How to Derive Skill from the Fractions Skill Score

BOBBY ANTONIO[a,b] AND LAURENCE AITCHISON[c]

[a] *Department of Physics, University of Oxford, Oxford, United Kingdom*
[b] *School of Geographical Sciences, University of Bristol, Bristol, United Kingdom*
[c] *Machine Learning and Computational Neuroscience Unit, University of Bristol, Bristol, United Kingdom*

(Manuscript received 12 June 2024, in final form 28 February 2025, accepted 14 March 2025)

ABSTRACT: The fractions skill score (FSS) is a widely used metric for assessing forecast skill, with applications ranging from precipitation to volcanic ash forecasts. By evaluating the fraction of grid squares exceeding a threshold in a neighborhood, the intuition is that it can avoid the pitfalls of pixelwise comparisons and identify length scales at which a forecast has skill. The FSS is typically interpreted relative to a "useful" criterion, where a forecast is considered skillful if its score exceeds a simple reference score. However, the typical reference score used is problematic, since it is not derived in a way that provides obvious meaning, does not scale with neighborhood size, and may not be exceeded by forecasts that have skill. We, therefore, provide a new method to determine forecast skill from the FSS, by deriving an expression for the FSS achieved by a random forecast, which provides a more robust and meaningful reference score to compare with. Through illustrative examples, we show that this new method considerably changes the length scales at which a forecast would be regarded as skillful and reveals subtleties in how the FSS should be interpreted.

SIGNIFICANCE STATEMENT: Forecast verification metrics are crucial to assess accuracy and identify where forecasts can be improved. In this work, we investigate a popular verification metric, the fractions skill score, and derive a more robust method to decide if a forecast has sufficiently high skill. This new method significantly improves the quality of insights that can be drawn from this score.

KEYWORDS: Forecast verification/skill; Diagnostics; Model errors; Model evaluation/performance; Numerical weather prediction/forecasting; Error analysis

1. Introduction

Assessing the performance of numerical weather prediction models is crucial for monitoring and guiding model development and is also extremely challenging, particularly for fields like precipitation that exhibit high spatial variability. One approach to address the double penalty issue that occurs for pixelwise comparisons (Wilks 2019) is to use aggregated quantities in a neighborhood around each grid point to assess the change in skill as the neighborhood size increases (Ebert 2008). A commonly used score in this category is the fractions skill score (FSS) (Roberts and Lean 2008; Roberts 2008), which evaluates the fractions of grid squares that exceed a threshold in a neighborhood surrounding each grid point. This score has been used to evaluate cutting edge machine learning weather prediction systems (Ayzel et al. 2020; Ravuri et al. 2021), convection permitting models (Woodhams et al. 2018; Weusthoff et al. 2010; Cafaro et al. 2021; Schwartz 2019), volcanic ash forecasts (Harvey and Dacre 2016), oil spill forecasts (Simecek-Beatty and Lehr 2021), and flood inundation forecasts (Hooker et al. 2022), as a loss function for training models (Ebert-Uphoff et al. 2021; Lagerquist et al. 2021; Lagerquist and Ebert-Uphoff 2022; Price and Rasp 2022), and has been proposed as a replacement for the equitable threat score in operational forecast verification (Mittermaier et al. 2013).

The fractions skill score is typically categorized as a "neighborhood" approach to forecast verification (Ebert 2008; Gilleland et al. 2009) for which the quality of forecasts are measured by comparing the neighbos around each grid square. The result of aggregating features over neighbos has the effect of blurring the forecast and observations and so was introduced as a way to mitigate the double penalty problem and assess the length scale at which the forecast becomes of high enough quality. Alternatively, the use of neighbos can be interpreted as a way to resample the probability distribution of forecasts and observations (Theis et al. 2005), which then motivates the use of probabilistic scores such as the Brier skill score (Brier 1950) and Brier divergence skill score (Stein and Stoop 2024). Other neighbo approaches include comparing ordered samples within neighbos (Rezacova et al. 2007), upscaling the data before comparison (Marsigli et al. 2008), and using neighbos to create contingency tables (Ebert 2008; Schwartz 2017).

The FSS was originally proposed in Roberts (2008) and Roberts and Lean (2008). We consider T forecast and observation pairs, defined over a domain of size $N_x \times N_y$. Here, T typically represents the time steps over which the forecast is being assessed, but it is not limited to this interpretation and

Denotes content that is immediately available upon publication as open access.

Corresponding author: Bobby Antonio, bobby.antonio@physics.ox.ac.uk

DOI: 10.1175/MWR-D-24-0120.1

refers to any collection of forecasts that are being aggregated. The fractions skill score is defined as

$$\mathrm{FSS}(n) = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}[f(n)_{ijt} - x(n)_{ijt}]^2}{\sum_{t=1}^{T}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}[f(n)_{ijt}^2 + x(n)_{ijt}^2]}, \tag{1}$$

$$\equiv \frac{\sum_{t=1}^{T}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}2f(n)_{ijt}x(n)_{ijt}}{\sum_{t=1}^{T}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}[f(n)_{ijt}^2 + x(n)_{ijt}^2]}, \tag{2}$$

where $f(n)_{ijt}$ [$x(n)_{ijt}$] is the fraction of forecast (observed) grid squares in the sample image at t that exceeds an event threshold within a square window of width $2n + 1$ centered at grid square i, j, where $0 \le n \le \max(N_x, N_y)$. Averaging over the T different samples is typically performed separately for the numerator and denominator before combining rather than taking an average of FSS values from different samples, since this reduces the possibility of comparing completely dry forecasts and observations, which result in an undefined score (Mittermaier 2021). Other variants exist whereby neighborhoods are constructed by aggregating in time instead of space [referred to as the localized FSS in Woodhams et al. (2018)] and using ensembles (Duc et al. 2013; Necker et al. 2023). For the most part, these variants differ in how data are aggregated to create a neighborhood and how the T samples are interpreted. For example, for ensemble forecasts, t can be used to indicate the ensemble members or the neighborhood could be constructed by averaging over ensemble members. Therefore, in these cases, the mathematical construction of the score still fits within the form of Eq. (2), and our results can be extended to these different scenarios. The exceptions are scores such as $\mathrm{FSS}_{\mathrm{single}}$ and $\mathrm{FSS}_{\mathrm{mean}}$ in Necker et al. (2023), for which the scores are calculated as an average over scores for different samples or ensemble members. However, we would still expect broadly similar results to occur from scores that have this modification.

A key part of interpreting the FSS is deciding on what level the FSS must reach such that a forecast is of high enough quality; this is referred to as "useful skill" in Roberts (2008) and Roberts and Lean (2008). In Roberts (2008), a method to interpret the skill from the FSS is provided, such that a forecast has useful skill if the FSS value exceeds a reference score of $[1 + x(0)]/2$, where $x(0)$ is the frequency with which the precipitation event is seen in the observations at the grid scale. The same reference score has also been proposed as a means to estimate the displacement of precipitation objects (Skok 2015; Skok and Roberts 2018), discussed further in section 3.

Despite its widespread use, there are two key problems with evaluating forecast skill by comparing with the reference score of $[1 + x(0)]/2$. First, it is known that forecasts that do not exceed this score can still have considerable skill (Nachamkin and Schmidt 2015; Mittermaier et al. 2013). Second, this reference score is derived at the grid scale, using inconsistent forecast definitions in the numerator and denominator (Skok 2015), such that it does not have a straightforward interpretation across all neighborhood sizes. With this as motivation, we present a much more robust method to assess forecast skill from the FSS, by deriving a baseline FSS for random forecasts. We demonstrate that this derivation aligns precisely with FSS results for actual random data and that it considerably changes how forecast skill is interpreted from the FSS.

This paper is laid out as follows: In section 2, we present a decomposition of the FSS in terms of summary statistics. In section 3, we explore existing approaches to derive skill from the FSS and present a new method based on comparison with the FSS of a random forecast. Concluding remarks are given in section 4.

