https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Fractions Skill Score across spatial scales: statistical significance and discrimination

Fabrizio Nerozzi[1] and Rodica Tomozeiu[1]

[1]Agenzia regionale per la prevenzione, l'ambiente e l'energia dell'Emilia-Romagna (Arpae), Bologna, Italy

Correspondence: Fabrizio Nerozzi (fnerozzi@arpae.it)

Abstract. The Fractions Skill Score (FSS) is usually interpreted through its magnitude, by comparison with reference values such as the useful-skill level $(1 + f_0)/2$ or the expected score of a random forecast. Controlled synthetic experiments with summer convective cells, in which the correspondence between forecast and observed cells is known by construction, are used here to separate the magnitude of the score, its statistical significance against forecasts made of independently placed cells, and its ability to discriminate corresponding from unrelated forecasts, for single field pairs and for seasons of 90 aggregated pairs. Three results follow. First, reference values are not significance tests: used as decision rules for one forecast, they are exceeded by 0–99 % of unrelated forecasts depending on width, threshold, domain and number of cells, whereas Monte Carlo critical values keep false positives near 5 %. For aggregated scores, the critical value converges at rate $n^{-1/2}$ to the asymptotic limit of the score under the null, and the random-forecast reference of Antonio and Aitchison is this limit for their Bernoulli null when the forecast probability equals the mean observed fraction: a central value, not a percentile. Second, the mean FSS of corresponding forecasts rises steadily with the width, whereas discrimination is largest at a width set by the position error, 5–15 grid points of 2.2 km for cells displaced by up to their radius, and declines beyond it; restricting the windows to the interior of the domain roughly doubles that decline. Third, seasonal aggregation multiplies the separation by about $\sqrt{n}$, and the most discriminating width of the aggregated score stays at or next to that of a single day: it ranks forecasting systems of different capability, but does not increase the evidence carried by the score of any single forecast. Critical values are specific to the population verified and to the null against which the score is judged; for seasons the delta method gives an asymptotic approximation to them in closed form from the daily sums of the score, without simulation. The seasonal score evaluates a forecasting system; the single-case analysis shows at which scale the score of each of its forecasts carries most evidence.

1 Introduction

High-resolution forecasts can place realistic convective cells a few grid points away from their observed locations, and pointwise scores then penalize them twice: for a miss where precipitation occurred and for a false alarm where it was predicted. Neighbourhood verification alleviates this double penalty by comparing properties aggregated over finite spatial windows (Ebert, 2008, 2009), and the intercomparison of spatial verification methods by Gilleland et al. (2009) showed that neighbourhood, scale-separation and object-based methods target different aspects of forecast quality. Within this family, the Fractions Skill Score (FSS) introduced by Roberts and Lean (2008) compares the observed and predicted fractions of neighbourhoods

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



that exceed a precipitation threshold, and it has become one of the most widely used scores for high-resolution precipitation forecasts. Its dependence on neighbourhood width is part of the design: larger windows tolerate greater spatial displacement, and the score tends to increase with the window, for unrelated as well as for corresponding forecasts.

For this reason raw FSS values at different widths are not compared directly, and the score is read against reference values. Roberts and Lean (2008) introduced the useful-skill reference $(1+f_0)/2$, where f_0 is the observed event frequency: a forecast is considered skilful at the widths where its score exceeds it, and the smallest such width, the skilful scale, is used as a diagnostic of the spatial scale at which the forecast becomes useful. Roberts (2008) used the FSS to assess the spatial and temporal variation of forecast skill, Mittermaier and Roberts (2010) to identify skilful spatial scales and examine the effects of bias, and Mittermaier et al. (2013) for a long-term assessment of precipitation forecasts; Mittermaier (2021) analysed the limiting behaviour of the score and showed how its aggregation over cases should be carried out. A score above the reference is thus taken as evidence that the forecast places precipitation approximately in the right location. Whether this reading is supported by the statistical properties of the score is the subject of this study.

Studies of random and displaced forecasts have clarified what FSS values can be expected without genuine skill. Skok (2015) examined displaced rainbands in finite domains. Skok and Roberts (2016) studied the FSS of random precipitation fields, including boundary effects and the dependence on spatial coverage, alongside ECMWF forecasts, and Skok and Roberts (2018) analysed how the geometry of precipitation objects affects displacement estimates. Necker et al. (2024) examined FSS formulations for ensemble verification, a setting distinct from the one considered here. Antonio and Aitchison (2025) derived the score of a random forecast whose grid points are independently wet with the observed frequency and proposed it as a reference for interpreting the FSS, and Antonio (2026) examined theoretically how the FSS responds to errors in the mean frequency and in the spatial structure of a forecast. These studies describe where the FSS of random, displaced or erroneous forecasts lies on average, and show that this value depends on domain, coverage and object geometry.

An expected value, however, describes where scores lie on average, not how widely they spread. It does not indicate how frequently an unrelated forecast reaches a given score, nor how effectively the score distinguishes unrelated from corresponding forecasts. Neither a high FSS nor a value above a reference therefore establishes that the score is statistically significant. Hypothesis testing has been applied to forecast verification, but mainly to a different question: Hamill (1999) developed hypothesis tests, including resampling methods, to assess differences in skill between competing precipitation forecasts, and Gilleland (2013) introduced the spatial prediction comparison test, which accounts for spatial correlation when comparing forecast losses. Both approaches concern the relative performance of competing forecasts.

The present study addresses instead two questions about the FSS itself: whether it exceeds what would be expected for unrelated forecasts, at a specified false-positive rate, and how effectively it discriminates between unrelated and corresponding forecasts. The former requires the upper tail of the FSS distribution under a null hypothesis of unrelated forecast and observed precipitation; the latter requires both the null and the alternative score distributions. The questions are asked at two levels. For one field pair, the situation of a forecaster judging one forecast, the null distribution has no simple form and is obtained by simulation. For a season of many field pairs, the situation of a verification report, the FSS aggregated by summing numerators and denominators (Mittermaier, 2021) has an asymptotically normal distribution around a limit, and this limit links the calibrated

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

test to the random-forecast reference: for the Bernoulli null of Antonio and Aitchison (2025) the reference is exactly the value towards which the seasonal critical value converges.

Controlled synthetic experiments have a long tradition in spatial verification: idealized geometric and perturbed cases with prescribed errors were central to the intercomparison of spatial verification methods (Ahijevych et al., 2009; Gilleland et al., 2009, 2020), and random precipitation fields have been used to derive reference values for the FSS (Skok and Roberts, 2016; Antonio and Aitchison, 2025). Their purpose is not to reproduce real forecasts, but to expose the properties of a score under conditions in which the correct answer is known. The mechanism examined here, by which wider neighbourhoods bring unrelated as well as corresponding precipitation into common windows, follows from the definition of the FSS; how strongly it acts, and from which width, depends on the spatial population, as the numerical critical values do.

Two explicitly defined populations of forecast–observation field pairs are compared. In the null population G_0, forecast cells are placed independently of the observed ones, although both fields contain spatially coherent, realistic convective cells. In the alternative population G_q, a prescribed fraction q of the forecast cells is generated from observed cells by a bounded displacement and a change of size, shape and intensity. Under G_0 an independently placed forecast cell can still overlap an observed one by chance, and under G_q a prescribed correspondence need not produce a high FSS. Because the correspondence is known by construction, the ability of the FSS to detect it can be measured directly. The statistical tools are standard: Monte Carlo calibration of critical values, ROC analysis, a maximum-statistic test across widths (Westfall and Young, 1993) and the delta method for the aggregated score. Along the way, two technical choices that affect every FSS computation are compared on the same field pairs: the treatment of the domain boundary, by windows restricted to the interior, zero padding, reflection or the true precipitation around the domain, and the way the score of several cases is computed, as single scores, their mean, or the aggregated score.

Three results emerge. First, reference values are not significance tests: used as decision rules for one forecast, their false-positive rates vary from almost zero to almost one, although in these experiments each lies close to the ROC curve of the score; for aggregated scores, the random-forecast reference is the asymptotic limit of the calibrated critical value under its own null. Second, the magnitude of the score and its discriminating ability follow different scale dependences: beyond a width set by the position error of corresponding cells, widening the neighbourhood raises the mean FSS while the separation between corresponding and unrelated forecasts declines. Third, aggregation over a season multiplies the separation by about the square root of the number of cases at every width: it ranks forecasting systems of different capability, but it does not change the width at which the score of each single forecast carries most evidence of correspondence. Together they lead to a few practical recommendations on how FSS values should be calibrated and reported, given in Sect. 6.

These results organize the rest of the paper. Section 2 describes the synthetic populations of convective cells. Section 3 defines the FSS and its treatment of the domain boundary, the modes of computation over several cases, the Monte Carlo calibration, the asymptotic distribution of the aggregated score and the measures of discrimination, and compares the boundary treatments and the modes of computation. Section 4 presents the results for single field pairs and for seasons. Section 5 discusses the meaning of reference values, the two uses of the FSS for a season and for a day, and the implications for verification practice. Section 6 summarizes the findings, Appendix A quantifies the accidental overlap of independently placed

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

cells, and Appendix B shows why a Gaussian quadratic-form approximation cannot replace the Monte Carlo calibration for a single field pair.

2 Controlled synthetic experiments

This section describes the synthetic populations to be compared. Their key property is that the correspondence between forecast and observed precipitation systems is known by construction, so the ability of the FSS to detect it can be measured rather than assumed.

Convective cells. The fields represent hourly precipitation from isolated summer convective cells on a grid of 2.2 km spacing. The verified domains are squares of $W \in \{100, 200\}$ grid points (220 and 440 km), and each field contains $N = 4$ or 8 cells. A cell is built in five steps. (i) The peak of its primary core is drawn from a lognormal distribution with median 7.5 mm h^{-1} and logarithmic standard deviation 0.55, capped at 30 mm h^{-1}; the equivalent diameter of the cell envelope grows with the peak, $D = 40\,\mathrm{km}\,(p/7.5\,\mathrm{mm\,h^{-1}})^{0.6}\,e^{0.25\varepsilon}$ with $\varepsilon \sim \mathcal{N}(0,1)$, limited to 10–110 km, so that weak showers are small and intense cells large. (ii) The envelope is an ellipse of aspect ratio between 1 and 4 and random orientation, whose radius is modulated by the harmonics 2–6 with random amplitudes up to 0.12, giving a ragged outline. (iii) A weak background of 0.8–4 mm h^{-1}, decaying towards the edge and modulated by a smooth texture, fills the envelope, and Gaussian cores are added on top of it: the primary one and a Poisson number of secondary ones (mean $1.2\,D/40\,\mathrm{km}$, peaks 0.4–1 times the primary), placed along the major axis, with widths that increase with $\sqrt{D}$; light satellite patches (0.1–0.5 mm h^{-1}) appear around 60 % of the cells. (iv) A smooth random perturbation of $\log(R + 0.5\,\mathrm{mm\,h^{-1}})$ makes the contours irregular. (v) A steepness limiter bounds the change of $\log(R+0.5)$ between neighbouring grid points to 0.39, so that the field never jumps across more than one of the usual colour classes of precipitation maps in one grid step; peaks are kept and their surroundings raised, so that cores widen instead of being cut. Values below 0.1 mm h^{-1} are set to zero, and where cells overlap the larger value is kept. The organized multicell systems of 100–300 km that a full summer climatology would contain are excluded, because a single one would span the whole 220-km domain. The verified events are the exceedances of 2 and 5 mm h^{-1}; a threshold of 10 mm h^{-1}, which is not reached in about 45 % of the observed fields with four cells, is reported in the Supplement only. Figure 1 shows three realizations with $W = 100$ and $N = 4$; they are placed on a map of northern Italy for illustration only, since neither the generator nor the analysis uses the geography.

Placement and visibility. Cell centres are drawn uniformly over the domain extended by a belt of 16 km (7 grid points), and a cell is accepted only if at least 10 % of its area above 1 mm h^{-1} lies inside the verified domain; rejected placements are redrawn. The same rule applies to observed cells and to all forecast cells. Each field is composed on a canvas that extends 20 grid points (44 km) beyond the domain on every side, half the largest neighbourhood width, so that the precipitation just outside the verified domain is known; Sect. 3.1 explains how it is used.

Correspondence. A corresponding forecast cell is not a copy of the observed one. A realistic forecast typically places a convective cell near, but not exactly at, its observed location, with a somewhat different size, shape and intensity; neighbourhood verification is designed precisely to tolerate such errors. A corresponding cell is built in two steps:

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

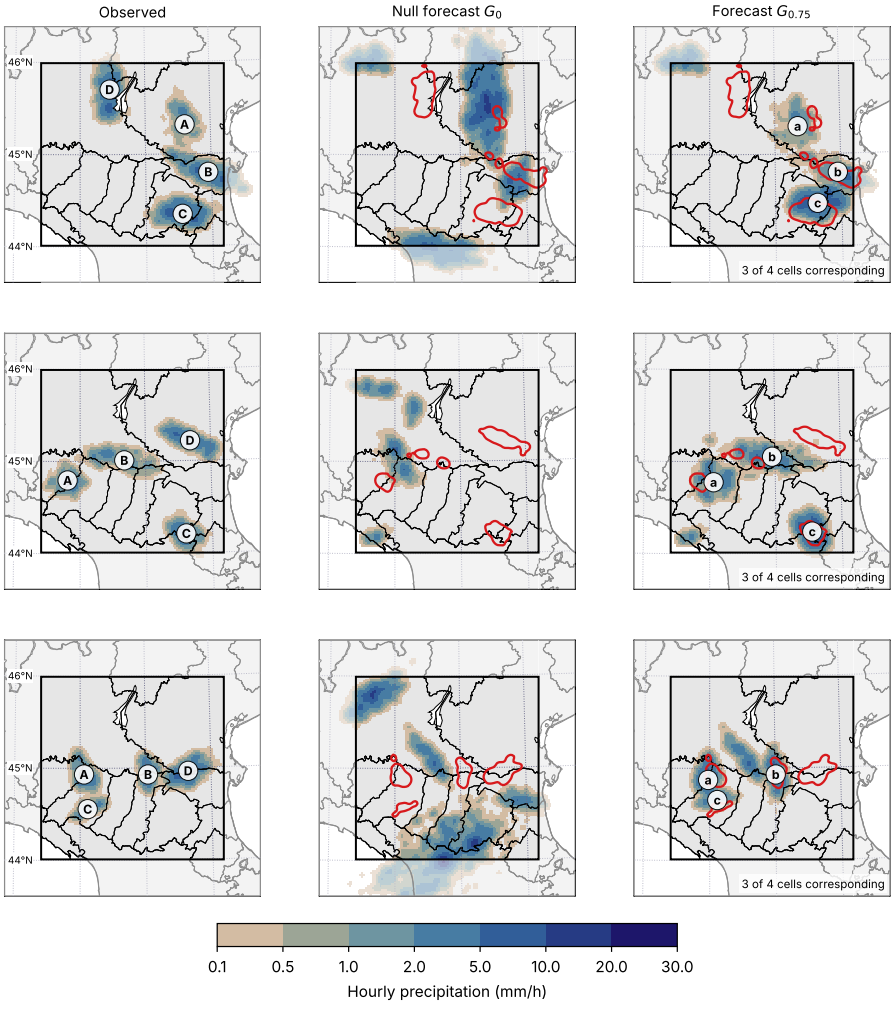


Figure 1. Three realizations of the generator in the main configuration (100×100 grid points of 2.2 km, four convective cells, maximum displacement $0.5\,D$), placed for illustration only on a map of northern Italy; the generator does not use the geography. Left: observed field; centre: forecast from the null population G_0, with all four cells independent; right: forecast from $G_{0.75}$, with three cells corresponding to observed ones, displaced and perturbed as described in Sect. 2, and one independent. The black square is the verified domain; outside it, the faded field is the belt of 20 grid points used by the belt treatment of the boundary. The red contour in the centre and right panels is the observed 2 mm h^{-1} isohyet. Observed cells are labelled A–D, at the centre of the cell, and the forecast cells corresponding to A, B and C are labelled a, b and c; independent forecast cells are not labelled. The three realizations were chosen among the first 300 seeds because their observed cells are separated and inside the domain, for clarity; in general cells may overlap or lie mostly outside the domain. The two forecasts of each realization share their random draws, so the independent cell on the right also appears in the centre panel. Colour map from Crameri et al. (2020). Land, lakes, coastline and national borders from Natural Earth (public domain); regional and provincial boundaries from ISTAT, via openpolis/geojson-italy (CC BY 4.0).

