https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

Spread/Error relationship and spatial error structure of precipitation ensemble nowcasting: Comparison of STEPS and generative AI

Martin Bonte[1], Lesley De Cruz[1, 2], Fabian Debal[1], and Stéphane Vannitsem[3]

[1]Royal Meteorological Institute of Belgium, Brussels, Belgium
[2]Vrije Universiteit Brussel, Brussels, Belgium
[3]School of Physical and Mathematical Sciences & The Asian School of the Environment, Nanyang Technological University, Singapore

Correspondence: Martin Bonte (martin.bonte@meteo.be)

Abstract. The predictability of the generative AI-based nowcasting model LDCast is evaluated over Belgium, together with the pysteps implementation of the nowcasting algorithm STEPS. Neither STEPS nor LDCast were fine-tuned for the Belgian region, so both models are evaluated under conditions in which they will most likely be used in practice at national weather offices. STEPS and LDCast are slightly underdispersive, but the ensemble spread provides an estimation of the error at almost all scales. Both models adapt the properties of their ensembles to the type of event, either convective or stratiform. The spatial scores of the STEPS and LDCast ensembles are compared with those of surrogate ensembles, revealing that both STEPS and LDCast have very little ability to spatially localise the error of the ensemble mean. This suggests that the content of STEPS and LDCast ensembles is informative in terms of statistics, but not in terms of dynamics.

1 Introduction

Many artificial intelligence (AI) models for weather and climate have been developed recently. However, detailed evaluations of these models remain limited and are often done from a statistical point of view, using general scores such as the RMSE between a single forecast (or the ensemble mean) and the observations. In order to be used by forecasters in operational activities, the consistency of the meteorological characteristics of their nowcasts must be assessed (Bouallègue et al., 2024; Radford et al., 2025; Bröcker et al., 2026).

Since the dawn of General Circulation Models (GCMs), there have been many in-depth analyses of their dynamics and predictability in the past 70 years (e.g. Kalnay (2003); Bauer et al. (2015)). Recently, Artificial Intelligence Weather Prediction (AIWP) models such as Pangu-Weather, GraphCast, and FourCastNet were developed, and their forecasts for various events of interest have been compared in detail with those from Numerical Weather Prediction (NWP) models (Olivetti and Messori, 2024; Charlton-Perez et al., 2024; Pasche et al., 2025; Hua et al., 2025). Baño-Medina et al. (2025) also conducted a sensitivity analysis of initial conditions on the Spherical Fourier Neural Operator (SFNO) in a storm environment, finding properties similar to an NWP model.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

Methods for estimating uncertainties in AI model forecasts are still under development. Bred vectors for AIWP models were constructed by Baño-Medina et al. (2025), Mahesh et al. (2025) and Almeida et al. (2025), and a similar method was developed by Pu et al. (2025). On the other hand, the uncertainty could also be represented naturally by generative models as illustrated in Lang et al. (2024). These models produce nondeterministic forecasts, which makes it easy to create ensembles. Generative models such as DGMR (Ravuri et al., 2021) and LDCast can also be trained to produce realistic forecasts that retain sharpness instead of being smoothed out due to the loss of predictability with increasing lead time.

However, the relevance of ensemble properties of generative models for operational activities has not been sufficiently investigated. Indeed, these model ensembles are primarily evaluated over the whole dataset using global scores (Lang et al., 2024) or global rank histograms (Price et al., 2023; Leinonen et al., 2023). Some models are also evaluated on a few selected events through a meteorologist's comparison and ranking of the outputs of different models (Ravuri et al., 2021; Zhang et al., 2023).

These evaluation methods are necessary first steps, but a deeper understanding of the characteristics of generative models is needed. This includes their dynamical properties, as well as their particular behavior and biases for different weather types or spatial scales. The central question of this paper is therefore to characterize the information contained in the ensembles generated by a generative nowcasting model, with a focus on its predictability properties.

This work aims to provide this characterization for LDCast (Leinonen et al., 2023), a generative model for rainfall nowcasting. It is evaluated using the Belgian radar composite RADCLIM (Goudenhoofdt and Delobbe, 2016; Journée et al., 2023), but it has not been retrained on this dataset. Therefore, the model weights are the original ones, and the model can be considered pre-trained for rainfall nowcasting. The pysteps implementation of the nowcasting algorithm STEPS (Bowler et al., 2006; Seed et al., 2013; Pulkkinen et al., 2019) is evaluated in parallel to clarify which approach is best in producing rainfall forecasts for both convective and stratiform rainfall cases.

The three main results for LDCast are the following:

1. The spread of LDCast ensembles saturates from small to large scales and provides an estimation of the current error at nearly all scales (Sec. 3.1).

2. Depending on the event type, LDCast can adapt the power spectra of the perturbation modes of the ensembles, as well as the distribution of the perturbation sizes (Sec. 3.2).

3. Ensembles of this version of LDCast are not better at spatially capturing the error than surrogate ensembles with the appropriate statistical properties, for the metrics used in this work (Sec. 3.3).

These results are also valid for STEPS, with the difference that STEPS ensembles seem to collapse, meaning that the members tend to align along a few directions in phase space.

LDCast was not retrained over Belgium, so this work does not show that generative AI is not able to spatially capture its mean error if it has been carefully fine-tuned for the domain at hand (see discussion in Sec. 4).

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

2 Data and methods

2.1 Data and preprocessing

The radar product on which STEPS and LDCast are evaluated in this work is RADCLIM (Goudenhoofdt and Delobbe, 2016; Journée et al., 2023): it is a quantitative precipitation estimation product based on radar measurements, which are merged with rain gauge measures. Several techniques exist for this merging: the product used in this study was obtained with the Kriging with External Drift (KED) technique. The time resolution is 5 minutes and the spatial resolution is 1 km $\times$ 1 km. The evaluation domain is a 320 px$\times$ 320 px square in the radar frame.

Ten stratiform and ten convective events were selected based on the convective precipitation and the mean large-scale precipitation derived from ERA5 (Hersbach et al., 2023). Events are considered convective if their mean convective precipitation was greater than 10^{-4} mm/h and their large-scale precipitation was less than 10^{-6} mm/h. Similarly, stratiform events were selected as those with a mean convective precipitation below 10^{-7} mm/h and a large-scale precipitation above $8.07 \cdot 10^{-4}$ mm/h.

As both STEPS and LDCast transform the rain rate to a logarithmic scale, all computations in this work are done with a similar scale. The transformation used in this work is

$$f(R) = \begin{cases} 10\log_{10}(R) & \text{for } R \geq 0.1 \text{ mm/h} \\ -15 & \text{for } R < 0.1\text{mm/h} \end{cases} \tag{1}$$

The unit after this transformation is dBR.

2.2 Models

The two nowcasting models considered in this work are STEPS and LDCast. STEPS is built on Lagrangian persistence (Zawadzki et al., 1994), which consists in advecting the rainfall field with a motion field estimated through an optical flow algorithm. In STEPS, the rainfall field is in addition decomposed into fields of different spatial scales via a cascade, and each of the levels of the cascade evolves according to an auto-regressive (AR) process, usually of order 2. There are two sources of perturbation in STEPS: the motion field is stochastically perturbed via the BPS method, and the field intensities are perturbed via the noise component of the AR(2) processes (Bowler et al., 2006). The noise itself was originally generated with a parametric method, but the nonparametric method developed in Seed et al. (2013) is now more commonly used. STEPS is implemented at the RMI with pysteps (Pulkkinen et al., 2019), and features blending with NWP forecasts (Imhoff et al., 2023).

In STEPS, there is no value for the pixels for which the value should come from the advection of pixel values out of the radar frame. To avoid as much as possible this problem, STEPS nowcasts were produced using the whole available radar domain of RADCLIM (see Sec. 2.1), and the analysis presented in this work was performed on a smaller 320 px $\times$ 320 px square extracted from this domain. There were still some pixels with missing values after doing this, and these were assigned the no-rain value of -15 dBR. In stratiform STEPS ensembles, the fraction of those pixels in an ensemble at a given lead time is at most 3% and it is smaller than 1% for convective STEPS ensembles. The effects of such missing values in STEPS nowcasts are therefore supposed to be negligible.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

LDCast is a latent diffusion model, meaning that a sequence of rainfall fields is first encoded to a latent space learned by the variational autoencoder of the model. The forecaster network then predicts the latent representation of future rainfall fields. Conditionally to this output, the denoiser stack produces different members, which are finally transformed back from the latent space to rainfall fields (Leinonen et al., 2023).

The LDCast nowcasts were first produced on a 416 px $\times$ 416 px area extracted from the radar domain. The nowcasts were then cropped to the 320 px $\times$ 320 px square used to evaluate the models.

