Response to Reviewer#1 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. Review: A neural network-based observation operator for weather radar data assimilation by Stefanelli et al.

In this paper, the observation operator for radar data assimilation is replaced by a neural network that maps state variables such as temperature, wind, and relative humidity at different vertical levels with the reflectivity. The machine-learning-based observation operator is coupled with a 3DVar data assimilation system, and observation impact experiments are performed in order to provide a preliminary evaluation of this observation operator.

The topic is novel and is aligned with the current trend of merging data assimilation with machine learning. However, there are aspects of the obtained results that are not very clear and that should be discussed in more detail. There are also some decisions taken in the design of the experiments and the methodology that deserve further discussion. My recommendation is that the paper should undergo major revisions before being considered for publication in GMD.

2. Major comments

1. *The authors selected 4 vertical levels from different state variables to be linked with the reflectivity. These levels correspond to low levels (i.e., approximately below 2 km). What is the motivation for considering only low levels? For example, deep convection can be strongly linked with upper-level winds and perturbations in temperature and relative humidity.*

   We thank the Reviewer for this comment. The reason for considering only low-level variables is that we used only the first radar sweep (0.5° elevation), which, after retaining only observations within the 160 km horizontal radius, represents the atmosphere mainly within the lowest ∼3 km, as reported in Section 2.1 (page 4, first paragraph). To ensure consistency between the model predictors and the observed atmospheric volume, we therefore selected model levels in the lower troposphere. We have clearly addressed the motivation for such choice in the revised paper, lines 91–95.
   We agree that deep convection is also influenced by upper-level winds and perturbations in

thermodynamic variables. However, to account for these processes consistently, we would have to include observations from additional radar elevation angles and corresponding higher model levels. However, in this proof-of-concept study, we prioritised simplicity over realism. We will explore applicability of our methodology to operational-like systems in the future work.

2. *The authors decided not to include hydrometeors as control variables, and because of this, these are not part of the observation operator. Since these are the variables more directly linked with radar reflectivity, some motivation should be provided. My impression is that choosing variables such as temperature and winds is an interesting approach, since this can better represent the links between radar reflectivity and the dynamic and thermodynamic fields (which are otherwise difficult to correct in a 3DVar framework). This can be an advantage of the selected approach and not a limitation, as the authors discuss at the end of the paper.*

   The exclusion of hydrometeors from the control vector was a deliberate design choice. Given that hydrometeors are the variables most directly related to radar reflectivity, many existing observation operators already focus on correcting them during the DA procedure. Our objective was to explore a complementary approach: using radar reflectivity to infer corrections in dynamical and thermodynamical variables, which are generally more difficult to constrain within a standard 3DVar framework.
   This choice was motivated by the expectation that corrections to variables such as winds, temperature, and humidity may provide a longer-lasting impact on the forecast. In contrast, direct hydrometeor adjustments are often short-lived in NWP applications due to rapid model adjustment (Fabry and Meunier, 2020). We agree with the Reviewer that this aspect of our methodology represents a significant potential advantage of the proposed approach. In line with Reviewer's inquiry, we have now discussed the motivation for such choice in the revised paper in lines 490–498.
   Nevertheless, we acknowledge that hydrometeors remain essential for describing the detailed microphysical structure of convection. Including them in the control vector would likely improve the representation of convective procesesses. We therefore consider this an important direction for future developments of this methodology.

3. *Since there are missing radar volumes (and maybe missing model outputs), it would be nice to provide more information about the actual sizes of the training, validation, and testing datasets before or after data augmentation.*

   We fully agree that providing the effective sizes of the training, validation, and testing datasets is important, especially given the intermittent missing radar volumes resulting in missing model outputs.
   The training dataset, constructed from the 2019–2022 period using only analysis times (00, 06, 12, and 18 UTC), contains 5706 samples in total. These were split as follows: 4564 samples for training, 571 for validation, and 571 for an initial test set used for a rapid check after training different models and model tuning. As described in the manuscript, only the precipitation events (reflectivity $\geq$ 13.5 dBZ) timesteps were augmented, corresponding to 3825. After data augmentation, the training dataset increased to 19 864 samples ($3825 * 4 + 4564$).
   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°.
   Finally, the independent test dataset for 2023, based on all forecast timesteps with hourly model output, contains 8725 samples.