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

(1) *Size, shape and intensity.* The diameter, the aspect ratio and the peaks of the cores of the observed cell are multiplied by independent log-uniform factors between $1/1.15$ and 1.15, the background by a factor between $1/1.10$ and 1.10, the orientation is rotated by a normal angle with standard deviation 15° and the cores are shifted along and across the axis; all factors have median 1, so that correspondence implies no systematic change of size or intensity. The cell is then rendered anew, with new outline, texture and perturbation, so that it resembles the observed one without copying it.

(2) *Position.* The whole cell is displaced rigidly from the observed centre by a vector uniformly distributed over a disk of radius κD and rounded to whole grid points. The main experiments use $\kappa = 0.5$ (mean displacement $0.33\,D$); Sect. 4.6 doubles it to $\kappa = 1$ (mean displacement $0.65\,D$).

The displaced cell must satisfy the visibility rule; otherwise a new displacement is drawn, and after 50 rejected positions a new perturbed cell. Independent forecast cells are drawn without reference to observed positions and without exclusion of accidental overlaps.

Null and alternative populations. For each configuration, G_0 denotes the population of field pairs in which every forecast cell is independent of the observed ones. Their placement does not depend on the observed cells, although rainfall within each field is spatially coherent, and physical or shared-neighbourhood overlap may occur by chance. The alternatives form the family $\{G_q : q \in (0,1]\}$, evaluated at $q \in \{0.25, 0.50, 0.75, 1\}$: exactly qN of the N forecast cells are corresponding and the remaining $(1-q)N$ independent, so that qN is a whole number for both $N = 4$ and $N = 8$. Thus q is a fraction of cells, not of precipitating area, and it describes how a pair of fields is generated: it is not estimated from the FSS and does not count cells subsequently identified as matching. Above a threshold the correspondence of the events can be smaller than q, because a cell and its corresponding cell need not both exceed the threshold. Over 8000 corresponding cells ($W = 100$, $N = 4$, displacement $0.5\,D$), 98 % of the observed cells with an area above $2\,\mathrm{mm\,h^{-1}}$ inside the domain have a corresponding forecast cell above the threshold. Above $5\,\mathrm{mm\,h^{-1}}$, 23 % of the corresponding cells have no core above the threshold in either field, 57 % have one in both and 19 % in one field only, in equal shares, so that 86 % of the observed cores have a forecast counterpart (84 % for the displacement D). The effective correspondence of the events above $5\,\mathrm{mm\,h^{-1}}$, the fraction of observed cores with a corresponding forecast core, is therefore about $0.86\,q$, about 0.64 for $q = 0.75$; the shortfall is symmetric and introduces no bias in the forecast coverage. The FSS scores the thresholded fields rather than identified objects, so it cannot distinguish planned pairings from accidental proximity, and a field pair with $q > 0$ need not receive a statistically significant score. The forecast cells have the same number and the same distributions of size and intensity as the observed ones in both populations, so the alternatives have essentially no frequency bias (Sect. 4.7). For each configuration, 5000 null pairs are generated for calibration, 5000 (for $W = 200$, 2000) for validation, and 5000 (2000) alternative realizations for each displacement.

A fixed number of cells. The number of cells is the same in the observed and in the forecast field and fixed within a configuration, as in a verification stratified by the number of convective cells of the day. A variable number, more realistic for a sample of summer days, mixes in one null population days whose scores have very different distributions; Sect. 4.11 shows the consequence with a Poisson number of cells of mean 4.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Accidental overlap under the null. In G_0 any positive score arises from chance coincidences between independently placed cells. Their frequency depends on the domain size and on the number of cells: the probability that two given cells intersect above $2\,\mathrm{mm\,h^{-1}}$ is 0.064 for $W = 100$ and 0.016 for $W = 200$, and doubling the number of cells in both fields multiplies the number of chance intersections by about four, from 1.0 to 3.9 per field pair for $W = 100$ (Appendix A). Both changes shift the whole score distribution of G_0, downwards for the larger domain and upwards for more cells. Two cells need not touch to contribute to the same neighbourhood, however: it is enough that they fall within a common window of width L, and such shared-neighbourhood coincidences become more frequent as L grows. This makes the distribution of the FSS under G_0 difficult to derive analytically for a single field pair; Sect. 3 estimates it by Monte Carlo simulation and derives its limit for many aggregated pairs.

3 Statistical verification

Given the null population G_0 and the alternatives G_q of Sect. 2, three questions are asked of the FSS at each neighbourhood width: whether the score is statistically significant against G_0; how often that test rejects H_0 when the field pairs are drawn from G_q; and how well the score separates the two populations over all possible critical values. The questions are asked for a single field pair, the situation of a forecaster judging one forecast, and for a season of n field pairs, the situation of a verification report. Section 3.1 defines the score and its treatment of the domain boundary, Sect. 3.2 the three ways of computing it over several cases, Sect. 3.3 the Monte Carlo calibration, Sect. 3.4 the asymptotic distribution of the aggregated score and its relation to the random-forecast reference of Antonio and Aitchison (2025), Sect. 3.5 power and ROC discrimination, and Sect. 3.6 the sensitivity experiments; Sect. 3.7 compares the boundary treatments and the modes of computation on the same field pairs, which motivates the choices used in the rest of the paper.

3.1 Fractions skill score and the domain boundary

For square neighbourhoods of width L, let o_k and f_k be the observed and forecast event fractions in the kth neighbourhood. For one field pair the FSS is

$$\mathrm{FSS}(L) = 1 - \frac{D}{S}, \qquad D = \sum_{k=1}^{M} (f_k - o_k)^2, \quad S = \sum_{k=1}^{M} (f_k^2 + o_k^2), \tag{1}$$

where D is the sum of squared differences of the fractions and S its value for forecast and observed fractions that never overlap. The score is undefined when neither field contains the event; such pairs are excluded at the threshold concerned.

The set of neighbourhoods matters near the domain boundary, and four conventions are compared. In the *valid* convention only windows entirely inside the domain are used, $M = (W - L + 1)^2$, so that grid points near the edge enter fewer windows than central ones. The other three use one window centred on every grid point of the domain, $M = W^2$ (for even L the window of point i spans $i - L/2, \ldots, i + L/2 - 1$), and differ in the values assumed outside the domain: zero (*zero padding*, the usual default of image filters), the mirror image of the domain field (*reflection*), or the true field (*belt*). The last is possible when observations and forecasts exist beyond the verified area, as for radar composites and limited-area models whose domains are

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

larger than the verification region, and it is the reference treatment of this study: every grid point of the domain is the centre of one window and no window contains invented values, although, as in the other conventions, points within $L/2$ of the edge enter fewer windows centred inside the domain than central points do. Section 3.7 compares the four conventions on the same field pairs.

3.2 The FSS of a set of cases

For n field pairs (days) with sums D_i and S_i, three summaries are compared: the score of each single day, $\mathrm{FSS}_i = 1 - D_i/S_i$; the mean of the daily scores, $n^{-1}\sum_i \mathrm{FSS}_i$; and the aggregated score

$$\mathrm{FSS}_n = 1 - \frac{\sum_{i=1}^{n} D_i}{\sum_{i=1}^{n} S_i}, \tag{2}$$

recommended by Mittermaier (2021), in which every day is weighted by its denominator, so that days with much precipitation count more. A season is represented by $n = 90$ days drawn at random, without replacement, from the field pairs of one population: for G_q every day of the season has the same capability q, because the purpose is to describe the seasonal score of a forecasting system with a given capability, not to reproduce a real season.

3.3 Monte Carlo calibration of the null

The critical value at width L for a configuration with $N \in \{4, 8\}$ cells and domain width $W \in \{100, 200\}$ is

$$\mathrm{FSS}_{\mathrm{crit}}(L; N, W) = Q_{0.95}\left[\mathrm{FSS}(L) \mid H_0, N, W\right], \tag{3}$$

where H_0 states that the pair is drawn from G_0. It is calibrated for a fully specified population, including number, sizes, intensities and placement rules of the cells; the frequency and organization of observed events are part of the null, and their variability between realizations is integrated into it. A pair is classified as significant at width L when its FSS strictly exceeds the critical value. For single days the critical value is the empirical 95th percentile of the 5000 calibration pairs, and the false-positive rate is checked on the independent validation pairs. For seasons the critical value is the 95th percentile of 2000 seasons drawn from the calibration pairs and the false-positive rate is estimated on 2000 seasons drawn from the validation pairs. Seasons drawn from the same finite pool share days, so that their mean deviates from the population mean by an amount that is not negligible for $n = 90$; the seasonal false-positive rate is therefore cross-validated, pooling calibration and validation pairs, splitting them at random into two halves 100 times and averaging the rates obtained with critical values from one half and seasons from the other.

The single-pair distribution. For one field pair, the distribution of the FSS under H_0 cannot be derived analytically: neighbouring windows share most of their grid points, so their squared differences are strongly dependent (Bretherton et al., 1999); numerator and denominator are computed from the same fractions; and the fractions are bounded, frequently zero and spatially clustered, far from Gaussian assumptions. In addition, the null distribution depends on number, size and placement of the cells and on the domain boundary (Sect. 2). Appendix B shows why a Gaussian quadratic-form approximation does not replace the simulation.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Joint test across widths. When several widths are inspected and a forecast is declared significant if any of them exceeds its critical value, false positives exceed 5 %, a multiple-testing problem emphasized more generally by Wilks (2016). A joint test based on the maximum of the per-width ranks (Westfall and Young, 1993) controls this rate: each score is expressed as its rank $U(L)$ among the null calibration scores at the same width, the statistic $T = \max_L U(L)$ records the most favourable width, and its 95th percentile over the calibration pairs is the joint critical value. The test assumes neither independence between widths nor a particular form of dependence.

3.4 The asymptotic distribution of the aggregated score

Unlike the single-pair score, the aggregated score of Eq. (2) has a simple limiting distribution, because the days of a season are independent. Let $\mu_D = \mathrm{E}[D_i]$ and $\mu_S = \mathrm{E}[S_i]$ under H_0, and $r = \mu_D/\mu_S$. By the law of large numbers $\mathrm{FSS}_n \to \mathrm{FSS}_\infty = 1-r$, the limit analysed by Mittermaier (2021). By the central limit theorem, the daily means $\overline{D}$ and $\overline{S}$ are jointly normal for large n, and their ratio is approximately normal (Hinkley, 1969); a first-order expansion around (μ_D, μ_S), the delta method, gives

$$\mathrm{FSS}_n \approx \mathrm{FSS}_\infty - \frac{1}{n}\sum_{i=1}^{n}\frac{D_i - rS_i}{\mu_S}, \qquad \mathrm{FSS}_{\mathrm{crit}}(n) \approx \mathrm{FSS}_\infty + z_{0.95}\,\frac{\sigma_1}{\sqrt{n}}, \qquad \sigma_1^2 = \frac{\mathrm{Var}(D_i - rS_i)}{\mu_S^2}, \tag{4}$$

with $z_{0.95} = 1.645$. The limit is not a χ^2 distribution: D_i and S_i are sums of squares within one day, but the seasonal score is a ratio of two means over independent days, and it is their joint normality that matters. The approximation can be refined with the second-order bias of the ratio, $-(r\,\mathrm{Var}\,S_i - \mathrm{Cov}(D_i, S_i))/(n\mu_S^2)$, and a Cornish–Fisher correction of $z_{0.95}$ for the skewness γ of the daily terms, $z \to z + (z^2-1)\gamma/(6\sqrt{n})$. The convergence is slow where the daily terms are strongly skewed, that is for intense thresholds and large windows, where few independent windows contribute.

Equation (4) has two consequences. First, a fixed threshold, such as a reference value, does not keep a fixed false-positive rate: as the season grows, the null seasonal scores concentrate around FSS_∞, so that a fixed threshold above FSS_∞ is exceeded by fewer and fewer null seasons and one below it by almost all. The calibrated critical value is not fixed: it decreases towards FSS_∞, approximately as $z_{0.95}\sigma_1/\sqrt{n}$, and keeps the false-positive rate at the nominal level for every n. The one-sided test is consistent: for any alternative whose limit exceeds FSS_∞ its power tends to one (for a limit below FSS_∞ it would tend to zero), so that for long seasons the significance of a system is nearly certain, and the relevant questions become the size of its departure from the null and the choice of the null itself, since a small misspecification of the null is also detected with near certainty. Second, the alternatives also converge to their own limits, $\mathrm{FSS}_\infty(q)$, with spreads of order $\sigma_1/\sqrt{n}$; the standardized separation between two populations grows as $\sqrt{n}$, and the AUC of the seasonal score tends to one when the limit of the alternative exceeds that of the null (to zero if it lies below). For the mean of the daily scores this multiplies the separation by $\sqrt{n}$ at every width and leaves the width of largest separation unchanged. The aggregated score weights days by their denominators, so that its width of largest separation need not coincide with that of a single day; in these experiments it is the same or one step of the widths away (Sect. 3.7).

The random-forecast reference as an asymptotic limit. Antonio and Aitchison (2025) derived, as the ratio of the expected numerator and denominator, the FSS of a random forecast whose grid points are independently wet with probability p, and

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

proposed

$$\mathrm{FSS}_{\mathrm{random}}(L) = \frac{2m^2}{2m^2 + p(1-p)/L^2 + s^2} \tag{5}$$

as a reference, where m and s^2 are the mean and variance of the observed fractions over all windows and p is set equal to the observed frequency. For a forecast independent of the observations with $\mathrm{E}[f_k] = p$ and $\mathrm{Var}(f_k) = p(1-p)/L^2$, the expectations of the sums give $\mathrm{FSS}_\infty = 1-(\mathrm{Var}f+p^2+\mathrm{E}[o^2]-2pm)/(\mathrm{Var}f+p^2+\mathrm{E}[o^2])$, which equals Eq. (5) when $p = m$. The reference of Antonio and Aitchison (2025) is therefore the asymptotic limit FSS_∞ of the aggregated score under their Bernoulli null. Two consequences deserve emphasis. First, it is the centre of the null distribution of a long aggregation, not one of its tails: as n grows, every percentile of the null seasonal scores, the 5th, the median and the 95th alike, converges to it, and the reference is the limit of the calibrated critical value only because all percentiles collapse onto the same value. At finite n the critical value lies above it by about $z_{0.95}\sigma_1/\sqrt{n}$: with valid windows, 2 mm h^{-1} and $L = 20$, the reference is 0.289, and the 95th percentile of the null scores is 0.434 for a single day and 0.315, 0.302 and 0.293 for seasons of 30, 90 and 365 days. Used as a threshold, the reference therefore defines a test whose level is not 5 % but depends on n, of the order of one half (Sect. 4.9). Second, being a ratio of expectations, it is not the expected score of a single case: under the same null the mean single-day FSS is 0.314. The reference thus describes a random forecasting system over many cases; it is neither a significance threshold nor the typical score of one random forecast. The identity, and these numbers, are obtained numerically by pairing every observed field of the null ensembles with two Bernoulli forecasts drawn on the whole extended canvas, one with constant p equal to the expected observed frequency of the configuration and one with p equal to the observed frequency of the day. The null of independent cells, G_0, has its own limit, lower than the Bernoulli one, because a forecast made of coherent cells has more variable fractions than a field of independent grid points (Sect. 4.9).