3 Results

Ensembles of nowcasts were produced with STEPS and with LDCast for the selected events (Sec. 2.1), in order to analyze their behavior in dynamically different situations. The members of the ensembles are denoted $\{x_i\}_{i=1,\dots,N}$ with $N = 50$. The time dependence is omitted in the notation for simplicity.

The ensemble mean is computed pixel-wise as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2}$$

The observations (radar images) are denoted by y, and the pixel-wise error is

$$e = y - \bar{x}. \tag{3}$$

The residual vectors v_i of the members with respect to the mean are defined as

$$v_i = x_i - \bar{x}. \tag{4}$$

The power spectrum of a field x, at a scale $1/k$, is

$$PS(x)_k = \langle |X_{\boldsymbol{k}}|^2 \rangle_{|\boldsymbol{k}|=k}, \tag{5}$$

where $X_{\boldsymbol{k}}$ is the Fourier coefficient of x at wavevector $\boldsymbol{k}$ and $\langle\cdot\rangle_{|\boldsymbol{k}|=k}$ denotes the average over the wavevectors with norm equal to k.

The analysis of the ensembles of nowcasts is presented below. Examples of nowcasts produced with STEPS and LDCast, as well as well a spatial representation of the mean error and of the spatial standard deviation in the ensembles, are provided in the Supplementary material (Figures S1, S2, S3 and S4).

3.1 Spectral error and spectral variance

The spectral error is computed as the power spectrum of the pixel-wise error e. It is a scale-by-scale decomposition of the mean squared error of the ensemble mean. On the other hand, the spectral variance is defined as

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N} PS(x_i - \bar{x})_k \tag{6}$$

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

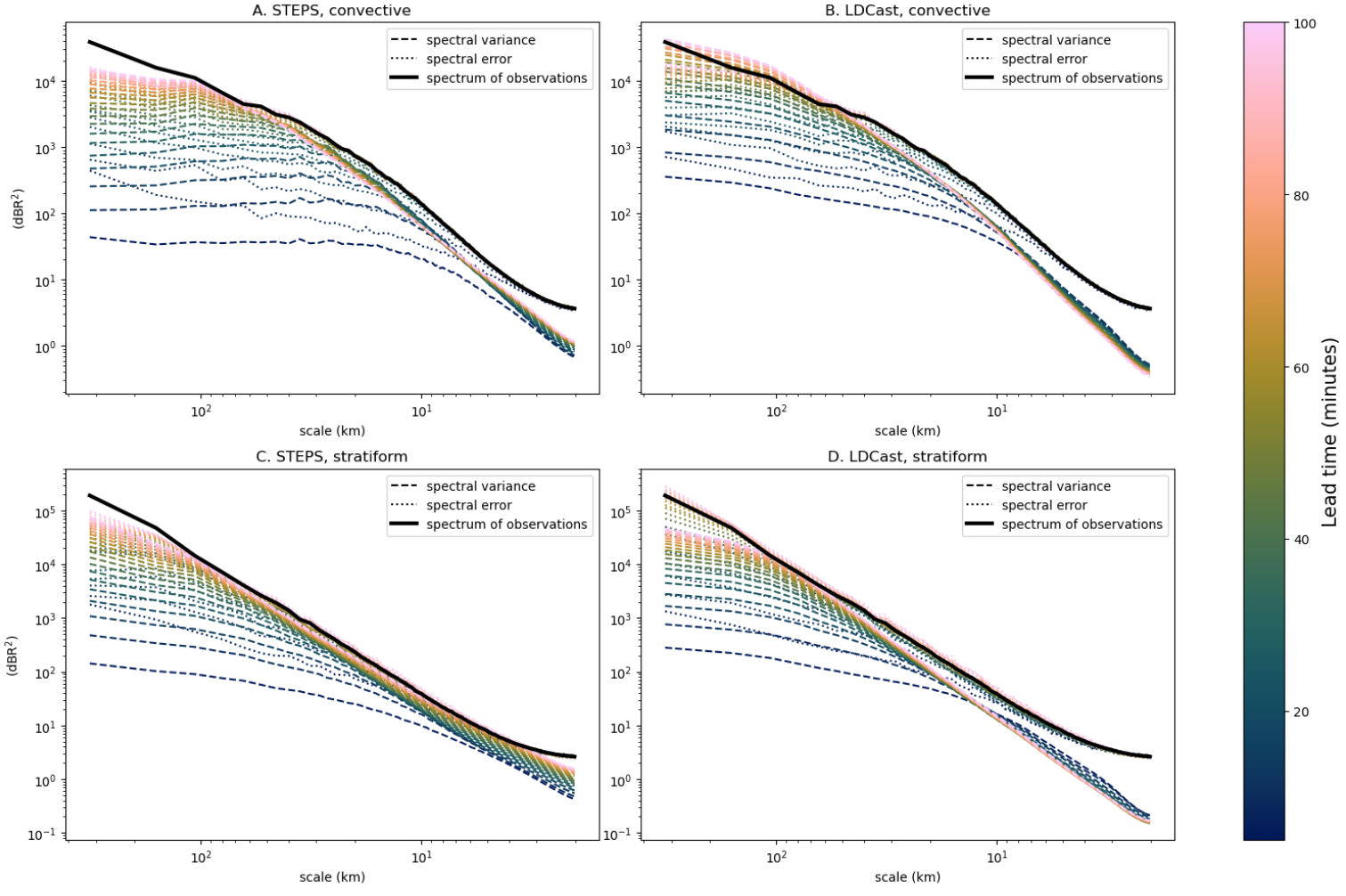

Figure 1. Spectral error and spectral variance, for STEPS and for LDCast, and for convective and for stratiform events (averaged over events). Dashed lines display the spectral variance and dotted lines display the spectral error. The color represents the lead time. The power spectra of observations are represented by thick black lines.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

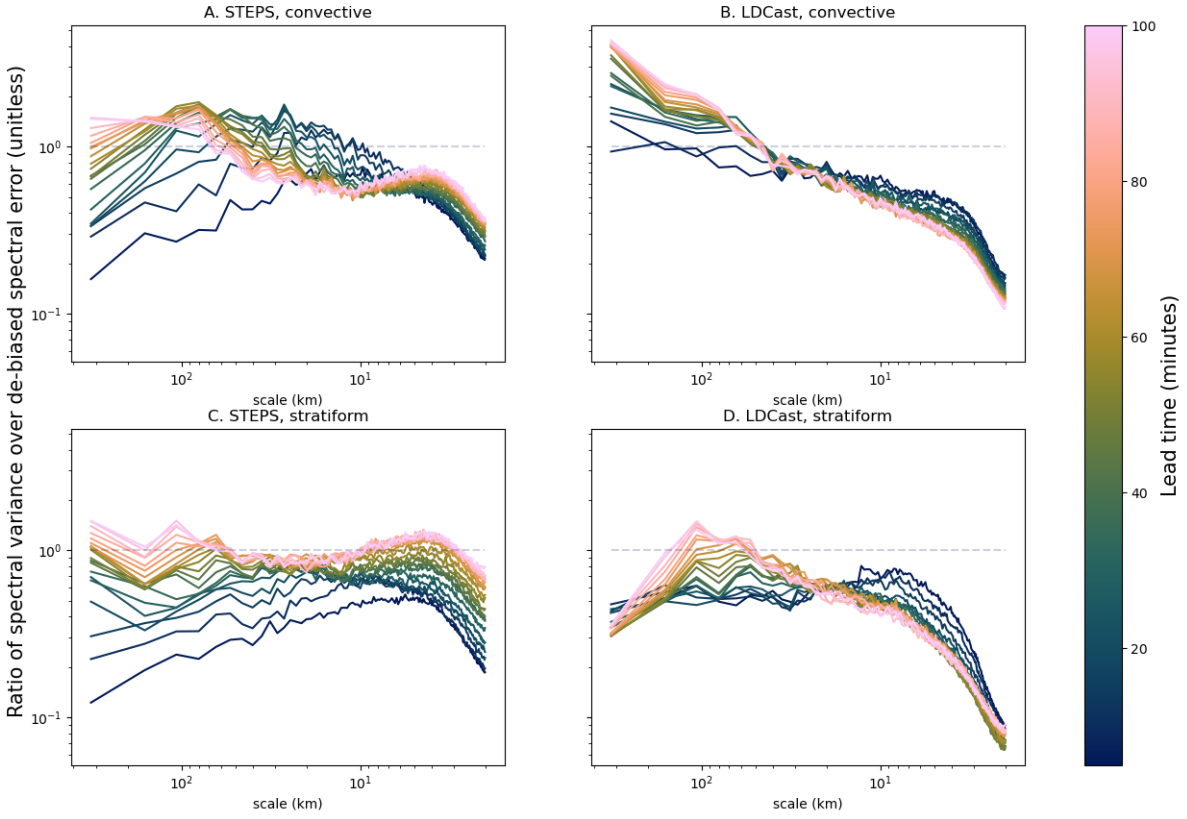

Figure 2. Ratios of the spectral variance over the de-biased spectral error, for STEPS and LDCast, and for convective and stratiform events (averaged over the events). The color represents the lead time. Horizontal dashed grey lines mark where the ratio is 1. Ratio < 1: underdispersion, ratio > 1: overdispersion.

and is often called the spread in meteorological applications.