We revised Section 2.4 (Training Setup) in the manuscript accordingly.

4. *Section 3.2. When the results in Figure 7 are described, there is no clear indication of the innovation that produced these analysis increments. This is said later in the discussion of Figure 8.*

   We have revised the manuscript (lines 393-394) to include a reference to the innovation figure (Fig. 9-d in the revised manuscript) in the discussion of the analysis increments figure.

5. *The authors should discuss in more detail the increments presented in Figure 6. There are some aspects that do not seem to be physically consistent. For example, the increment in V changes signs in vertical levels that are very close to each other. This also happens in r, where some levels show a dominant moistening effect while others (850 hPa) show a dominant drying effect. Also, the increment is not restricted or bounded to the circle, particularly for low-level temperature and relative humidity. In the wind field, it is difficult to see a dipole consistent with low-level convergence (something that would be expected if convection intensity is enhanced within the circle). Figure A.13 also shows some of these aspects, with changing signs in the wind increment at different vertical levels and no dipole structure in the wind increments.*

   We believe that Reviewer's comment most likely refers to Figure 9 (Figure 10 in the revised version) rather than Figure 6. Our response is therefore based on that assumption.
   The Reviewer raises an important point regarding the apparent lack of physical consistency of some of the increments across nearby vertical levels. This behavior is primarily related to the simplified construction of the background-error covariance matrix **B**. As described in Section 2.2, **B** was intentionally specified as a simple univariate, flow-independent model, with an isotropic recursive filter in the horizontal used to represent spatial autocorrelation. No vertical cross-level correlations were included. As a result, increments at adjacent vertical levels are not explicitly constrained to vary consistently, which can lead to sign changes between nearby levels in variables such as wind and relative humidity. This is now explicitly mentioned in the revised manuscript in lines 446–448.
   We adopted this simplified formulation deliberately in order to keep the 3DVar configuration lightweight, robust, and suitable for a proof-of-concept assessment of the neural-network-based observation operator. In this sense, the present setup was designed to test the feasibility of the proposed approach rather than to provide a fully optimized multivariate data assimilation framework. We agree, however, that a more realistic specification of **B**, including vertical and multivariate correlations, would likely improve the physical consistency of the resulting increments. This is a natural next step of this work, and we also note that recent studies suggest that neural-network-based approaches may offer promising directions for more advanced **B** modelling (Melinc and Zaplotnik, 2024; Melinc et al., 2026).
   Regarding the fact that the increments are not restricted to the circle, this is an expected consequence of the horizontal recursive filter used in the minimization. As also noted in the manuscript, the filter spreads the initially localized increment horizontally according to a Gaussian-like structure whose extent is determined by the prescribed correlation radius. Therefore, the analysis increments are not expected to remain sharply confined to the area of the imposed innovation, particularly for variables such as low-level temperature and relative humidity.
   Specifically, this is explained in Section 3.4:
   "... the localisation mask is defined as 1 over all grid points inside the radar disc and 0 elsewhere to ensure that the observation information is not spread with the gradient of

$\mathcal{H}$ outside the radar disc. Nevertheless, the **B**-matrix spreads the impact of observations slightly across the disc boundary (Fig. 10). As a result, the analysis increment pattern reflects the combined action of the RF-based background autocovariances and the spatially extended, nonlinear mapping provided by NN-based $\mathcal{H}$. The properties of the applied **B**-matrix are illustrated using a single observation experiment in Appendix A.2."
Also the difficulty to see in the wind field a dipole consistent with low-level convergence we believe is linked to the simplified setting of the **B** matrix.

We revised Section 3.4 in the manuscript to make these points clearer and to better distinguish which features arise from the proof-of-concept design of the **B** matrix and which are physically interpretable aspects of the analysis response.