3.5 Power and ROC discrimination

For a given width L, let $Y_{0,L}$ denote the score of a field pair (or season) drawn from G_0 and $Y_{q,L}$ that of one drawn from G_q, with cumulative distribution functions $\mathcal{F}_{0,L}$ and $\mathcal{F}_{q,L}$. The critical value of Eq. (3) at nominal level $\alpha = 0.05$ is $c_\alpha(L) = \mathcal{F}^{-1}_{0,L}(1-\alpha)$, estimated from the calibration pairs (or seasons) as $\widehat{c}_\alpha(L)$, their empirical $(1-\alpha)$ quantile. The power of the test is

$$\pi_L(q) = \Pr\{Y_{q,L} > c_\alpha(L)\}, \tag{6}$$

estimated as the rejection rate, the fraction of alternative pairs whose score exceeds $\widehat{c}_\alpha(L)$; computed on null validation pairs, the same fraction estimates the false-positive rate. Rejecting H_0 for a pair drawn from G_0 is a Type I error, whose probability the critical value fixes at α; failing to reject it for a pair drawn from G_q is a Type II error, with probability $1-\pi_L(q)$, the fraction of forecasts that do correspond to the observations but are not recognized as such. The detection rate refers to whole field pairs: it tends to one for every $q > 0$ as discrimination improves, and it is not an estimate of q, which counts cells. The area under the ROC curve,

$$\mathrm{AUC}_L(q) = \Pr(Y_{q,L} > Y_{0,L}) + \tfrac{1}{2}\Pr(Y_{q,L} = Y_{0,L}), \tag{7}$$

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

summarizes discrimination over all false-positive levels; 0.5 is chance level, unity perfect ranking. The AUC is a property of the FSS used as a test statistic for the specified null and alternative, not a measure of forecast quality; its uncertainty is assessed with 500 bootstrap resamples of field pairs, each pair's seven-width vector resampled as a unit. When the AUC saturates, as for seasons, discrimination is measured by the standardized separation of the two marginal distributions, $(\overline{Y}_q - \overline{Y}_0)/(s_q^2 + s_0^2)^{1/2}$. The same quantities computed between two alternatives, $G_{q'}$ against G_q with $q' > q$, measure how well the score separates forecasting systems of different capability; since the realizations of G_q and $G_{q'}$ share their observed fields, the paired difference $Y_{q'} - Y_q$, whose variance includes $-2\,\mathrm{Cov}(Y_{q'}, Y_q)$, is also reported.

Dependence across q. For each observed cell the generator draws a corresponding candidate and an independent one; the field for a given q takes the corresponding candidate for the first qN cells and the independent one for the others, and all five fields are compared with the same observed field. This common-random-number design makes the scores at different q dependent; it does not affect the results for any single q, estimated against independently generated null pairs, and it makes the comparison between capabilities a paired one.

Population comparability. Correspondence should not change how much precipitation is forecast, otherwise differences in FSS could reflect forecast coverage rather than correspondence. For each alternative realization the forecast coverage is compared with that of the independent ($q = 0$) forecast of the same realization.

3.6 Sensitivity experiments

Position error. The alternatives of all configurations are generated twice, with $\kappa = 0.5$ and $\kappa = 1$, from the same random seeds, so that the two sets share their observed fields. The null ensembles do not change.

Variable number of cells. All the experiments above use a fixed number of cells. Only as a sensitivity experiment, reported in Sect. 4.11, the $W = 100$ configuration is repeated with a number of cells $N_o \sim \mathrm{Poisson}(4)$, the same in the observed and in the forecast field, to show what happens when a single calibration is applied to days with different numbers of cells. For q the number of corresponding cells is $\lfloor qN_o \rfloor$ or $\lfloor qN_o \rfloor + 1$, the latter with probability $qN_o - \lfloor qN_o \rfloor$ (one uniform number per realization, common to all q), so that the expected fraction of corresponding cells is q for every N_o.

Cell morphology. For $W = 100$ and $N = 4$, two further ensembles of 1500 calibration, 1500 validation and 1500 alternative pairs (displacement $0.5\,D$) change one property of the cells at a time: the amplitude of the harmonics that make the outline ragged, from 0.12 to 0.25, and the amplitude of the random perturbation of $\log(R + 0.5)$ that makes the contours irregular, from 0.35 to 0.70. The ensembles use their own seeds and are compared with the reference ensemble by bootstrap, with 1000 independent resamples of the two samples.

Boundary treatments. All four conventions of Sect. 3.1 are computed for every field pair, so that their comparison is paired; differences of AUC are tested with 1000 paired bootstrap resamples.

Implementation check. The FSS computed with summed-area tables was compared with two established implementations, the `scores` package (Leeuwenburg et al., 2024), for valid windows, and `pysteps` (Pulkkinen et al., 2019), for zero padding at odd widths, on 40 regenerated field pairs: the largest absolute difference is 4.4×10^{-16} in 1120 and 480 comparisons respectively. Regenerating sampled blocks of every configuration from their seeds reproduces the stored arrays exactly.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

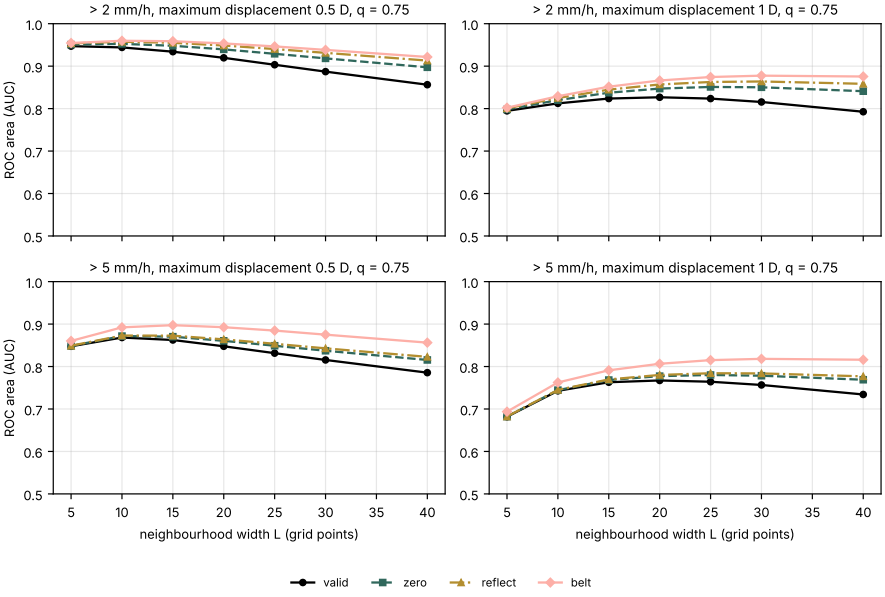


Figure 2. AUC of the single-day FSS versus neighbourhood width for the four treatments of the domain boundary (valid windows, zero padding, reflection and belt), computed on the same field pairs ($W = 100$, $N = 4$, $q = 0.75$), for 2 and 5 mm h^{-1} (rows) and maximum displacements $0.5\,D$ and D (columns).

3.7 Boundary treatment and mode of computation compared

Boundary treatment. The four conventions of Sect. 3.1 were compared on the same field pairs of the $W = 100$, $N = 4$ configuration (Fig. 2). For corresponding forecasts with $q = 0.75$, the AUC is ordered valid $<$ zero padding $<$ reflection $<$ belt at every width, threshold and displacement, and 167 of the 168 paired differences (six contrasts, seven widths, two thresholds, two displacements) are significant at the 5 % level. The differences grow with the width: between belt and valid windows, from 0.007 at $L = 5$ to 0.065 at $L = 40$ for 2 mm h^{-1}, and from 0.013 to 0.071 for 5 mm h^{-1} ($\kappa = 0.5$); with $\kappa = 1$ they reach 0.083 and 0.082. Valid windows lose discrimination at large widths because the windows that remain are increasingly central, so that cells near the edge, which a forecast places right or wrong as often as central ones, are underweighted; zero padding fills the windows centred near the edge with assumed dry points, and reflection with mirror images of the cells inside, whereas the belt uses the precipitation that is actually there. The choice also moves the width of maximum discrimination: with valid windows the AUC of $q = 0.75$, $\kappa = 0.5$, 2 mm h^{-1} is largest at $L = 5$ (0.947) and falls to 0.856 at $L = 40$; with the belt it is largest at $L = 10$ (0.960) and falls only to 0.922. In the 200×200 domain, where the edge is a smaller part of the domain, the differences are two to three times smaller and keep the same ordering, although 12 of the 168 differences, all at 5 mm h^{-1} and $L \leq 15$ between valid, zero and reflected windows, are not significant. Part of the loss of discrimination at large widths reported with valid windows is therefore an effect of the boundary, not of the score; the remaining results use the belt.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Mode of computation. Figure 3 compares the separation between corresponding forecasts and the null for single days, for the mean of 90 daily scores and for the aggregated score of 90 days, the latter two divided by $\sqrt{90}$. The mean of the daily scores reproduces the single-day separation multiplied by $\sqrt{n}$ almost exactly, as it should for a mean of independent terms, and has the same dependence on width. The aggregated score has the same shape but a smaller separation: weighting the days by their denominator gives more weight to days with much precipitation, so that a season of 90 days carries the information of 39–77 days with equal weights, the fewer for the higher threshold. The width of largest separation is the same for the mean of the daily scores as for a single day in 60 of the 64 combinations of configuration, threshold (2 and 5 mm h^{-1}), displacement and capability $q > 0$, and one step of the widths away in the other four, a sampling effect. For the aggregated score, whose weights differ between days, it is the same in 46 of the 64 combinations and one step away, in either direction, in the other 18; evaluating the aggregated score at the single-day width loses at most 5.1 % of its largest separation. The separation curves are flat near their maxima, so that these shifts are small, but the invariance is exact only for the mean of the daily scores. What a season does is multiply the separation by about $\sqrt{n}$ (Sect. 3.4), so that for $n = 90$ the AUC of every capability $q \geq 0.25$ against the null is practically one.

4 Results

The results are given first for single field pairs (Sects. 4.1–4.6), then for seasons of aggregated pairs (Sects. 4.9–4.10). Unless stated otherwise they refer to the belt treatment of the boundary, the 100×100 domain with four cells, the threshold of 2 mm h^{-1} and the maximum displacement $0.5\,D$; the other configurations are summarized in the text and shown in the Supplement.

4.1 Critical values and their validation

Under G_0 the FSS of a single field pair is far from negligible and grows with the width (Fig. 4, dotted line): its mean rises from 0.07 at $L = 5$ to 0.31 at $L = 40$ for 2 mm h^{-1}, and the critical value from 0.27 to 0.66. At 5 mm h^{-1} the null score is zero for 71 % of the pairs at $L = 5$, because no window contains both events, and for 5 % at $L = 40$; the critical value rises from 0.09 to 0.53. Applied to the independent validation pairs (Table 1), the critical values give false-positive rates of 4.6–5.9 % for $W = 100$ (5000 pairs) and 3.7–6.6 % for $W = 200$ (2000 pairs, binomial standard error 0.5 percentage points), at both thresholds and all widths. Under the null, testing every width and declaring significance at any of them gives 6.6–10 % false positives; the joint test restores 3.9–6.6 %.

4.2 Published references used as decision rules

Table 2 applies to the same validation pairs the reference values used in the literature: the observed frequency f_0 (Roberts and Lean, 2008), the useful-skill reference $(1 + f_0)/2$, its small-f_0 form 0.5 (Mittermaier and Roberts, 2010; Skok and Roberts, 2016), and the random-forecast reference of Eq. (5) (Antonio and Aitchison, 2025), computed with the frequency and the fractions of each observed field. None of them controls the false-positive rate. Exceeding f_0 is reached by 38–93 % of the independent forecasts for $W = 100$, $N = 4$, and by up to 99 % for $N = 8$. The useful-skill reference is conservative at small

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

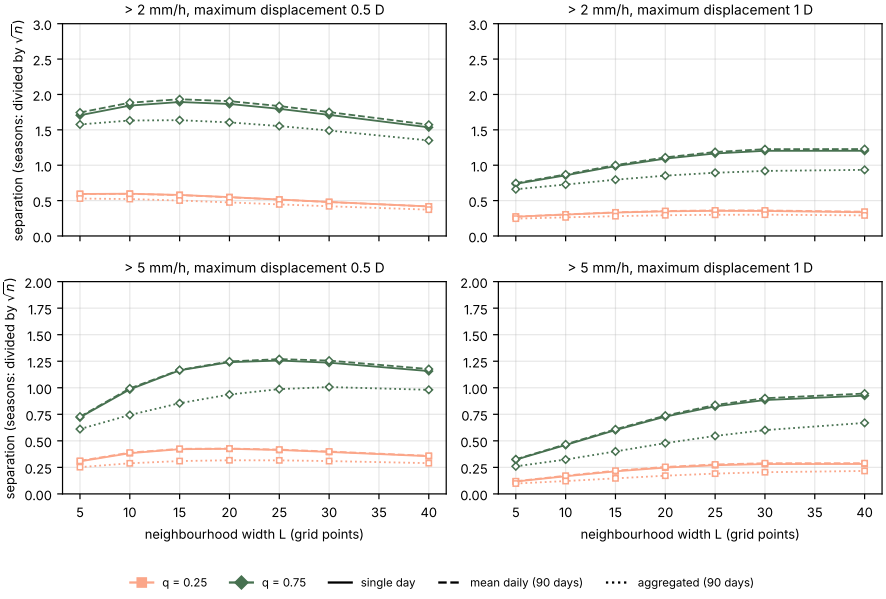


Figure 3. Standardized separation between corresponding forecasts ($q = 0.25, 0.75$) and the null versus neighbourhood width, for three ways of computing the FSS ($W = 100$, $N = 4$, belt). Solid: separation between the distributions of the FSS of single days. Dashed: separation between the distributions of the mean of the daily FSS over seasons of 90 days, divided by $\sqrt{90}$. Dotted: the same for the aggregated score of 90 days, Eq. (2), divided by $\sqrt{90}$. A mean of 90 independent days has a spread $\sqrt{90}$ times smaller than one day, so that its separation is $\sqrt{90}$ times larger: once divided by $\sqrt{90}$, the dashed curve coincides with the solid one, which shows that averaging over days multiplies the separation by $\sqrt{n}$ at every width and leaves its dependence on width unchanged. The dotted curve lies below because the aggregated score weights days by their denominator; the squared ratio of the dotted to the solid curve is the equivalent number of equally weighted days, 39–77 of the 90.

widths and permissive at large ones: its false-positive rate is 0.1 % at $L = 5$ and 14.1 % at $L = 40$ for $W = 100$, $N = 4$, 31.6 % at $L = 40$ for $N = 8$ (Supplement Table S4), but below 2 % at every width in the 200×200 domain. A low false-positive rate, however, comes with a low detection rate (Table 2): at $L = 5$ the useful-skill reference and 0.5 recognize only 20–26 % of the forecasts with $q = 0.75$ at 2 mm h^{-1} and 4.1–4.3 % at 5 mm h^{-1}, few false alarms but also few hits, whereas at $L = 40$ they recognize 84–86 % and 61.8–62.2 % at false-positive rates of 14–17 % and 7.1–7.4 %. The random-forecast reference gives 5–19 % at each width and 15–25 % when any width is accepted, in all configurations.

