Thank you for your comment. We took all the suggestions into account, as detailed below.

Referee's comment 1: *The introduction does not make the objective of this research entirely clear. In line 45, the authors state that the central question of the paper is to characterize the information contained in the ensembles generated by the nowcasting algorithms. Moreover, the authors prematurely present their results in this section. I recommend formulating clear research questions instead, and omitting the results here.*

To clarify the question addressed in this work, we propose to remove "with a focus on its predictability properties." at line 36, and replace it by

> with a focus on the error representation in the ensembles (both in size and in localization).

The lines presenting the results in the introduction (lines 43-54) will be removed since the results are summarized in the conclusion.

Referee's comment 2: *Several of the diagnostic tools utilized in this paper would benefit from additional context. In particular, the eigenvalue method based on the ensemble covariance matrix requires further explanation. Did the authors develop this method specifically for this study, or are there existing applications of this approach for similar problems in the literature?*

We felt it was quite natural to look at eigenvalues and eigenvectors of the covariance matrix and looking now to the literature, we indeed see that this tool has already been used in this context. Thank you very much for drawing our attention to that.

We propose to add the following at the beginning of Sec. 3.2:

> The ensemble covariance matrix and its eigenvalues and eigenvectors are used in this Section to investigate the morphology of the ensembles. If the ensemble is considered as a cloud of points in phase space, the eigenvectors of the covariance matrix point in the directions in which the ensemble is the most extended (as measured by the variance in these directions, which are given by the associated eigenvalues). This tool was originally introduced by Patil et al. (2001) and then by Oczkowski et al. (2005) to estimate the dimension of ensemble of forecasts ("E dimension"), reducing the information in the covariance matrix to one number. Other works (Satterfield and Szunyogh, 2011; Ono, 2025) followed in that direction. On the other hand, the full eigenvalue spectrum of the covariance of ensemble members was used by Snyder et al. (2003) in an idealized model.
>
> Some of the authors who developed the "E dimension" have also later used the information in the eigenvectors of the ensemble covariance matrix, by using the fraction of error explained by the eigenvectors as a metric to evaluate ensembles of forecasts (Szunyogh et al., 2005; Satterfield and Szunyogh, 2010).

This browsing of the literature brought up additional references, which are relevant for Sec. 3.3, so that we will change the sentence

> The first is $\cos\gamma(e,p)$ is the cosine of the angle between the error e and its projection p on the subspace spanned by the residual vectors v_i.

at lines 190-191 to

> The first is the Perturbation versus Error Correlation Analysis (PECA) score Wei and Toth (2003), which is called $\cos\gamma(e,p)$ here in order to emphasize its geometric interpretation in phase space. Indeed, it is the cosine of the angle between the error e and its projection p on the subspace spanned by the residual vectors v_i.

The sentence

> A similar quantity was considered in Uboldi and Trevisan (2015) to estimate the ability of the breeding vectors of a realistic model to capture the error growth in a convective situation.

at lines 199-200 will also be changed into

> This metric is indeed equal to the PECA score, except that the error and the perturbations with respect to a control run when computing the PECA score. Apart from that, minimizing the L_2-norm of $e - \sum_i \alpha_i v_i$ amounts to construct the projection of e on the space spanned by the v_i.
>
> The same quantity was also introduced independently by Szunyogh et al. (2005); Satterfield and Szunyogh (2010) as the Explained Variance (EV) to evaluate the efficiency of data assimilation schemes to provide initial conditions leading, under the perfect model hypothesis, to ensembles of forecasts capturing the error. Finally, Uboldi and Trevisan (2015) also considered that metric to estimate the ability of breeding vectors of a realistic model to capture the error growth in a convective situation.

Referee's comment 3: *To my knowledge, the original LDCast method was trained using rainfall rates (the MeteoSwiss RZC product). In contrast, this study uses 5-minute rainfall sums resulting from a rain-gauge adjustment using Kriging with External Drift. Previous research has demonstrated that rain-gauge adjustment via geostatistics tends to smooth spatial rainfall fields. While this may not significantly impact the overall results, a brief discussion regarding this potential effect should be included in the manuscript.*
Thank you for pointing this out. We will add at line 86 the following remark:

> LDCast was trained on the radar product with code RZC, which uses only data from the radar themselves, while the RADCLIM dataset used in this work includes a merging with rain gauges. Merging with rain gauges is known to smooth out small scales (Ochoa-Rodriguez et al., 2019), so that a difference between the nowcasts of LDCast and the radar observations at small scales can be expected.