2. Decomposing the FSS

We begin by rewriting the FSS in Eq. (2) in a novel way that reveals the underlying factors that drive the score and makes constructing a reference score possible. We use angle bracket notation $\langle x \rangle$ to indicate the sample mean calculated over all grid points. Explicitly, it is defined as

$$\langle x \rangle := \frac{1}{TN_xN_y}\sum_{t=1}^{T}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}x_{ijt}. \tag{3}$$

Using this notation, the FSS equations in Eqs. (1) and (2) can be written as

$$\mathrm{FSS}(n) = 1 - \frac{\langle [f(n) - x(n)]^2 \rangle}{\langle f(n)^2 + x(n)^2 \rangle}$$

$$= 1 - \frac{\langle f(n)^2 \rangle + \langle x(n)^2 \rangle - 2\langle x(n)f(n) \rangle}{\langle f(n)^2 \rangle + \langle x(n)^2 \rangle}, \tag{4}$$

$$= \frac{2\langle x(n)f(n) \rangle}{\langle f(n)^2 \rangle + \langle x(n)^2 \rangle}, \tag{5}$$

where $\langle x(n) \rangle$ and $\langle f(n) \rangle$ are the sample neighborhood frequency for observations and forecast, respectively, calculated over all square neighborhoods of width $2n + 1$. We define $s_{x,n}$ and $s_{f,n}$ as the (uncorrected) sample standard deviations for observations and forecast:

$$s_{x,n}^2 := \langle [x(n) - \langle x(n) \rangle]^2 \rangle = \langle x(n)^2 \rangle - \langle x(n) \rangle^2, \tag{6}$$

$$s_{f,n}^2 := \langle [f(n) - \langle f(n) \rangle]^2 \rangle = \langle f(n)^2 \rangle - \langle f(n) \rangle^2. \tag{7}$$

Note that these are biased estimates of the true standard deviations, since we are dividing by TN_xN_y, rather than $(TN_xN_y - 1)$ (Von Storch and Zwiers 2002). Here, we choose to use the biased estimator since it ensures that all terms have consistent denominators, and we assume that the domain has $N_x > 10$, $N_y > 10$ such that the biased and unbiased estimates will be very similar. The r_n is defined as the sample Pearson correlation coefficient between the forecast and observed fractions:

$$r_n := \frac{1}{s_{f,n}s_{x,n}}\langle[x(n) - \langle x(n)\rangle][f(n) - \langle f(n)\rangle]\rangle,$$
$$= \frac{1}{s_{f,n}s_{x,n}}[\langle x(n)f(n)\rangle - \langle x(n)\rangle\langle f(n)\rangle]. \quad (8)$$

With these definitions, we are now in a position to express Eq. (5) in terms of the sample statistics. A rearrangement of Eq. (8) gives an expression for the numerator term:

$$\langle x(n)f(n)\rangle = s_{x,n}s_{f,n}r_n + \langle x(n)\rangle\langle f(n)\rangle. \quad (9)$$

Rearranging Eq. (6) gives

$$\langle x^2(n)\rangle = s^2_{x,n} + \langle x(n)\rangle^2, \quad (10)$$

and similarly for $\langle f^2(n)\rangle$, so that the denominator can be written as

$$\langle f^2(n)\rangle + \langle x^2(n)\rangle = s^2_{x,n} + \langle x(n)\rangle^2 + s^2_{f,n} + \langle f(n)\rangle^2. \quad (11)$$

Inserting Eqs. (11) and (9) into Eq. (5), we arrive at a decomposed version of the FSS:

$$\mathrm{FSS}(n) = \frac{2\langle x(n)\rangle\langle f(n)\rangle + 2s_{x,n}s_{f,n}r_n}{\langle x(n)\rangle^2 + \langle f(n)\rangle^2 + s^2_{x,n} + s^2_{f,n}}. \quad (12)$$

Despite the simplicity of this derivation, this expression of the FSS has not to the authors' knowledge been shown in existing literature, although decompositions of the mean-square error in this way are common (e.g., Murphy 1988), and a similar decomposition is arrived at in the context of the intensity scale skill score (Casati et al. 2023). If we limit ourselves to the case where the data at the grid scale have no spatial correlations, then $f(n)_{ijt}$ and $x(n)_{ijt}$ are independent and drawn from a binomial distribution, and we recover the results in Skok and Roberts (2016).

To show the explicit properties of the neighborhood terms, we can arrive at expressions for $\langle x(n)\rangle$, $\langle f(n)\rangle$, $s_{f,n}$, and $s_{x,n}$ in terms of quantities calculated at the grid scale and the spatial autocorrelations (see the appendix). The effect of zero padding used to perform square convolutions at the edges [as is done for a standard implementation of the FSS (Pulkkinen et al. 2019)] makes the derivation of such expressions slightly more complicated. When using percentile thresholds to remove intensity biases, we observe that the neighborhood frequency can still be reasonably different between forecast and observations when using zero padding, in contrast to using a scheme that pads with data from within the image, such as reflective padding. For this reason and because it allows much simpler expressions for the neighborhood mean and standard deviation later in this section, we calculate the FSS with reflective padding in this work. Another option is to not use padding, so that the number of grid cells to be compared shrinks as the neighborhood size grows; we do not consider this in our work; however, a similar analysis would still apply with different definitions of how $\langle x(n)\rangle$, $\langle f(n)\rangle$, $s_{x,n}$, $s_{f,n}$, and r_n are calculated.

Derivations of neighborhood mean and standard deviation under the assumption of reflective padding are given in the appendix. The neighborhood mean is equal to that on the grid scale, i.e., $\langle x(n)\rangle = \langle x(0)\rangle$ and $\langle f(n)\rangle = \langle f(0)\rangle$. The neighborhood standard deviation $s_{x,n}$ can be written as

$$s^2_{x,n} = \frac{\langle x(0)\rangle[1 - \langle x(0)\rangle]}{(2n+1)^2}\left[1 + \sum_{d=1}^{(2n+1)} \nu_x(d)\Omega_d(n)\right], \quad (13)$$

and similarly for $s_{f,n}$, where $\nu_x(d)$ $[\nu_f(d)]$ is an estimate of the spatial autocorrelation between two grid squares a distance d apart within the observations (forecasts), and $\Omega_d(n)$ accounts for the number of pairs of points within a neighborhood that are separated by distance d. Thus, $s_{x,n}$ and $s_{f,n}$ depend on $\langle x(0)\rangle$, the size of the neighborhood, and the spatial autocorrelation.

3. How to interpret the FSS

In this section, we begin by summarizing and clarifying previous results on how to interpret the FSS, before establishing a more robust method to assess forecast skill based on comparison to random forecasts.

a. *Summary of existing approaches*

We begin by summarizing previously derived approaches for defining the no-skill to skill transition point from the FSS. In previous works, this has been defined as the point where the FSS for a forecast exceeds that of a simple reference score.

The first reference score is described in Roberts and Lean (2008) as "the FSS that would be obtained from a random forecast with the same fractional coverage over the domain as . . . the base rate, $[\langle x(0)\rangle]$." In other words, the score for a forecast sampled from a Bernoulli distribution at the grid scale, with the Bernoulli probability set to $\langle x(0)\rangle$. This is given in Roberts and Lean (2008) as

$$\mathrm{FSS}_{\mathrm{random}} = \langle x(0)\rangle. \quad (14)$$

However, this reference score is only accurate for a neighborhood size of 1 (i.e., at the grid scale), and we shall show later in this section how it may be derived more rigorously. Because Eq. (14) scales with $\langle x(0)\rangle$, it is typically too small to be informative and so does not appear to be used in the literature.

The most widely used reference score is defined as "the FSS that would be obtained at the grid scale . . . from a forecast with a fraction equal to $[\langle x(0)\rangle]$ at every point" (Roberts and Lean 2008), defined as

$$\mathrm{FSS}_{\mathrm{uniform}} = \frac{1}{2} + \frac{\langle x(0)\rangle}{2}. \quad (15)$$

However, as noted in Skok (2015), Eq. (15) does not result from the description given in Roberts and Lean (2008) and in fact results from setting $f(0)_{ijt} = \langle x(0)\rangle$ in the numerator and using a random binary forecast with mean $\langle f(0)\rangle = \langle x(0)\rangle$ in

the denominator. We can verify this by inserting these definitions into Eq. (1):

$$\begin{aligned}\mathrm{FSS}_{\mathrm{uniform}} &= 1 - \frac{\langle[\langle x(0)\rangle - x(0)]^2\rangle}{\langle f(0)^2 + x(0)^2\rangle} = 1 - \frac{s_{o,0}^2}{\langle f(0)^2\rangle + \langle x(0)^2\rangle},\\ &= 1 - \frac{\langle x(0)\rangle[1 - \langle x(0)\rangle]}{2\langle x(0)\rangle} = \frac{1}{2} + \frac{\langle x(0)\rangle}{2}, \end{aligned} \tag{16}$$

where we have used the fact that $\langle x(0)^2\rangle = \langle x(0)\rangle$, and similarly, $\langle f(0)^2\rangle = \langle f(0)\rangle = \langle x(0)\rangle$, since the data are binary at the grid scale.