Figure 5 places each criterion on the ROC curve of the FSS, the curve traced by constant thresholds on the score. Except for 0.5, the references depend on the observed field, so the rules they define need not lie on that curve; in these experiments, however, the pair (false-positive rate, detection rate) of every reference lies within 4 percentage points of it, and a constant critical value with the same false-positive rate detects about as many corresponding forecasts (Table 2). At $L = 20$, the useful-skill reference detects 68.5 % of the forecasts with $q = 0.75$ at a false-positive rate of 2.6 %, and the random-forecast reference 92.1 % at 15.8 %; the calibrated test, by construction at 5 %, detects 78.2 %. The information that the references take from

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

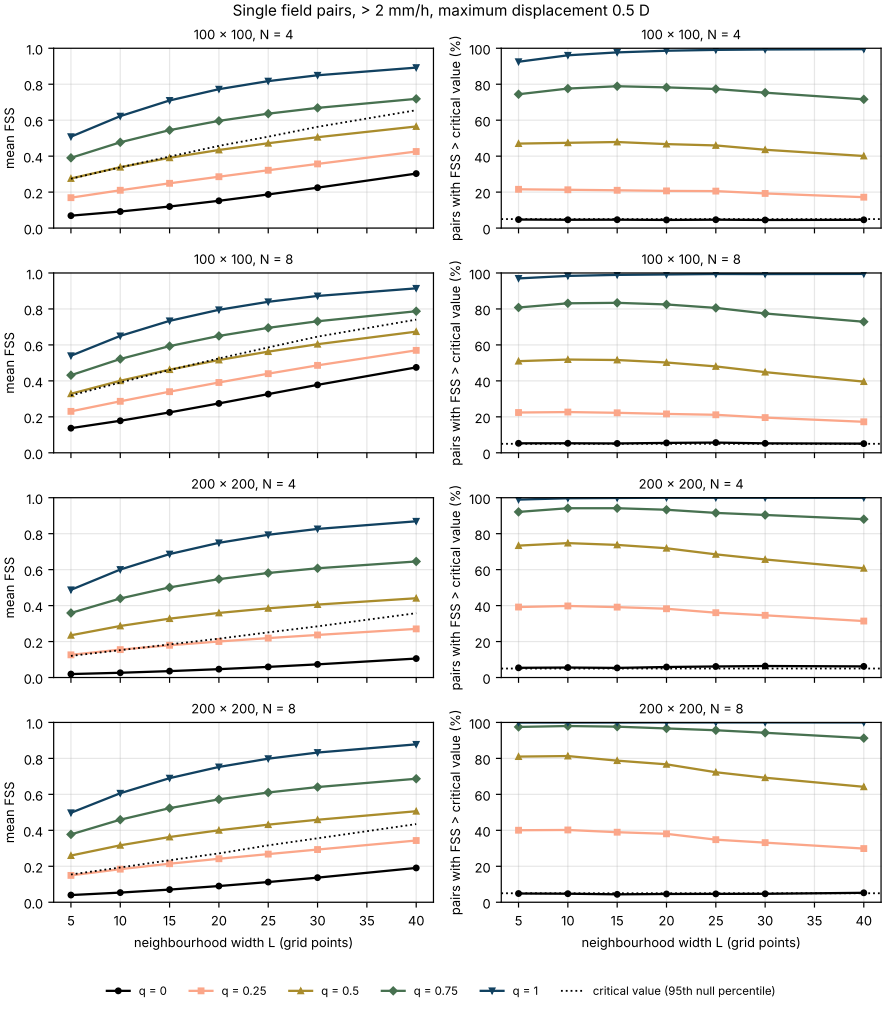


Figure 4. Mean FSS (left) and rejection rate (right) of single field pairs across neighbourhood widths at 2 mm h^{-1}, in the four configurations (rows); belt, maximum displacement $0.5\,D$. Colours and markers identify the fraction q of corresponding cells; black, with circles, the separately generated $q = 0$ null diagnostic. Left: the dotted line is the critical value (95th percentile of the null FSS); a mean below it can coexist with a large fraction of significant pairs. Right: percentage of field pairs whose FSS exceeds the critical value (rejection rate): the power, i.e. the fraction of corresponding forecasts recognized as significant, for $q > 0$, and the false-positive rate for $q = 0$; the dotted line marks 5 %. For $q > 0$ the complement of the plotted percentage is the Type II error rate, the fraction of corresponding forecasts not recognized. The same figure for 5 mm h^{-1} is Supplement Fig. S1.

the observed field adds little discrimination: the AUC of the differences $\mathrm{FSS} - r(O)$, which use it fully, differs from that of the FSS by between -0.06 and $+0.01$ for the three observation-dependent references, both thresholds, both displacements, $q = 0.25$ and 0.75 and all widths, and it has the same dependence on width (Supplement Table S5). A higher point on the figure

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table 1. Critical values of the FSS at $L = 20$ (belt) and their validation. Single days: 95th percentile of 5000 null calibration pairs, with a 95 % percentile-bootstrap interval (1200 resamples), false-positive rate (FPR) on the independent validation pairs and power for $q = 0.75$; seasons of 90 days: critical value of the aggregated score, cross-validated FPR and power for $q = 0.25$. Maximum displacement $0.5\,D$; rates in per cent.

		Single day				Season of 90 days		
Configuration	Threshold	Critical value	95 % interval	FPR	Power $q = 0.75$	Critical value	FPR	Power $q = 0.25$
100×100, $N = 4$	$> 2\,\mathrm{mm\,h^{-1}}$	0.457	[0.441, 0.473]	5.0	78.2	0.193	5.9	100.0
	$> 5\,\mathrm{mm\,h^{-1}}$	0.322	[0.300, 0.338]	5.3	61.4	0.111	6.0	99.8
100×100, $N = 8$	$> 2\,\mathrm{mm\,h^{-1}}$	0.525	[0.515, 0.538]	5.4	82.5	0.314	5.3	100.0
	$> 5\,\mathrm{mm\,h^{-1}}$	0.392	[0.378, 0.407]	5.2	62.8	0.169	5.5	100.0
200×200, $N = 4$	$> 2\,\mathrm{mm\,h^{-1}}$	0.216	[0.204, 0.228]	6.4	93.3	0.064	5.5	100.0
	$> 5\,\mathrm{mm\,h^{-1}}$	0.114	[0.103, 0.132]	6.0	79.8	0.033	5.2	100.0
200×200, $N = 8$	$> 2\,\mathrm{mm\,h^{-1}}$	0.272	[0.264, 0.282]	4.0	96.7	0.110	5.7	100.0
	$> 5\,\mathrm{mm\,h^{-1}}$	0.191	[0.181, 0.205]	4.2	84.3	0.058	5.3	100.0

Table 2. Reference values used as decision rules for single field pairs ($W = 100$, $N = 4$, belt): false-positive rate under G_0 (FPR), detection rate of corresponding forecasts with $q = 0.75$ and maximum displacement $0.5\,D$ (Det.), and detection rate of a calibrated critical value with the same false-positive rate (Cal.), in per cent, at widths $L = 5$, 20 and 40.

		$L = 5$			$L = 20$			$L = 40$		
Threshold	Criterion	FPR	Det.	Cal.	FPR	Det.	Cal.	FPR	Det.	Cal.
$> 2\,\mathrm{mm\,h^{-1}}$	Calibrated test	4.9	74.4	–	5.0	78.2	–	4.9	71.6	–
	f_0	38.3	99.3	99.2	67.9	99.8	99.7	92.7	99.9	99.9
	$(1 + f_0)/2$	0.1	20.3	16.6	2.6	68.5	68.0	14.1	84.2	84.4
	0.5	0.2	26.0	22.1	3.4	72.6	71.7	16.7	85.7	86.4
	$\mathrm{FSS}_{\mathrm{random}}(L)$	20.3	94.5	95.2	15.8	92.1	92.3	14.2	82.4	84.5
$> 5\,\mathrm{mm\,h^{-1}}$	Calibrated test	5.8	46.6	–	5.3	61.4	–	5.8	59.0	–
	f_0	17.6	78.8	79.7	54.5	95.6	95.1	86.0	97.5	97.0
	$(1 + f_0)/2$	0.2	4.1	4.8	1.5	37.0	38.1	7.1	61.8	64.1
	0.5	0.2	4.3	5.3	1.5	38.0	38.5	7.4	62.2	64.9
	$\mathrm{FSS}_{\mathrm{random}}(L)$	10.3	65.0	66.2	13.7	78.6	80.0	12.7	71.1	72.0

is therefore not a better criterion: f_0 and the random-forecast reference detect more corresponding forecasts only because they also accept far more unrelated ones (18–93 % and 10–20 % of the null pairs at the widths and thresholds shown), and criteria

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

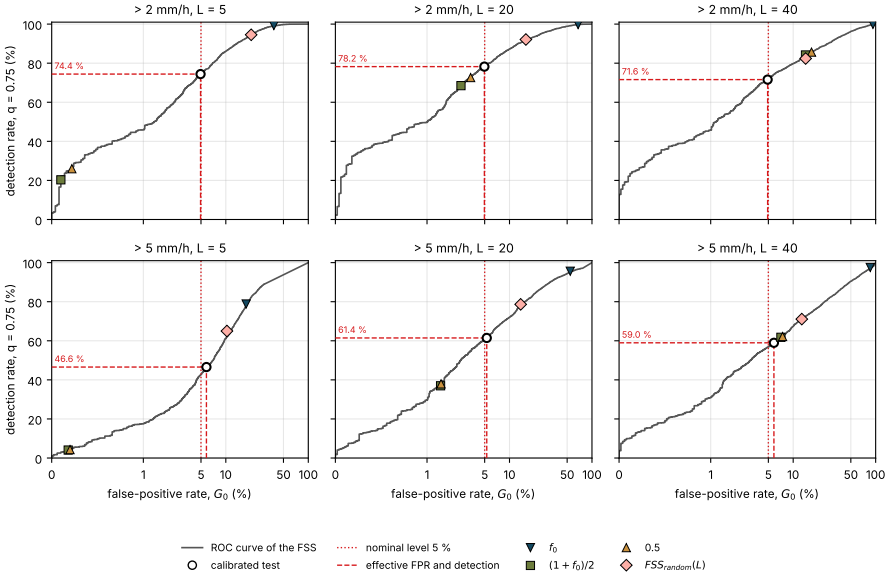


Figure 5. ROC curves of the single-day FSS for $q = 0.75$ against G_0 ($W = 100$, $N = 4$, belt, maximum displacement $0.5\,D$) at three widths and two thresholds, with the false-positive and detection rates of the reference values of the literature used as decision rules: the observed frequency f_0, the useful-skill reference $(1 + f_0)/2$, 0.5, and the random-forecast reference of Antonio and Aitchison (2025), and of the calibrated test at 5 %. Criteria are compared at the same false-positive rate, along the curve: a higher detection rate obtained at a higher false-positive rate, as for f_0 and $\mathrm{FSS}_{\mathrm{random}}(L)$, is not a better discrimination. The red dotted line marks the nominal level of the test, 5 %; the red dashed segments mark the effective false-positive rate of the calibrated test on the validation pairs and its detection rate, given in red. The horizontal axis is linear below 1 % and logarithmic above.

can be compared only at the same false-positive rate, that is, by their position along the curve. What the references lack, in these populations, is not power but a known error rate, which moreover varies with width, threshold, domain size and number of cells.

4.3 Magnitude, power and discrimination

The mean FSS of corresponding forecasts increases monotonically with the width for every q (Fig. 4); for $q = 0.75$ from 0.39 at $L = 5$ to 0.72 at $L = 40$. Discrimination does not follow it. The AUC of $q = 0.75$ is largest at $L = 10$ (0.960) and decreases to 0.922 at $L = 40$, and that of $q = 0.25$ decreases from 0.759 at $L = 5$ to 0.659 (Fig. 6, Table 3). The pattern is the same in all four configurations: for $q \leq 0.75$ and the displacement $0.5\,D$, the width of maximum AUC is 5–10 grid points at $2\,\mathrm{mm\,h^{-1}}$ and 10–15 at $5\,\mathrm{mm\,h^{-1}}$, and the decrease from the maximum to $L = 40$ is 0.01–0.13 in AUC, the largest for weak correspondence. For $q = 1$, with every cell corresponding, the AUC increases to 0.96–1.00 at $L = 20$–40 and does not decrease appreciably beyond. Power at the 5 % level follows the AUC more loosely, because the null distribution changes shape with the width: for $q = 0.75$ it is 74.4 % at $L = 5$, 78.2 % at $L = 20$ and 71.6 % at $L = 40$. Its complement is the Type II error rate: at $L = 20$,

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

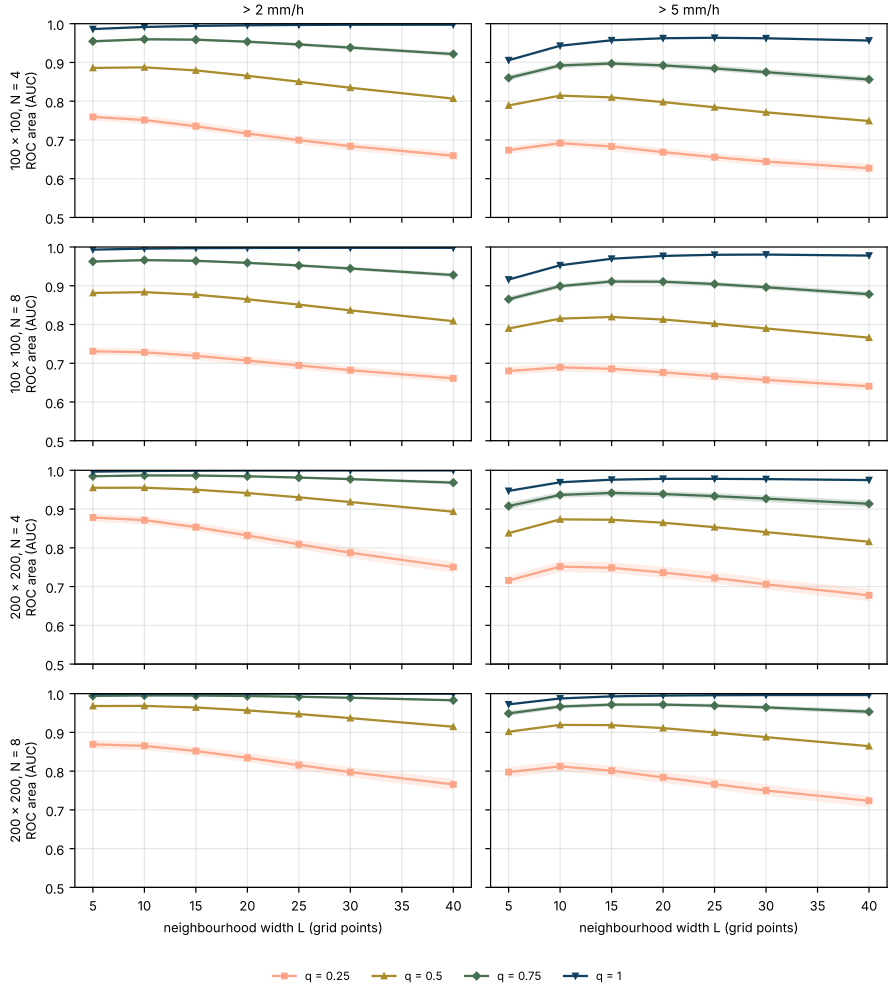


Figure 6. AUC of the single-day FSS versus neighbourhood width for every q in the four configurations (rows) and two thresholds (columns); belt, maximum displacement $0.5\,D$. Shading: 95 % bootstrap intervals for $q = 0.25$ and 0.75; the black line marks chance level.

