RMSE between observations and $\mathcal{H}(\mathbf{x}_b)$, computed over the radar-disc domain and stratified by reflectivity thresholds (13.5, 20, 32, 44, and 50 dBZ), included to quantify the seasonal evolution errors (Fig. 5).
4. Monthly mean spatial standard deviation of observations and $\mathcal{H}(\mathbf{x}_b)$, included to assess whether the derived observation operator reproduces not only the mean reflectivity signal but also its spatial variability (Fig. 6).
5. Threshold-dependent categorical verification scores, including the probability of detection (POD), critical success index (CSI), false alarm ratio (FAR) (Fig. 7), and fractions skill score (FSS) (Fig. 8), were added to complement the continuous metrics and quantify skill as a function of event intensity. For POD, CSI, and FAR analysis, given the reflectivity threshold, only those timesteps which have at least 1% of the valid radar-disc grid points exceeding that threshold are considered. To the fields in these timesteps the thresholding is applied which produceses the corresponding binary fields from which counts of hits (O), misses (M), and false alarms (F) are computed. These quantities are then used to compute the scores: POD $= O/(O+M)$, CSI $= O/(O+M+F)$, and FAR $= F/(O+F)$. From these binary fields also the FSS is computed using different sizes of spatial neighborhood.

This expanded analysis shows that the trained observation operator captures several first-order features of the observed reflectivity field at low and moderate thresholds, but also exhibits

clear limitations for stronger echoes. In particular, the results indicate that:

- echo coverage decreases with threshold in both observations and $\mathcal{H}(\mathbf{x}_b)$, but $\mathcal{H}(\mathbf{x}_b)$ generally underestimates observed coverage for stronger echoes,
- Bias and RMSE increase with threshold,
- the spatial standard deviation of $\mathcal{H}(\mathbf{x}_b)$ is systematically smaller than that of observations, indicating that $\mathcal{H}(\mathbf{x}_b)$ tends to produce less variability than radar observations,
- skill metrics such as POD, CSI, and FSS decrease substantially at high thresholds while FAR increase.

We believe that these findings provide the critical assessment that the Reviewer requested.

We agree that a comparison with a more traditional radar observation operator would be valuable. However, implementing and validating such a benchmark operator is beyond the scope of the present revision and could not be completed robustly within the available time. We therefore focused on providing the longer-term statistical evaluation suggested by the Reviewer, which we believe is already a substantial and important improvement to the manuscript. In the revised manuscript we included and discussed, in a separate subsection (3.2) titled "Statistical evaluation", Fig. 2, Fig. 7 and Fig. 8 considering that all togeter give the necessary information about the trained observation operator.

We appreciate the Reviewer's encouragement and believe that the revised manuscript now places much greater emphasis on the actual performance of the proposed NN-based observation opeartor construction, which is indeed the core methodological contribution of the work.