Note that if we take the same definitions for the $f(0)_{ijt}$ values but instead start from the rearranged form of the FSS in Eq. (2), we arrive at a different result, since the numerator and denominator are not consistent with one another:

$$\mathrm{FSS}_{\mathrm{uniform}} = \frac{2\langle x(0)\rangle\langle x(0)\rangle}{\langle f(0)^2\rangle + \langle x(0)^2\rangle} = \frac{2\langle x(0)\rangle^2}{2\langle x(0)\rangle} = \langle x(0)\rangle. \tag{17}$$

Since this reference score is derived using different forecast definitions on the numerator and denominator and is only derived at the grid scale, it has no obvious interpretation and does not necessarily scale properly with neighborhood size. Previous work has also demonstrated that forecasts not exceeding this reference score can still have considerable skill (Nachamkin and Schmidt 2015).

A derivation of a similar reference score is shown in Skok and Roberts (2016), when the forecasts and observation events are independent and are assumed to be drawn from a Bernoulli distribution at the grid scale. Under these idealized assumptions, the FSS is shown to be equal to the reference score when the average number of rainy grid squares within the neighborhood equals 1. While there is a more solid mathematical derivation to this, it is not clear why this is a sensible reference score with which to assess the skill of a forecast. It also relies on the assumption that there are no spatial correlations, which is clearly not true for real data.

The point at which the FSS reaches $\mathrm{FSS}_{\mathrm{uniform}}$ is also motivated as a means of estimating the displacement of forecast objects. Intuitively, increasing the FSS length scale reduces the effects of position errors in the forecast, and the point at which the FSS meets a critical point contains information about the displacement of precipitation objects. It can be shown that for idealized narrow vertical rainbands and distant sets of circular rainfall patterns (Roberts and Lean 2008; Skok 2015; Skok and Roberts 2018):

$$\mathrm{FSS}(n) = 1 - \frac{d}{2n + 1}, \tag{18}$$

where d is the displacement between forecast and observation. This motivates the comparison between the FSS and $\mathrm{FSS}_{\mathrm{uniform}}$ as a means to estimate forecast displacement (where the resulting estimate in displacement is typically denoted dFSS). Skok and Roberts (2018) analyze the behavior of dFSS from real forecasts compared to reanalysis data, finding that dFSS appears to correlate well with the actual displacement but that the measure is sensitive to frequency biases and the relative sizes of the precipitation objects. A subsequent comparison of several different distance measures (Gilleland et al. 2020) found that although in many cases, the dFSS produces scores that correlate well with displacement, it is often undefined for pathological cases where one or both of the forecast and observations are 0, is sensitive to the positioning and orientation of the precipitation objects within the domain, and becomes problematic for very large biases in the event frequency.

b. An improved method to interpret the FSS

Having summarized previous results, we now present a more meaningful method to interpret the FSS. We do this by comparing the FSS to the score that would be achieved by a random forecast, where the random forecast is constructed by sampling from a Bernoulli distribution at the grid scale, with the Bernoulli probability set to $\langle x(0)\rangle$. Forecasts that achieve an FSS exceeding this baseline are then interpreted as having skill relative to that reference. This aligns with the standard concept of skill as defined in, e.g., Wilks (2019) and also appears to be the original intention in Roberts and Lean (2008) and Roberts (2008), where they refer to useful skill.

Note the difference from the work in Skok and Roberts (2016), in which both forecast and observation are assumed to be sampled from Bernoulli distributions at the grid scale, whereas here crucially only the reference forecast is. In Skok and Roberts (2016), the authors use these simplified random forecasts and observations to make the FSS mathematically tractable to study its properties, whereas here we are using random forecasts as a baseline to compare against. Using random observations for this application is, therefore, inappropriate, as it would provide an unrealistic reference score.

We note that other definitions of useful are possible and that in general these different definitions will give rise to different reference scores. This appears to be the case for using the FSS to estimate forecast displacement (as discussed in the previous subsection). We regard estimation of distance errors using the FSS as a separate problem and refer the reader to Skok and Roberts (2018) and Gilleland et al. (2020) for detailed insights of how the FSS can be used for this task.

We now show how skill relative to a random forecast can be derived for the FSS. Despite being named as a skill score, the FSS differs from other skill scores in that the reference forecast used is dependent on the forecast itself [and in fact may not be achievable by any forecast (Mittermaier 2021)]. This means that, unlike conventional skill scores, it is not straightforward to see whether or not a forecast has skill, which necessitates the following derivation.

We start with Eq. (12) and consider the most straightforward situation where the random forecast is not correlated with observations, so $r_n = 0$. The random binary forecast at each grid square is constructed by sampling from a Bernoulli distribution, with probability $\langle x(0)\rangle$, so that $\langle f(0)\rangle = \langle x(0)\rangle$, and therefore, as discussed in section 2, $\langle f(n)\rangle = \langle x(n)\rangle$. For the standard deviation term $s_{f,n}$, we use the expression for the neighborhood standard deviation given by Eq. (13) with $\nu_d = 0$ since these forecast data are sampled independently at

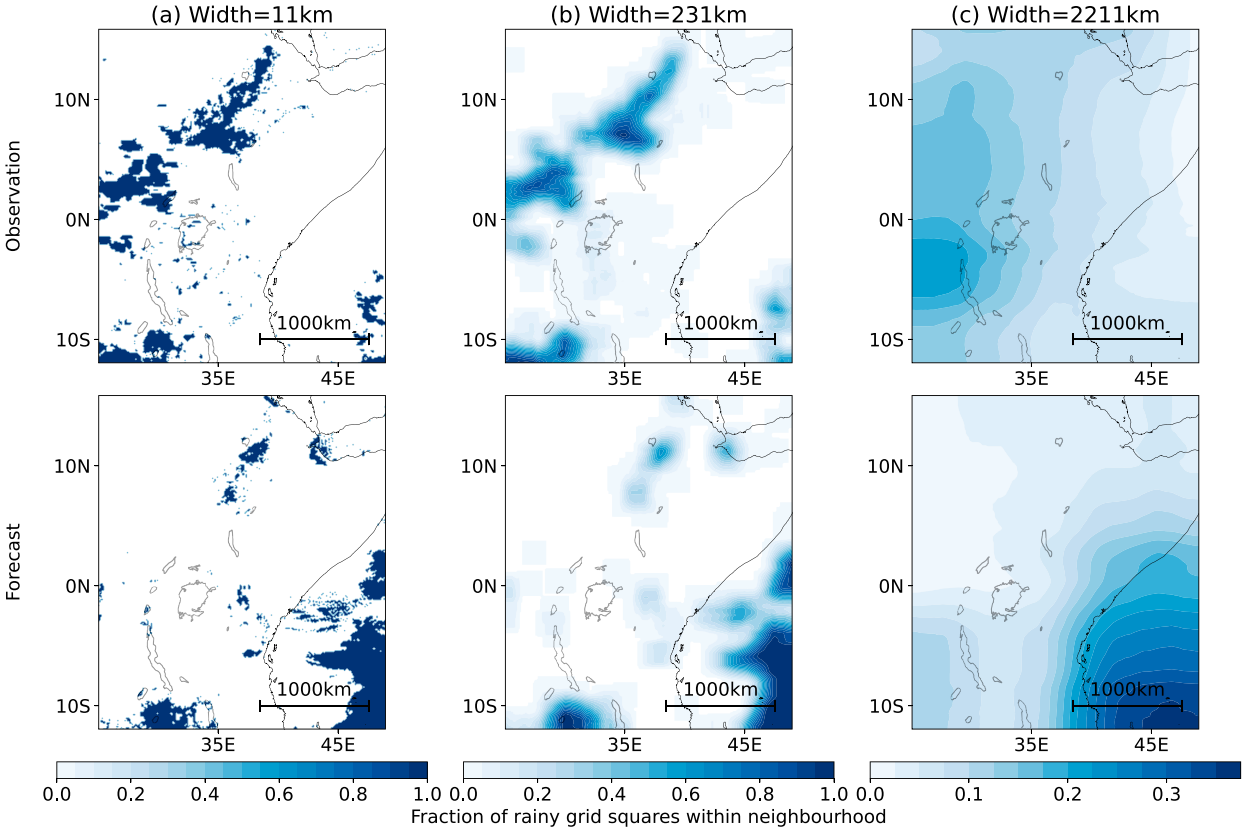

FIG. 1. Images of the fraction of neighboring grid squares at different neighborhood widths for the first example, corresponding to the scores shown in Fig. 2. Each column shows the result of converting (top) observations and (bottom) forecasts to a binary mask by applying a 90th percentile threshold and then calculating fractions of rainy pixels in a square neighborhood around each pixel, with neighborhood width given at the top of each column. (a) Fractions with a neighborhood width of 11 km, (b) fractions with a neighborhood width of 231 km, and (c) fractions with a neighborhood width of 2211 km (around the point where the neighborhood correlation is maximally negative).