The spectral error and spectral variance for STEPS and for LDCast, together with the power spectra of observations, are displayed in Fig. 1 (average over the selected events). The color of the curves represents the lead time at which the spread and the error are computed.

Both the spectral error and the spectral variance increase with lead time and eventually saturate to the power spectrum of the observation (see Appendix B). The error at some scales is immediately saturated (Pulkkinen et al., 2019): depending on the model and on the event type, the largest saturated scale after 5 minutes is between 5 and 10 km.

For a well-calibrated ensemble, the error and the spread should be equal, provided the ensemble mean is not biased (Fortin et al., 2014). Whether STEPS and LDCast ensembles are well-calibrated is assessed more precisely with Fig. 2, where the

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

scale-by-scale ratio of the spectral variance over the de-biased spectral error is displayed. A ratio smaller than 1 indicates underdispersion, while there is overdispersion when it is larger than 1. The spread/error ratio in Fig. 2 was computed with the de-biased error, meaning that a small bias of the ensemble mean was removed in Eq. (3).

Overall, both models have variance/error ratios close to 1 for scales above 5 km. LDCast is slightly underdispersive, except for the larger scales in convective events, where it is rather overdispersive. On the other hand, for stratiform events, STEPS ensembles are underdispersive for short lead times but become well calibrated for longer lead times. In convective cases, STEPS is either underdispersive or overdispersive depending on the lead time and the scale.

For scales below 5 km, all ensembles seem to be underdispersive. 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), 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.

3.2 Morphology of STEPS and LDCast ensembles

The morphology of the ensembles is investigated through their covariance matrix. This matrix is largely singular since the ensemble members have $320^2 = 102400$ components, while the ensembles have only 50 members. Actually, the covariance matrix has only 49 non-zero eigenvalues because it is constructed out of the residual vectors v_i, which satisfy $\sum_i v_i = 0$ (and it was supposed that this is the only relationship between them). The eigenvalues of the covariance matrix are denoted λ_i, and the eigenvectors u_i are also called perturbation modes (PM).

Panels A. and B. of Fig. 3 display the eigenvalues λ_i of the covariance matrix for both models, averaged over events. A few eigenvalues largely dominate the others, especially in convective STEPS ensembles (panel A.). To investigate the geometry of these ensembles, the residual vectors v_i of the members with respect to the ensemble mean were considered (see Eq. (4)). There are roughly speaking two different geometries that could explain this: a) the v_i have comparable norms but they align along the directions of the first eigenvectors of the covariance matrix or b) the v_i are each pointing in different directions but some of them have much larger norms. In order to determine which of the two geometries is actually realized in this case, the cosine c_{ij} of the angle between the residual vector v_i and u_j was computed for all i and j. It is computed as

$$c_{ij} = \frac{v_i \cdot u_j}{|v_i|} \tag{7}$$

since the eigenvector u_j has a unit norm.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

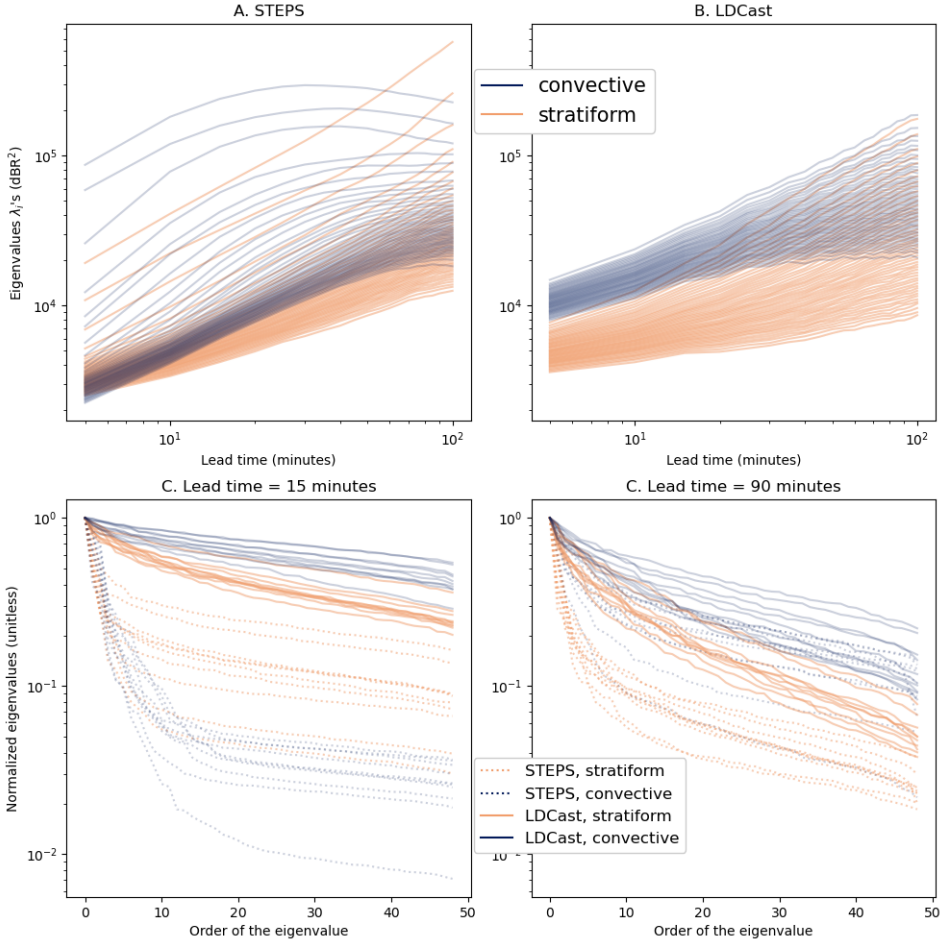

Figure 3. Eigenvalues λ_i of the covariance matrix. Evolution of the eigenvalues for convective and stratiform events for STEPS (A.) and LDCast (B.). Normalized eigenvalues for convective and stratiform events, for STEPS and LDCast, after 15 minutes (C.) and after 90 minutes (D.) (each line is the eigenvalue spectrum for one event).

Figure 4 represents the histograms of values of c_{ij} for $j = 1, 2, 10$ and 15, for a lead time of 15 minutes. In convective STEPS ensembles (panel A.), the distributions of c_{i1} and c_{i2} contain much higher values than the distributions of c_{i10} and c_{i15}, meaning that the v_i project much more on the first and the second eigenvectors than on the 10th and the 15th. This is also the case for stratiform STEPS ensemble, albeit to a lesser extent. The conclusion is that the v_i in these ensembles align along the first eigenvectors (option a)).