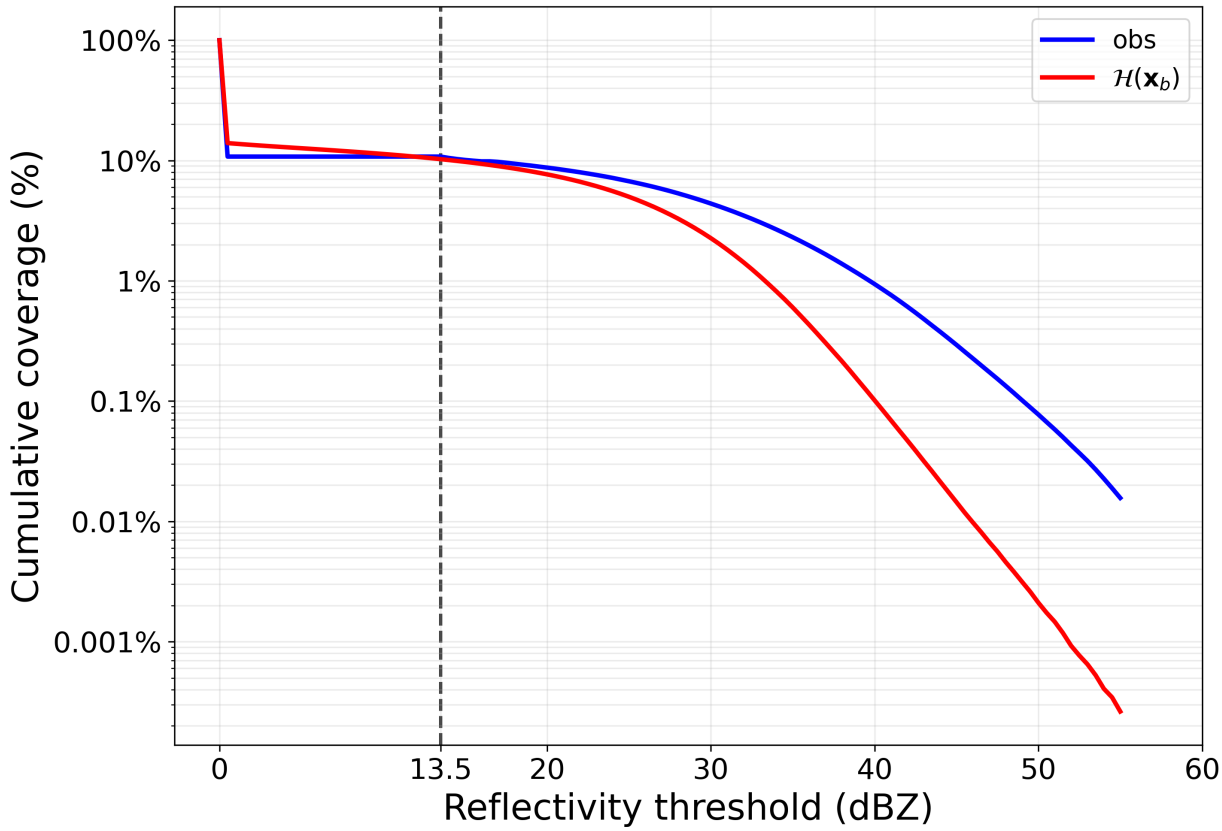


Figure 2: Cumulative coverage of reflectivity for the observations (blue) and $\mathcal{H}(\mathbf{x}_b)$ (red), computed over all radar disc pixels for all timesteps in 2023 shown as a function of reflectivity threshold. The tresholds span the interval [0, 55] dBZ with a step of 0.5 dB(Z). The coverage at a certain threshold is defined as the fraction of pixels (in all timesteps) whose reflectivity is equal or larger than the threshold. Because the metric is cumulative, the curve decreases monotonically with increasing threshold and summarizes the exceedance frequency of progressively stronger echoes. The logarithmic y-axis enhances the visibility of differences in the low-coverage regime and in the upper-reflectivity tail. The dashed vertical line indicates the 13.5 dBZ threshold which is the minimum non-zero reflectivity (used to separate precipiation events from dry and cloud-only events). The figure indicates that the observations and $\mathcal{H}(\mathbf{x}_b)$ have similar cumulative coverage at low reflectivity thresholds, with both curves close to 10% up to 20 dBZ. For larger thresholds, $\mathcal{H}(\mathbf{x}_b)$ decays more rapidly than the observations, revealing a deficit of moderate and intense echoes. The discrepancy becomes marked above 30 dBZ and is strongest in the high-reflectivity tail, where the logarithmic scale emphasizes the substantially lower exceedance frequency in $\mathcal{H}(\mathbf{x}_b)$. This suggests that the observation operator reflectivity captures the lower-reflectivity coverage reasonably well but underestimates the occurrence of stronger radar returns. This is an expected behaviour considering that high intensity events are rare in the dataset and consequently hard to be learned by the trained observation opeartor. However, this behaviour is in line with the purpose of the observation operator whose aim is not to perfectly reproduce the observation but to map the model space into the observation space without correcting the model errors.