21.8 % of the forecasts with three corresponding cells out of four, and 79.3 % of those with one out of four, score below the critical value and are indistinguishable from unrelated forecasts. Three causes combine (Fig. 7): the critical value, placed at the 95th percentile of the null to keep false positives at 5 %, also cuts the lower part of the alternative distribution; corresponding cells, displaced by up to half their diameter and perturbed, may share few windows with the observed ones; and the independent cells of the same forecast contribute to the score as in the null. Power therefore increases with q and with the domain size, where the null is concentrated nearer zero, and changes little with the width, because the critical value rises with the width together with the scores of the alternatives.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table 3. AUC of the single-day FSS (belt) for $q = 0.25$ and $q = 0.75$ and maximum displacements $0.5\,D$ and D: at $L = 5$, width of maximum AUC ($L_{\max}$), maximum, and at $L = 40$.

		Maximum displacement $0.5\,D$								Maximum displacement D							
		$q = 0.25$				$q = 0.75$				$q = 0.25$				$q = 0.75$			
Configuration	Threshold	L=5	$L_{\max}$	max	L=40	L=5	$L_{\max}$	max	L=40	L=5	$L_{\max}$	max	L=40	L=5	$L_{\max}$	max	L=40
100×100, $N = 4$	$> 2\,\mathrm{mm\,h^{-1}}$	0.759	5	0.759	0.659	0.955	10	0.960	0.922	0.641	15	0.653	0.633	0.802	30	0.878	0.876
	$> 5\,\mathrm{mm\,h^{-1}}$	0.674	10	0.692	0.627	0.860	15	0.897	0.856	0.574	15	0.623	0.609	0.694	30	0.818	0.816
100×100, $N = 8$	$> 2\,\mathrm{mm\,h^{-1}}$	0.731	5	0.731	0.661	0.963	10	0.966	0.928	0.620	25	0.643	0.633	0.792	30	0.867	0.867
	$> 5\,\mathrm{mm\,h^{-1}}$	0.680	10	0.689	0.641	0.865	15	0.911	0.878	0.583	30	0.623	0.619	0.695	40	0.819	0.819
200×200, $N = 4$	$> 2\,\mathrm{mm\,h^{-1}}$	0.878	5	0.878	0.750	0.985	10	0.987	0.968	0.768	10	0.790	0.727	0.913	30	0.951	0.948
	$> 5\,\mathrm{mm\,h^{-1}}$	0.716	10	0.752	0.677	0.908	15	0.942	0.914	0.610	20	0.704	0.666	0.753	25	0.900	0.892
200×200, $N = 8$	$> 2\,\mathrm{mm\,h^{-1}}$	0.869	5	0.869	0.766	0.994	10	0.995	0.983	0.720	25	0.750	0.736	0.904	40	0.963	0.963
	$> 5\,\mathrm{mm\,h^{-1}}$	0.798	10	0.812	0.723	0.949	15	0.972	0.953	0.642	25	0.709	0.701	0.791	40	0.922	0.922

Discrimination improves with domain size; the number of cells matters less. In the 200×200 domain the AUC of $q = 0.25$ is 0.88 at $L = 5$ instead of 0.76, because independent cells meet less often by chance and the null scores are lower; with eight cells instead of four, more chance coincidences raise the null scores, but more corresponding cells raise the alternative scores: the AUC changes by less than 0.03 at $2\,\mathrm{mm\,h^{-1}}$ and increases by up to 0.08 at $5\,\mathrm{mm\,h^{-1}}$ in the large domain.

4.4 Distributional mechanism

The decline of discrimination at large widths arises from the null distribution (Fig. 7). Between $L = 20$ and $L = 40$, the mean null score doubles (0.15 to 0.31) and its standard deviation increases from 0.15 to 0.19, whereas the mean score for $q = 0.75$ rises less (0.60 to 0.72) and its standard deviation does not change (0.19); the difference of the means decreases from 0.44 to 0.41 and the standardized separation from 1.87 to 1.54. Wider windows bring unrelated cells into common neighbourhoods more often than they improve the match of corresponding ones, once the window exceeds the displacement of the corresponding cells. At small widths the null is concentrated at zero: at $5\,\mathrm{mm\,h^{-1}}$ and $L = 5$, 71 % of the null pairs score zero, so that the critical value is low and any overlap of cores is significant, but most corresponding forecasts with a displacement comparable with the core size also score near zero. This is why the most discriminating width is larger for the higher threshold. It also explains why the detection rate of the calibrated test is lower at 5 than at $2\,\mathrm{mm\,h^{-1}}$ at every width, 47–61 % against 72–78 % for $q = 0.75$ (Table 2, Fig. 5). Three properties of the cores combine. The areas above $5\,\mathrm{mm\,h^{-1}}$ are the cores, a few grid points across, so that a displacement of up to half the cell diameter often moves a corresponding core away from the observed one: at $L = 5$, 10.8 % of the forecasts with $q = 0.75$ score exactly zero, none at $2\,\mathrm{mm\,h^{-1}}$. A third of the observed cells have no core above $5\,\mathrm{mm\,h^{-1}}$ inside the domain, so that the fields contain fewer events at this threshold. And the perturbations of size, shape and core peaks, up to 15 %, can bring a corresponding core above the threshold in one field but not in the other, so that the effective correspondence of the events is about 0.64 instead of 0.75 (Sect. 2); for comparison, at $2\,\mathrm{mm\,h^{-1}}$ and $L = 5$ detection falls from 74.4 % for $q = 0.75$ to 47.0 % for $q = 0.5$, so that a reduction of the effective correspondence from 0.75 to 0.64 alone accounts for about ten percentage points. Widening the window recovers part of the displaced cores, which is why

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

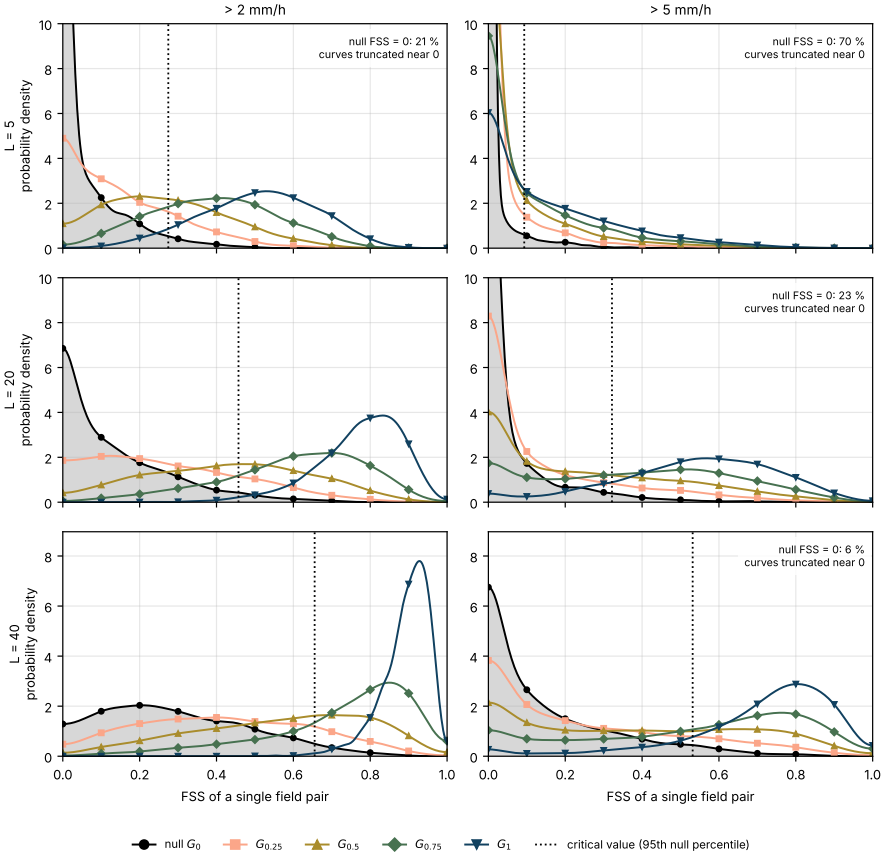


Figure 7. Empirical distributions of the single-day FSS in the main configuration (100×100 domain, four cells, belt, maximum displacement $0.5\,D$) for the null population G_0 (black, shaded, circles) and the alternatives G_q with $q = 0.25$, 0.50, 0.75 and 1 (colours and markers as in Figs. 4 and 6), at widths of 5, 20 and 40 grid points (rows) and for 2 and 5 mm h^{-1} (columns). Vertical dotted lines are the critical values: the area under the null density to their right is the false-positive rate, the area under each G_q density to their left is the Type II error rate. Curves are kernel-density estimates (Scott's bandwidth, reflected at 0 and 1) used for display only; all significance and AUC calculations use the unsmoothed scores. Panels in the same row share a vertical scale; the percentage of null pairs with an FSS of exactly zero is given where it exceeds 5 %, and curves are truncated near zero where marked. The other three configurations are shown in Supplement Figs. S3–S5.

425 detection rises from 46.6 % at $L = 5$ to 61.4 % at $L = 20$. In the 200×200 domain the null scores are concentrated even more near zero (59 % of the null pairs score zero at $L = 5$ and 2 mm h^{-1}, against 21 % in the smaller domain), because independent cells meet less often by chance, and the alternatives separate from it more clearly, which is why discrimination is higher there (Supplement Figs. S3–S5).

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

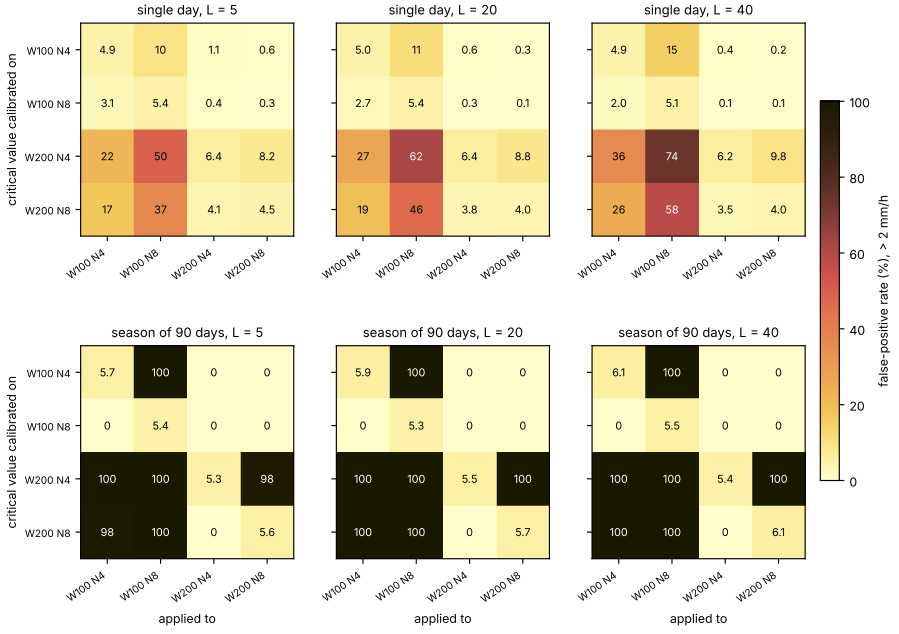


Figure 8. False-positive rates (%) of critical values calibrated in one configuration (rows) and applied to the validation pairs of another (columns), at 2 mm h^{-1} and three widths, for single days (top) and seasons of 90 days (bottom); belt. The diagonal of the seasonal panels gives the cross-validated rate.

4.5 Transfer of critical values

Critical values depend on the configuration and do not transfer (Fig. 8, top). Calibrated in the 200×200 domain and applied to the 100×100 one, they give 17–36 % false positives when the small domain has four cells and 37–74 % when it has eight; calibrated with four cells and applied to eight in the same domain, 8–15 %. In the opposite directions the test becomes conservative, with false-positive rates of 2.0–3.1 % from eight to four cells and below 1.1 % from the small to the large domain. A critical value is thus a property of the population, not of the score.

4.6 Position error

Doubling the maximum displacement to D lowers discrimination at small widths and moves its maximum to larger ones (Fig. 9). For $q = 0.75$, the AUC at $L = 5$ falls from 0.955 to 0.802, and the width of maximum AUC moves from 10 to 30 grid points (2 mm h^{-1}) and from 15 to 30 (5 mm h^{-1}); in all configurations and for $q \leq 0.75$ it lies at 10–40 grid points, and for $q = 1$ at 40. With a larger position error, the windows must be wider before corresponding cells share them, and over the range examined the gain outweighs the increase of chance coincidences. The most discriminating width is therefore set mainly by the position error of the forecast relative to the size of the cells, not by the score.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

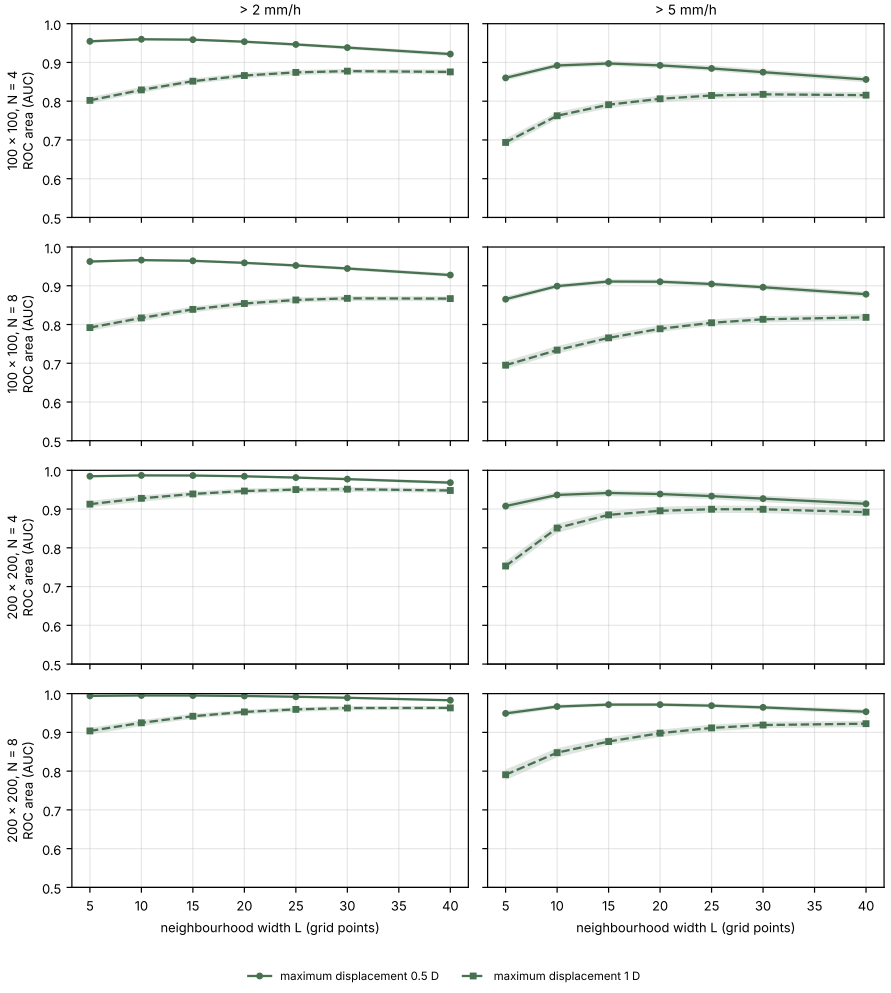


Figure 9. AUC of the single-day FSS for $q = 0.75$ versus neighbourhood width for maximum displacements $0.5\,D$ (solid, circles) and D (dashed, squares), in the four configurations and two thresholds; belt. Shading: 95 % bootstrap intervals.

