1.10 A NEW SPATIAL SCALE DECOMPOSITION OF THE BRIER SCORE FOR THE VERIFICATION OF PROBABILISTIC LIGHTNING FORECASTS

B. Casati * L. J. Wilson

Meteorological Service of Canada, Meteorological Research Branch,
Recherche en Prevision Numerique, Dorval, Quebec, Canada

Abstract

A new scale decomposition of the Brier score for evaluating spatial probabilistic forecasts is presented. The technique is illustrated on the Canadian Meteorological Center (CMC) lightning probabilistic forecasts. Probability forecasts of lightning rate for 3 hour time windows and 22 km spatial resolution are verified against lightning frequencies from the North American Lightning Detection Network (NALDN) on a domain encompassing Canada and the northern United States. Verification is performed for lightning rates exceeding different thresholds, to evaluate the forecast performance both for modest and intense lightning activity. Forecast and observation are decomposed into the sum of components on different spatial scales by performing a 2D Haar wavelet decomposition. Evaluation at different spatial scales is then performed by evaluating Brier score and skill score for each spatial scale component.

1 INTRODUCTION

Verification is a key component of weather forecasting. In fact, verification not only allows one to monitor and compare the performance of weather forecasts, but also to analyze the nature of the forecast error. A *diagnostic* verification can help to detect the forecast weaknesses and systematic errors in Numerical Weather Prediction (NWP) models. Therefore, a diagnostic verification provides guidances for forecasters and NWP modelers which leads to new development and improvements. This work introduces a new diagnostic verification technique for probabilistic forecasts defined on a spatial domain.

Weather phenomena are characterized by the presence of features on different scales. Phenomena on different scales are often driven by different physical processes. Verification on different spatial scales can therefore provide useful insight into the NWP model representation of different physical processes, and indicate which of these processes might need further development. The verification technique introduced in this work aims to provide feedback on the performance of a probabilistic forecast on different spatial scales. For studying the forecast predictability scale limits, it is desirable to establish at which scale there is a transition from negative to positive skill. Moreover, the technique aims to provide feedback on the capability of the forecast to reproduce the scale structure of the observation.

Few techniques for the verification of spatial deterministic forecasts can be found in the literature: Briggs and Levine (1997) introduced a wavelet-based verification method on different spatial scales based on continuous verification statistics (e.g. MSE); Casati *et al.* (2004) developed an intensity-scale verification technique based again on 2D wavelet decomposition and on a categorical verification approach; Zepeda-Arce and Foufoula-Georgiou (2000) and Harris and Foufoula-Georgiou (2001) assess the forecast capability of reproducing the observation spatio-temporal and multi-scale spatial structure of precipitation fields. De Elia *et al.* (2002) and Denis *et al.* (2003) evaluate the forecast timescale predictability limits as a function of the scale for high resolution regional models. The verification technique introduced in this work is specifically designed for the verification on different scales of probabilistic forecasts. Forecast probability and corresponding observed frequency images are decomposed into the sum of components on different spatial scales by performing a 2D Haar wavelet decomposition. Brier score and skill score are eval-

* *Corresponding author address*: Dr. B. Casati, RPN, 2121 Trans-Canadian Highway, Dorval, QC, H9P 1J3, Canada; e-mail: *barbara.casati@ec.gc.ca*

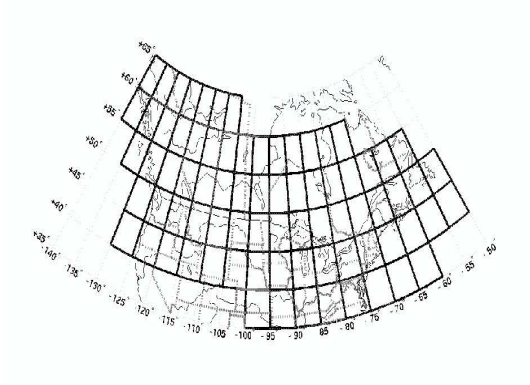

Figure 1: Domain of $5^o \times 5^o$ latitude-longitude sectors on which statistical regression models were developed to produce the CMC lightning probability forecast (image kindly provided by W. Burrows).

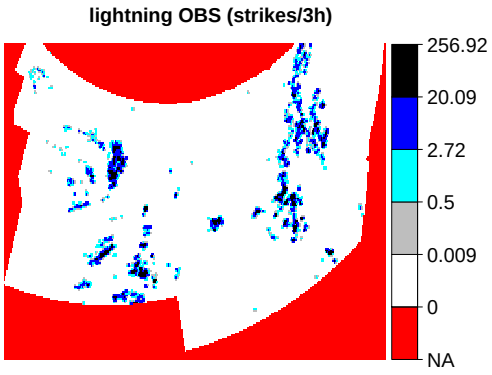

Figure 2: Observed lightning occurrence in the three hour window from 21:00 to 24:00 UTC on the 17^{th} July 2004.

uated on each scale component along with the squared energy bias. The scale structure representation is assessed by the ratio of forecast and observation percentage of energy that each scale exhibits.

The technique is illustrated on a representative case study of the CMC lightning probabilistic forecasts. Section 2 reviews the general features of the CMC lightning probabilistic forecasts and introduce the case study. The verification method is fully described in Section 3. Interpretation of the verification results for the case study analyzed are given along with the verification method description. Finally, in section 4, some conclusions are given.

2 THE CMC LIGHTNING PROBABILITY FORECAST

Lightning probabilistic forecasts are produced operationally at the Canadian Meteorological Center (Burrows *et al.*, 2005). The probability of lightning occurrence exceeding specific thresholds in 3 hour time windows is forecast on a domain of approximately 20 km resolution encompassing Canada and the northern United States, with a time projection up to 48 hours. The forecast is produced by a tree structured regression model: individual statistical regression models were developed for each $5^o \times 5^o$ latitude-longitude sector (see Figure 1), for each of the months from May to September. Predictors to construct the regression equations were derived from the 24 km resolution Global Environment Multiscale (GEM) NWP model output (Côté *et al.*, 1997). In the year 2004 the GEM NWP model has been updated to 15 km resolution: predictors for the lightning statistical model are currently obtained from the higher resolution GEM model output and interpolated on the old 24 km resolution GEM grid before applying the regression equations. Predictands are lightning flash reports from the NALDN distributed by Vaisala Inc. In order to match the predictors, the predictands have been gridded on the 24 km resolution GEM domain. Each flash has been assigned a weight of one if within a distance of 10 km from the grid point, and a weight decreasing linearly from one to zero as the distance from the grid point increases from 10 to 20 km. Predictors and predictands of the summers of 2000 and 2001 have been used as training data set to construct the regression model.