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

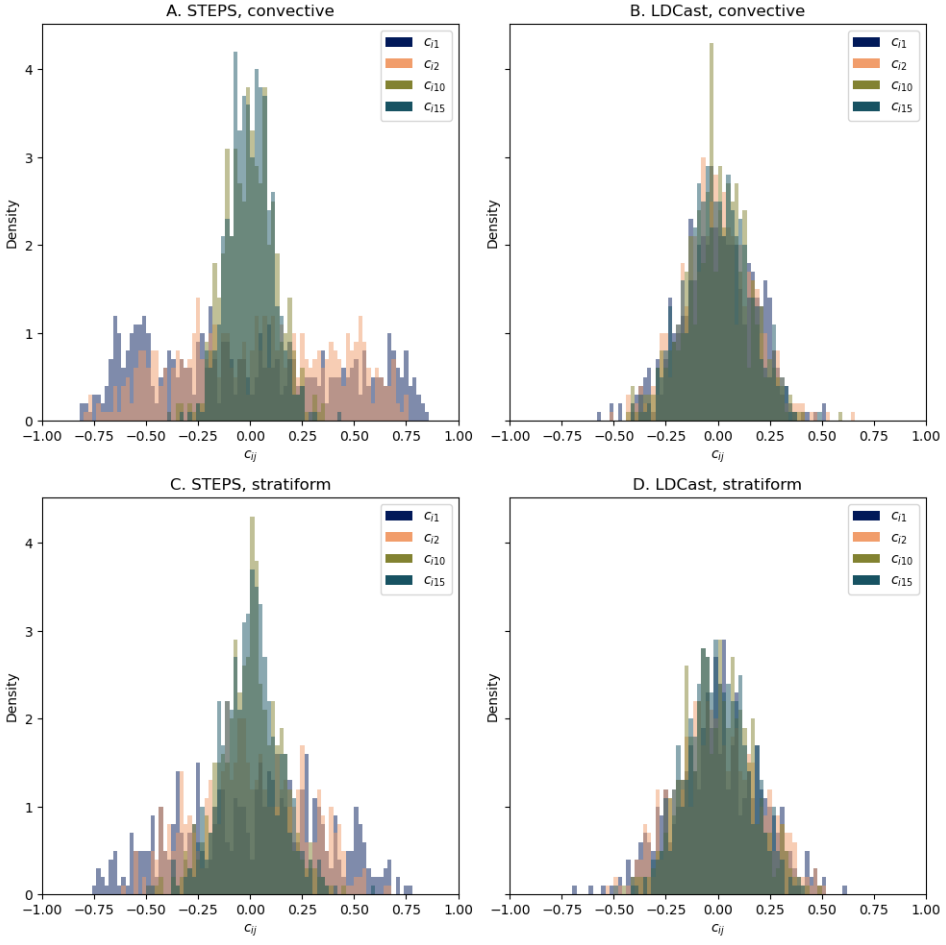

Figure 4. Histograms of the c_{ij} for $j = 1, 5, 10$ and 15 (for all events) after 15 minutes.

The evolution of eigenvalues λ_i of the covariance matrix for stratiform events is comparable for STEPS and LDCast: the eigenvalues increase steadily following power laws of the lead time (Fig. 3, panels A. and B.). This is also the case for convective LDCast ensembles. These power laws should be connected with the theoretical time that an error, initially located at small scales, needs to reach a given larger scale. Indeed, on a theoretical basis, it is expected that the highest scale contaminated by the error evolves as an exponential of the lead time 2D turbulent fluids (as is seen when working on synoptic scales), while the highest scale reached by the error increases as a power law of the lead time for 3D turbulent fluids (Vallis, 2017). Given the scales of nowcasting, the relevant physical processes are those of 3D turbulence, so that there is a qualitative agreement between these theoretical considerations and the scaling on the error observed in Fig. 3.

Panels C. and D. of Fig. 3 show, for two different lead times, the spectrum of the normalized eigenvalues. These are the eigenvalues divided by the highest one in the ensemble at that lead time, and they allow to better understand the relative sizes of the perturbation modes. Consistent with the panels A. and B., it appears that LDCast ensembles have generally more

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

homogeneous eigenvalue spectra (i.e. have more eigenvalues of similar size) than STEPS ensembles. Convective LDCast ensembles exhibit more homogeneous eigenvalue spectra than stratiform LDCast ensembles, especially at long lead times. As already noted from panel A., it is also obvious that STEPS ensembles are largely dominated by a few eigenvalues: convective STEPS ensembles are dominated by ~ 10 eigenvalues, while stratiform STEPS ensembles are dominated by ~ 5 eigenvalues. However, the relative size of higher-order eigenvalues (order > 5 in stratiform ensembles, order > 10 in convective ensembles) is larger in stratiform STEPS ensembles than in convective STEPS ensembles.

The power spectra of the eigenvectors of the covariance matrix are displayed in Fig. 5, for different lead times, for stratiform and convective events. The color represents the associated eigenvalues λ_i. Overall, the perturbation modes gradually develop over larger scales, while the spectrum profile remains the same for small scales. This shows that the directions of perturbation evolve in phase space with the lead time: at long lead times, the perturbations are vectors pointing in directions corresponding to larger scales than at the beginning of the nowcast.

It also appears that eigenvectors with higher eigenvalues have more power at large scales. This leads to a hierarchy: perturbations with larger amplitudes affect the large scales more. This is in agreement with the fact that, for a given scale, there is a maximum size for the error (see Fig. 1), so that once a perturbation reaches the saturation value at a given scale, it propagates to larger scales.

For both STEPS and LDCast, the eigenvectors of stratiform ensembles (especially the eigenvectors associated with the highest eigenvalues) have more power at larger scales than those of convective ensembles. This means that the perturbations contained in stratiform ensembles are at larger scales than those in convective ensembles.

The curves standing out in the top row (A.) of Fig. 5 for short and intermediate lead times are those of the power spectra of the eigenvectors that largely dominate the STEPS ensembles in convective cases (panel A. of Fig. 3). Similar power spectra are also present in stratiform STEPS ensembles.

Spectra at small scales decrease over time. This is because the eigenvectors have unit norm, so that the sum over the scales of each power spectrum is equal to 1.

3.3 Spatial skill of STEPS and LDCast ensembles

This section presents results assessing the skill of ensemble members to spatially represent the error. The focus is particularly on evaluating whether the ensembles contain any spatial information beyond that present in the ensemble mean.

Two metrics are considered. 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. Using the fact that the eigenvectors of the covariance matrix are orthonormal and span the same subspace as the v_i, p can be written as

$$p = \sum_i (u_i \cdot e) u_i \tag{8}$$

and its norm is $|p| = \sqrt{\sum_i (u_i \cdot e)^2}$. So $\cos\gamma(e,p)$ can be computed in terms of the scalar product between e and p as

$$\cos\gamma(e,p) = \frac{e \cdot p}{|e||p|}, \tag{9}$$

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

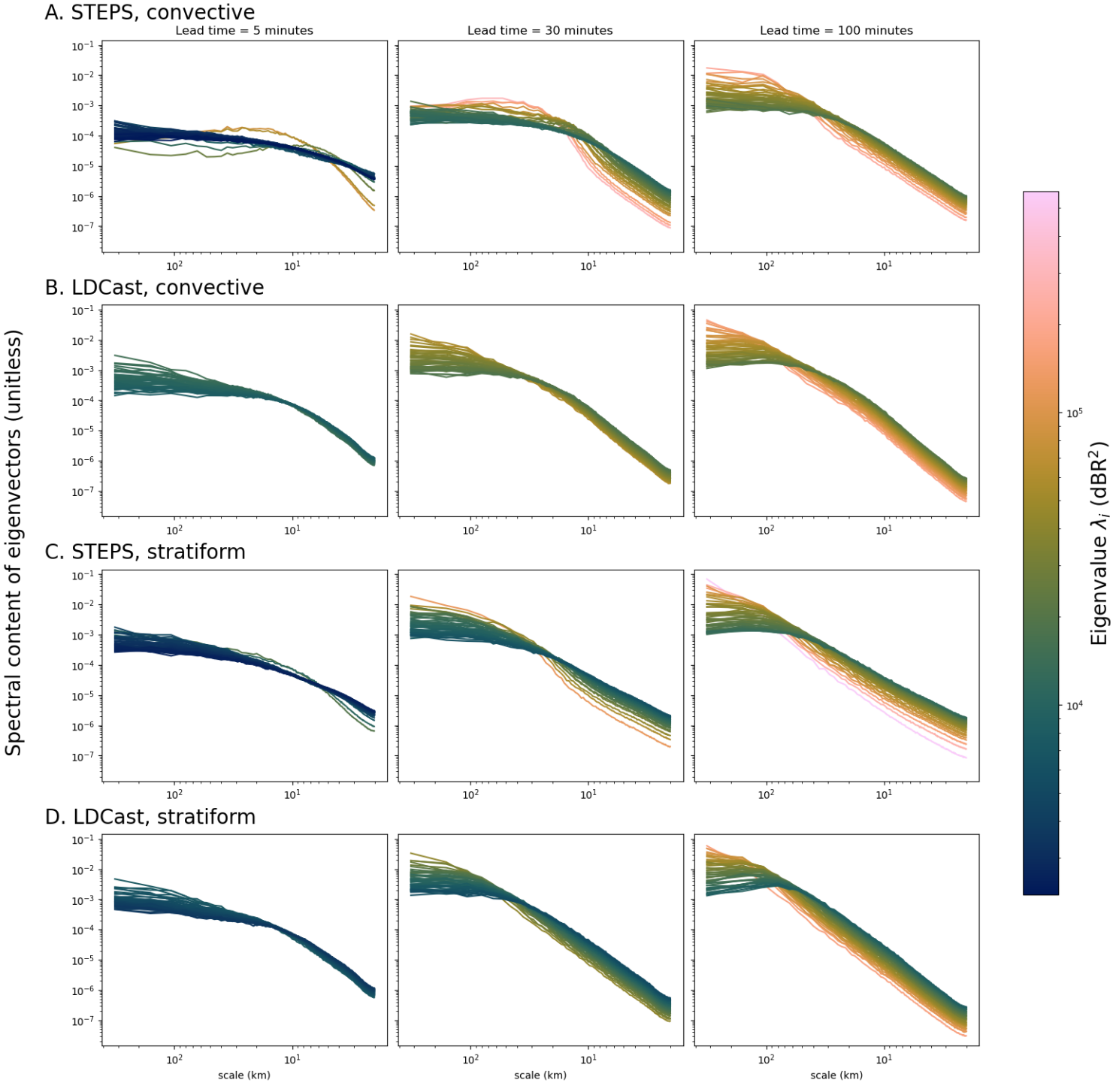

Figure 5. Evolution of the spectra of the eigenvectors for convective and stratiform events, for LDCast and STEPS ensembles. The spectra are averaged over all events. The color represents the corresponding eigenvalue.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