4.7 Population comparability

Correspondence does not change how much precipitation is forecast: the forecast coverage of the alternatives differs from that of the independent forecasts of the same realizations by less than 2.2 % at $2\,\mathrm{mm\,h^{-1}}$ and less than 3.6 % at $5\,\mathrm{mm\,h^{-1}}$, in all configurations and for both displacements, within about two standard errors of zero.

4.8 Sensitivity to cell morphology

Rougher outlines and rougher fields leave the dependence of discrimination on width unchanged (Supplement Fig. S10 and Table S8). For both variants and both thresholds, the width of maximum AUC is the same as in the reference or one step (5 grid

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

points) larger, and the decrease from the maximum to $L = 40$ is 0.03–0.04 for $q = 0.75$ and 0.05–0.08 for $q = 0.25$, against 0.04 and 0.07–0.10 in the reference. The AUC itself changes little: for $q = 0.75$ by less than 0.02, never significantly except at $L = 5$ and 5 mm h^{-1} with rougher outlines; for $q = 0.25$ it is 0.02–0.03 lower at small widths, significantly so up to $L = 15$–20 at 2 mm h^{-1}, except for rougher fields at 5 mm h^{-1}, where it does not change. The critical values change with the morphology of the null fields: with rougher outlines they are within 0.025 of the reference, but rougher fields raise them by 0.03–0.07, presumably because fragmented areas above the threshold meet by chance more often. The validation false-positive rates are 4.9–6.8 % with 1500 validation pairs. As for domain size and number of cells, the critical values belong to the population, whereas the relation between width and discrimination does not depend on these details of the cells.

4.9 The asymptotic limit of the seasonal critical value

Figure 10 shows how the critical value of the aggregated score approaches its limit as the number of aggregated days grows, for the null of independent cells and the two Bernoulli nulls (Table 4). In every case it decreases towards FSS_∞ at the rate $n^{-1/2}$ of Eq. (4): under G_0, at 2 mm h^{-1} and $L = 20$, from 0.457 for one day to 0.214, 0.191 and 0.174 for 30, 90 and 365 days, against a limit of 0.159. The normal approximation reproduces the empirical critical values within 0.005 for $n \geq 30$ at 2 mm h^{-1}, and its cross-validated false-positive rate is 3.4–7.6 % for $n = 30$ and 90 at both thresholds and all widths; the Cornish–Fisher correction improves it where the daily terms are most skewed, the null of independent cells at small widths and 5 mm h^{-1}, and is of little use elsewhere. For a single day neither approximation applies.

Under the Bernoulli null with constant probability, the limit coincides with the exact expression of Sect. 3.4 to within 6×10^{-4} at every threshold and width, and with the reference of Antonio and Aitchison (2025) when the mean observed fraction equals the probability of the random forecast. With valid windows, for which the reference was derived, the two agree within 0.004 at every width and threshold (for example 0.285 and 0.289 at $L = 20$ and 2 mm h^{-1}, 0.499 and 0.503 at $L = 40$; Supplement Table S11); with the belt the mean observed fraction over the windows differs from the expected frequency, because cells outside the domain enter the windows near its edge, and the reference falls below the limit by up to 0.024 at $L = 40$, whereas the exact expression still agrees. Used as a threshold under its own null, the reference is exceeded by 46–75 % of single days and by 44–85 % of seasons of 30 days; for longer aggregations the rate moves towards 0 or 100 % according to whether the reference lies slightly above or below the limit (23–70 % for 365 days with valid windows). It is a central value of the null, not a percentile.

The three nulls have different limits. The limit under G_0 is 0.54–0.62 times the Bernoulli limit: a forecast made of coherent cells has more variable fractions than a field of independent grid points, its denominator is larger and its expected score lower. With the daily probability, a random forecast always has the observed frequency of the day and its limit is higher, 0.17 instead of 0.14 at $L = 5$ and 0.57 instead of 0.48 at $L = 40$ (2 mm h^{-1}). The reference of Antonio and Aitchison (2025), applied to seasons of independent cells, is therefore exceeded by fewer than 1 % of null seasons of 90 days at 2 mm h^{-1}, whereas applied to the daily-probability Bernoulli forecasts it is exceeded by almost all of them. For single days under G_0, with m and s^2 pooled over all days, it is exceeded by 10–22 % of the null pairs, close to the 10–19 % obtained with the statistics of each observed field (Sect. 4.2).

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

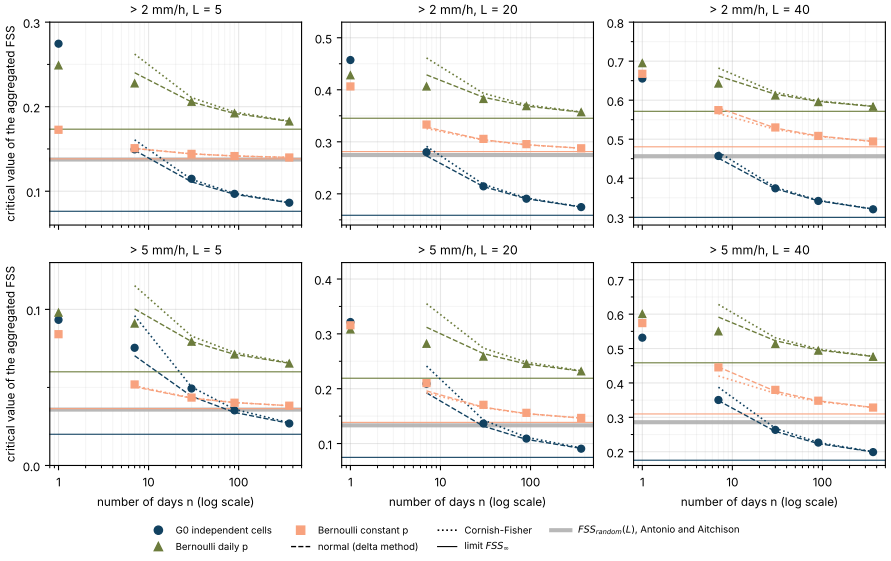


Figure 10. Critical value of the aggregated FSS versus the number of aggregated days n, for the null of independent cells G_0 and the Bernoulli nulls with constant and daily probability ($W = 100$, $N = 4$, belt), at three widths and two thresholds. Markers: empirical 95th percentile, for $n = 1$, 7, 30, 90 and 365 days (logarithmic axis); dashed: normal approximation by the delta method, Eq. (4); dotted: with Cornish–Fisher correction (both shown for $n \geq 7$); thin horizontal lines: asymptotic limits FSS_∞; thick grey line: random-forecast reference of Antonio and Aitchison (2025), Eq. (5).

4.10 Seasons: discrimination between capabilities

For seasons of 90 days, the critical value of the aggregated score is close to the mean null score (Fig. 11, top): at $2\,\mathrm{mm\,h^{-1}}$ it is 0.098 at $L = 5$ and 0.344 at $L = 40$, against null means of 0.078 and 0.301, and the cross-validated false-positive rate is 5.1–6.1 % in all configurations. Corresponding forecasts lie well above it: for $\kappa = 0.5$ the power is at least 98 % at both thresholds and every width, in all configurations; for $\kappa = 1$ it is at least 95 % at $2\,\mathrm{mm\,h^{-1}}$, but only 46–88 % for $q = 0.25$ at $5\,\mathrm{mm\,h^{-1}}$ in the 100×100 domain, increasing with the width. As anticipated in Sect. 3.4, the season does not test whether a system has some skill, which it detects almost always, but allows forecasting systems of different capability to be ranked.

The separation between the seasonal score distributions of adjacent capabilities, $q + 0.25$ against q, is 3.2–5.1 standard deviations at $2\,\mathrm{mm\,h^{-1}}$ and 1.3–3.2 at $5\,\mathrm{mm\,h^{-1}}$ (Fig. 11, bottom), and its dependence on width differs with q. Between weak capabilities ($q = 0.25$ against 0, and 0.5 against 0.25) it is largest at small widths, as for single days; between strong ones ($q = 1$ against 0.75) it increases with the width up to $L = 40$. These separations compare the marginal distributions of the seasonal scores; because the seasons of different capabilities share their observed fields, the paired comparison is even sharper, and in every pair of seasons with the same days the system of higher capability has the higher score. A width that is good for showing that a system is better than chance is thus not necessarily the width that best ranks two good systems. At $5\,\mathrm{mm\,h^{-1}}$ all separations are largest at $L = 15$–40.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table 4. Critical value of the aggregated FSS as a function of the number of aggregated days n ($W = 100$, $N = 4$, belt): asymptotic limit FSS_∞, random-forecast reference of Antonio and Aitchison (2025) (AA, Eq. 5), empirical 95th percentiles, normal approximation for $n = 90$ (Eq. 4), and cross-validated false-positive rates (%) of the normal critical value and of the AA reference for $n = 90$.

Threshold	Null	L	FSS_∞	AA	$n = 1$	7	30	90	365	Normal, 90	FPR	FPR AA
$> 2\,\mathrm{mm\,h^{-1}}$	G_0	5	0.076	0.138	0.275	0.149	0.115	0.097	0.086	0.096	6.3	0.0
		20	0.159	0.275	0.457	0.280	0.214	0.191	0.174	0.191	5.4	0.0
		40	0.300	0.456	0.655	0.457	0.374	0.342	0.320	0.342	5.2	0.0
	Bernoulli, constant p	5	0.138	0.137	0.173	0.151	0.144	0.142	0.140	0.142	5.4	65.3
		20	0.281	0.275	0.406	0.333	0.306	0.295	0.287	0.294	6.0	84.7
		40	0.480	0.456	0.668	0.575	0.530	0.508	0.494	0.508	5.6	95.4
	Bernoulli, daily p	5	0.173	0.138	0.249	0.228	0.206	0.192	0.182	0.192	5.8	100.0
		20	0.345	0.275	0.428	0.406	0.382	0.368	0.357	0.369	5.0	100.0
		40	0.572	0.456	0.695	0.643	0.612	0.595	0.583	0.597	5.5	100.0
$> 5\,\mathrm{mm\,h^{-1}}$	G_0	5	0.020	0.036	0.093	0.075	0.049	0.035	0.027	0.034	7.0	5.5
		20	0.075	0.133	0.322	0.209	0.137	0.109	0.091	0.108	7.1	0.7
		40	0.175	0.286	0.532	0.350	0.264	0.227	0.199	0.224	5.7	0.1
	Bernoulli, constant p	5	0.036	0.036	0.084	0.052	0.043	0.040	0.038	0.040	4.6	64.2
		20	0.139	0.133	0.316	0.210	0.171	0.156	0.147	0.155	6.8	79.6
		40	0.310	0.286	0.574	0.445	0.380	0.348	0.328	0.348	6.4	91.7
	Bernoulli, daily p	5	0.060	0.036	0.098	0.091	0.079	0.071	0.065	0.071	5.4	100.0
		20	0.219	0.133	0.308	0.282	0.259	0.245	0.232	0.245	4.8	100.0
		40	0.459	0.286	0.600	0.550	0.513	0.494	0.476	0.496	4.8	100.0

The seasonal critical values are even more specific to the population than the single-day ones (Fig. 8, bottom). Because their spread is small, a critical value from another configuration is either always exceeded or never: transferred from the 200×200 to the 100×100 domain, it produces 98–100 % false positives; in the opposite direction, none.

4.11 Variable number of cells

With a number of cells $N_o \sim \mathrm{Poisson}(4)$ in the 100×100 domain (mean 4.02, 1.8 % of the days without cells, excluded), the critical value calibrated on all days keeps the overall false-positive rate at 5.1–5.9 % for single days and at 5.5–5.9 % (cross-validated) for seasons. The rate is not uniform across days, however (Table 5): at $2\,\mathrm{mm\,h^{-1}}$ it is 1.8–2.6 % on days with one cell, 3.1–5.2 % on days with two to five, 7.2–8.0 % on days with six or seven and 12–16 % on days with eight or more; at $5\,\mathrm{mm\,h^{-1}}$ the range is similar, from 1.0–2.3 % to 9–13 %. On days with many cells, independent cells meet by chance more often and the null scores are higher, so that a critical value calibrated on the mixture is too low for them and too high for days with few cells. Discrimination is slightly lower than with exactly four cells, because the mixture widens both null and alternative distributions (AUC of $q = 0.75$ 0.929 instead of 0.960 at $L = 10$ and $2\,\mathrm{mm\,h^{-1}}$), and its dependence on width is the same: largest at $L = 10$ ($2\,\mathrm{mm\,h^{-1}}$) and 15 ($5\,\mathrm{mm\,h^{-1}}$) for the displacement $0.5\,D$, at $L = 30$ for D.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

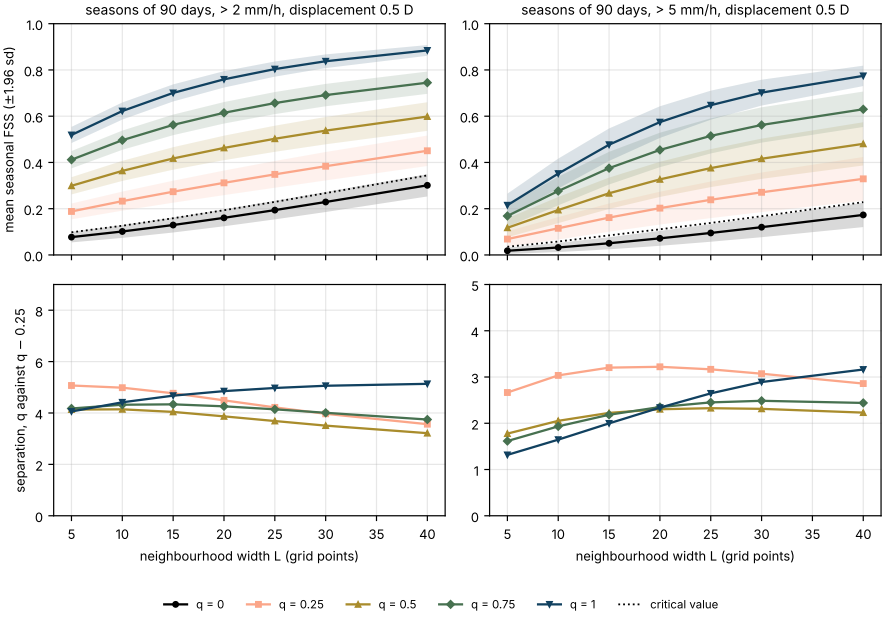


Figure 11. Seasons of 90 aggregated days ($W = 100$, $N = 4$, belt, maximum displacement $0.5\,D$). Top: mean seasonal FSS for every q, with ± 1.96 standard deviations of the seasons (shading), and the seasonal critical value (dotted). Bottom: standardized separation between seasons of adjacent capabilities, q against $q - 0.25$, with colour and marker of q, for 2 (left) and 5 mm h^{-1} (right).