The paragraph at lines 128-132 will be changed to

> The spectral error and the power spectrum of observations flatten out for scales smaller than 5km. This is however not the case for the spectral variance of STEPS and LDCast, so that the ensembles seem to be underdispersed. However, the power spectrum of radar

images is known to flatten due to white noise contamination at these scales due to non-meteorological sources (Seed et al., 2013). Merging with rain gauge, as is done for the radar data used in this work, is also known to smooth the small scales (Ochoa-Rodriguez et al., 2019), and that explains the shape of the corresponding part of the spectral error. Therefore, the flattening of the power spectrum of radar images can be considered as not physical, and this underdispersion should not be identified as a weakness of the ensembles. This occurs at scales for which the error is immediately saturated, so that this is anyway irrelevant from a forecasting point of view.

It is not clear why the spectral variance of STEPS is not as large as the spectral error, since STEPS mainly extrapolates radar observations. On the other hand, LDCast was trained on a radar product without rain gauge merging, so that it can be expected that its nowcasts behave differently at small scales than the radar observations used in this work.

Referee's comment 4: *The use of surrogate ensembles based on MAAFT is central to reaching one of the paper's main conclusions: namely, that neither STEPS nor LDCast ensembles contain dynamical information regarding the spatial localization of the error. The authors should incorporate an explanation of this method into Section 3, rather than relegating it almost entirely to the appendix.*

Thank you for the comment, the different types of ensembles are indeed not properly introduced. We propose to replace the lines 206-208 by

> The two metrics are computed for STEPS and LDCast ensembles and compared to those of three types of surrogate ensembles. Surrogate ensembles are generated as follows:
>
> – HIST ensembles: for each member x_i in the original ensemble, a HIST surrogate z_i is generated by shuffling all the pixel values of x_i (x_i and z_i have thus the same histogram of values);
>
> – SPEC ensembles: for each x_i (that we decompose as $\bar{x} + v_i$), a SPEC surrogate is generated as $\bar{x} + w_i$, where w_i has the same power spectrum than v_i but with random Fourier phases (SPEC surrogates do not have the same histogram values of values than original members);
>
> – MAAFT ensembles: for each x_i, a MAAFT surrogate is built in such a way that it has the same histogram of pixel values than x_i and the same power spectrum for its residual vector than v_i. The name stands for Modified Amplitude Adjusted Fourier Transform.
>
> MAAFT surrogates are inspired from the Iterated Amplitude-Adjusted Fourier Transform (IAAFT) (Schreiber and Schmitz, 1996). The construction of SPEC and MAAFT ensembles is detailed in Appendix A.

In order to ease the reading of the core part of the manuscript, we propose to keep the detailed steps of the construction of SPEC and MAAFT ensembles in Appendix A, which would read:

The construction of the surrogates used in this work are detailed in this Appendix.

Surrogate ensembles are generated member by member and lead time by lead time. As explained in the text, HIST ensembles are generated by shuffling the pixel values of each original member x_i. The original member and its surrogate have exactly the same distribution of values.

SPEC surrogates are generated for each original member x_i by

1. initializing a surrogate z_i by shuffling all the values of x_i (as for a HIST surrogate);
2. computing the residual of z_i with respect to the mean as $w_i = z_i - \bar{x}$;
3. adjusting the power spectrum of w_i to that of $v_i = x_i - \bar{x}$: the Fourier coefficients $W_{i,\boldsymbol{k}}$ of w_i are replaced by $|V_{i,\boldsymbol{k}}|\phi_{\boldsymbol{k}}$, where $V_{i,\boldsymbol{k}}$ are the Fourier coefficients of v_i and $\phi_k = W_{i,\boldsymbol{k}}/|W_{i,\boldsymbol{k}}|$ contains the complex phase of $W_{i,\boldsymbol{k}}$;
4. constructing the new version of the surrogate as $z_i = w_i + \bar{x}$.

The residual vectors of SPEC surrogates have thus the same power spectra than the v_i, but the distributions of values are not the same.

MAAFT surrogates are built in order to share both the distribution of values of the original members and the power spectrum of the residual vectors. The construction starts as for SPEC surrogates and adds a fifth step:

5. adjust the distribution of values of z_i to that of x_i: the highest value in z_i is replaced by the highest value in x_i, the second highest in z_i by the second highest in x_i, and so on.

Steps 2 to 5 are then cycled through for 30 iterations.

This technique to generate surrogates is inspired from the Iterated Amplitude-Adjusted Fourier Transform (IAAFT) (Schreiber and Schmitz, 1996), and the only difference is that the power spectrum and the distribution adjustments are not done on the same quantity: the distribution is adjusted on the surrogate member z_i itself, while the power spectrum is adjusted on the residual $w_i = z_i - \bar{x}$ of the surrogate with respect to the ensemble mean. This allows to construct surrogate ensembles with the same power spectrum as the v_i and the same distribution of values for each member while keeping the information of the mean in the ensemble.