6. *Following from the previous comment, Figure 9 includes the effect of spatial covariances. However, the vertical structure is still unclear. For r, some levels indicate that the observations produce a moistening (where the main precipitation band is located), while a drying effect is observed at other levels (975 and 800 hPa). Similar changes in signs are observed in the increments in V, so although the increments are aligned with the rainband, there is no consistent increase in low-level convergence (which would be the expected behavior).*

   The considerations raised here are closely related to those discussed in the previous point, and the same explanation applies. In particular, the unclear vertical structure and the sign changes between nearby levels are mainly a consequence of the simplified specification of the background-error covariance matrix B, which in the present proof-of-concept setup does not include vertical correlations or multivariate coupling between variables. We revised Section 3.4 in the manuscript to clarify this point.

7. *I could not follow figure 8. Panel d shows the innovation. But if the innovation is defined only within the small circle, why are there shades outside the circle? And why are the values outside the circle so different from the values inside the circle?*

   We thank the Reviewer for pointing this out. The main goal of this Figure 8 is to demonstrate how the observation operator responds in a deliberately extreme configuration, with a highly localized observation embedded in a much broader model-equivalent reflectivity pattern, and to show how $\mathcal{H}(\mathbf{x}_a)$ is modified in response to the assimilated signal. Therefore, the innovation is not defined only in the small circle, but over the entire radar disc.
   Figure 8 illustrates the effect of assimilating a very localized set of observations, namely a compact cluster of 250 radar pixels, while evaluating the NN-based observation operator over the full radar domain. In this experiment, panels (b) and (c) show the model-equivalent reflectivity computed from background and analysis over the entire radar disk, that is, without applying any spatial masking outside the assimilated cluster.
   For this reason, panel (d) should not be interpreted as the localized innovation used in the minimization only at the 250 observed pixels. Rather, it is shown as a diagnostic full-domain difference field, OBS $- \mathcal{H}(\mathbf{x}_b)$, in order to illustrate the mismatch between the localized observed signal and the model-equivalent reflectivity produced by the observation operator over the whole domain. This is why shaded values also appear outside the small observed circle. Those values are different from those inside the circle because outside the assimilated cluster there are no local reflectivity observations constraining the field, whereas $\mathcal{H}(\mathbf{x}_b)$ still produces nonzero reflectivity structures over the full radar disk.
   We revised the manuscript (lines 393–403) to make this aspects clearer.

3. Minor points

1. *The quality-controlled radar observations are summed into 1-hour intervals to match the resolution of the model output. However, the model output consists of instantaneous fields, not time-averaged or aggregated quantities (at least for the variables used in the observation operator). Using a time-averaged reflectivity, on the other hand, can help to smooth out small-scale details in the radar data that are not well resolved by the model.*

   We thank the Reviewer for this important comment. 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.
   We revised lines 122-135 in the manuscript to make our rationale clear.

2. *Radar data is interpolated to the model grid using nearest neighbor interpolation. This may not be the best approach if the resolution of the radar data is larger than that of the model. In that case, a box-averaging approach can be better.*

   We agree that when radar data are available at higher spatial resolution than the model grid, an area-based remapping method (such as box averaging) is generally more appropriate than nearest-neighbor interpolation. In our application, the radar data are indeed of higher spatial resolution than the model grid. However, implementing a different remapping strategy at this stage would require reprocessing the full dataset and retraining the neural network. Optimising it is beyond the scope of the present proof-of-concept study. For this reason, we maintain the current approach. However, we're very thankful to the Reviewer for pointing this out and will consider box averaging and related area-based remapping strategies in future work when we move closer to operational tests with an improved neural network.

3. *Figure 10 is mentioned before Figure 9.*

We corrected this accordingly in the revised manuscript.

4. *RMSE instead of RM in the caption of Figure 8.*

   We corrected the typo in the caption of Figure 8.

Bibliography

F. Fabry and V. Meunier. Why are radar data so difficult to assimilate skillfully? Monthly Weather Review, 148(7):2819–2836, 2020.

B. Melinc and Ž. Zaplotnik. 3d-var data assimilation using a variational autoencoder. Quarterly Journal of the Royal Meteorological Society, 150(761):2273–2295, 2024.

B. Melinc, U. Perkan, and Ž. Zaplotnik. A unified neural background-error covariance model for midlatitude and tropical atmospheric data assimilation. Journal of Advances in Modeling Earth Systems, 18(1):e2025MS005360, 2026. doi: https://doi.org/10.1029/2025MS005360. e2025MS005360 2025MS005360.