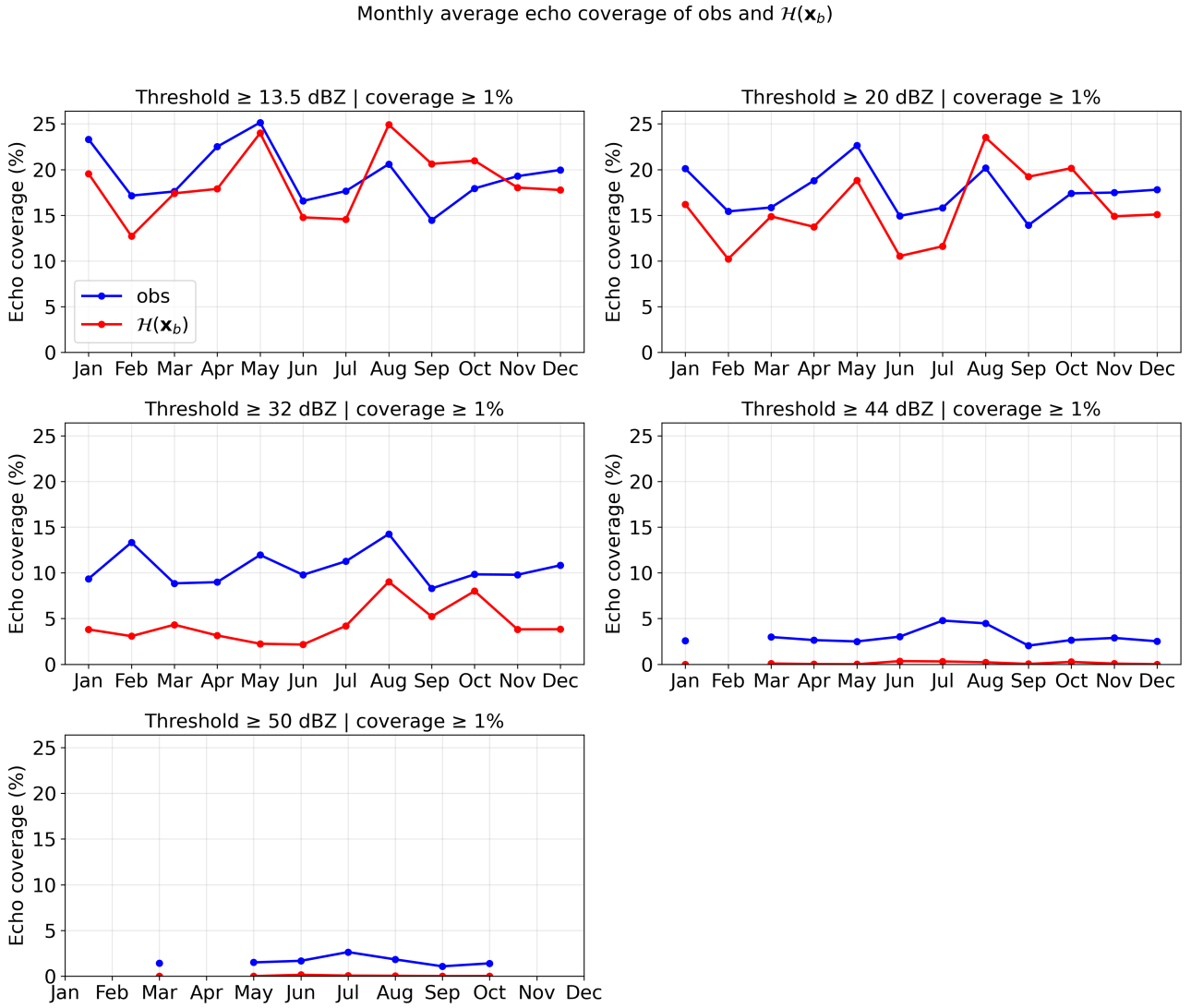


Figure 3: Monthly mean echo coverage (%) for observations (blue) and NN-based observation operator reflectivity $\mathcal{H}(\mathbf{x}_b)$ (red), computed over the radar disc domain for thresholds of 13.5, 20, 32, 44, and 50 dBZ in 2023. For each threshold, the script counts, at every timestep, the number of pixels exceeding the threshold in both observations and model space. A timestep-selection filter is applied before computing the average. A timestep is kept only if the observations or the model reach a minimum echo coverage of 1% of the radar disc. This isolates timesteps with meaningful precipitation echoes. Each panel corresponds to one reflectivity threshold, and monthly values represent the spatial fraction grid points exceeding that threshold, averaged over all radar disc and selected times within each month. A common y-axis scale is used across panels to enable direct comparison of variability and systematic differences between obs and $\mathcal{H}(\mathbf{x}_b)$ as a function of threshold. Overall, echo coverage decreases with increasing threshold, while $\mathcal{H}(\mathbf{x}_b)$ generally underestimates observed coverage at higher reflectivity thresholds, especially for 32 dBZ and above.

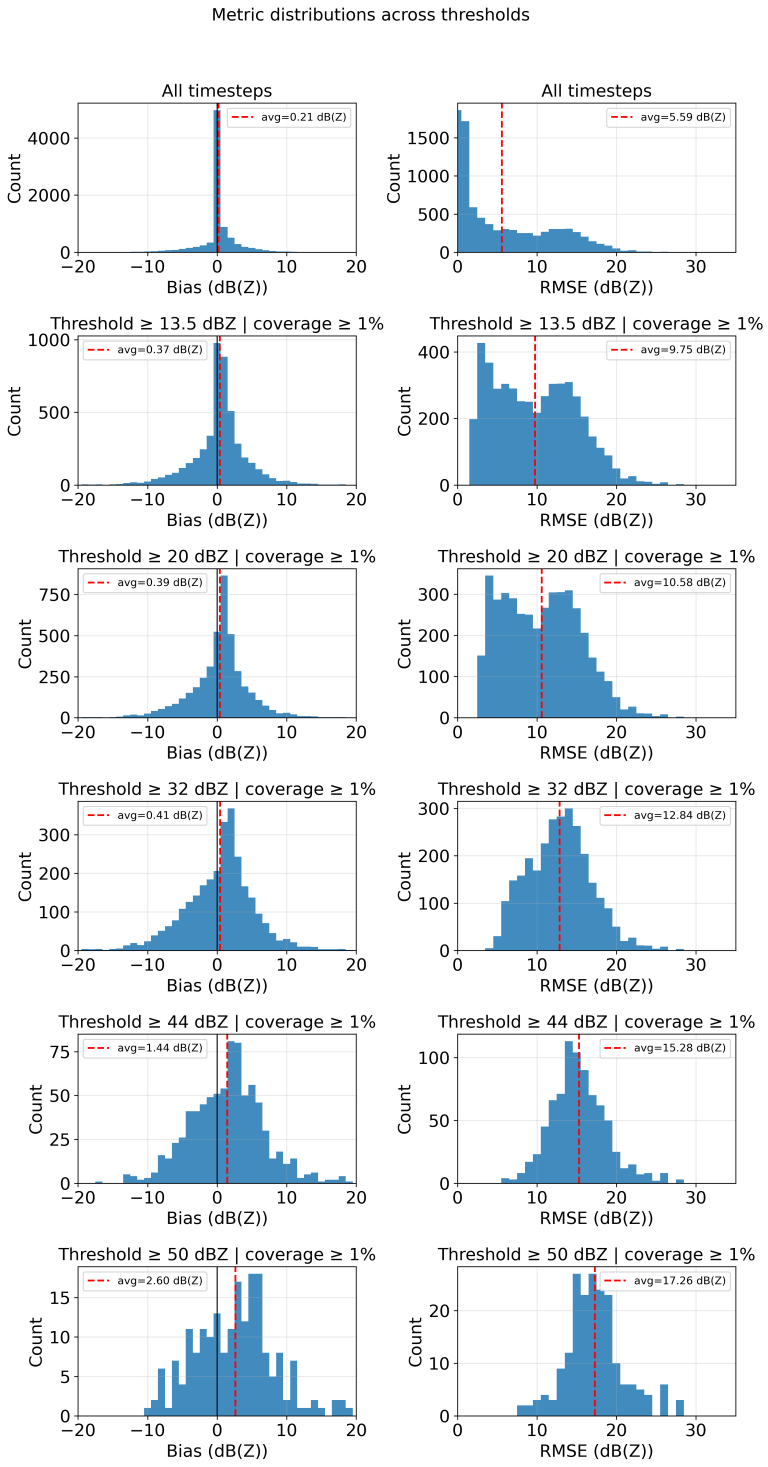


Figure 4: Distributions of spatial innovation bias (left column) and RMSE (right column) computed over the radar disc domain considering all the available timesteps (first row) and for thresholds of 13.5, 20, 32, 44, and 50 dBZ in 2023. For each threshold, the script counts, at every timestep, the number of pixels exceeding the threshold in both observations and model space. A timestep-selection filter is applied before computing the average, a timestep is kept only if the observations or the model reach a minimum echo coverage of 1% of the radar disc. This isolates times with meaningful precipitation echoes. Each row corresponds to one threshold-based sample selection, from all selected times to cases with observed reflectivity exceeding 13.5, 20, 32, 44, and 50 dBZ. For each threshold, histograms of bias and RMSE are plotted, together with vertical lines marking the mean values. The bins are fixed to 1 dB(Z), centered on integer values and identical for all thresholds, which makes the metric distributions directly comparable across rows of the figure. As the reflectivity threshold increases, the sample size decreases. The bias and RMSE distributions shift toward larger values at higher thresholds.