or in terms of the ratio between the norm of p and the norm of e:

$$\cos\gamma(e,p) = \frac{|p|}{|e|} \tag{10}$$

since $e \cdot p = \sum_i (u_i \cdot e)^2 = |p|^2$. These expressions show that $\cos\gamma(e,p)$ indicates the extent to which the error is captured by the ensemble members. 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.

The second metric considered is the Fraction Skill Score (FSS), which was introduced by Roberts and Lean (2008) as a metric for the spatial accuracy of a deterministic forecast. With respect to the mean square error, it mitigates displacement errors by comparing the forecast and the observation over neighborhoods of different scales. Necker et al. (2024) recently compared different versions of the FSS for ensemble verification, and recommended the 'probabilistic FSS' proposed in Schwartz et al. (2010), which is the one used in the current work.

The two metrics are computed for STEPS and LDCast ensembles and compared to those of MAAFT ensembles (see Appendix A), a baseline surrogate nowcast ensemble with the same distribution of rainfall intensities and whose residual vectors have the same power spectra.

Using $\cos\gamma(e,p)$

The cosine of the angle between the error e and its projection p on the STEPS and LDCast residual vectors v_i is represented by solid lines in the panels of Fig. 6. The thin lines depict this cosine for each event, and the thick line shows the mean of this quantity over the events. The dotted lines represent the same quantity for the MAAFT ensembles constructed from the original STEPS and LDCast ensembles.

The $\cos\gamma(e,p)$ values of the surrogate MAAFT ensembles are very close to those of the corresponding STEPS and LDCast ensembles. The reason for this is the shape of the Fourier spectra of the perturbations. Indeed, it was verified that SPEC ensembles (surrogates generated without the constraint on the distribution of values, only imposing the power spectra of residual vectors, see Sec. A) lead to similar values for $\cos\gamma(e,p)$ (Fig. S5 of Supplementary material). In particular, the spatial localization of the perturbations contained in the v_i of STEPS and LDCast ensembles does not explain the values of $\cos\gamma(e,p)$ of the error with these ensembles, since the same values are reached for surrogate ensembles with random residual vectors where no constraint is imposed on the spatial localization.

One way to understand this is by expressing $\cos\gamma(e,p)$ in terms of Fourier modes: if $E_{\boldsymbol{k}}$ and $V_{\boldsymbol{k}}$ are respectively the Fourier coefficients of e and of a residual vector v (of a STEPS, LDCast or MAAFT ensemble) and $\phi^e_{\boldsymbol{k}}$ and $\phi^v_{\boldsymbol{k}}$ the complex phases of these coefficients, the terms in the projection of e on the v_i is proportional to the scalar product of e and v, which is of the form $|E_{\boldsymbol{k}}||V_{\boldsymbol{k}}|\,(\phi^e_{\boldsymbol{k}}\phi^{v*}_{\boldsymbol{k}} + \phi^{e*}_{\boldsymbol{k}}\phi^v_{\boldsymbol{k}})$. The moduli $|V_{\boldsymbol{k}}|$ are imposed when constructing MAAFT ensembles and the complex phases $\phi^v_{\boldsymbol{k}}$ are precisely random. The fact that the values of $\cos\gamma(e,p)$ for a STEPS or an LDCast ensemble are the same as with the corresponding MAAFT ensemble shows that the complex phases $\phi^v_{\boldsymbol{k}}$ are also random for STEPS and LDCast ensembles. Since the spatial localization of the structures of a field is contained in the complex phases of its Fourier coefficients, the fact that

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

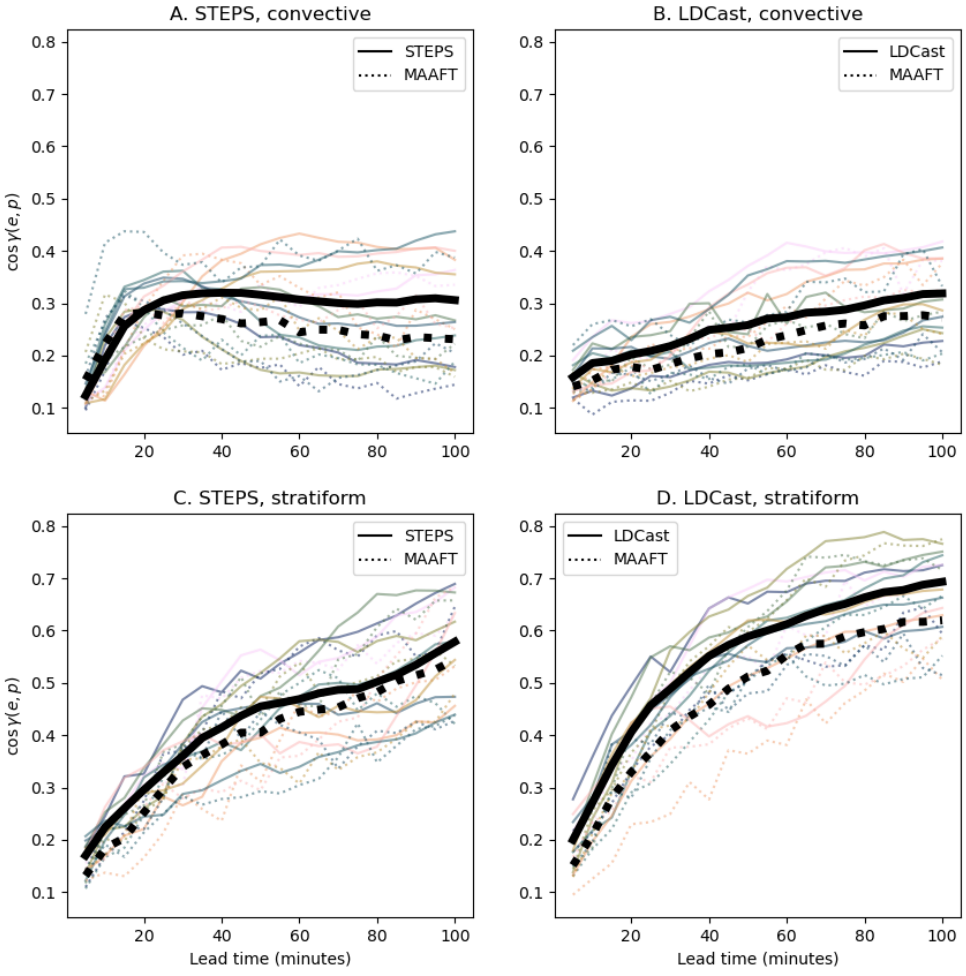

Figure 6. Comparison of $\cos\gamma(e,p)$ of the error and its projection on STEPS/LDCast ensembles (solid lines) with the $\cos\gamma(e,p)$ for the corresponding MAAFT ensembles (dotted lines), for all lead times. Thin lines depict $\cos\gamma(e,p)$ for each event and thick lines represent the event average. The MAAFT ensembles whose scores are presented in panels A. and C. are built from the corresponding STEPS ensembles, while those whose scores are presented in panels B. and D. are built from the corresponding LDCast ensembles.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

the complex phases for the Fourier coefficients of the v_i are random indeed means that the localization of the structures they contain is random.

The value of $\cos\gamma(e,p)$ increases with the lead time and this can be explained by the evolution of the power spectra of the vectors v_i. When the lead time increases, these vectors progressively represent larger perturbations. Because of this, it can be expected that it becomes easier for ensembles to cover the whole domain with independent perturbations of a given scale, and thus to spatially capture the actual instability.