the grid scale, and so there are no spatial correlations by construction. Substituting these into Eq. (12) gives

$$\begin{aligned}\mathrm{FSS}_{\mathrm{random}}(n) &= \frac{2\langle x(n)\rangle\langle f(n)\rangle}{\langle x(n)\rangle^2 + \langle f(n)\rangle^2 + s_{f,n}^2 + s_{x,n}^2} \\ &= \frac{2\langle x(n)\rangle^2}{2\langle x(n)\rangle^2 + s_{f,n}^2 + s_{x,n}^2}, \\ &= \frac{2\langle x(n)\rangle^2}{2\langle x(n)\rangle^2 + \dfrac{1}{(2n+1)^2}\langle x(0)\rangle[1-\langle x(0)\rangle] + s_{x,n}^2}.\end{aligned} \tag{19}$$

We can see that for a neighborhood width of 1 where $n = 0$ and $s_{o,0} = \langle x(0)\rangle[1 - \langle x(0)\rangle]$, we recover the reference score in Eq. (14) from Roberts and Lean (2008) as expected:

$$\begin{aligned}\mathrm{FSS}_{\mathrm{random}}(0) &= \frac{2\langle x(0)\rangle^2}{2\langle x(0)\rangle^2 + \langle x(0)\rangle[1-\langle x(0)\rangle] + \langle x(0)\rangle[1-\langle x(0)\rangle]} \\ &= \langle x(0)\rangle.\end{aligned} \tag{20}$$

It is straightforward to use this formula to explore more nuanced reference scores, such as where there is nonzero correlation between the random forecast and observations. However, we take the correlation to be zero to explore the simplest case and in lieu of an obvious choice of correlation value to choose. It is also possible to use this formula to represent an approximate FSS for reference forecasts such as climatology or persistence, for which we would expect there to be nonzero correlation between observations and forecast. However, this would require an accurate estimate of this correlation, and therefore, it may be more accurate to simply calculate the FSS for a climatological forecast empirically rather than using a formula.

We now examine how comparing to the reference score in Eq. (19) changes the interpretation of the FSS by plotting some illustrative examples on real data, chosen to highlight particularly interesting behaviors. For observations, we use data collected by the Global Precipitation Measurement (GPM) satellites, processed using the Integrated Multi-satellitE Retrievals for GPM (IMERG) algorithm (Huffman et al. 2022). For forecasts, we use data from the European Centre for Medium-Range

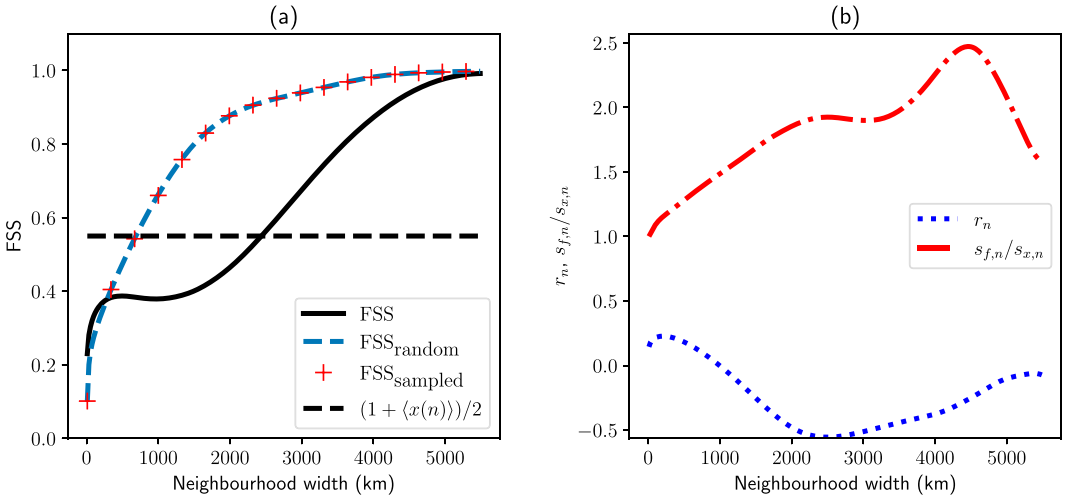

FIG. 2. An example of the FSS, using 90th percentile thresholds, for 6-h accumulated rainfall between 18 and 24 h on 15 Mar 2019. (a) The FSS (solid black line), the standard reference score for the FSS (black dashed line), the improved reference score based on random forecasts (FSS$_{\text{random}}$), and the FSS achieved from a Bernoulli forecast with the same frequency as the observations (FSS$_{\text{sampled}}$). (b) The neighborhood correlation r_n and relative sizes of the neighborhood quantities $s_{x,n}$ and $s_{f,n}$ that contribute to the FSS. Note that $\langle f(n) \rangle = \langle x(n) \rangle$ since we are using percentile thresholds.

Weather Forecasts (ECMWF) Integrated Forecasting System (IFS). Both datasets are regridded to 0.1° × 0.1° resolution and hourly time steps, over the time period October 2018–June 2019 and over equatorial East Africa 12°S–15°N 25°–51°E; this is an area that has problems with extreme rainfall and drought and in which standard rainfall parameterization schemes typically struggle to perform well due to the dominance of convective rainfall (Woodhams et al. 2018). For all examples, we use 90th percentile thresholds to remove frequency biases, as is the recommended way to calculate the FSS in Roberts and Lean (2008) and Skok and Roberts (2018).

To validate that our derivation is accurate, we plot the results alongside the FSS for an actual forecast constructed by sampling from a Bernoulli distribution at each grid square independently, with probability equal to the observed event frequency at the grid scale. The score achieved by this sampled forecast is labeled as FSS$_{\text{sampled}}$ in the figures. Due to the domain size, the variability in score between sampled forecasts is small; hence, only one sample is shown here. Alongside each plot of the FSS, we also show the values of $s_{f,n}/s_{x,n}$ and the neighborhood correlation r_n, to illustrate the factors underpinning the scores [note that $\langle f(n) \rangle = \langle x(n) \rangle$ since we are using percentile thresholds].

Maps of forecast and observation fractions for the first example, calculated over three different neighborhoods, are shown in Fig. 1. We can see that at a neighborhood width of 231 km (Fig. 1b), the fields are slightly blurred but retain most of the structure, and at a much higher neighborhood width of 2211 km (201 grid points, Fig. 1c), the fields are very smooth, with highest fractions occurring in different parts of the domain for the forecast and observations.

The FSS curves for this example are shown in Fig. 2, with plots showing fractions skill scores on the left-hand side and plots showing the behavior of the correlation and neighborhood standard deviation on the right-hand side. For this and plots in Figs. 4 and 5, the solid black line shows the FSS for the actual forecast and observation data, while the dashed blue line shows the FSS calculated from Eq. (19). The dashed black line shows the standard useful threshold, from which it is common to assess a forecast as being useful when it crosses that line. The red crosses show the FSS realized from an actual random forecast, created by sampling from a Bernoulli distribution at the grid scale. In the right-hand plots, the neighborhood standard deviation is expressed as the ratio $s_{f,n}/s_{x,n}$, so that higher values indicate where the forecast has higher neighborhood standard deviation than the observations.

Our first observation of the curves in Fig. 2a is that the newly derived reference score is barely distinguishable from the FSS achieved from the random forecasts sampled from a Bernoulli distribution, FSS$_{\text{sampled}}$, confirming that the new reference score is a good approximation to that achieved by the random forecast we are considering in this work. In contrast, the standard reference score bears no resemblance to it, including at the grid scale. The same is true of the FSS curves in Figs. 4 and 5.

The results in Figs. 2a and 2b show particularly striking behavior, in that the FSS curve meets the standard reference score (black dashed line) at a neighborhood width of around 2500 km, at which point the neighborhood correlation between the forecast and observations is substantially negative (as shown in Fig. 2b). This is reflected in Fig. 1c, where the anticorrelation between the forecast and observation fractions is clear. In contrast, the newly derived reference score FSS$_{\text{random}}(n)$ is larger than the FSS curve within this range, correctly identifying this region of negative neighborhood correlation as unskillful; only at low neighborhoods (less than around 200 km) is the forecast better than the random benchmark.