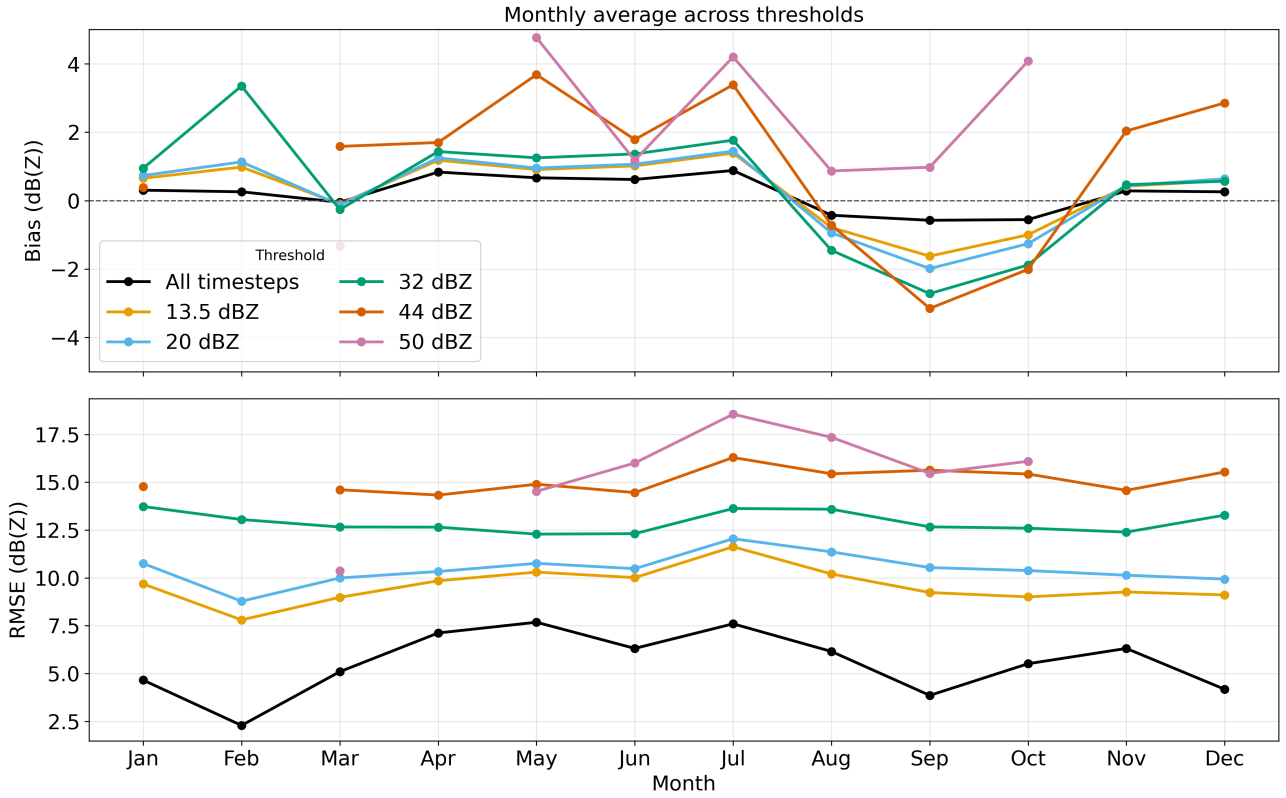


Figure 5: Innovation monthly mean bias (top) and RMSE (bottom) as a function of month computed over the radar disc domain for all the available timesteps and for thresholds of 13.5, 20, 32, 44, and 50 dBZ in 2023. For each threshold, the script counts, at every timestep, the number of pixels exceeding the threshold in both observations and model space. A timestep-selection filter is applied before computing the average, a timestep is kept only if the observations or the model reach a minimum echo coverage of 1% of the radar disc. This isolates times with meaningful precipitation echoes. Colored lines denote the different threshold-based sample selections. The dashed horizontal line in the upper panel marks zero bias. Bias exhibits a clear dependence on both season and threshold, with generally small values for all valid times and larger positive or negative departures for higher reflectivity thresholds. RMSE increases systematically with threshold, indicating larger disagreement between observations and $\mathcal{H}(\mathbf{x}_b)$ for more intense echo regimes, and also shows pronounced month-to-month variability, with enhanced errors during the summer season months.