The verification method introduced in this work has been tested on forty-six case studies of the CMC lightning probabilistic forecast for the summer 2004. In this work we illustrate the verification method on one representative case. Three categories of lightning probabilities are considered: 1) probability of *any lightning*, which is defined as the probability that the lightning occurrence in the three hour time window is greater than zero; 2) probability of *occasional to extreme lightning*, i.e. the probability that the lightning occurrence in the three hour time window exceeds the threshold of 0.5; 3) probability of *frequent lightning*, which is defined as the probability that the lightning occurrence in the three hour time window exceeds the threshold of $e^3 \simeq 20.085$. The forecast probabilities are verified against lightning flash reports from the NALDN. The observed occurrence of lightning in three hour time windows is gridded on the forecast domain. Observations are processed in the same fashion used for the predictands when developing the statistical regression model, i.e. each flash has been assigned a weight of one if within a

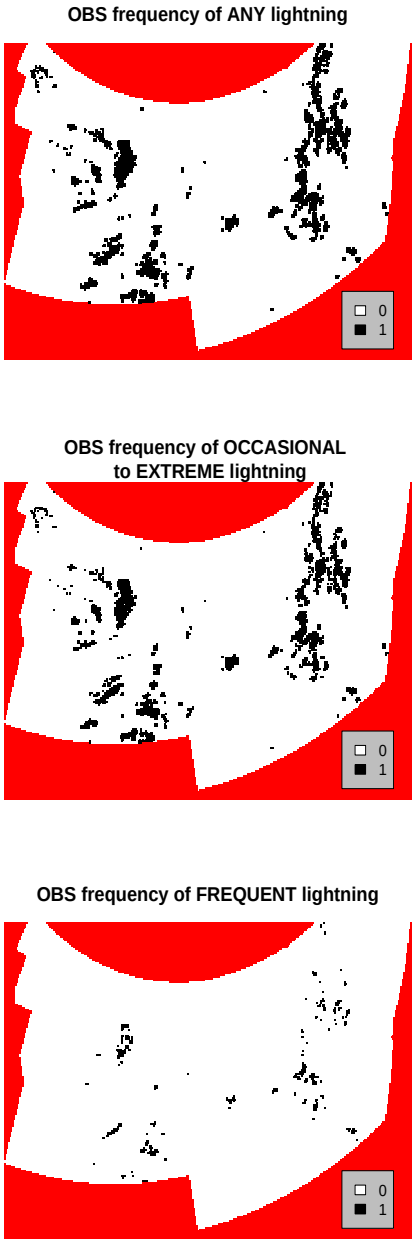

Figure 3: Observed lightning frequency in the three hour window from 21:00 to 24:00 UTC on the 17^{th} July 2004 obtained from the observed lightning occurrences shown in Figure 2.

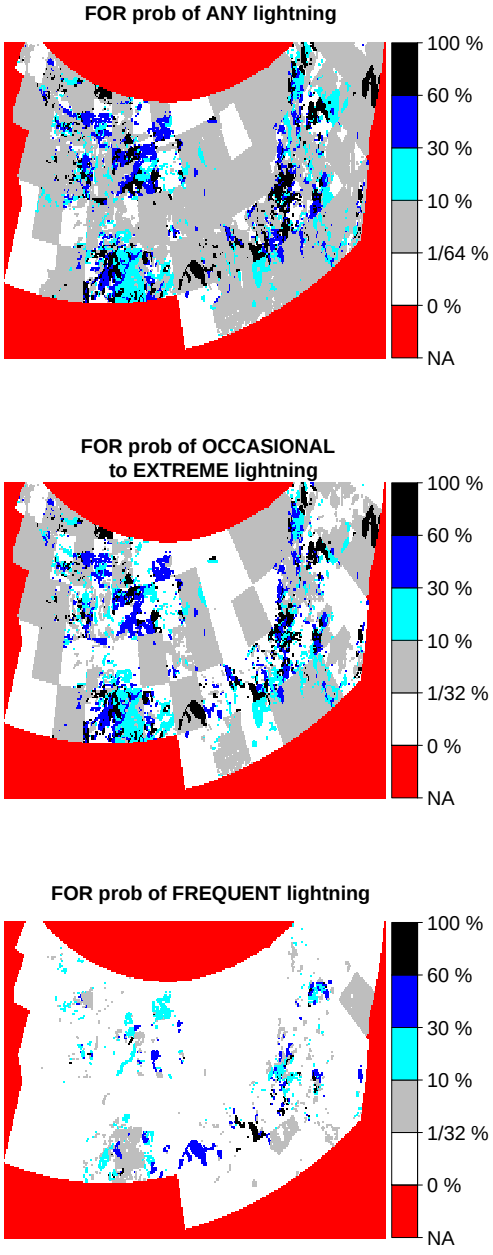

Figure 4: 21 hour lead time probabilistic forecasts for the three categories of any lightning, occasional to extreme lightning and frequent lightning valid in the three hour window from 21:00 to 24:00 UTC on the 17^{th} July 2004.

distance of 10 km from the grid point, and a weight decreasing linearly from one to zero as the distance from the grid point increases from 10 to 20 km. Forecasts probabilities of any lightning, occasional to extreme lightning and frequent lightning are verified against observed frequencies of the same three categories, i.e. binary images equal to one where the observed occurrence of lightning exceeds the corresponding category threshold, and equal to zero elsewhere. Figure 2 shows the observed lightning occurrence in the three hour window from 21:00 to 24:00 UTC on the 17^{th} of July 2004; Figure 3 shows the corresponding observed lightning frequency, obtained from the observed lightning occurrences by thresholding with the appropriate category threshold; Figure 4 shows the 21 hour lead time probabilistic forecasts for the three above-mentioned categories valid at the same time. This case shows a typical synoptic situation: the lightning activity on the east side of the domain is related to a large frontal system; the lightning activity on the west side of the domain is mainly related to small scale convective activity located in the region of the Rocky Mountains.