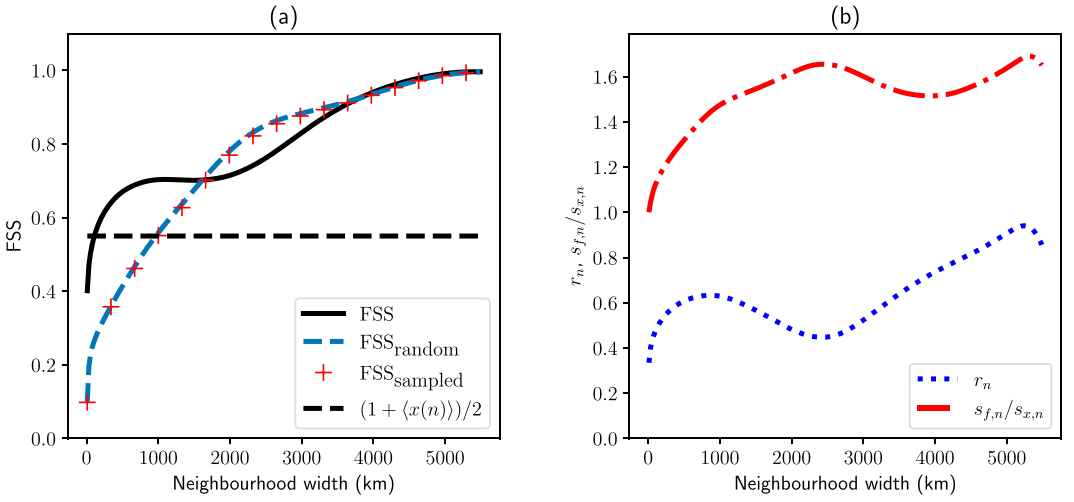

FIG. 3. As in Fig. 2, but for 6-h accumulated rainfall between 18 and 24 h on 16 Mar 2019.

Similar but not as extreme behavior is seen in Fig. 3; the dip in neighborhood correlation and increase in $s_{f,n}/s_{x,n}$ in Fig. 3b coincide with the FSS dip below FSS$_{\text{random}}$ around 1000 km, before rising again at around 3500 km. This and the previous example highlight that, contrary to the typical interpretation, there is a spatial scale beyond which the forecast is useful and there are in fact ranges of spatial scales for which the forecast has skill.

In Fig. 4, the FSS exceeds FSS$_{\text{random}}$ over all neighborhood widths and exceeds the standard reference score at around 100 km. This highlights how the standard reference score can set much too high a bar at low neighborhood sizes and, in some instances, erroneously labels forecasts at the grid scale as unskillful. Notice that the increase in bias $s_{f,n}/s_{x,n}$ seen above a neighborhood width of 4000 km does not affect the score, since at this point $s_{x,n}$ and $s_{f,n}$ are much lower than $\langle x(n) \rangle$ and $\langle f(n) \rangle$.

In contrast to the example in Fig. 4, the example in Fig. 5 shows a case where the FSS does not exceed FSS$_{\text{random}}$ for any length scale, despite crossing the standard reference score line at a width of around 1500 km. For this example, the bias in the neighborhood standard deviation $s_{f,n}/s_{x,n}$ rises as the neighborhood correlation does, with a net effect of no skill. This highlights the tradeoffs that are being made between different forecast errors. Further insight for this example can be obtained from the plots of fractions in Fig. 6. From Eq. (13), we can see that any differences in $s_{f,n}$ and $s_{x,n}$ must be due to the spatial autocorrelation, since we are using percentile thresholds which remove frequency biases. This is indeed what is seen at a neighborhood width of 1551 km in Fig. 6; the forecast fraction is more densely concentrated and so has larger spatial autocorrelations at ranges up to about 1000 km, whereas the observations show a more diffuse pattern with lower short-range spatial correlations. While the standard reference score would not make

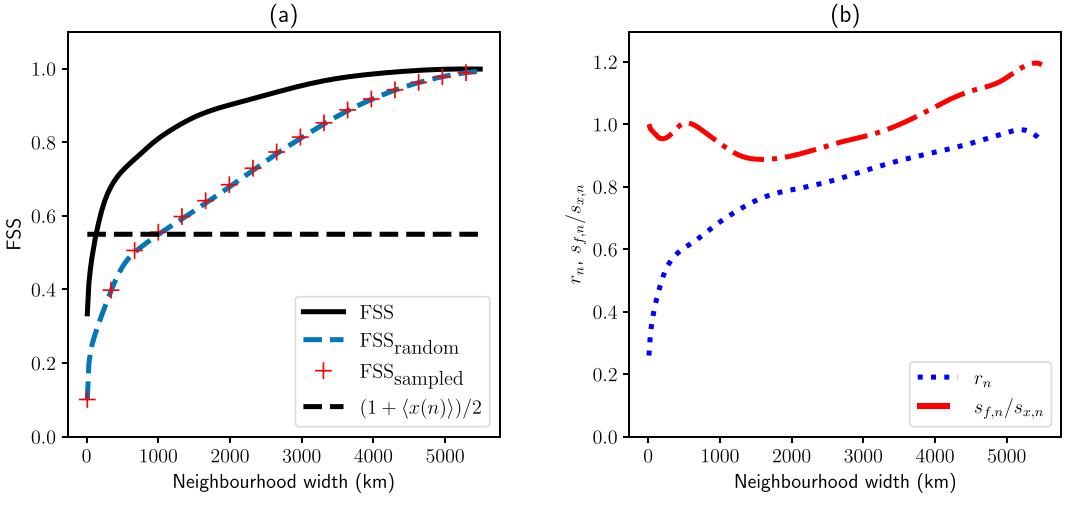

FIG. 4. As in Fig. 2, but for 6-h accumulated rainfall between 18 and 24 h on 1 Mar 2019.

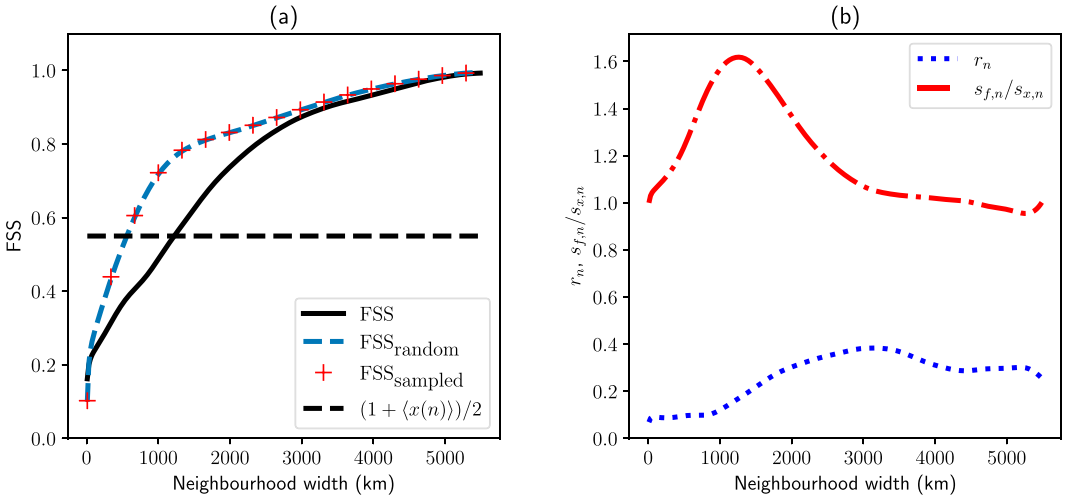

FIG. 5. As in Fig. 2, but for 6-h accumulated rainfall between 0 and 6 h on 31 May 2019.

this region of low skill visible, the large gap between the calculated FSS values and FSS$_{\text{random}}$ highlights more clearly which neighborhood lengths are problematic, in a way that also agrees with the underlying differences in $s_{f,n}$ and $s_{x,n}$.