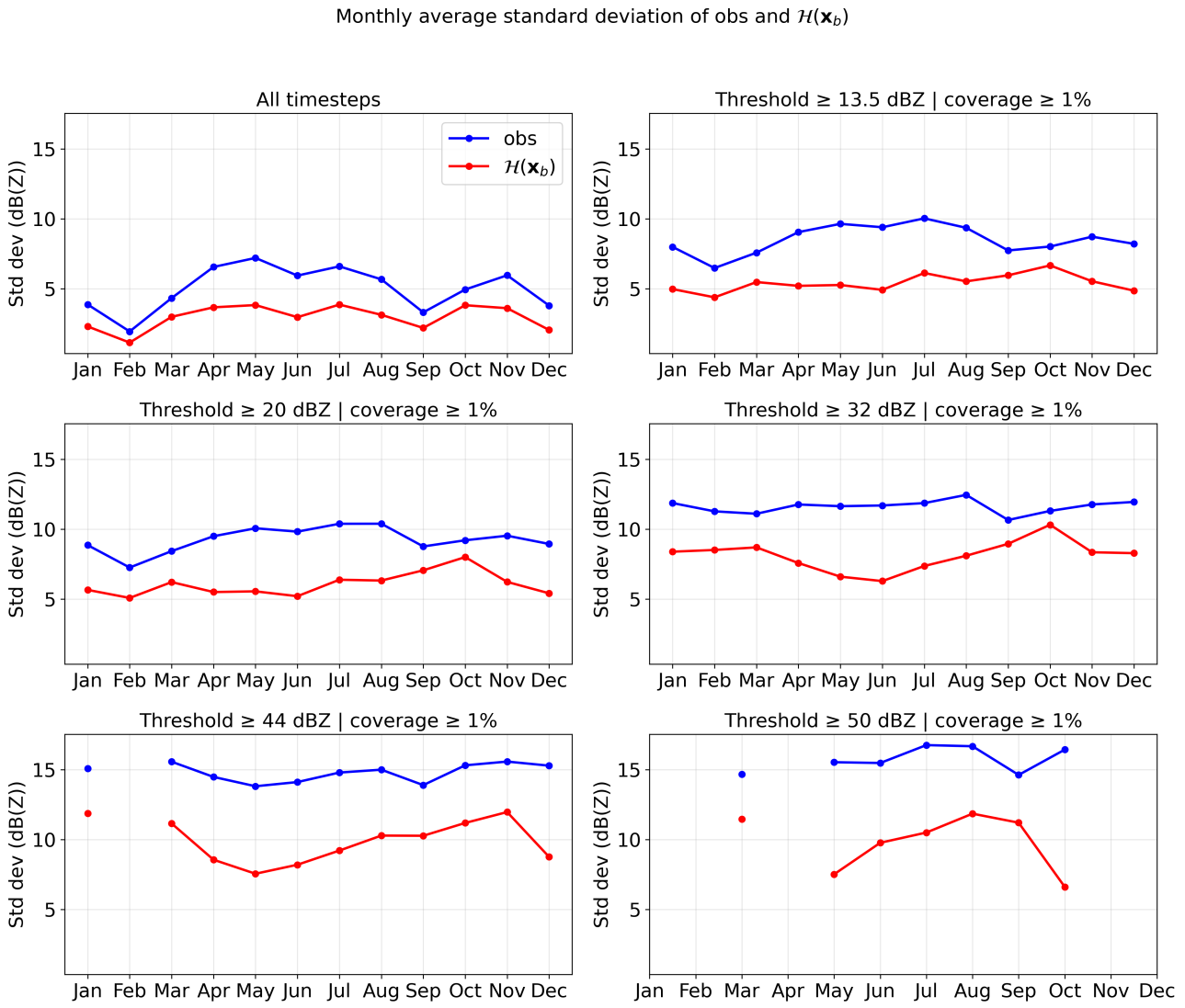


Figure 6: Monthly mean spatial standard deviation for the observations (blue) and $\mathcal{H}(\mathbf{x}_b)$ (red) computed over the radar disc domain for all the available timesteps and for thresholds of 13.5, 20, 32, 44, and 50 dBZ in 2023. For each threshold, the script counts, at every timestep, the number of pixels exceeding the threshold in both observations and model space. A timestep-selection filter is applied before computing the average, a timestep is kept only if the observations or the model reach a minimum echo coverage of 1% of the radar disc. This isolates times with meaningful precipitation echoes. Each panel corresponds to one threshold-based sample selection. A common y-axis scale is used across panels to facilitate comparison of variability across thresholds and seasons. The spatial standard deviation generally increases with threshold, indicating stronger variability in more intense echo regimes. For all thresholds and most months, observations exhibit larger spatial variability than $\mathcal{H}(\mathbf{x}_b)$.