3 THE VERIFICATION METHOD AND INTERPRETATION OF THE VERIFICATION RESULTS

3.1 Verification domain

The forecast and gridded observation domain is a polar stereographic grid of 295×183 pixels with pole coordinates at $N = (81.5, 309.5)$, a distance of 20 km between 2 grid points at 60 N and an angle of 21^o between the Greenwich meridian and the x axis, positive counter-clockwise. Since the regression model was not developed for covering such a domain, some of the more external grid points have missing value (red pixels in Figures 3, 4). The domain considered in this work for verification purposes is a reduced rectangular domain of 256×128 pixels embedded in the 295×183 pixel domain so that the number of missing values is minimized. The dimensions of the verification domain chosen are integer powers of 2 ($256 = 2^8; 128 = 2^7$) to have a dyadic domain, appropriate for performing the 2D discrete wavelet transform (see Appendix). Note that the rectangular verification sub-domain is the union of two squared sub-domains of $2^7 \times 2^7$ pixels. The wavelet decomposition is performed on the east and on the west squared sub-domains, separately, as described in the Appendix. Then, the union of the east and west wavelet components on each scale is considered and, where appropriate, the average of the verification statistics for the east and west sub-domains is evaluated. When performing the wavelet decomposition, the missing values within the rectangular verification sub-domain are assigned the average of the non-missing values either of the west or east sub-domain, depending on which of these square sub-domains they belong to. Note that this substitution does not affect the statistic behaviors since this value is the largest scale father wavelet component value evaluated on the non-missing values (see Appendix).

3.2 Images decomposition on different spatial scales

Forecast probability image (Y) and observed frequency image (X) for the three categories of any lightning, occasional to extreme lightning and frequent lightning are decomposed on different scales by a 2D discrete Haar wavelet decomposition (Daubechies (1992); Mallat (1989); see Appendix). Each image is expressed as the sum of image components on different spatial scales:

$$Y = \sum_{j=1}^{J} Y_{j,m} + Y_{J,f}, \tag{1}$$

$$X = \sum_{j=1}^{J} X_{j,m} + X_{J,f}, \tag{2}$$

where $Y_{j,m}$ and $X_{j,m}$ are the mother wavelet components of forecast and observation images on the scale j and $Y_{J,f}$ and $X_{J,f}$ are the father wavelet components of forecast and observation images on the largest scale $J = 7$. The resolution of the mother wavelet components for $j = 1, \ldots, 7$ is equal to 2^{j-1} pixels, corresponding approximately to 20, 40, 80, 160, 320, 640, 1280 km. The largest scale father wavelet components $Y_{J,f}$ and $X_{J,f}$ are obtained from the average of the forecast and observed values on the east and west squared sub-domains. The resolution of the father wavelet components on the largest scale $J = 7$ is 2^7 pixels, corresponding approximately to 2560 km.

3.3 Energy and energy bias on different scales

The squared energy of forecast probability image and observed frequency image is defined as

$$En^2(Y) = \overline{Y^2}, \qquad En^2(X) = \overline{X^2}, \tag{3}$$

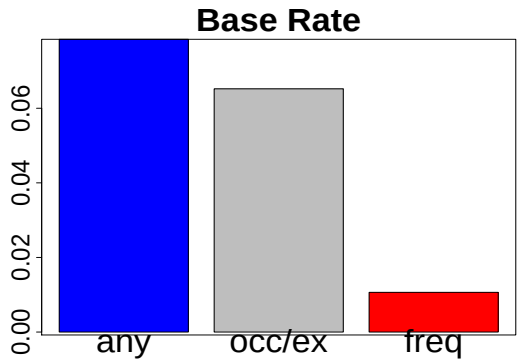

Figure 5: Base rate of the observed frequency images obtained from the case study shown by Figures 3, 4 for the three categories of any lightning, occasional to extreme lightning and frequent lightning.

where the over-bar indicates hereafter averaging over the pixels in the east and west squared subdomains. The squared energy provides feedback on the quantity of events present in an image. The squared energy of an image with zero mean is equal to the image variance. The squared energy of binary images (e.g. the observed frequencies image X) is equivalent to the image mean and to the image sample climatology (or base rate):

$$En^2(X) = \overline{X^2} = \overline{X}. \tag{4}$$

Figure 5 shows the base rate of the observed frequency images obtained from the case study shown by Figures 3, 4 for the three categories of any lightning, occasional to extreme lightning and frequent lightning. Note that the frequency images for the three categories are defined for increasing thresholds. Therefore, the amount of events and the base rate associated to the three categories decreases as the threshold increases. The category of frequent lightning exhibits the smaller base rate, the category of any lightning exhibits the largest base rate.

The squared energy spatial scale components of forecast probability image and observed frequency image are defined as:

$$\begin{aligned} En^2(Y_{j,m}) &= \overline{Y_{j,m}^2}, \quad En^2(X_{j,m}) = \overline{X_{j,m}^2}, \\ En^2(Y_{J,f}) &= \overline{Y_{J,f}^2}, \quad En^2(X_{J,f}) = \overline{X_{J,f}^2}, \end{aligned} \tag{5}$$

where $Y_{j,m}$, $X_{j,m}$, $Y_{J,f}$ and $X_{J,f}$ are the forecast and observation mother and father wavelet components (see Eqns. (1) and (2)). The squared energy spatial scale components provide feedback on the quantity of events present in each image at each different spatial scale. Figure 6 shows the squared energy scale components for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning. Note that the squared energy of the category of frequent lightning are visibly smaller on all the scales than the squared energy of the other two categories, for both forecast and observation; this is due to the presence of less events in the category of frequent lightning, since defined by a higher thresholding on the lightning occurrence. The same argument applies to the categories of occasional to extreme lightning and any lightning, however the differences in their behavior are less remarkable.

Figure 6 shows that for both the probability forecast and the observed frequency the smallest scale exhibits the largest squared energy. Then, as the scale increases, the squared energy decreases. This indicates that both in the forecast probability image and in the observed frequency image there is a large number of small scale events and then, as the scale increases, the number of events decreases. The largest father wavelet component (scale 8 in Figure 6) exhibits a larger squared energy with respect to the immediately preceding large scales: this scale provides a measure of the average value of the forecast and observed images over the entire domain. From the comparison of the largest scale components of the squared energy of the forecast and observation it can be seen that the forecast overall average value is significantly larger than the observed one. This is due to the overall over-forecasting clearly shown in Figures 3, 4. The squared energy components on the intermediate scales 3,4 and 5 are visibly larger in the observation than in the forecast, showing an under-forecast of features on the 80, 160, 320 km scales.