To summarize, in this section, we have presented a more rigorous derivation of a reference score for the FSS, such that if the FSS exceeds this score, the forecast can be seen as superior to a random forecast with event frequency equal to that

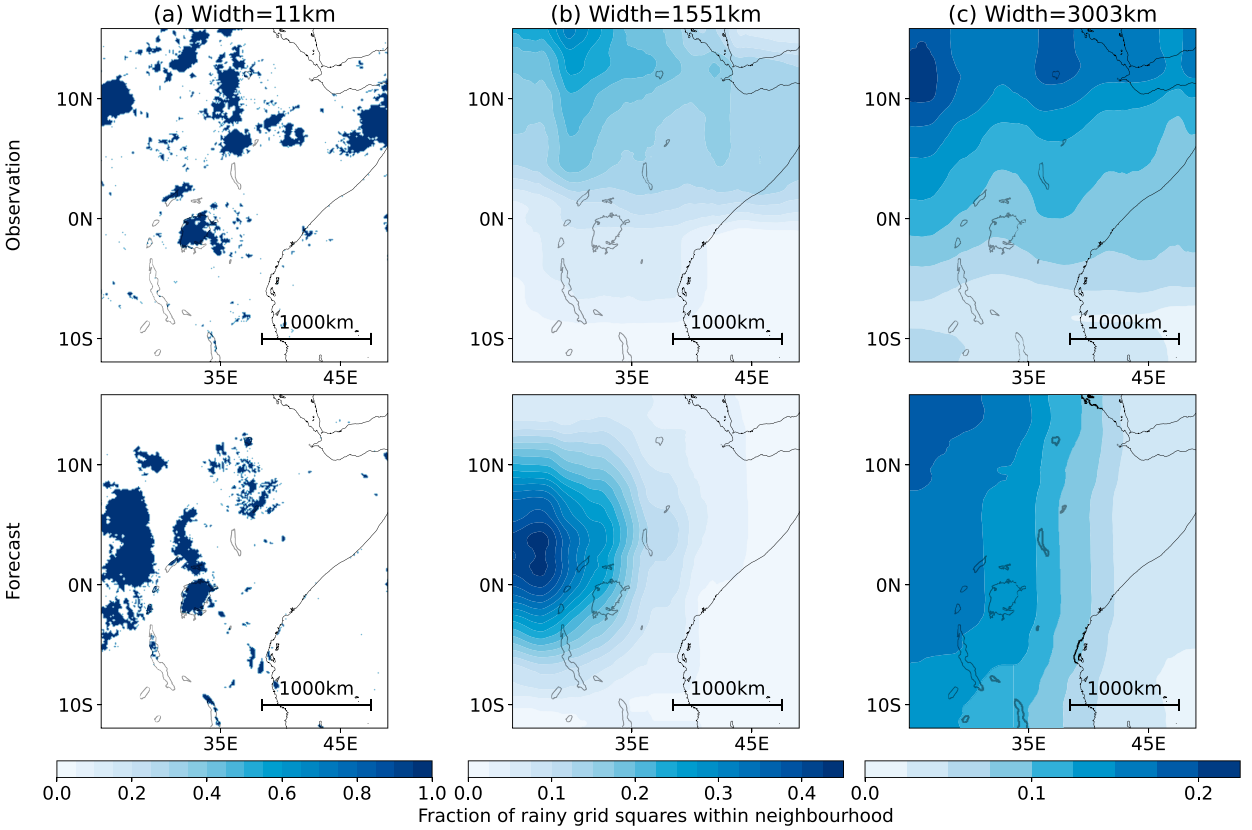

FIG. 6. Images of the fraction of neighboring grid squares at different neighborhood widths, for the case in Fig. 5. Each column shows the result of converting (top) observations and (bottom) forecasts to a binary mask by applying a 90th percentile threshold and then calculating fractions of rainy pixels in a square neighborhood around each pixel, with neighborhood width given at the top of each column. (a) Fractions with a neighborhood width of 11 km, (b) fractions with a neighborhood width of 1551 (around the peak in $s_{f,n}/s_{x,n}$), and (c) fractions with a neighborhood width of 3003 km.

of the observations. In contrast to the existing reference score, which is derived at the grid scale only and uses inconsistent terms in the numerator and denominator, this new reference score scales appropriately with the neighborhood size and is mathematically consistent. This is verified by comparing both reference scores to the average FSS achieved for actual random forecasts, for which the new reference score is a precise match.

Through illustrative examples, we have also demonstrated how this new reference score significantly changes the interpretation of FSS results. One particularly striking example is that the FSS can exceed the standard reference score while being substantially negatively correlated with observations, even when other neighborhood biases are small. In contrast, the newly derived reference score correctly identifies this as a region of no skill. These examples also show that it is more accurate to say that there are ranges of spatial scales that are skillful, instead of the typical interpretation that there is a spatial scale beyond which the forecast is skillful.

4. Discussion and conclusions

In this work, we have provided a new method for interpreting skill from the fractions skill score (FSS), by deriving a new reference score corresponding to the score achieved by a random forecast; a score that exceeds this new reference score can be said to have skill relative to the random forecast. In contrast to the standard reference score, which is derived at the grid scale and has unclear meaning due to the inconsistent use of terms in the derivation, this reference score aligns precisely with the FSS achieved for actual random data and has a clear interpretation. It also considerably alters how the FSS would be interpreted in many situations and, therefore, presents a significant improvement to the insights that can be drawn from the FSS. One particularly interesting example shows that a forecast can exceed the standard reference score when the neighborhood correlation between forecasts and observations is substantially negative. In contrast, the FSS for this situation does not exceed the newly derived reference score, demonstrating that interpreting results relative to this new reference score aligns more closely with our intuitions of skill. Therefore, we recommend that FSS results should be assessed relative to the improved reference score presented in this work in place of the conventional approach or else directly compared to other simple baselines, such as climatology or persistence.

We stress that this work focuses on the use of the FSS to assess the skill of a forecast relative to a random baseline and not for other purposes such as estimating forecast displacement, as is done in Skok and Roberts (2018). For the applicability of the standard reference score for estimating displacement errors, we direct the reader to the extensive analysis in Skok and Roberts (2018) and Gilleland et al. (2020).

Acknowledgments. The authors are grateful to Fenwick Cooper and Llorenç Lledó for comments on an earlier version of this work, and to the reviewers for their thoughtful and constructive comments. David MacLeod regridded the IFS forecast data used to illustrate the results.

Data availability statement. The Python code and data used to create the plots in this work can be found at https://github.com/bobbyantonio/fractions_skill_score.

APPENDIX

Mean and Variance of Neighborhood Fractions

In this appendix, we derive expressions for the sample mean and variance of a fraction produced by a square convolution over binary data. Define $\mathcal{D}$ as the domain of grid squares over which the neighborhood mean and standard deviation are to be calculated. Each location in this domain will be indexed by a single integer, to make the notation in this section easier to follow. The binary values of the forecast (after a threshold has been applied) are denoted as z_i.

The fraction calculated over this neighborhood at the location i, denoted as $y(n)_i$, is given by the summation of values around the central point i up to a distance of n grid cells [i.e., $y(n)_i$ is a placeholder for either the observed fraction $x(n)_i$ or the forecast fraction $f(n)_i$]. We denote $W_n(i)$ as the set of all coordinate indices that are within the neighborhood of width $2n + 1$ centered at point i. Then, $y(n)_i$ is

$$y(n)_i = \frac{1}{(2n+1)^2} \sum_{j \in W_n(i)} z_j. \tag{A1}$$

The mean fraction is the average of y_i over all sites; intuitively, we can see that, since the averaging is a linear operation, the sample average $\langle y(n) \rangle$ will be approximately equal to the sample average of the individual sites excluding padding, $\langle z \rangle$. The complicating factor is the padding used to compensate for the finite domain size; however, it can be shown that with reflective padding, this relationship is in fact an equality. To show this, we first explicitly write out the sample average of $y(n)$:

$$\begin{aligned} \langle y(n) \rangle &= \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \frac{1}{(2n+1)^2} \sum_{j \in W_n(i)} z_j, \\ &= \frac{1}{|\mathcal{D}|} \frac{1}{(2n+1)^2} \sum_{i \in \mathcal{D}} \sum_{\{j : i \in W_n(j)\}} z_i, \end{aligned} \tag{A2}$$

where in the last line, we have simply rearranged the summation to be in terms of a sum over all neighborhoods that contain the point i (which is only possible because we are performing a summation over all points in the domain). An illustration of this rearrangement is provided for a simplified one-dimensional case in Fig. A1.