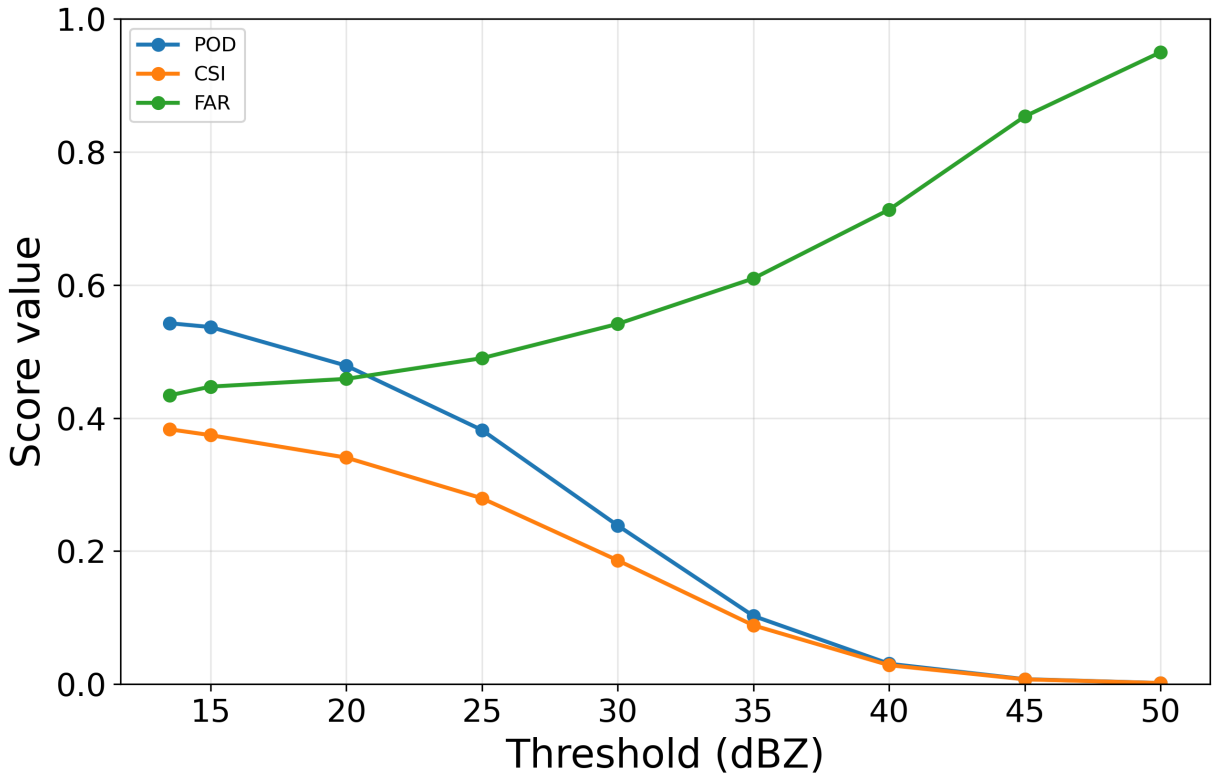


Figure 7: Categorical verification scores as a function of reflectivity threshold for NN-based observation operator $\mathcal{H}(\mathbf{x}_b)$ relative to observations computed over the radar disc domain considering all the available timesteps and for thresholds of 13.5, 15, 20, 25, 30, 35, 40, 45, and 50 dBZ in 2023. For each threshold, the script counts, at every timestep, the number of pixels exceeding the threshold in both observations and model space. A timestep-selection filter is applied before computing the average, a timestep is kept only if the observations or the model reach a minimum echo coverage of 1% of the radar disc. This isolates times with meaningful precipitation echoes. The plotted scores include the probability of detection (POD), critical success index (CSI) and false alarm ratio (FAR). The categorical verification scores show a clear degradation of model skill as the reflectivity threshold increases. POD decreases steadily from about 0.54 at 13.5 dBZ to nearly zero above 40 dBZ, indicating that the model progressively misses a larger fraction of observed high-reflectivity events. CSI follows the same behavior, dropping from about 0.38 at low threshold to almost zero at the highest thresholds, which reflects an overall loss of event agreement between model and observations for intense echoes. In contrast, FAR increases monotonically with threshold, from roughly 0.44 at 13.5 dBZ to about 0.95 at 50 dBZ, showing that most modeled exceedances at high thresholds are not confirmed by observations. Together, these results indicate that the model retains moderate skill for weak-to-moderate reflectivity but performs poorly for strong convective cores, where events become both harder to detect and increasingly dominated by false alarms.