The ratio of the squared energy scale components for forecast probabilities and observed frequencies provides a measure of the bias on different spatial scales. If the scale component squared energy bias is greater than one, it indicates over-forecasting on such a scale; if it is smaller than one it indicates under-forecasting. The bottom panel of Figure 6 shows the ratio of the squared energy scale components for forecast probabilities and observed frequencies for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning. The large value of the squared energy bias component on the largest scale indicates the overall over-forecast for all three categories. The forecast for the categories of occasional to extreme lightning and frequent lightning underestimate all the smaller scales, in particular the scales 3,4 and 5 corresponding to features of 80, 160, 320 km resolution. The fore-

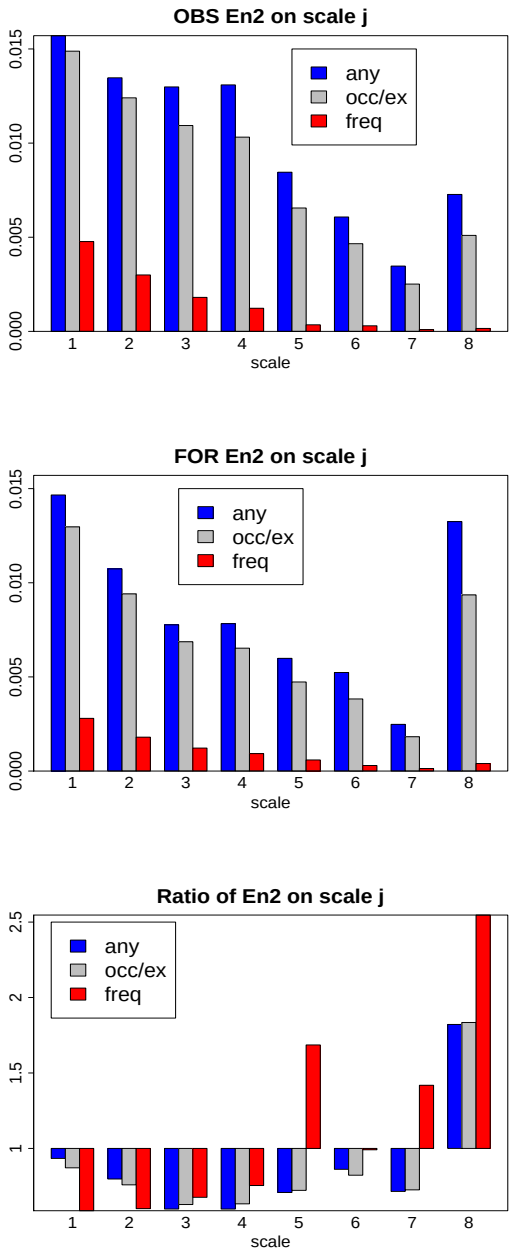

Figure 6: Squared energy scale components and their ratio for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning.

cast for the category of frequent lightning underestimates small scales (scales 1 to 4) and overestimates some of the larger scales (in particular scale 5): in fact it can be seen in Figures 3, 4 that the forecast is smoother than the observed field and is characterized the presence of features of 320 km or larger, not present in the observed frequency image.

3.4 Scale structure representation

The sum of the squared energy components on different scales defined by Eqn. (5) is equal to the total squared energy defined in Eqn. (3). In fact, since the wavelet components obtained from a discrete wavelet filter are orthogonal, it follows that

$$\begin{array}{ll} \overline{Y_{j,m}Y_{k,m}} = 0, & \overline{X_{j,m}X_{k,m}} = 0, \quad \forall j \neq k; \\ \overline{Y_{j,m}Y_{J,f}} = 0, & \overline{X_{j,m}X_{J,f}} = 0. \end{array} \tag{6}$$

From this result and Eqns. (3), (1), (2) and (5) the forecast and observation squared energy can be written as

$$\begin{array}{l} En^2(Y) = \overline{Y^2} = \overline{(\sum_{j=1}^{J} Y_{j,m} + Y_{J,f})^2} = \\ \overline{(\sum_{j=1}^{J} Y_{j,m} + Y_{J,f})(\sum_{j=1}^{J} Y_{j,m} + Y_{J,f})} = \\ \sum_{j=1}^{J} \overline{Y_{j,m}^2} + \overline{Y_{J,f}^2} = \\ \sum_{j=1}^{J} En^2(Y_{j,m}) + En^2(Y_{J,f}); \\ \\ En^2(X) = \overline{X^2} = \overline{(\sum_{j=1}^{J} X_{j,m} + X_{J,f})^2} = \\ \overline{(\sum_{j=1}^{J} X_{j,m} + X_{J,f})(\sum_{j=1}^{J} X_{j,m} + X_{J,f})} = \\ \sum_{j=1}^{J} \overline{X_{j,m}^2} + \overline{X_{J,f}^2} = \\ \sum_{j=1}^{J} En^2(X_{j,m}) + En^2(X_{J,f}). \end{array} \tag{7}$$

This result enables one to evaluate the fraction with which each scale contributes to the total squared energy:

$$\begin{array}{l} En^2_{\%}(Y_{j,m}) = En^2(Y_{j,m})/En^2(Y); \\ En^2_{\%}(Y_{J,f}) = En^2(Y_{J,f})/En^2(Y); \\ En^2_{\%}(X_{j,m}) = En^2(X_{j,m})/En^2(X); \\ En^2_{\%}(X_{J,f}) = En^2(X_{J,f})/En^2(X). \end{array} \tag{8}$$

The percentage of squared energy on each scale provides feedback on the partition of the image total amount of events on the different scales, and therefore on the scale structure of forecast and observation images. Figure 7 shows the forecast and observation percentages of the squared energy scale components for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning. Most of the conclusions drawn for the squared energy (Figure 6), can be deduced also from Figure 7: for both forecast and observation, the smallest

scale exhibits the largest fraction of squared energy (events), and then, as the scale increases, the fraction of squared energy (events) decreases; the forecast largest scale component (overall average value) contributes in a significantly larger proportion to the total squared energy than the observed one, due to the already diagnosed overall over-forecasting; the observed larger percentage of the squared energy components on the intermediate scales 3,4 and 5 diagnoses the under-forecast of features on the 80, 160, 320 km scales.