For a point that lies away from the edges, it is straightforward to see that this point will be contained in the neighborhood of $(2n + 1)^2$ points (including its own neighborhood). For points near the edges and corners, however, this is not as obvious. Consider a point lying near an edge (as shown in Fig. A2a); if the point is a distance $d < n$ from the edge, then without padding it is no longer contained within the

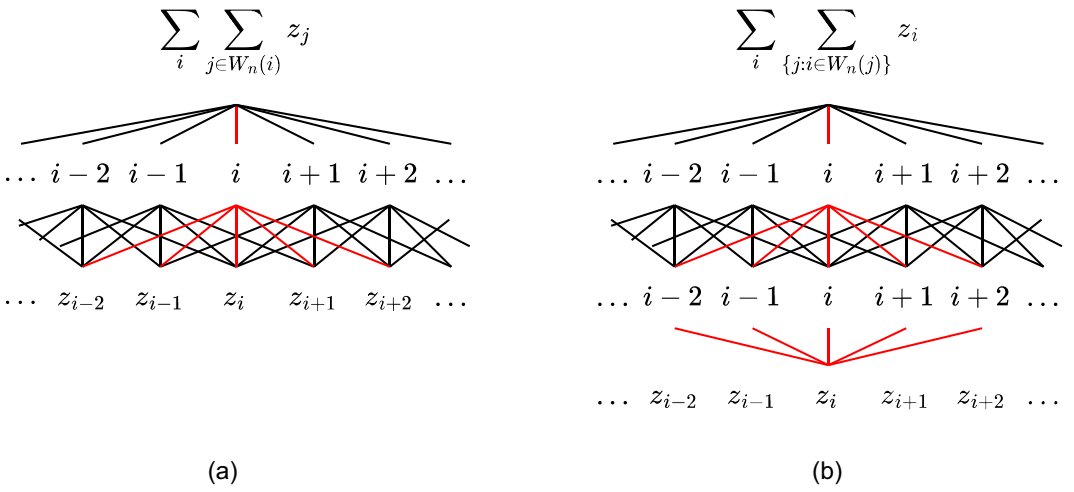

FIG. A1. An illustration of how it is possible to rearrange the sum of terms in Eq. (A2) from $\sum_i \sum_{j \in W_n(i)} z_j$ to $\sum_i \sum_{\{j:i \in W_n(j)\}} z_i$, for a simple one-dimensional case, where the neighborhood width is 5. Each edge indicates a term that contributes to the summation shown at the top. Red lines indicate all of the terms that contribute to the sum for the ith index. (a) The original sum, where the terms from the ith index are made up of sites in the neighborhood of i (including site i itself). (b) The rearrangement of the sum, where this time the terms from the ith index contain multiple copies of z_i, with multiplicity given by the number of items in the neighborhood of i (5 in this case).

neighborhoods of $(2n + 1)(n - d)$ points (i.e., any points that lie within a distance of $n - d$ from the edge). With reflective padding, however, this point is included in several neighborhoods twice; the number of such neighborhoods is equal to the number of points lying within a distance $n - d$ from the edge, which is also $(2n + 1)(n - d)$. Thus, each point lying along an edge is also contained within $(2n + 1)^2$ neighborhoods.

The same can be seen for corner cases. Consider a point situated in a corner a distance $d_y < n$ from the top edge and $d_x < n$ from the side edge, as illustrated in Fig. A2b. Without reflective padding, it is only contained within $(n + d_x + 1)(n + d_y + 1)$ neighborhoods. With the inclusion of reflective padding, however, this site is included in several neighborhoods multiple times; each point in the singly hatched area in Fig. A2b includes point i one additional

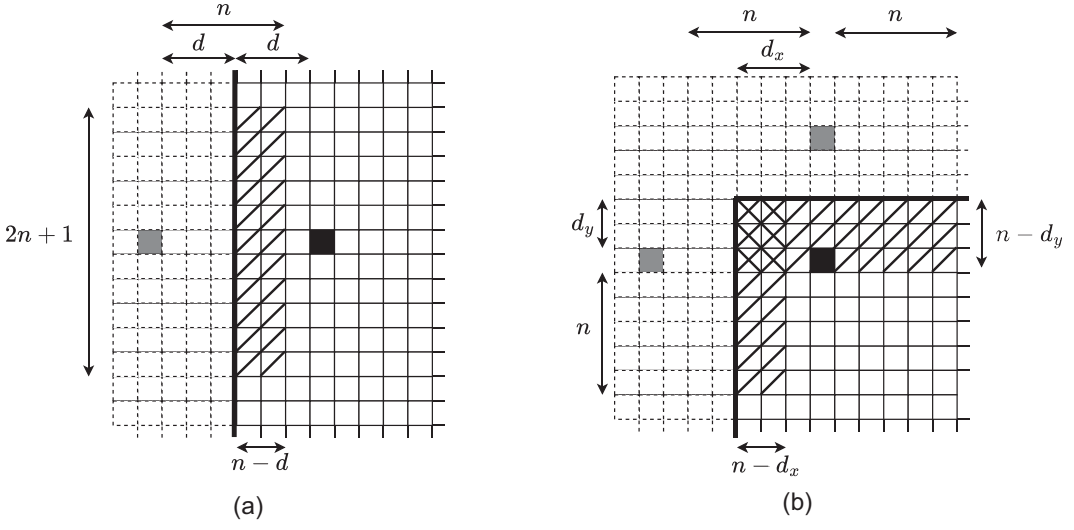

FIG. A2. Diagram illustrating how averaging over neighborhoods behaves at the edges when reflective padding is used (a) for the case where a point is located a distance $d < n$ from an edge and (b) for the case where a point is located a distance $d_x < n$ from a vertical edge and $d_y < n$ from a horizontal edge. For both images, the point of interest is represented as a filled black square, the region of dashed lines represents the reflective padding, and the filled gray squares are where the original point is reflected to. Points that contain the reflected point once and twice are represented as single and double hatching, respectively.

time, whereas each point in the doubly hatched area includes point i two additional times. The area of the single hatched areas plus twice the doubly hatched areas is just $(n - d_y)(n + d_x + 1) + (n - d_x)(n + d_y + 1)$. This then brings the total number of neighborhoods that i is included in up to $(2n + 1)^2$.

Applying this to Eq. (A2), we therefore see that, with reflective padding,

$$\langle y(n)\rangle = \langle y(0)\rangle = \langle z\rangle. \tag{A3}$$

The (biased estimate of the) sample variance calculated over all fractions y_i, denoted as s_n^2, can be written as

$$\begin{aligned}
s_n^2 &= \langle y(n)^2\rangle - \langle y(n)\rangle^2, \\
&= \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\left[\frac{1}{(2n+1)^2}\sum_{j\in W_n(i)} z_j\right]^2 - \langle z\rangle^2, \\
&= \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\frac{1}{(2n+1)^4}\sum_{j\in W_n(i)}\sum_{k\in W_n(i)}(z_j z_k - \langle z\rangle^2), \\
&= \frac{1}{|\mathcal{D}|(2n+1)^4}\sum_{i\in\mathcal{D}}\sum_{j\in W_n(i)}(z_j^2 - \langle z\rangle^2) \\
&\quad + \frac{1}{|\mathcal{D}|(2n+1)^4}\sum_{i\in\mathcal{D}}\sum_{j\in W_n(i)}\sum_{\substack{k\in W_n(i)\\ k\neq j}}(z_j z_k - \langle z\rangle^2), \\
&= \frac{1}{|\mathcal{D}|(2n+1)^4}\sum_{i\in\mathcal{D}}\sum_{\{j:i\in W_n(j)\}}(z_i^2 - \langle z\rangle^2) \\
&\quad + \frac{1}{|\mathcal{D}|(2n+1)^4}\sum_{i\in\mathcal{D}}\sum_{j\in W_n(i)}\sum_{\substack{k\in W_n(i)\\ k\neq j}}(z_j z_k - \langle z\rangle^2),
\end{aligned} \tag{A4}$$

where in the last line, we have once again rearranged the sum to be in terms of the number of neighborhoods containing point i, rather than the number of points in the neighborhood of i. Using the same argument as above, the first term contains $(2n + 1)^2$ copies of each summand, and so this simplifies to

$$\begin{aligned}
s_n^2 &= \frac{1}{(2n+1)^2}(\langle z^2\rangle - \langle z\rangle^2) \\
&\quad + \frac{1}{|\mathcal{D}|(2n+1)^4}\sum_{i\in\mathcal{D}}\sum_{j\in W_n(i)}\sum_{\substack{k\in W_n(i)\\ k\neq j}}(z_j z_k - \langle z\rangle^2), \\
&= \frac{\langle y(0)\rangle[1 - \langle y(0)\rangle]}{(2n+1)^2} \\
&\quad \times\left[1 + \frac{1}{|\mathcal{D}|(2n+1)^2}\sum_{i\in\mathcal{D}}\sum_{j\in W_n(i)}\sum_{\substack{k\in W_n(i)\\ k\neq j}}\frac{(z_j z_k - \langle z\rangle^2)}{s_y^2}\right],
\end{aligned} \tag{A5}$$

where in the last line, we have rewritten $\langle z^2\rangle - \langle z^2\rangle = \langle y(0)^2\rangle - \langle y(0)^2\rangle = \langle y(0)\rangle[1 - \langle y(0)\rangle]$ [because the data $y(0)$ are binary].