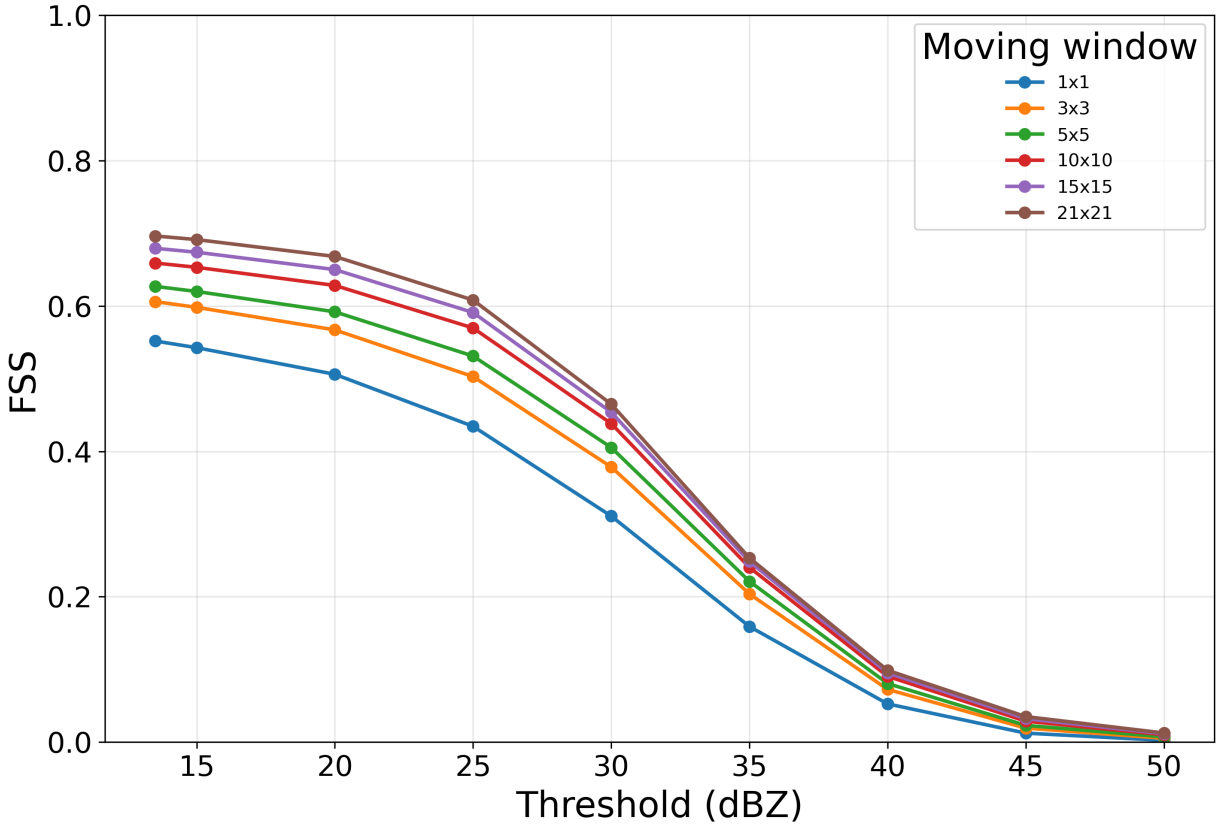


Figure 8: Fractions Skill Score (FSS) versus reflectivity threshold for different square neighbourhood sizes computed over the radar disc domain considering all the available timesteps and for thresholds of 13.5, 15, 20, 25, 30, 35, 40, 45, and 50 dBZ in 2023. For each threshold, the script counts, at every timestep, the number of pixels exceeding the threshold in both observations and model space. A timestep-selection filter is applied before computing the average, a timestep is kept only if the observations or the model reach a minimum echo coverage of 1% of the radar disc. This isolates times with meaningful precipitation echoes. For each threshold, the observed and NN-based observation operator $\mathcal{H}(\mathbf{x}_b)$ reflectivity fields were thresholded into binary event masks, and neighbourhood event fractions were computed using sqare-shaped moving windows of different sizes that ranged from 1 to 21 grid points. The FSS was then calculated from the spatial mismatch between the observed and model fraction fields over all valid radar-covered pixels. The figure shows that FSS decreases monotonically with increasing reflectivity threshold for all neighbourhood sizes, indicating that the spatial agreement between the observations and $\mathcal{H}(\mathbf{x}_b)$ becomes progressively poorer for more intense echoes. At the lowest thresholds (13.5–20 dBZ), the FSS values indicate at least a moderate skill, with clear improvement as the neighbourhood size increases, which indicates that part of the mismatch is associated with small spatial displacements. At intermediate thresholds (30 dBZ), the FSS value drops substantially for all window sizes, revealing reduced skill in reproducing the location and extent of stronger precipitation features. At the highest thresholds (45–50 dBZ), the FSS approaches zero regardless of neighbourhood size, showing that the most intense echoes are only rarely matched in space and time. Overall, the figure suggests that the observation operator reflectivity captures the broader-scale distribution of weak-to-moderate echoes reasonably well, but has limited skill for the most intense convective cores.

2. Specific comments

(i) *The title does not reveal the key novelty, namely that you are devising and evaluating a "measurement-based observation operator" as opposed to the more traditional "simulation-based observation operator" (I find myself having to invent new terminology to express what you have designed; if you can find a better way to express this, please do so!). One could devise a NN-based OO by having it learn to imitate a simulation-based OO, and your current title would apply to such a work; but this is not what you did. It would be better if your title would better express or describe your unusual approach.*

We thank the Reviewer for this helpful suggestion and agree that the current title does not sufficiently convey the key novelty of our study. Following the Reviewer's comment, we propose to revise the manuscript title to:
"A neural-network model-measurement-based observation operator for weather radar reflectivity assimilation."
We believe this revised title more clearly reflects our novel approach and reduces potential ambiguity with studies in which a neural-network operator is trained to solely model simulated fields.

(ii) *Methodology, 2nd sentence: I propose the following edit, if you believe it is still correct: "The NN observation operator is trained to determine the expected reflectivity averaged over the previous hour from fields of temperature (t), horizontal wind components (u and v), relative humidity (r), at four pressure levels (975 hPa, 925 hPa, 850 hPa, and 800 hPa) and three surface variables, 2 m temperature (t2m), 2 m relative humidity (r2m), and mean sea level pressure (msl) from the ALADIN numerical weather prediction model (hereafter, 'ALADIN model outputs' refers to the listed fields)."*

We thank the Reviewer for this helpful suggestion. We agree that the original sentence was not sufficiently clear. We revised it accordingly. However, our target variable is not the reflectivity averaged over the previous hour, but the reflectivity summed over 1-hour intervals, as described in the original manuscript (Section 2.1, page 4, last paragraph): "The quality-controlled radar observations were subsequently summed into 1-hour intervals to match the temporal resolution of the ALADIN model outputs."

To improve clarity while preserving the correct meaning, we revised the sentence in the manuscript as follows:

"The NN observation operator is trained to determine the expected reflectivity summed over the previous hour from short-range forecasts of temperature (t), horizontal wind components (u and v), relative humidity (r), at four pressure levels (975 hPa, 925 hPa, 850 hPa, and 800 hPa) and three surface variables, 2 m temperature ($t2m$), 2 m relative humidity ($r2m$) and mean sea level pressure (msl) from the ALADIN numerical weather prediction model (hereafter 'ALADIN model outputs' refers to the listed fields)"

(iii) *Last paragraph of 2.1, also related to (ii): "The quality-controlled radar observations were subsequently summed into 1-hour intervals to match the temporal resolution of the ALADIN model outputs". First, this choice of using hourly averages or sums of reflectivity is interesting per se, because you could have chosen to use the radar data closest to the hour, and the temporal resolution would still be matched. Why that choice? I personally believe it is a good idea, but I believe it needs to be articulated here. Then, a few more details are required:*

(a) How is the sum done (in dBZ, in Z, in R)?

(b) Are the radar maps shown in all figures the hourly sums, the hourly averages, or something else?

(c) Is the 13.5-dBZ thresholding done before or after the summing or averaging?

(d) How do you handle the NaN resulting from that thresholding in the summing (if the thresholding is done before) and in the training of the machine?

We thank the Reviewer for this insightful comment. We agree that the choice of aggregating radar reflectivity over 1-hour intervals, rather than using the observation closest to the analysis time, should be better justified in the manuscript. The revised lines are 122–135.