Table 5. Variable number of cells ($W = 100$, $N_o \sim$ Poisson(4), belt): false-positive rate (%) of the single-day critical value calibrated on all days, on the validation days grouped by observed number of cells N_o.

		> 2 mm h^{-1}			> 5 mm h^{-1}		
N_o	Days	$L = 5$	20	40	$L = 5$	20	40
1	342	2.0	2.6	1.8	1.9	1.6	1.0
2	713	4.8	4.5	3.1	2.8	3.1	3.2
3	987	5.2	3.5	3.6	4.1	4.4	4.6
4	966	4.2	4.0	3.5	5.1	4.9	5.2
5	805	4.8	5.2	5.1	5.8	5.6	6.3
6-7	797	7.5	7.9	8.0	11.2	9.5	10.0
>=8	296	12.5	13.5	15.5	12.8	9.5	11.5

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

5 Discussion

5.1 Spatial scale, score magnitude and discriminating ability

A broader neighbourhood does two things at once. It tolerates the position error of corresponding cells, which raises their scores; this is the intended behaviour of the FSS. It also brings unrelated precipitation into common windows, which raises the scores of independent forecasts. As long as the window is smaller than the position error, the first effect dominates and widening improves discrimination; once the window exceeds it, the first effect saturates while the second continues, and discrimination declines although the mean score keeps rising (Sects. 4.3–4.4). The width of maximum discrimination is therefore set mainly by the position error relative to the size of the cells (Sect. 4.6), and it is a property of the forecast, not of the score: for a given forecasting system there is a scale beyond which widening the neighbourhood adds tolerance but no evidence. For convective cells displaced by up to their radius this scale is 5–15 grid points (11–33 km), the larger values for the higher threshold; for displacements up to their diameter it moves to 10–40 grid points.

How strongly the decline appears depends on the treatment of the domain boundary. With windows restricted to the interior, the decline is two to two and a half times as large and starts at smaller widths, because the remaining windows increasingly ignore the cells near the edge (Sect. 3.7). Studies that use valid windows, or zero padding, therefore overstate the loss of discrimination at large widths, and the belt, where the data allow it, should be preferred. With the belt the decline is smaller but remains in every configuration for weak and moderate correspondence; for forecasts in which every cell corresponds it disappears within the widths examined.

The ROC curves compare samples from specified generative mechanisms; they do not recover corresponding objects from a single FSS realization. A high AUC for $q = 0.75$ means that the score tends to rank those synthetic field pairs above independent ones, and the label q is a property of data generation, not a skill statistic.

5.2 Reference values and significance

The reference values of the literature answer a different question from a test. The observed frequency f_0 is the score of a forecast that is uniformly correct on average; the useful-skill reference $(1 + f_0)/2$ is half-way between f_0 and a perfect score (Roberts and Lean, 2008), and 0.5 is its form for rare events (Mittermaier and Roberts, 2010; Skok and Roberts, 2016); the random-forecast reference of Antonio and Aitchison (2025) is the expected score of a random forecast. They say where the score lies, not how often an unrelated forecast exceeds it. Used as decision rules for one forecast, they have false-positive rates that vary from almost zero to almost one with the width, the threshold, the domain and the number of cells (Sect. 4.2). Their weakness, in these experiments, is not a lack of power. Rules that depend on the observed field could in principle discriminate better than any constant threshold on the score, but here they lie close to the ROC curve of the FSS, and even the differences $\mathrm{FSS} - r(O)$, which exploit the observed field fully, discriminate no better than the score itself (Sect. 4.2). This is an empirical result for the populations considered, not a general property. Calibration does not improve the score; it fixes the error rate at a chosen value.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

The asymptotic analysis gives the random-forecast reference a precise place in this picture (Sect. 4.9). Under the Bernoulli null for which it was derived, it is the limit towards which the calibrated critical value of the aggregated score converges as the number of aggregated cases grows: the mean seasonal score of a random forecast, not an upper percentile. Used as a threshold for long aggregations under its own null it is exceeded by about half of the random forecasts, and for a single case its distance from the critical value is large. Under other nulls it has no fixed relation to the critical value at all. A forecast made of independently placed coherent cells has a lower limit than a field of independent grid points, so that for long seasons the reference is never exceeded by such forecasts; a random forecast with the observed frequency of each day has a higher limit, and the reference is then exceeded by almost all of them. The reference of Antonio and Aitchison (2025) is thus the right yardstick for the question it was designed for, the average skill of a forecasting system relative to a random one. It is the limit of the critical value under the Bernoulli null, not a critical value for any finite number of cases: the critical value exceeds it by about $z_{0.95}\sigma_1/\sqrt{n}$.

5.3 Two uses of the FSS: a season and a day

Aggregating many cases, as recommended by Mittermaier (2021), answers whether a forecasting system is skilful on average, and how it compares with another. For that question the season is powerful: with 90 days, every capability considered here is significant against the null of independent cells, and systems that differ by a quarter of corresponding cells are separated by several standard deviations (Sect. 4.10). The critical value converges to the mean null score at rate $n^{-1/2}$, so that the significance test becomes a comparison with a limit, and the question shifts from whether a system has skill to how much. For the same reason the choice of the null matters more as the season grows: the limits of the three nulls examined here differ by far more than the seasonal spread, so that a season judged against the null of independent cells, against a random field with the climatological frequency or against one with the frequency of each day can lead to opposite conclusions (Sect. 4.9). A seasonal significance statement is therefore only as meaningful as the null it refers to. This is the use of the FSS for model evaluation and development, and the results support it, with two qualifications: the aggregated score weights days by their precipitation and carries the information of fewer days than it contains (Sect. 3.7), and the width that best ranks two systems depends on their capability, small for weak systems and large for good ones.

The seasonal result does not, however, describe what the score says about each forecast. A system with $q = 0.25$ is significant over a season in almost every case, yet on a single day its forecast is recognized as related to the observations in only one case out of five, and the AUC of its daily score is 0.76 at best. The $\sqrt{n}$ growth of the separation comes from averaging over days, not from a better forecast on any of them: a season reveals the mean behaviour of the system, but the forecaster confronts one day at a time. For this operational question the single-pair analysis is the relevant one, and its conclusion concerns the evidence carried by the score. On each day, the FSS separates related from unrelated forecasts best at a width set by the position error of the system, and widening the neighbourhood beyond it increases the score without increasing the evidence of correspondence. This width is a property of the forecasting system and of the null against which it is judged, and it can be estimated. It is not a threshold of usefulness: at smaller widths discrimination is lower but may remain high (for $q = 0.75$ the AUC is 0.955 at $L = 5$

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

and 0.960 at $L = 10$), and the AUC measures how well the score ranks populations, not how accurately a particular forecast places its cells.

5.4 Calibration and its transfer

A critical value belongs to the population of fields being verified. Transferring single-day critical values between domain sizes or numbers of cells changes the false-positive rate from 5 % to anywhere between less than 1 % and 74 %, and seasonal critical values, whose spread is small, do not transfer at all: they are always or never exceeded (Sect. 4.5). The same holds within one population when it mixes days of different character. With a variable number of cells, a critical value calibrated on all days controls the false-positive rate on average, but not for days with few or many cells (Sect. 4.11); a verification system should therefore stratify its calibration by the variables that change the null distribution most, such as the number of convective cells or the observed coverage. The calibrated false-positive rate is an average over the observed fields of the null population, $P_{G_0}(\text{reject}) = \mathrm{E}_O[P_{G_0}(\text{reject} \mid O)]$; a test conditioned on each observed field would remove this variation and is a natural extension for operational use.

5.5 Implications for verification practice

Five conclusions bear on the interpretation of FSS values. First, a higher FSS at a broader neighbourhood is not, by itself, stronger evidence of spatial correspondence, because widening the window also raises the scores of unrelated forecasts. Second, reference values are not significance tests: they correspond to unknown, variable error rates, and the random-forecast reference is the asymptotic limit of the critical value of an aggregation under the Bernoulli null, not a critical value itself. Third, inspecting several widths and reporting the most favourable one inflates the false-positive rate; the joint test accounts for the selection. Fourth, critical values belong to the population being verified and must be calibrated for it. Fifth, the aggregated seasonal FSS and the FSS of a single case answer different questions, the average skill of a system and the evidence that one forecast corresponds to the observations, and should not be read for each other.

The procedure does not require a synthetic generator. A null population with realistic structure can be formed from a forecast archive by pairing each forecast with an observation from a different date in the same season and region, excluding dates close enough to share the same weather regime and grouping the pairs by number of cells or observed coverage, since pooling cases with different climatological frequencies can create apparent skill (Hamill and Juras, 2006). The 95th percentile of the FSS of such pairs, width by width, gives critical values matched to the domain and the forecasting system, and the width of maximum discrimination can be estimated in the same way from pairs of forecasts and observations of the same date. A further extension would turn the same distributions into a probabilistic statement about position. Given the FSS of a day, the simulated distributions $f(\text{FSS} \mid q)$ and a prior on the capability q give, by Bayes' theorem, the posterior probability of each capability; combined with the distribution of the position error and the probability of a chance coincidence, it gives the probability that an observed storm lies within a distance r of a forecast one, and hence the radius within which this probability reaches, say, 95 %. The width of maximum discrimination is a natural first choice for such a statement, but it is not necessarily the width at which the posterior is sharpest, which depends on the whole likelihood, on the prior and on the observed score; the radius depends

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

moreover on the distribution of the position error and on the prior, which must be estimated from a forecast archive; in these experiments the most discriminating width is about twice the mean displacement of the corresponding cells.

5.6 Scope and limitations

The generator represents isolated summer convective cells with realistic size, intensity, structure and gradients, and independently positioned cells that may coincide by chance. Its principal restrictions are a fixed number of cells within a configuration, the exclusion of organized mesoscale systems, alternatives without frequency bias, rigid displacement with bounded perturbations of size, shape and intensity, and no temporal evolution, lead time or model-specific bias. The fields are not conditioned on actual observations, and the null of independent cells is one of several possible nulls; the Bernoulli nulls show how much the critical values and their limits depend on that choice. Seasons are formed from independent days of the same capability; real seasons mix days of different predictability, and serial correlation between days would reduce the effective number of cases. The near-perfect AUCs obtained for strong alternatives and for seasons should not be extrapolated to the closer populations expected in real forecast verification. The synthetic critical values are specific to the populations studied here and should not be applied to real fields; what carries over is the procedure, calibration on a null population matched to the verification setting, independent validation and joint testing across widths, and the distinction between the significance of a season and the reliability of a forecast.

6 Conclusions

The FSS is usually read through its magnitude, against reference values that describe where the score of a random or useless forecast lies on average. This study has asked instead how often an unrelated forecast reaches a given score and how well the score separates related from unrelated forecasts, for one field pair and for a season of aggregated pairs. Synthetic summer convective cells, in which the correspondence between forecast and observed cells is known by construction, were analysed in four configurations of domain size and number of cells, with two magnitudes of position error, two variants of cell morphology, four treatments of the domain boundary, independently calibrated and validated nulls, and three null populations. Five conclusions follow, each of which changes how an FSS value should be read.

First, reference values are not significance tests, and their error rates are unknown. Used as decision rules for one forecast, the observed frequency, the useful-skill reference $(1 + f_0)/2$, its form 0.5 and the random-forecast reference of Antonio and Aitchison (2025) are exceeded by between almost none and almost all unrelated forecasts, depending on width, threshold, domain and number of cells. They are not less powerful than a calibrated test, since they lie close to the ROC curve of the score; what they lack is a known false-positive rate, so that a skilful scale identified with them may reflect chance coincidences as much as forecast quality, and a change of skilful scale between configurations may reflect only a change of error rate.

Second, for aggregated scores the link between reference values and significance becomes exact. The critical value of the aggregated FSS converges at rate $n^{-1/2}$ to the asymptotic limit of the score under the null, and the delta method gives an asymptotic approximation to it in closed form, $\mathrm{FSS}_\infty + 1.645\,\sigma_1/\sqrt{n}$, which in these experiments is within about 0.01 of the

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Monte Carlo value from about 30 cases; for shorter aggregations its accuracy depends on the skewness of the daily scores and on the null. Under the Bernoulli null, FSS_∞ is the random-forecast reference of Antonio and Aitchison (2025): that reference is the centre of the null distribution of a long aggregation, not a percentile, and used as a threshold it corresponds to a test at a level of the order of one half. Under the null of independently placed cells the limit is about 0.6 times the Bernoulli one, so that the same seasonal score can be significant against one null and not against another; for long seasons the choice of the null decides the conclusion.

Third, magnitude and discrimination follow different scale dependences. The mean FSS of corresponding forecasts rises steadily with the width, whereas discrimination is largest at a width set by the position error of the forecast relative to the size of the cells, 5–15 grid points of 2.2 km for cells displaced by up to their radius and 10–40 for displacements up to their diameter, and declines beyond it for weak and moderate correspondence. Beyond that width a broader neighbourhood raises the score without adding evidence of correspondence. The decline is twice as large when the windows are restricted to the interior of the domain, an effect of the boundary rather than of the score.

Fourth, a season and a day answer different questions. Aggregation multiplies the separation between populations by about $\sqrt{n}$ at every width: over 90 days every capability considered is significant and systems that differ by a quarter of corresponding cells are separated by several standard deviations, whereas on a single day the forecast of a system with one corresponding cell out of four is recognized in only one case out of five. The seasonal score evaluates and ranks forecasting systems; it does not increase the evidence carried by the score of any single forecast.

Fifth, critical values belong to the population verified. Transferred between domain sizes or numbers of cells, single-day critical values give false-positive rates from below 1 % to 74 %, seasonal ones 0 or 100 %, and a single calibration applied to days with different numbers of cells controls the error rate only on average, from about 2 % on days with one cell to 12–16 % on days with eight or more.

For verification practice these results translate into five recommendations: state the null against which a score is judged; calibrate critical values on a population matched to the verification setting, for instance by pairing forecasts and observations of different dates of an archive, and stratify the calibration by the number of convective cells or the observed coverage; fix the width in advance or use a joint test across widths; use the precipitation around the verified domain where it is available; and, for seasonal or longer aggregations, compute the critical value from the delta method with the daily numerators and denominators of the null pairs, which requires no simulation of seasons, and check it by resampling the null pairs when the aggregation is short. Magnitude, significance, power and discrimination of the FSS are related but distinct properties, and should be reported as such.

Code and data availability. All experiments use synthetic fields; no observed meteorological dataset is analysed. Generator code, Monte Carlo score arrays for the four boundary treatments and the Bernoulli forecasts, calibrated, joint-scale, seasonal and asymptotic results, sensitivity experiments, figures, random seeds, environment information and file checksums are compiled in a reproducibility archive on

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Zenodo (https://doi.org/10.5281/zenodo.23499240, version 2.2.0 or later), together with two notebooks that regenerate every table and figure of this paper and of its Supplement. Code and notebooks are licensed MIT; synthetic data and documentation are licensed CC BY 4.0.