Interestingly, Feng et al. (2024) argued that the error at convective scales always has a negligible projection onto the perturbation modes of an ensemble. While this is always the case at short lead times in Fig. 6, the projection is largely non-negligible for intermediate and long lead times, especially for stratiform events. However, this is because the perturbation modes develop large-scale components that eventually cover the entire domain.

Using the FSS

Figure 7 depicts the FSS averaged over events, for STEPS/LDCast ensembles and for the corresponding MAAFT ensembles, by solid and dotted lines, respectively. The thick black lines (HIST lines) in this Figure represent the FSS of the surrogate HIST ensemble (surrogates constructed by simply shuffling the pixel values, without power-spectrum adjustment, see Sec. A). This score remains relatively constant over time, so that only the time average is represented in Fig. 7. This FSS is the one of a random prediction given the distribution of values in the nowcast. For the smallest scales, this FSS is analogous to the random FSS in Roberts and Lean (2008). For the largest scale, it corresponds to their Asymptotic Fraction Skill Score (AFSS) (Roberts and Lean, 2008).

The LDCast and the corresponding MAAFT scores saturate with lead time to a value close to the HIST value. Interestingly, the LDCast and STEPS values of the FSS saturate more slowly than the spectral variance in Fig. 1: even at the smallest scales, the FSS saturates after 40 minutes, whereas the spectral variance is immediately saturated at those scales. The other intriguing feature in this Figure is the behavior of LDCast for the highest threshold: for the highest scales for which it is computed, the FSS scores increase with lead time. At the highest scale, the FSS score merely assesses whether there is a correct number (over the whole radar frame) of values above the chosen threshold. In this case, the number of high-intensity values is better estimated as the lead time increases.

The MAAFT FSSs are quite close to STEPS and LDCast FSSs, showing again that the FSS scores of STEPS and LDCast ensembles are mainly due to their statistical properties, but not to the spatial localization of the perturbations in the ensembles.

4 Conclusions

This work investigates the spectral and spatial properties of STEPS ensembles and a pre-trained version of LDCast. The spreads of ensembles of both models saturate from small to large scales and overall, they provide an estimation of the error for most scales even though the models were not fine-tuned for Belgium. The latter observation extends the conclusion of Leinonen et al. (2023) in the case of LDCast. Depending on the event type, STEPS and LDCast adapt the perturbation modes of their

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

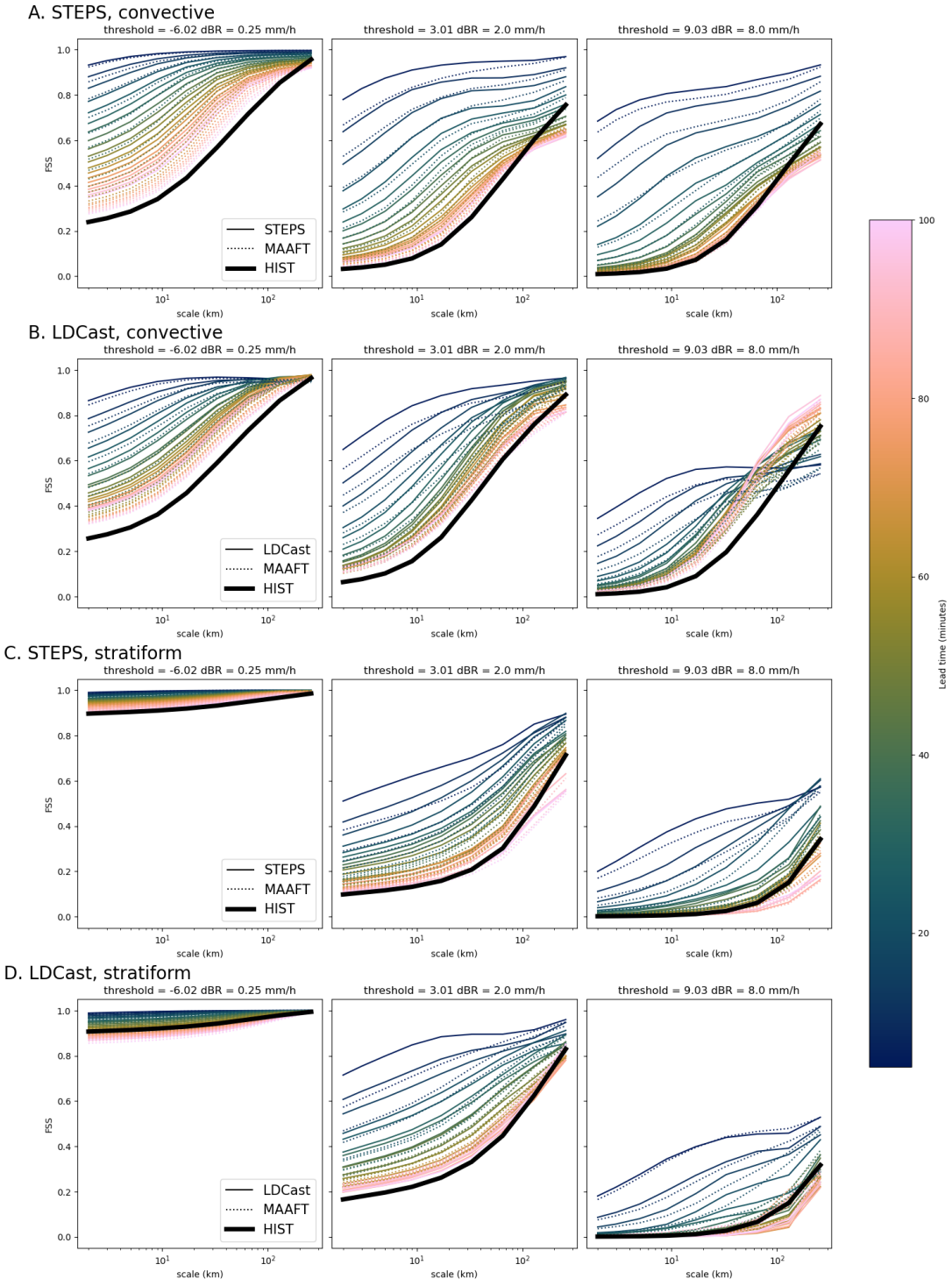

Figure 7. Fraction Skill Scores (FSSs) as functions of scale for a) convective events and b) stratiform events, with the time dependence represented by the color. The thin solid lines represent the LDCast FSSs, the dotted lines represent the MAAFT FSSs while the thick solid line represents the HIST FSSs (no time dependence shown). The x-axis represents the scale for which the FSS was computed, and is reversed with respect to the previous Figures.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

ensembles, as well as the distribution of perturbation sizes. STEPS ensemble members, however, tend to align along a few directions in phase space. This implies that STEPS ensembles have only a few modes of variability, and that the members are quite similar to each other.

Spatial scores were computed to assess STEPS's and LDCast's ability to capture the error in the ensembles. The surrogate MAAFT ensembles were designed to have similar statistical properties to those of the original ensembles, allowing to challenge the scores of the models. MAAFT ensembles and original ensembles have very close scores, demonstrating that the skill of STEPS and this version of LDCast is largely due to the statistical properties of the ensembles, and not to the localization of the error and spread. Appropriate power spectra are sufficient to reproduce the values of the cosine of the angle between the error and original ensembles, while an adequate distribution of pixel values is also needed to reproduce the FSS.

According to the metrics used in this work ($\cos\gamma(e,p)$ and the FSS), these results show that the ensembles of STEPS and of this version of LDCast do not have any dynamical information on the spatial localization of the error. The results could be different for a version of LDCast which is specifically trained for the Belgian domain. The results in this work are of practical interest, since most AI models will very probably be used without retraining. Indeed, proper retraining requires time and an archive of data on which to fine-tune the model, so users (as forecasters from national weather offices) will likely use the outputs of several non-retrained models for operational activities.

The domain on which the models were evaluated was chosen to be smaller than the region over which the nowcasts were computed (see Sec. 2.2) in order to avoid having missing values in STEPS nowcasts. However, the fact that STEPS has missing values in its nowcasts can also be seen as an advantage of the model: these missing values explicitly show where information was missing to produce the nowcast. In contrast, it is more difficult to spot where a generative model like LDCast has completely invented the values.

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. It will in particular be interesting to see how much this improves the quality of the nowcasts, and whether the model learned features specific to the Belgian region (such as orography-induced rainfall). 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. Other generative models for nowcasting should also be evaluated. 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. The version of pysteps blended with NWP forecasts (Imhoff et al., 2023) would also be interesting to evaluate.