However, the information provided by the squared energies shown in Figure 6 is conceptually different from the information provided by the fraction of the squared energies shown in Figure 7. The former provides information on the quantity of events on each scale. Therefore they are dependent on the total amount of events present in the images (or sample climatology). This is the reason why in Figure 6 the frequent lightning category exhibits smaller energies than the other two categories, because of its smaller sample climatology (see Figure 5). The behavior of the squared energy on different scales for the category of frequent lightning is barely noticeable in Figure 6, because its magnitude is so small when compared with the other two categories. On the other hand, Figure 7 provides information on the fraction for which each scale contributes to the total energy. Such a fraction is independent from the sample climatology (the sums of the fractions of the squared energies shown in Figure 7 for the three categories are equal: the statistics for the three categories are more directly comparable). The behavior of the fraction of the squared energy on different scales for the category of frequent lightning is better shown in Figure 7 and is more directly comparable with the one of the other two categories. The fraction for which each scale contributes to the total squared energy provides feedback on the scale structure of the image. From the comparison of the forecast and observation percentages of squared energy on different scales shown in Figure 7 for the category of frequent lightning it can be noticed that the forecast energy is more concentrated on large scales and the observation energy is slightly more concentrated on small scales: the forecast is smoother than the observation.

Note that for both forecast and observation the fraction of squared energy on small scales for the frequent lightning category is larger than for the other two categories. This is due to the presence of a larger number of small scale events in the frequent lightning category, isolated by thresholding the lightning occurrence with a higher threshold (e^3) than the

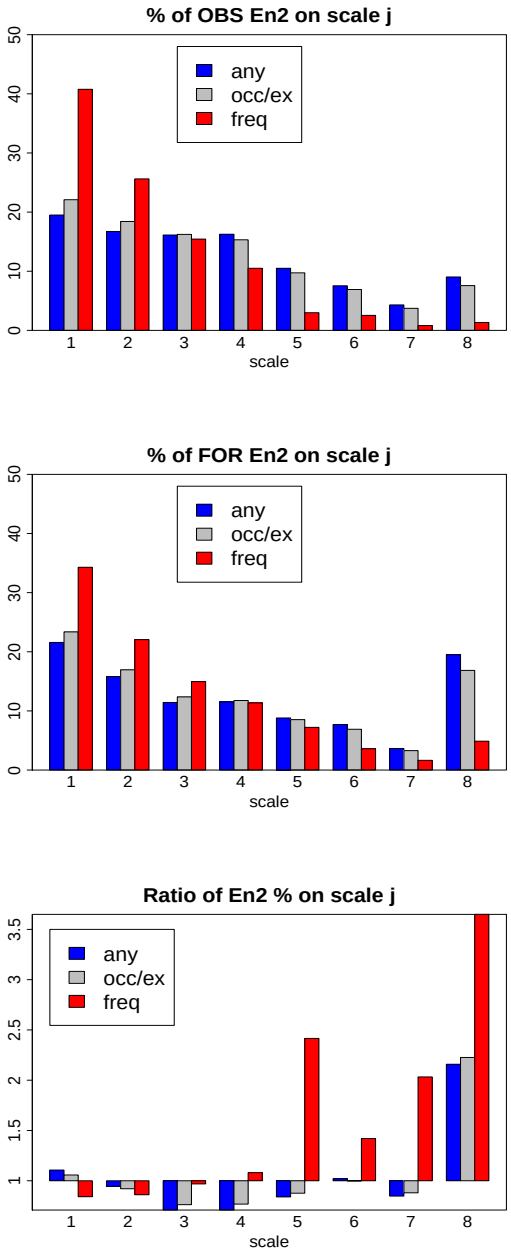

Figure 7: Percentages of the squared energy scale components and percentages ratio for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning.

other two categories (0 and 0.5). Vice-versa, as the scale increases, the fraction of squared energy on large scales for any lightning or occasional to extreme lightning becomes larger than for the frequent lightning category. This is due to the presence of large scale features in the probability forecast and observation frequency images for this two categories because defined from a low threshold on lightning occurrence. The same argument applies when explaining the similar behavior of the percentage of the squared energy for the categories of occasional to extreme lightning and any lightning. However, the difference in the statistical behaviors for these two categories are less remarkable than the one between these and the frequent lightning category.

The ratio of the fraction of the squared energy scale components for forecast and observation measures the differences in the scale structure representation of the forecast probability and observation frequency images. Such a ratio is shown in the bottom panel of Figure 7 for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning. All the three categories exhibit a large value at the largest scale due, as already explained, to the overall over-forecast. For the two categories of any lightning and from occasional to extreme lightning, the observed scale structure are not too badly reproduced (scales 3,4 and 5 are slightly under-forecast, as previously diagnosed, and the other scales are slightly over- or under-forecast, but the ratio does not depart too much from one, indicating good representation of the scale structure). For the category of frequent lightning, small scales (scales 1, 2 and 3, corresponding to 20, 40 and 80 km scale features) in the forecast contribute in smaller proportion to the total squared energy than in the observation. On the other hand, larger scales (scales 4, 5, 6, 7 and 8, corresponding to 160, 320, 640, 1280 and 2560 km scale features) contribute in a larger proportion to the total squared energy than in the observation.This shows that the forecast for the category of frequent lightning is smoother than the observation.

3.5 Brier Score decomposition on different scales

The Brier Score (BS; see Brier, 1950) for each lightning probabilistic forecast versus its corresponding observed frequency image is given by:

$$\mathrm{BS} = \overline{(Y-X)^2} = \overline{Z^2}, \tag{9}$$

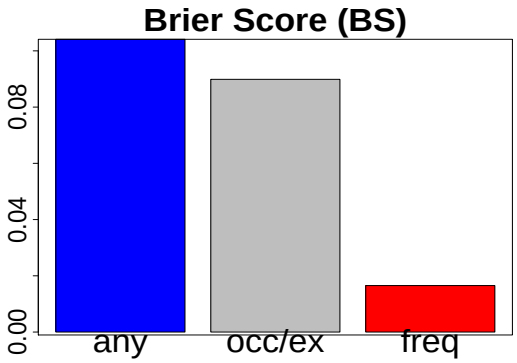

Figure 8: Brier score for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning.

where $Z = Y - X$ is the probability error image. The Brier score measures the forecast error. The Brier score for a perfect forecast is equal to zero. Figure 8 shows the Brier score for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning. The Brier score is larger for the categories of any lightning and from occasional to extreme lightning and smaller for the category of frequent lightning. This implies that the former categories have a larger error than the latter. However, the Brier score is highly dependent on the sample climatology, i.e. the error is proportional to the amount of events present in the forecast and observation. Therefore the apparent better performance of the frequent lightning category deduced from its smaller error in Figure 8 is in reality due to the fact that this category has a smaller sample climatology with respect to the other two categories, as shown in Figure 5. The Brier score is not suitable to compare the performance of forecasts with significantly different sample climatologies; a more fair comparison can be provided by the Brier skill score (see following sections).