Appendix A: Accidental overlap under the null

In the null population G_0 forecast and observed cells are placed independently, so any positive score arises from chance coincidences between them. This appendix quantifies the simplest such coincidence, the physical intersection of two cells above a threshold, for two purposes: to show how its probability depends on domain size, number of cells and placement rule, and hence why the score distribution under G_0 is specific to each configuration; and to show why this geometric calculation cannot by itself provide that distribution.

Let $V_W = [0, W)^2$ be the verified square and $R_W = [-b, W+b)^2$ the square over which cell centres are proposed, with $b = 7$ grid points (16 km). Let Ω denote the area of a cell above 1 mm h^{-1} in local coordinates and $\Omega^{(t)}$ its area above the threshold t. The accepted centre set of a cell is

$$R_W^*(\Omega) = \{x \in R_W : |(x+\Omega) \cap V_W|/|\Omega| \geq \eta\}, \qquad \eta = 0.10, \tag{A1}$$

the set of centre positions at which the cell passes the visibility rule. Because proposals are drawn uniformly over R_W and rejected proposals are redrawn, an accepted centre is uniformly distributed over R_W^* for a given cell. For two independent cells with given shapes and accepted sets R_o^* and R_f^*, the probability that they intersect above t inside the domain is

$$p_{\text{pair}} = \frac{1}{|R_o^*||R_f^*|} \int_{R_o^*} \int_{R_f^*} \mathbf{1}\left\{(x+\Omega_o^{(t)}) \cap (y+\Omega_f^{(t)}) \cap V_W \neq \varnothing\right\} \mathrm{d}y\,\mathrm{d}x, \tag{A2}$$

the fraction of admissible pairs of centre positions at which the two cells touch; for grid coordinates the integrals become sums, and the overall probability is the average over the shapes of accepted cells. Far from the boundaries, with centres uniformly distributed over a common region of area $|R|$,

$$p_{\text{pair}} \simeq \frac{|\Omega_o^{(t)} \oplus (-\Omega_f^{(t)})|}{|R|}, \qquad \mathrm{E}[n_{\text{pairs}}] \simeq N_o N_f\, p_{\text{pair}}, \tag{A3}$$

where the Minkowski sum $\Omega_o^{(t)} \oplus (-\Omega_f^{(t)})$ is the set of relative positions at which the two areas touch and n_{pairs} is the number of intersecting pairs among the $N_o N_f$ observed–forecast pairs of a field pair.

Table A1 gives these quantities for the generator, from 16 000 placement proposals, 8000 pairs of independent cells and 2000 field pairs for each configuration. With the belt of 7 grid points, 95 % of the proposals are accepted for $W = 100$ and 97.5 % for $W = 200$, giving admissible areas of about 12 300 and 44 600 grid points, a ratio of 3.6. The simulated probability that two independent cells intersect above 2 mm h^{-1} is 0.064 for $W = 100$ and 0.016 for $W = 200$, a ratio of 4.0; above 5 mm h^{-1}, where only the cores count, it is 0.007 and 0.003. The contact zone of two cells covers on average about 930 grid points above 2 mm h^{-1} and 125 above 5 mm h^{-1}, and Eq. (A3) overestimates the simulated probability by about a fifth (18 and 27 % at 2 mm h^{-1}, the latter with a sampling uncertainty of about 10 %): near the edge of R_W^* part of the contact zone falls outside

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table A1. Accidental intersection of independently placed cells (Appendix A): acceptance of placement proposals, admissible centre area $|R_W^*|$ (grid points), simulated probability that two independent cells intersect inside the domain above the threshold (standard error in parentheses), mean contact zone $|\Omega_o^{(t)} \oplus (-\Omega_f^{(t)})|$ (grid points) and the approximation of Eq. (A3); for field pairs with N observed and N forecast cells, simulated mean number of intersecting pairs, its approximation $N^2 p_{\mathrm{pair}}$, and probability of at least one intersection.

							$N = 4$			$N = 8$	
W	Threshold	Accepted	$\lvert R_W^* \rvert$	p_{pair}	Contact zone	Eq. (A3)	Mean	$N^2 p$	$P(\geq 1)$	Mean	$N^2 p$	$P(\geq 1)$
100	$> 2\,\mathrm{mm\,h^{-1}}$	94.9 %	12 335	0.0644 (0.0027)	936	0.0759	1.01	1.03	0.63	3.89	4.12	0.98
	$> 5\,\mathrm{mm\,h^{-1}}$	94.9 %	12 335	0.0073 (0.0009)	126	0.0102	0.13	0.12	0.12	0.49	0.46	0.37
200	$> 2\,\mathrm{mm\,h^{-1}}$	97.5 %	44 634	0.0163 (0.0014)	926	0.0207	0.31	0.26	0.27	1.17	1.04	0.69
	$> 5\,\mathrm{mm\,h^{-1}}$	97.5 %	44 634	0.0027 (0.0006)	124	0.0028	0.04	0.04	0.04	0.15	0.18	0.14

the admissible area, and the large cells, whose contact zones dominate the mean, lose more of it. Doubling the number of cells in both fields acts in the opposite direction from enlarging the domain: the per-pair probability does not depend on N, so the expected number of intersections rises fourfold as $N_o N_f$ goes from 16 to 64, from 1.0 to 3.9 per field pair for $W = 100$ at 2 mm h^{-1}, and the probability of at least one intersection from 0.63 to 0.98. Because in G_0 a positive score arises only from such coincidences, both changes shift the whole score distribution of G_0: downwards for the larger domain, upwards for more cells.

Two limits of this calculation should be noted. First, the expected number of intersections does not give the probability that at least one occurs, which must be simulated. Second, and more important for the FSS, two cells need not touch to affect the same neighbourhood: it is enough that they fall within a common window of width L. Such shared-neighbourhood coincidences become more frequent as L grows, and their effect on the score depends on the fractions inside the windows. These coincidences make the distribution of the FSS under G_0 difficult to derive analytically for one field pair; Sect. 3 therefore estimates it by Monte Carlo simulation under the same placement rules.

Appendix B: Gaussian quadratic form and the moment approximation

This appendix shows why a Gaussian quadratic-form (chi-squared) approximation cannot supply the null quantiles of the FSS, which are therefore obtained by Monte Carlo simulation (Sect. 3.1).

The neighbourhood fraction error at location k is $e_k = f_k - o_k$. With M valid neighbourhoods, the numerator of the FSS is the mean squared error

$$\overline{D} = \frac{1}{M} \sum_{k=1}^{M} e_k^2 = \frac{\mathbf{e}^{\mathrm{T}} \mathbf{e}}{M}. \tag{B1}$$

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

If, purely as a mathematical reference, $\mathbf{e}$ is jointly Gaussian with zero mean and covariance matrix Σ_L, diagonalizing Σ_L gives

$$M\overline{D} \overset{d}{=} \sum_{j=1}^{M} \lambda_j Z_j^2, \tag{B2}$$

where λ_j are covariance eigenvalues and the Z_j are independent standard normal variables, a weighted sum of chi-squared variables (Imhof, 1961). Since $\mathrm{E}(Z_j^2) = 1$ and $\mathrm{Var}(Z_j^2) = 2$,

$$\mu_D = \mathrm{E}(\overline{D}) = \frac{\mathrm{tr}(\Sigma_L)}{M}, \qquad v_D = \mathrm{Var}(\overline{D}) = \frac{2\,\mathrm{tr}(\Sigma_L^2)}{M^2}. \tag{B3}$$

Approximating $\overline{D}$ by $h\chi^2_{\nu_{\mathrm{eff}}}$ and equating its first two moments gives

$$\nu_{\mathrm{eff}} = \frac{2\mu_D^2}{v_D} = \frac{[\mathrm{tr}(\Sigma_L)]^2}{\mathrm{tr}(\Sigma_L^2)}, \qquad h = \frac{v_D}{2\mu_D}. \tag{B4}$$

If K independent values $\overline{D}_r$ are available, moment-equivalent degrees of freedom can be estimated as $\widehat{\nu}_{\mathrm{eff}} = 2\widehat{\mu}_D^2/\widehat{v}_D$ from their sample mean and variance. More generally,

$$\mathrm{Var}(\overline{D}) = \frac{1}{M^2} \sum_{i,j} \mathrm{Cov}(e_i^2, e_j^2), \tag{B5}$$

which depends on fourth-order moments; the Gaussian identity $\mathrm{Cov}(e_i^2, e_j^2) = 2\,\mathrm{Cov}(e_i, e_j)^2$ requires joint Gaussianity and zero means. The synthetic fractions are bounded, spatially clustered and frequently zero, so changes in $\widehat{\nu}_{\mathrm{eff}}$ cannot be ascribed uniquely to overlapping windows. More importantly, the approximation describes only the numerator. Writing $\overline{C} = \frac{1}{M}\sum_k f_k o_k$, the identity $\overline{D} = \overline{S} - 2\overline{C}$ gives $\mathrm{FSS} = 2\overline{C}/\overline{S}$ whenever $\overline{S} > 0$: numerator and denominator depend on the same neighbourhood fractions. Even an accurate approximation to the distribution of $\overline{D}$ therefore does not determine the distribution, or the critical values, of the ratio $\overline{D}/\overline{S}$, which depend on the joint distribution of numerator and denominator. All critical values are therefore obtained directly from Monte Carlo simulations of the complete score.

Author contributions. FN conceived the study, designed and carried out the experiments and the analysis, and wrote the manuscript. RT contributed to the design of the study and to the interpretation of the results, and reviewed and edited the manuscript.

Competing interests. No competing interests are present.

Acknowledgements. Use of generative AI. ChatGPT (OpenAI, GPT-6) and Claude (Anthropic, Claude Opus 5.5) were used as interactive research-support tools to explore methodological alternatives, examine statistical assumptions, generate Python code for simulations, analyses and figures, and assist in drafting and revising the manuscript in English. DeepSeek was also used to review manuscript drafts.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

AI-generated suggestions and simulated referee assessments were critically evaluated by the authors, who challenged, modified or rejected recommendations when they were not scientifically justified. The authors formulated the research question, directed the experimental design and the methodological choices, and interpreted the results. The numerical implementation was checked against two independent FSS packages and by reproducing the stored scores from their seeds. The authors reviewed all text and results and take full responsibility for the study and its conclusions.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

References

Ahijevych, D., Gilleland, E., Brown, B. G., and Ebert, E. E.: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts, Weather Forecast., 24, 1485–1497, https://doi.org/10.1175/2009WAF2222298.1, 2009.

Antonio, B.: Investigating how the Fractions Skill Score and Brier Divergence Skill Score reflect forecast error, Mon. Weather Rev., https://doi.org/10.1175/MWR-D-26-0095.1, 2026.

Antonio, B. and Aitchison, L.: How to derive skill from the fractions skill score, Mon. Weather Rev., 153, 1021–1033, https://doi.org/10.1175/MWR-D-24-0120.1, 2025.

Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M., and Bladé, I.: The effective number of spatial degrees of freedom of a time-varying field, J. Climate, 12, 1990–2009, https://doi.org/10.1175/1520-0442(1999)012<1990:TENOSD>2.0.CO;2, 1999.

Crameri, F., Shephard, G. E., and Heron, P. J.: The misuse of colour in science communication, Nat. Commun., 11, 5444, https://doi.org/10.1038/s41467-020-19160-7, 2020.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorol. Appl., 15, 51–64, https://doi.org/10.1002/met.25, 2008.

Ebert, E. E.: Neighborhood verification: a strategy for rewarding close forecasts, Weather Forecast., 24, 1498–1510, https://doi.org/10.1175/2009WAF2222251.1, 2009.

Gilleland, E.: Testing competing precipitation forecasts accurately and efficiently: the spatial prediction comparison test, Mon. Weather Rev., 141, 340–355, https://doi.org/10.1175/MWR-D-12-00155.1, 2013.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of spatial forecast verification methods, Weather Forecast., 24, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1, 2009.

Gilleland, E., Skok, G., Brown, B. G., Casati, B., Dorninger, M., Mittermaier, M. P., Roberts, N., and Wilson, L. J.: A novel set of geometric verification test fields with application to distance measures, Mon. Weather Rev., 148, 1653–1673, https://doi.org/10.1175/MWR-D-19-0256.1, 2020.

Hamill, T. M.: Hypothesis tests for evaluating numerical precipitation forecasts, Weather Forecast., 14, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2, 1999.

Hamill, T. M. and Juras, J.: Measuring forecast skill: is it real skill or is it the varying climatology?, Q. J. Roy. Meteor. Soc., 132, 2905–2923, https://doi.org/10.1256/qj.06.25, 2006.

Hinkley, D. V.: On the ratio of two correlated normal random variables, Biometrika, 56, 635–639, https://doi.org/10.1093/biomet/56.3.635, 1969.

Imhof, J. P.: Computing the distribution of quadratic forms in normal variables, Biometrika, 48, 419–426, https://doi.org/10.1093/biomet/48.3-4.419, 1961.

Leeuwenburg, T., Loveday, N., Ebert, E. E., et al.: scores: a Python package for verifying and evaluating models and predictions with xarray, Journal of Open Source Software, 9, 6889, https://doi.org/10.21105/joss.06889, 2024.

Mittermaier, M. P.: A “meta” analysis of the fractions skill score: the limiting case and implications for aggregation, Mon. Weather Rev., 149, 3491–3504, https://doi.org/10.1175/MWR-D-18-0106.1, 2021.

Mittermaier, M. and Roberts, N.: Intercomparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score, Weather Forecast., 25, 343–354, https://doi.org/10.1175/2009WAF2222260.1, 2010.

https://doi.org/10.5194/egusphere-2026-6860
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Mittermaier, M., Roberts, N., and Thompson, S. A.: A long-term assessment of precipitation forecast skill using the Fractions Skill Score, Meteorol. Appl., 20, 176–186, https://doi.org/10.1002/met.296, 2013.

Necker, T., Wolfgruber, L., Kugler, L., Weissmann, M., Dorninger, M., and Serafin, S.: The fractions skill score for ensemble forecast verification, Q. J. Roy. Meteor. Soc., 150, 4457–4477, https://doi.org/10.1002/qj.4824, 2024.

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), Geosci. Model Dev., 12, 4185–4219, https://doi.org/10.5194/gmd-12-4185-2019, 2019.

Roberts, N.: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model, Meteorol. Appl., 15, 163–169, https://doi.org/10.1002/met.57, 2008.

Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Mon. Weather Rev., 136, 78–97, https://doi.org/10.1175/2007MWR2123.1, 2008.

Skok, G.: Analysis of Fraction Skill Score properties for a displaced rainband in a rectangular domain, Meteorol. Appl., 22, 477–484, https://doi.org/10.1002/met.1478, 2015.

Skok, G. and Roberts, N.: Analysis of Fractions Skill Score properties for random precipitation fields and ECMWF forecasts, Q. J. Roy. Meteor. Soc., 142, 2599–2610, https://doi.org/10.1002/qj.2849, 2016.

Skok, G. and Roberts, N.: Estimating the displacement in precipitation forecasts using the Fractions Skill Score, Q. J. Roy. Meteor. Soc., 144, 414–425, https://doi.org/10.1002/qj.3212, 2018.

Westfall, P. H. and Young, S. S.: Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment, Wiley, New York, 1993.

Wilks, D. S.: "The stippling shows statistically significant grid points": how research results are routinely overstated and overinterpreted, and what to do about it, B. Am. Meteorol. Soc., 97, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1, 2016.