The specific structure of STEPS ensembles should also be further investigated to understand the origin of the alignment of the v_i and whether some correction should be applied. This work focused on the spatial localization of the perturbations, but did not evaluate the spatial correlations in the fields themselves. It is important for some downstream applications (such as hydrological simulations) to have nowcasts correctly representing the current anisotropy of the rainfall fields in the ensemble members. The nonparametric method for generating noise (Seed et al., 2013) in STEPS ensembles allows to generate members with the appropriate spatial correlation, but it is not known how LDCast members are in this regard. The results in this study point to the fact that, in the case where the spatial anisotropy is correctly represented in the nowcasts (which is the case in

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

STEPS nowcasts when the nonparametric method is used), the structures are not better localized for the members than for the ensemble mean.

The mechanism proposed in Sec. 3.3 to explain the increase of $\cos\gamma(e,p)$ over time (Fig. 6) seems to imply that there is a relationship between the ensemble size, the domain size and the scale at which $\cos\gamma(e,p)$ saturates. It would be interesting to quantify this relationship more precisely by varying the ensemble size.

Code and data availability. The RADCLIM dataset and the nowcasts will be made available upon reasonable requests to the authors. The code used in this work is available at the Zenodo repository https://doi.org/10.5281/zenodo.18341086.

Appendix A: Surrogate ensembles

This study introduces the MAAFT technique to construct surrogate ensembles. It is inspired by the Iterated Amplitude-Adjusted Fourier Transform (IAAFT) (Schreiber and Schmitz, 1996), which is a method to construct random surrogate time series with the same distribution of values and the same autocorrelation function as the original time series. The Modified Amplitude-Adjusted Fourier Transform (MAAFT) produces, for each original ensemble, a surrogate ensemble whose members have a) distributions of values prescribed by the corresponding member in the original ensemble and b) residual vectors with respect to the mean with the same power spectra as the v_i.

The MAAFT ensembles are constructed member by member as follows. For each ensemble member x_i, a surrogate z_i is initialized by shuffling all its pixel values. The surrogate is then iteratively refined by repeating the following steps:

1. compute the residual of z_i with respect to the mean: $w_i = z_i - \bar{x}$
2. adjust 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 $\phi_{\boldsymbol{k}} = W_{i,\boldsymbol{k}}/|W_{i,\boldsymbol{k}}|$ contains the complex phase of $W_{i,\boldsymbol{k}}$
3. construct the new version of the surrogate as $z_i = w_i + \bar{x}$
4. 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.

These four steps were repeated 30 times to construct the MAAFT ensembles in this work.

The only difference from the original IAAFT algorithm 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.

The power spectrum adjustment technique is already used in nowcasting contexts, for example to create the noise of the AR(2) processes in the STEPS nowcasting algorithm (Seed et al., 2013).

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

Another type of surrogate ensemble considered in this work is one where only the power spectrum of the residual vectors is preserved (SPEC ensembles). They are constructed by initializing z_i as in MAAFT, and then performing only once steps 1., 2. and 3. of the MAAFT construction.

The last type of surrogate ensembles (HIST ensembles) used in this work are produced simply by shuffling the pixel values member by member. This produces ensemble members with exactly the same histogram of values, but without any spatial information. The MAAFT ensembles can be seen as ensembles combining the properties of both SPEC and HIST ensembles.

Appendix B: Error saturation

Both the spectral error and spectral variance in Fig. 1 reach a maximum scale-dependent value during the nowcast. This saturation value is close to the power spectrum of the rain field itself.

The spectral error is close to the power spectrum of observations because the ensemble mean has much less power at saturation than the observations (often one order of magnitude smaller, cf. Fig. B1). This means that the Fourier coefficients of the ensemble mean are small with respect to those of the observation, so that the latter provide the main contribution to the error: $e \approx y$ (cf. Eq. (3)).

For the spectral variance, at saturation, the power spectrum of the ensemble mean is small with respect to the power spectra of ensemble members, which themselves remain similar to that of the observation. All the terms in the sum in Eq. (6) are therefore also of the same order as the power spectrum of the observation.

Note that, when the error is computed as the difference between any two ensemble members, it is on average equal to twice the variance:

$$\frac{1}{N(N-1)}\sum_{i,j=1}^{N} PS(x_i - x_j)_k = \frac{1}{N(N-1)}\sum_{i,j=1}^{N} \langle |(X_{i,k} - \bar{X}_k) - (X_{j,k} - \bar{X}_k)|^2 \rangle_{|\boldsymbol{k}|=k} \tag{B1}$$

$$= \frac{2}{N-1}\sum_{i=1}^{N} \langle |X_{i,k} - \bar{X}_k|^2 \rangle_{|\boldsymbol{k}|=k} - \frac{2}{N(N-1)}\sum_{i,j=1}^{N} \langle (X_{i,k} - \bar{X}_k)^* (X_{j,k} - \bar{X}_k) \rangle_{|\boldsymbol{k}|=k} \tag{B2}$$

$$= 2\sigma_k^2 \tag{B3}$$

The second term of the second line vanishes because $\sum_{i=1}^{N}(X_{i,k} - \bar{X}_k) = 0$. Therefore, if the error is taken to be the difference between two members (and not the difference between a member and the mean), it saturates in average to $2PS(x)_k$ (since σ_k^2 saturates to $PS(x)_k$).

Author contributions. M.B., L.D.C. and S.V. designed the study. M.B. wrote the code and performed the computations. M.B., L.D.C. and S.V. interpreted the results and wrote the manuscript. F.D. helped in the selection of convective and stratiform events.

Competing interests. The authors have no competing interests.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

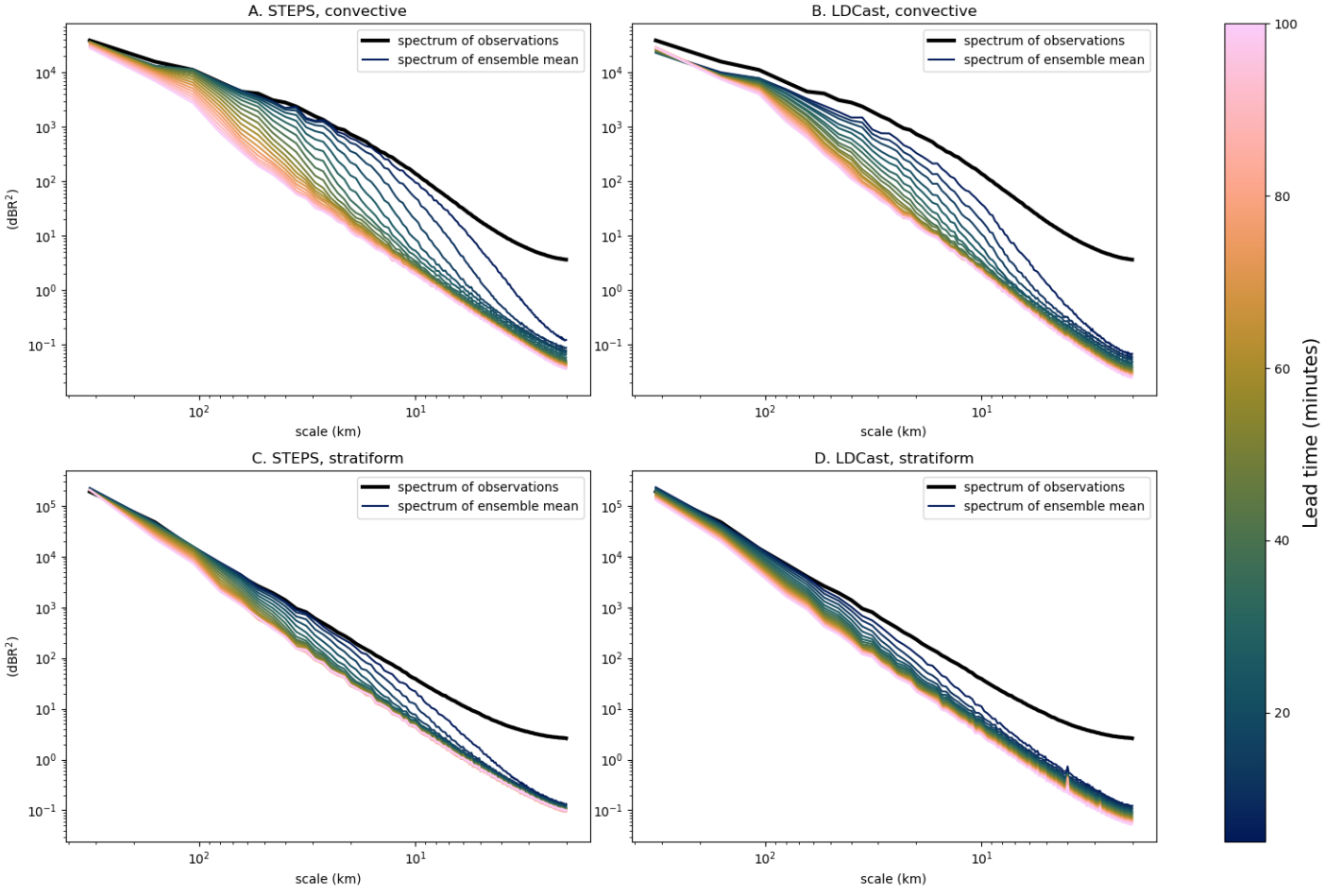

Figure B1. Comparison of power spectra of ensemble means averaged over events (thin colored lines) with the mean power spectrum of observations (thick black lines) for convective and stratiform cases, and for STEPS and LDCast. The color represents the lead time.