The Brier score on each spatial scale is defined as the Brier score of the scale component of each lightning probabilistic forecast and its corresponding scale component of the observed frequency image:

$$\begin{aligned} \mathrm{BS}_{j,m} &= \overline{(Y_{j,m} - X_{j,m})^2}, \quad j = 1, J \\ \mathrm{BS}_{J,f} &= \overline{(Y_{J,f} - X_{J,f})^2}. \end{aligned} \tag{10}$$

The Brier score components on each spatial scale provide feedback on the amount of forecast error that each scale exhibits, separately. Figure 9 shows

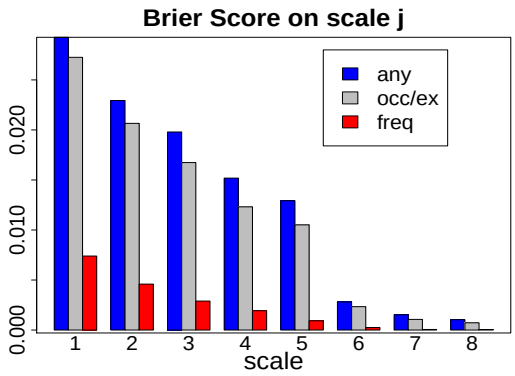

Figure 9: Brier score scale components for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning.

the Brier score scale components for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning. As for the squared energy (Figure 6), small scales exhibit the largest error and then the error decreases as the scale increases. Note again that the category of frequent lightning exhibits a smaller error on all the scales with respect to the other two categories; this is not due to better performance, but to the presence of less events in the forecast and observation for this category (in fact this category exhibits smaller sample climatology and smaller squared energy on all the scales, as shown in Figures 5 and 6). The same argument applies to the categories of occasional to extreme lightning and any lightning, however the differences in their behavior are less remarkable. The error is proportional to the proportion of events on each scale.

3.6 Brier score percentage for each scale

The sum of the Brier score components on different scales defined by Eqn. (10) is equal to the total Brier score given by Eqn. (9). In fact, the probability error image can be decomposed as the sum of components on different scales by using the 2D discrete Haar wavelet filter:

$$Z = \sum_{j=1}^{J} Z_{j,m} + Z_{J,f} \tag{11}$$

(the notation used here is the same as for Eqns. (1) and (2) for the wavelet decomposition of forecast probability and observed frequency images). The wavelet decomposition is a linear operator, therefore

$$\begin{aligned} Z_{j,m} &= (Y-X)_{j,m} = Y_{j,m} - X_{j,m}, \\ Z_{J,f} &= (Y-X)_{J,f} = Y_{J,f} - X_{J,f}, \end{aligned} \tag{12}$$

and so it follows

$$\mathrm{BS}_{j,m} = \overline{Z_{j,m}^2}, \qquad \mathrm{BS}_{J,f} = \overline{Z_{J,f}^2}. \tag{13}$$

Moreover, the wavelet components obtained from a discrete wavelet transform are orthogonal, therefore

$$\begin{aligned} \overline{Z_{j,m} Z_{k,m}} &= 0, \quad \forall\, j \neq k; \\ \overline{Z_{j,m} Z_{J,f}} &= 0. \end{aligned} \tag{14}$$

From Eqns. (9), (11), (14) and (13) it is shown that the Brier score is equal to the sum of its components on different spatial scales:

$$\begin{aligned} \mathrm{BS} = \overline{(Z)^2} = \overline{\left(\sum_{j=1}^{J} Z_{j,m} + Z_{J,f}\right)^2} = \\ \overline{\left(\sum_{j=1}^{J} Z_{j,m} + Z_{J,f}\right)\left(\sum_{j=1}^{J} Z_{j,m} + Z_{J,f}\right)} = \\ \sum_{j=1}^{J} \overline{Z_{j,m}^2} + \overline{Z_{J,f}^2} = \sum_{j=1}^{J} \mathrm{BS}_{j,m} + \mathrm{BS}_{J,f}. \end{aligned} \tag{15}$$

This result enables one to evaluate the percentage for which each scale contributes to the total Brier score:

$$\begin{aligned} \mathrm{BS}^{\%}_{j,m} &= \mathrm{BS}_{j,m}/\mathrm{BS}; \\ \mathrm{BS}^{\%}_{J,f} &= \mathrm{BS}_{J,f}/\mathrm{BS}. \end{aligned} \tag{16}$$

The percentage of Brier scores on each scales provides feedback on the fraction of error that each scale carries. Figure 10 shows the percentages of the Brier score scale components for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning. As for the Brier score (Figure 9), small scales exhibit the largest percentage of error and then the error decreases as the scale increases. However, the behavior of the frequent lightning category is better shown by Figure 10 than Figure 9. In fact, the Brier score components shown in Figure 9 do not reveal really the frequent lightning category error behavior, since this error is so small when compared to the other two category errors (because of the Brier score dependence on the sample climatology) that it is barely noticeable. On the other hand, the Brier score percentages shown in Figure 10 show how much each scale contributes to the total error as a fraction of the total, which is equal for all three categories. These percentages are independent from the total error itself or from the sample climatology and enable a more direct comparison of the error structure for the three probability categories.