To simplify this further, we will group the terms inside the sum according to the L_1 distance (or taxicab norm) between them, where $L_1(i, j)$ denotes this distance (chosen since it is a natural metric for square neighborhoods, but this could be substituted for other distance metrics with only slight modifications to the following derivation):

$$\begin{aligned}
s_n^2 &= \frac{\langle x(0)\rangle[1 - \langle x(0)\rangle]}{(2n+1)^2} \\
&\quad \times\left[1 + \frac{1}{|\mathcal{D}|(2n+1)^2}\sum_{d=1}^{(2n+1)}\sum_{i\in\mathcal{D}}\sum_{j\in W_n(i)}\sum_{\substack{k\in W_n(i)\\ L_1(j,k)=d}}\frac{(z_j z_k - \langle z\rangle^2)}{s_y^2}\right].
\end{aligned} \tag{A6}$$

Within a neighborhood of size $(2n + 1) \times (2n + 1)$, we denote the number of points separated by a distance d as $\Omega_d(n)$. With this notation, we define ν_d as an estimate of the spatial autocorrelation for points a distance d apart, for a given neighborhood size n:

$$\nu_d := \frac{1}{|\mathcal{D}|(2n+1)^2\Omega_d(n)}\sum_{i\in\mathcal{D}}\sum_{j\in W_n(i)}\sum_{\substack{k\in W_n(i)\\ L_1(j,k)=d}}\frac{(z_j z_j - \langle z\rangle^2)}{s_y^2}. \tag{A7}$$

The ν_d will contain biases due to the reflective padding; near the edges, the correlation will be artificially inflated since any reflected points will be perfectly correlated with one other point in the neighborhood. However, for our analysis, where the spatial autocorrelation term is only required to qualitatively understand what influences the value of s_n^2, this bias is acceptable.

Using the definition in Eq. (A7), we can then rewrite Eq. (A6) as

$$s_n^2 = \frac{\langle y(0)\rangle[1 - \langle y(0)\rangle]}{(2n+1)^2}\left[1 + \sum_{d=1}^{(2n+1)}\Omega_d(n)\nu_d\right]. \tag{A8}$$

In the absence of spatial autocorrelation (i.e., $\nu_d = 0$), s_n^2 is equal to the standard deviation for a Binomial distribution divided by $(2n + 1)^2$ (since values are expressed as fractions).

REFERENCES

Ayzel, G., T. Scheffer, and M. Heistermann, 2020: RainNet v1.0: A convolutional neural network for radar-based precipitation nowcasting. *Geosci. Model Dev.*, **13**, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Cafaro, C., and Coauthors, 2021: Do convection-permitting ensembles lead to more skillful short-range probabilistic rainfall forecasts over tropical East Africa? *Wea. Forecasting*, **36**, 697–716, https://doi.org/10.1175/WAF-D-20-0172.1.

Casati, B., C. Lussana, and A. Crespi, 2023: Scale-separation diagnostics and the symmetric bounded efficiency for the intercomparison of precipitation reanalyses. *Int. J. Climatol.*, **43**, 2287–2304, https://doi.org/10.1002/joc.7975.

Duc, L., K. Saito, and H. Seko, 2013: Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18171, https://doi.org/10.3402/tellusa.v65i0.18171.

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

Ebert-Uphoff, I., R. Lagerquist, K. Hilburn, Y. Lee, K. Haynes, J. Stock, C. Kumler, and J. Q. Stewart, 2021: CIRA guide to custom loss functions for neural networks in environmental sciences – version 1. arXiv, 2106.09757v1, https://doi.org/10.48550/arXiv.2106.09757.

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

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

Harvey, N. J., and H. F. Dacre, 2016: Spatial evaluation of volcanic ash forecasts using satellite observations. *Atmos. Chem. Phys.*, **16**, 861–872, https://doi.org/10.5194/acp-16-861-2016.

Hooker, H., S. L. Dance, D. C. Mason, J. Bevington, and K. Shelton, 2022: Spatial scale evaluation of forecast flood inundation maps. *J. Hydrol.*, **612**, 128170, https://doi.org/10.1016/j.jhydrol.2022.128170.

Huffman, G., D. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, and P. Xie, 2022: Integrated Multi-Satellite Retrievals for GPM (IMERG), V06B. NASA's Precipitation Processing Center, accessed 1 October 2022, https://arthurhouhttps.pps.eosdis.nasa.gov/text/gpmallversions/V06/YYYY/MM/DD/imerg/.

Lagerquist, R., and I. Ebert-Uphoff, 2022: Can we integrate spatial verification methods into neural network loss functions for atmospheric science? *Artif. Intell. Earth Syst.*, **1**, e220021 https://doi.org/10.1175/AIES-D-22-0021.1.

——, J. Q. Stewart, I. Ebert-Uphoff, and C. Kumler, 2021: Using deep learning to nowcast the spatial coverage of convection from Himawari-8 satellite data. *Mon. Wea. Rev.*, **149**, 3897–3921, https://doi.org/10.1175/MWR-D-21-0096.1.

Marsigli, C., A. Montani, and T. Paccagnella, 2008: A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. *Meteor. Appl.*, **15**, 125–143, https://doi.org/10.1002/met.65.

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

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

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424, https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

Nachamkin, J. E., and J. Schmidt, 2015: Applying a neighborhood fractions sampling approach as a diagnostic tool. *Mon. Wea. Rev.*, **143**, 4736–4749, https://doi.org/10.1175/MWR-D-14-00411.1.

Necker, T., L. Wolfgruber, L. Kugler, M. Weissmann, M. Dorninger, and S. Serafin, 2023: The fractions skill score for ensemble forecast verification. *Quart. J. Roy. Meteor. Soc.*, **150**, 4457–4477, https://doi.org/10.1002/qj.4824.

Price, I., and S. Rasp, 2022: Increasing the accuracy and resolution of precipitation forecasts using deep generative models. arXiv, 2203.12297v1, https://doi.org/10.48550/arXiv.2203.12297.

Pulkkinen, S., D. Nerini, A. A. Pérez Hortal, C. Velasco-Forero, A. Seed, U. Germann, and L. Foresti, 2019: 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.

Ravuri, S., and Coauthors, 2021: Skilful precipitation nowcasting using deep generative models of radar. *Nature*, **597**, 672–677, https://doi.org/10.1038/s41586-021-03854-z.

Rezacova, D., Z. Sokol, and P. Pesice, 2007: A radar-based verification of precipitation forecast for local convective storms. *Atmos. Res.*, **83**, 211–224, https://doi.org/10.1016/j.atmosres.2005.08.011.

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

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

Schwartz, C. S., 2017: A comparison of methods used to populate neighborhood-based contingency tables for high-resolution forecast verification. *Wea. Forecasting*, **32**, 733–741, https://doi.org/10.1175/WAF-D-16-0187.1.

——, 2019: Medium-range convection-allowing ensemble forecasts with a variable-resolution global model. *Mon. Wea. Rev.*, **147**, 2997–3023, https://doi.org/10.1175/MWR-D-18-0452.1.

Simecek-Beatty, D., and W. J. Lehr, 2021: Oil spill forecast assessment using Fractions Skill Score. *Mar. Pollut. Bull.*, **164**, 112041, https://doi.org/10.1016/j.marpolbul.2021.112041.

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

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

——, and ——, 2018: Estimating the displacement in precipitation forecasts using the Fractions Skill Score. *Quart. J. Roy. Meteor. Soc.*, **144**, 414–425, https://doi.org/10.1002/qj.3212.

Stein, J., and F. Stoop, 2024: Evaluation of probabilistic forecasts of binary events with the neighborhood Brier divergence skill score. *Mon. Wea. Rev.*, **152**, 1201–1222, https://doi.org/10.1175/MWR-D-22-0235.1.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic

approach. *Meteor. Appl.*, **12**, 257–268, https://doi.org/10.1017/S1350482705001763.

Von Storch, H., and F. W. Zwiers, 2002: *Statistical Analysis in Climate Research.* Cambridge University Press, 496 pp.

Weusthoff, T., F. Ament, M. Arpagaus, and M. W. Rotach, 2010: Assessing the benefits of convection-permitting models by neighborhood verification: Examples from MAP D-PHASE. *Mon. Wea. Rev.*, **138**, 3418–3433, https://doi.org/10.1175/2010MWR3380.1.

Wilks, D. S., 2019: Forecast verification. *Statistical Methods in the Atmospheric Sciences*, 4th ed. Elsevier, 369–483, https://doi.org/10.1016/b978-0-12-815823-4.00009-2.

Woodhams, B. J., C. E. Birch, J. H. Marsham, C. L. Bain, N. M. Roberts, and D. F. A. Boyd, 2018: What is the added value of a convection-permitting model for forecasting extreme rainfall over tropical East Africa? *Mon. Wea. Rev.*, **146**, 2757–2780, https://doi.org/10.1175/MWR-D-17-0396.1.