This choice was motivated by several considerations. First, ALADIN outputs precipitation at time t as 1-hour accumulation of precipitation between t - 1 hr and t. Our approach would allow a qualitative comparison between radar reflectivity simulated from model precipitation output at time t (through the use of a Z–R relation) and radar reflectivity simulated with our neural-network observation operator from outputs of basic model variables at time t. Second, hourly aggregation yields smoother reflectivity fields (small-scale features are filtered through aggregation/averaging), which are generally easier for a neural network to learn than noisy fields with sharp gradients. Finally, the observed reflectivity field exhibits larger spatio-temporal variability than the model output (the effective temporal resolution of 4.4 km ALADIN is approximately 15 mins due to numerical diffusion), so we assume that the instantaneous model state at time t is representative of a broader range of precipitation observations within the corresponding hourly window.

For these reasons, we chose to sum the quality-controlled radar observations over 1-hour periods. The aim was not to make them strictly equivalent to instantaneous model state variables, but rather to define a temporally more representative observational target at the same nominal time resolution as the model output. We considered this to be a reasonable compromise for the presented proof-of-concept study.

Despite that, we acknowledge that alternative temporal matching strategies may be more consistent, for example, relating the instantaneous model output at time t to radar observations over [t - 30 min, t + 30 min] with time t at the center of the interval, or relating an average of model outputs at t - 1 hr and t to radar reflectivity accumulated over [t - 1 hr, t]. However, implementing such alternatives would require retraining the network and recomputing all results. We therefore considered the present setup adequate for the purposes of this proof-of-concept study.

Following are the specific questions replies:

(a) The sum is done in Z.

(b) The radar maps shown in all figures are the hourly sum.

(c) The 13.5 dBZ thresholding is done before the summing.

(d) The thresholding is applied before the hourly accumulation by masking all reflectivity values $\leq$ 13.5 dBZ to NaN. During the hourly sum, those NaN values are ignored with `np.nansum` so only values above the threshold contribute. For the NN training, the pixels that has NaN after the summing operation are then replaced by 0. This means that, for the model, a zero represents "no valid signal above threshold or better no precipitation" rather than a physically measured zero reflectivity.

(iv) *Second paragraph of 2.4: "All such timesteps are extracted and augmented (Aggarwal et al., 2018) by rotating both the input ALADIN fields and the corresponding radar reflectivity by 90°, 180°, and 270° around the vertical axis". I presume you changed what was u-winds and v-winds in response to that rotation. More importantly, this rotation will hamper your radar field statistics (average, standard deviations) and make it less representative by combining points occurring at the right geographical location with three times*

more points occurring at another geographical location and having different averages and standard deviations, introducing location-dependent biases. Can you justify this choice? Wouldn't a simple repetition (or weight increase in the loss function) or a minor displacement of a model pixel (4.3 km) in each direction be a better choice?

We thank the Reviewer for this careful and technically important comment regarding the rotation-based augmentation.
In our implementation, the augmentation is applied consistently to the a reflectivity selected set of predictors and the radar target: the ALADIN input fields and the corresponding radar reflectivity are rotated together by 90°, 180°, and 270°. The training does not assume a fixed one-to-one correspondence between a given pixel index and a specific geographic location across samples (i.e., the model is not trained as a location-anchored mapping on a fixed grid). Moreover, the preprocessing uses min–max normalization rather than position-dependent mean/standard deviation normalization. For these reasons, we do not expect the proposed rotation to introduce the type of location-dependent bias described by the Reviewer through altered per-location statistics; rather, the augmentation increases the effective sample size and reduces orientation-specific overfitting while preserving the internal spatial relationships between predictors and target within each augmented sample.
We fully acknowledge, however, the Reviewer's point that the horizontal wind components are vector quantities. In the current version, u and v are treated as separate image-like channels that are rotated together with the other predictors. While this preserves input–output consistency for each augmented pair (the rotated predictors correspond to the rotated reflectivity), it does not explicitly enforce a strict vector-component transformation of (u, v) under rotation. We agree that incorporating a proper vector rotation (i.e., transforming (u, v) consistently with the rotation angle) would be a more physically rigorous treatment and may further improve the realism of the augmentation. We clarified this aspect in the revised manuscript.
We also acknowledge that the description of the augmentation procedure in the manuscript was imprecise. The augmentation was not applied directly to all samples only through rotation. Rather, the resulting samples of precipitation events were first duplicated, and then rotated of 90°, 180°, and 270°. We revised Section 2.4 (Training Setup) in the manuscript accordingly.

We also plan to evaluate a vector-consistent wind rotation, as well as alternative imbalance-mitigation strategies suggested by the Reviewer (e.g., sample reweighting or small spatial displacements), in future work.

(v) *Figure A.11 (which should be A.1) could use units to x and y. Kilometers? Grid points?*

We thank the Reviewer for noticing this. We corrected the appendix figure numbering (Figure A.11 has been renumbered to Figure A1) in the revised manuscript. The axes in Figure A11 (now A1) represent grid points, not kilometers. We added the appropriate axis labels/units to make this explicit. We noticed that also Figure A12 (now A2) needed the same correction.

(vi) *Errors, differences, bin sizes, and standard differences in reflectivity should have units of "dB" or "dB(Z)" (better choice), but not "dBZ". I counted 16 corrections to make in the text and figure captions, plus a few more on the figures themselves.*

We thank the Reviewer for this careful and helpful observation. We agree that quantities such as errors, differences, bin sizes, and standard deviations of reflectivity should be

expressed in units of dB (more precisely dB(Z)) rather than dBZ, which is reserved for the reflectivity quantity itself. We corrected the unit notation accordingly throughout the manuscript, including the text, figure captions, and the labels within the figures.