Note that, as for the percentage of the squared energy (Figure 7), the fraction of error on small

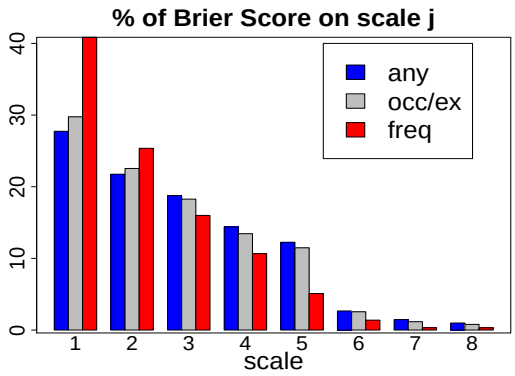

Figure 10: Percentages of the Brier score scale components for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning.

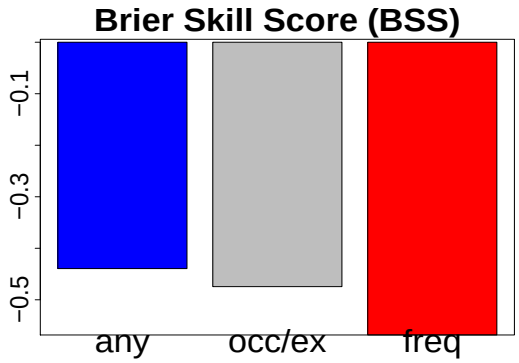

Figure 11: Brier skill score for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning.

scales for the frequent lightning category is larger than for the other two categories. This is due to the presence of a larger number of small scale events in the frequent lightning category, isolated by thresholding the lightning occurrence with a higher threshold (e^3) than the other two categories (0 and 0.5). Viceversa, as the scale increases, the fraction of error on large scales for any lightning or occasional to extreme lightning is larger than for the frequent lightning category. This is due to the presence of large scale features in the probability forecast and observed frequeency images for this two categories because defined from a low threshold on ligthning occurrence. The error is proportional to the proportion of events on each scale. The same argument applies when explaining the similar behaviour of the percentage of the Brier Score for the categories of occasional to extreme lightning and any lightning. However, the differences in the statistical behaviour of these two categories are less remarkable than the one between these and the frequent lightning category, because of the less remarkable difference in the amount of events (squared energy) present on each scale (see Figures 6 and 7).

3.7 Scale decomposition of the Brier skill score

To assess the lightning probability forecast skill, the Brier skill score versus climatology (Jolliffe and Stephenson, 2003, chapter 7) is evaluated:

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{\mathrm{BS}_{\mathrm{perf}} - \mathrm{BS}_{\mathrm{ref}}} = 1 - \frac{\mathrm{BS}}{\sigma_X^2}, \qquad (17)$$

where $\mathrm{BS}_{\mathrm{perf}} = 0$ is the Brier score for a perfect forecast, and $\mathrm{BS}_{\mathrm{ref}} = \mathrm{BS}_{\mathrm{clim}}$ is the Brier score one would obtain by forecasting for each pixel in the forecast image the sample climatology ($Y = \overline{X}$). Note that the climatological forecast Brier score is equal to the observation variance:

$$\mathrm{BS}_{\mathrm{clim}} = \overline{(\overline{X} - X)^2} = \sigma_X^2. \qquad (18)$$

The Brier skill score for a perfect forecast is equal to one; when the skill is positive the forecast performs better than the climatological forecast, when negative the forecast performs worse. Figure 11 shows the Brier skill score for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning. The skill is negative for all the categories, indicating that the lightning probabilistic forecast performs worse than the climatological forecast $Y = \overline{X}$. The skill for the category of frequent lightning is worse than for the other two categories.

Variance and squared energy are related by

$$\sigma_X^2 = \overline{(\overline{X} - X)^2} = \overline{X^2} - \overline{X}^2 \qquad (19)$$

so that the variance is equal to the difference between the squared energy and the square of the average field value. The average field value is equal to the largest father wavelet component (see Appendix), therefore

$$\overline{X}^2 = X_{J,f}^2 = \overline{X_{J,f}^2} = En^2(X_{J,f}). \qquad (20)$$

From this result and Eqns. (19), (3), (7) and (5) we can express the variance as a sum of components on different scales

$$\sigma_X^2 = \overline{X^2} - \overline{X}^2 = En^2(X) - En^2(X_{J,f}) = \sum_{j=1}^{J} En^2(X_{j,m}) = \sum_{j=1}^{J} \overline{X_{j,m}^2}. \qquad (21)$$

where each scale component of the variance is equal to the mother wavelet component of the squared energy:

$$\sigma^2_{X,j} = \overline{X^2_{j,m}} = En^2(X_{j,m}). \qquad (22)$$

Note that mother wavelets have zero integral; therefore, each mother wavelet component $X_{j,m}$ has zero mean; the variance of a field with zero mean is equal to its squared energy; therefore, the definition of scale component of the variance given in Eqn. (22) is identical to defining the scale component of the variance as the variance of each spatial scale component of the field:

$$\sigma^2_{X,j} = \sigma^2_{X_{j,m}}. \qquad (23)$$

The scale component of the variance corresponding to the largest father wavelet component $X_{J,f} = \overline{X}$ is zero, since it is the variance of a constant field.

The Brier skill score components on different spatial scales are defined from the Brier score components on different scales given by Eqn. (10) and the observation variance components on different scales given by Eqn. (22):

$$\mathrm{BSS}_j = 1 - \frac{\mathrm{BS}_{j,m}}{\sigma^2_{X,j}}, \qquad j = 1, \ldots, J. \qquad (24)$$

The Brier skill score components on the different scales measure the skill of the forecast at each scale; BSS_j is equal to one for perfect skill; BSS_j is positive when the forecast performs better than the climatological forecast, and it is negative when the forecast performs worse than the climatological forecast (no skill). Figure 12 shows the Brier skill score components on the different scales for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning. For all three categories, the skill is negative on small and intermediate scales (1 to 5, corresponding to 20 to 320 km scale features), and it becomes positive only on very large scales (640 km and larger features). The negative skill is due mainly to feature displacements. Positive skill on scales larger than 640 km indicates that large scale features, such as frontal systems, are detected by the lightning probabilistic forecast. The category of frequent lightning exhibits a particularly negative skill at the scale 5, corresponding to 320 km features. This is due partially to the smoothing and overforecasting of such features and to dispacement error on this scales, whcih can be noticed directly from Figures 3, 4.