Acknowledgements. This research has been supported by the Belgian Federal Science Policy Office (BELSPO) under contract number B2/233/P2/PRECIP-PREDICT and through the FED-tWIN programme (Prf-2020-017).

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

References

Almeida, R., Otero, N., Ángel Fernández-Torres, M., and Ma, J.: On the Predictive Skill of Artificial Intelligence-based Weather Models for Extreme Events using Uncertainty Quantification, https://arxiv.org/abs/2511.17176, 2025.

Baño-Medina, J., Sengupta, A., Doyle, J. D., Reynolds, C. A., Watson-Parris, D., and Monache, L. D.: Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia, npj Climate and Atmospheric Science, 8, 92, 2025.

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, 2015.

Baño-Medina, J., Sengupta, A., Watson-Parris, D., Hu, W., and Delle Monache, L.: Toward Calibrated Ensembles of Neural Weather Model Forecasts, Journal of Advances in Modeling Earth Systems, 17, e2024MS004 734, https://doi.org/https://doi.org/10.1029/2024MS004734, 2025.

Bouallègue, Z. B., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F.: The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context, Bulletin of the American Meteorological Society, 105, E864 – E883, https://doi.org/10.1175/BAMS-D-23-0162.1, 2024.

Bowler, N. E., Pierce, C. E., and Seed, A. W.: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP, Quarterly Journal of the Royal Meteorological Society, 132, 2127–2155, https://doi.org/https://doi.org/10.1256/qj.04.100, 2006.

Bröcker, J., Driscoll, S., Necker, T., Rodríguez, J., Dacre, H., Harvey, N., and Ben Bouallègue, Z.: Verification of AI–based environmental forecasting systems: What can we do, what do we need to do, and what are the challenges?, Journal of the European Meteorological Society, 4, 100 032, https://doi.org/https://doi.org/10.1016/j.jemets.2026.100032, 2026.

Charlton-Perez, A. J., Dacre, H. F., Driscoll, S., Gray, S. L., Harvey, B., Harvey, N. J., Hunt, K. M., Lee, R. W., Swaminathan, R., Vandaele, R., et al.: Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán, npj Climate and Atmospheric Science, 7, 93, 2024.

Feng, J., Toth, Z., Zhang, J., and Peña, M.: Ensemble forecasting: A foray of dynamics into the realm of statistics, Quarterly Journal of the Royal Meteorological Society, 150, 2537–2560, https://doi.org/https://doi.org/10.1002/qj.4745, 2024.

Fortin, V., Abaza, M., Anctil, F., and Turcotte, R.: Why should ensemble spread match the RMSE of the ensemble mean?, Journal of Hydrometeorology, 15, 1708–1713, 2014.

Goudenhoofdt, E. and Delobbe, L.: Generation and Verification of Rainfall Estimates from 10-Yr Volumetric Weather Radar Measurements, Journal of Hydrometeorology, 17, 1223 – 1242, https://doi.org/10.1175/JHM-D-15-0166.1, 2016.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present., https://doi.org/10.24381/cds.adbb2d47, 2023.

Hua, Z., Hakim, G., and Anderson-Frey, A.: Performance of the Pangu-Weather deep learning model in forecasting tornadic environments, Geophysical Research Letters, 52, e2024GL109 611, 2025.

Imhoff, R. O., De Cruz, L., Dewettinck, W., Brauer, C. C., Uijlenhoet, R., van Heeringen, K.-J., Velasco-Forero, C., Nerini, D., Van Ginderachter, M., and Weerts, A. H.: Scale-dependent blending of ensemble rainfall nowcasts and numerical weather prediction in the open-source pysteps library, Quarterly Journal of the Royal Meteorological Society, 149, 1335–1364, https://doi.org/https://doi.org/10.1002/qj.4461, 2023.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

Journée, M., Goudenhoofdt, E., Vannitsem, S., and Delobbe, L.: Quantitative rainfall analysis of the 2021 mid-July flood event in Belgium, Hydrology and Earth System Sciences, 27, 3169–3189, https://doi.org/10.5194/hess-27-3169-2023, 2023.

Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511802270, standard textbook on NWP and data assimilation, 2003.

Lang, S., Alexe, M., Clare, M. C., Roberts, C., Adewoyin, R., Bouallègue, Z. B., Chantry, M., Dramsch, J., Dueben, P. D., Hahner, S., et al.: AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score, arXiv preprint arXiv:2412.15832, 2024.

Leinonen, J., Hamann, U., Nerini, D., Germann, U., and Franch, G.: Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification, arXiv preprint arXiv:2304.12891, 2023.

Mahesh, A., Collins, W. D., Bonev, B., Brenowitz, N., Cohen, Y., Elms, J., Harrington, P., Kashinath, K., Kurth, T., North, J., O'Brien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., and Willard, J.: Huge ensembles – Part 1: Design of ensemble weather forecasts using spherical Fourier neural operators, Geoscientific Model Development, 18, 5575–5603, https://doi.org/10.5194/gmd-18-5575-2025, 2025.

Necker, T., Wolfgruber, L., Kugler, L., Weissmann, M., Dorninger, M., and Serafin, S.: The fractions skill score for ensemble forecast verification, Quarterly Journal of the Royal Meteorological Society, 150, 4457–4477, 2024.

Olivetti, L. and Messori, G.: Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather, and GraphCast, Geoscientific Model Development, 17, 7915–7962, 2024.

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., and Engelke, S.: Validating deep learning weather forecast models on recent High-Impact extreme events, Artificial Intelligence for the Earth Systems, 4, e240 033, 2025.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al.: Gencast: Diffusion-based ensemble forecasting for medium-range weather, arXiv preprint arXiv:2312.15796, 2023.

Pu, J., Mu, M., Feng, J., Zhong, X., and Li, H.: A fast physics-based perturbation generator of machine learning weather model for efficient ensemble forecasts of tropical cyclone track, npj Climate and Atmospheric Science, 8, 128, 2025.

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., and Foresti, L.: Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0), Geoscientific Model Development, 12, 4185–4219, https://doi.org/10.5194/gmd-12-4185-2019, 2019.

Radford, J. T., Ebert-Uphoff, I., Stewart, J. Q., Musgrave, K. D., DeMaria, R., Tourville, N., and Hilburn, K.: Accelerating community-wide evaluation of AI models for global weather prediction by facilitating access to model output, Bulletin of the American Meteorological Society, 106, E68–E76, 2025.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al.: Skilful precipitation nowcasting using deep generative models of radar, Nature, 597, 672–677, 2021.

Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Monthly Weather Review, 136, 78–97, 2008.

Schreiber, T. and Schmitz, A.: Improved Surrogate Data for Nonlinearity Tests, Phys. Rev. Lett., 77, 635–638, https://doi.org/10.1103/PhysRevLett.77.635, 1996.

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., Coniglio, M. C., and Wandishin, M. S.: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership, Weather and Forecasting, 25, 263–280, 2010.

https://doi.org/10.5194/egusphere-2026-1460
Preprint. Discussion started: 23 April 2026

Seed, A. W., Pierce, C. E., and Norman, K.: Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme, Water Resources Research, 49, 6624–6641, 2013.

Uboldi, F. and Trevisan, A.: Multiple-scale error growth in a convection-resolving model, Nonlinear Processes in Geophysics, 22, 1–13, https://doi.org/10.5194/npg-22-1-2015, 2015.

Vallis, G. K.: Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulation, Cambridge University Press, 2 edn., 2017.

Zawadzki, I., Morneau, J., and Laprise, R.: Predictability of Precipitation Patterns: An Operational Approach, Journal of Applied Meteorology and Climatology, 33, 1562 – 1571, https://doi.org/10.1175/1520-0450(1994)033<1562:POPPAO>2.0.CO;2, 1994.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J.: Skilful nowcasting of extreme precipitation with NowcastNet, Nature, 619, 526–532, 2023.