Referee's comment 5: *It would be highly valuable to include the authors' perspective on how the results of this paper might translate to other AI nowcasting methods. It seems unlikely that alternative approaches, such as DGMR, would be capable of providing true dynamic forecast uncertainty as defined in this study. The authors are encouraged to elaborate on this point in the discussion.*

Referee's comment 6: *The same applies to retraining. Do the authors expect to obtain significantly different results when LDCast is retrained using the Belgian RADCLIM dataset? Furthermore, what are* *the implications of the chosen training strategy on the expected results? To my knowledge, LDCast was trained using radar crops of different sizes while also applying data augmentation techniques such as random rotations and flipping.*

Thank you for these comments ! We propose to replace the paragraph at lines 280-287 by

> There is a number of possible directions open for future work. First, it should include the same type of analysis for a version of LDCast that has been fine-tuned over the Belgian region. Huge improvements are not necessarily expected from this retraining, since Leinonen et al. (2023) found that evaluating LDCast outside of its training region did not significantly degrade its global scores. If the model is retrained over a fixed region of fixed size without data augmentation, some improvement in the quality of the nowcasts could come from the fact that the model learned features specific to the training region (such as orography-induced rainfall), if it is evaluated and used over that region. A more promising approach is to inform or blend the nowcasting model with NWP forecasts, such as the NWP-blended version of pysteps (Imhoff et al., 2023).
>
> Other generative models for nowcasting should also be evaluated: for example, DGMR produces less diverse members than LDCast (Leinonen et al., 2023) so that its ensemble spread might be not as good at estimating the error of the ensemble mean. Less diverse members in ensembles could also lead to a smaller projection of the error on the ensemble (smaller $\cos\gamma(e,p)$). It would be interesting to see if the ensembles also have similar properties in Fourier space. This might not be the case for all models, depending on their architectures. Without additional context from local geography and atmospheric variables, we do not expect other pretrained generative AI model ensembles to dynamically capture the spatial error patterns.

References

Ochoa-Rodriguez, S., Wang, L.-P., Willems, P., and Onof, C.: A Review of Radar-Rain Gauge Data Merging Methods and Their Potential for Urban Hydrological Applications, Water Resources Research, 55, 6356–6391, https://doi.org/https://doi.org/10.1029/2018WR023332, 2019.

Oczkowski, M., Szunyogh, I., and Patil, D. J.: Mechanisms for the Development of Locally Low-Dimensional Atmospheric Dynamics, Journal of the Atmospheric Sciences, 62, 1135 – 1156, https://doi.org/10.1175/JAS3403.1, 2005.

Ono, K.: Assessing the Predictability of Mesoscale Moist Airflow Using the Ensemble Dimension, Weather and Forecasting, 40, 2601 – 2620, https://doi.org/10.1175/WAF-D-25-0042.1, 2025.

Patil, D. J., Hunt, B. R., Kalnay, E., Yorke, J. A., and Ott, E.: Local Low Dimensionality of Atmospheric Dynamics, Phys. Rev. Lett., 86, 5878–5881, https://doi.org/10.1103/PhysRevLett.86.5878, 2001.

Satterfield, E. and Szunyogh, I.: Predictability of the Performance of an Ensemble Forecast System: Predictability of the Space of Uncertainties, Monthly Weather Review, 138, 962 – 981, https://doi.org/10.1175/2009MWR3049.1, 2010.

Satterfield, E. and Szunyogh, I.: Assessing the Performance of an Ensemble Forecast System in Predicting the Magnitude and the Spectrum of Analysis and Forecast Uncertainties, Monthly Weather Review, 139, 1207 – 1223, https://doi.org/10.1175/2010MWR3439.1, 2011.

Snyder, C., Hamill, T. M., and Trier, S. B.: Linear Evolution of Error Covariances in a Quasi-geostrophic Model, Monthly Weather Review, 131, 189 – 205, https://doi.org/10.1175/1520-0493(2003)131<0189:LEOECI>2.0.CO;2, 2003.

Szunyogh, I., Kostelich, E. J., Gyarmati, G., Patil, D. J., Hunt, B. R., Kalnay, E., Ott, E., and Yorke, J. A.: Assessing a local ensemble Kalman filter: perfect model experiments with the National Centers for Environmental Prediction global model, Tellus A, 57, 528–545, https://doi.org/https://doi.org/10.1111/j.1600-0870.2005.00136.x, 2005.

Wei, M. and Toth, Z.: A New Measure of Ensemble Performance: Perturbation versus Error Correlation Analysis (PECA), Monthly Weather Review, 131, 1549 – 1565, https://doi.org/10.1175//1520-0493(2003)131<1549:ANMOEP>2.0.CO;2, 2003.