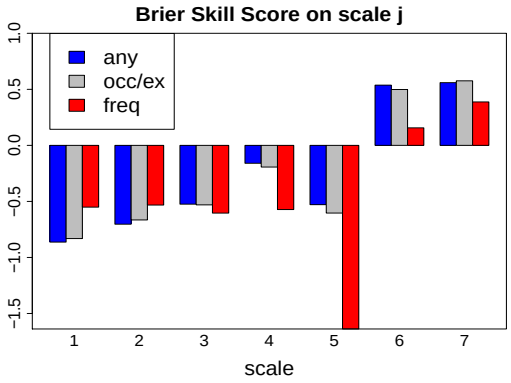

Figure 12: Brier skill score components on the different scales for the case study illustrated in Figures 3, 4 for the three categories of any lightning, from occasional to extreme lightning and frequent lightning.

4 DISCUSSION AND CONCLUSIONS

A new diagnostic verification technique for probabilistic forecasts defined on a spatial domain has been introduced in this work. The method provides feedback on the forecast performance and nature of the forecast error on different scales. It measures error, skill and bias on different scales. Moreover, the method is capable of verifying the ability of the forecast to reproduce the observed scale structure.

The verification technique has been tested using a case study of the CMC lightning probabilistic forecasts. Three probability categories for any lightning, occasional to extreme lightning and frequent lightning have been assessed. Forecast and observation are decomposed into the sum of components on different spatial scales by performing a 2D Haar wavelet decomposition. Brier score and Brier skill score are then evaluated on each scale, along with squared energy percentages and bias.

The decomposition of the Brier score on different scales revealed that the largest error is associated with the smallest scales. As the scale increases the error decreases. The fraction of the error for more intense lightning activity is larger on small scales and smaller on large scales compared to the fraction of the error for less intense lightning activity. The error is strongly related to the number of events present in the forecast and observation image on each scale.

The Brier skill score on different scales shows that only very large scales (larger than 640 km) have positive skill. This indicates that only very large scale features (such as fronts) are correctly forecast. In the case study considered, the forecast for intense

lightning activity exhibits a particular negative skill at the 320 km resolution scale, due to overforecasting (caused by smoothing) and displacement of features on this scale.

The squared energy bias and the its percentage on each scale show that the forecast exhibits an overall over-forecast, detected by the larger (father wavelet) scale component. The categories of any lightning and the occasional to extreme lightning under-forecast on all the scales, in particular on the 80, 160 and 320 resolution scales. The ratio of the squared energy percentages reveals that the scale structure is well represented by these two forecast categories. The forecast of frequent lightning underforecast small scales (features of 20 to 160 km resolution) and overforecast large scales (in particular the 320 km resolution scale). The ratio of the squared energy percentages reveals that the forecast is smoother than the observation.

The technique still needs further development. Reliability images can be obtained from the observed frequency image and the forecast probability image. Brier score and skill score reliability and resolution components can then be decomposed on different scales. Furthermore the technique needs to be tested on a larger number of case studies and on monthly runs. Verification statistics will be provided with their associated confidence intervals. The verification technique should be tested also on different probabilistic forecasts.

The technique enables the comparison of different forecasts on different spatial scales. This includes also forecasts on different resolutions. It is very well known that high resolution forecasts, when assessed with traditional verification techniques, perform dramatically worse than low resolution forecasts, due to their intrinsic high variability. The Brier skill score decomposition introduced in this work is defined on each scale by a normalization of the Brier score by the scale variance. This enables a more fair verification of forecasts on different scale resolutions.

APPENDIX: THE 2D DISCRETE HAAR WAVELET FILTER

Wavelets are real function characterised by a location and a scale (Daubechies, 1992; Mallat, 1989). Similar to Fourier transforms, wavelets can be used to represent a function as a sum of components on different spatial scales, and therefore they can be used to analyze the frequency structure of a signal or the scale structure of a field. Because of their local properties, wavelets are more suitable than Fourier series for representing spatially discontinuous fields such as lightning. Fourier expansions can describe smooth functions, but when used for discontinuous fields they lead to problematic Gibbs' phenomena. Moreover, because of their locality, wavelets are more efficient than Fourier components at representing sparse images containing few non-zero values. For these reasons, in this study wavelets are used rather than Fourier expansions.

Different types of wavelets exist. Each wavelet type is defined by a mother and a father wavelet, characterised by different shapes and mathematical properties (e.g. smoothness, symmetry, etc.). When performing a wavelet decomposition, it is often desirable to select a certain wavelet so as to gain from the characteristics of the wavelet itself. As an example, the wavelet chosen could be the one that minimises the number of significant wavelet coefficients describing the decomposed function. The choice of the "optimal" type of wavelet to be used to perform a wavelet decomposition depends on the characteristics of the function to be decomposed. In this work, Haar wavelets are used, because of their square shape which best deal with sharp discontinuities. Figure 13 shows the one- and two- dimensional Haar wavelets. Note that the two-dimensional wavelets are generated simply as the Cartesian product of one-dimensional wavelets.

A discrete wavelet family is a set of wavelets of the same type generated from the mother and father wavelets by a deformation and a translation. The deformation characterises the scale j of the wavelet: it stretches the domain of the wavelet by a factor of 2^j and reduces its amplitude by a factor of $2^{-j/2}$ (this is to maintain its L^2 norm[1] equal to one). Therefore, wavelets on the scale j have a domain which is twice as large as the domain of the wavelets on the spatial scale $j-1$, i.e. as the scale increases wavelets are stretched by a factor of 2. The translation determines the location of the wavelet in the domain: wavelets of scale j is translated by a multiple of 2^j units. For Haar wavelets this implies that wavelets of the same spatial scale cover the whole domain and their supports do not overlap.

A discrete wavelet family is an orthonormal basis for $L^2(\mathbb{R}^n)^*$. Therefore, any function belonging to $L^2(\mathbb{R}^n)$ (e.g. any function with finite values defined on a discrete finite grid, such as any signal or field with finite values stored in a computer data set) can be expressed as a linear combination of

*$L^2(\mathbb{R}^n)$ is defi ned as the set of functions f defi ned on $\mathbb{R}$ that satisfy $\int |f|^2 < \infty$; the L^2 norm of a function f belonging to $L^2(\mathbb{R}^n)$ is $(\int |f|^2)^{1/2}$; the inner product of two functions f and g belonging to $L^2(\mathbb{R}^n)$ is $(\int |fg|)^{1/2}$.

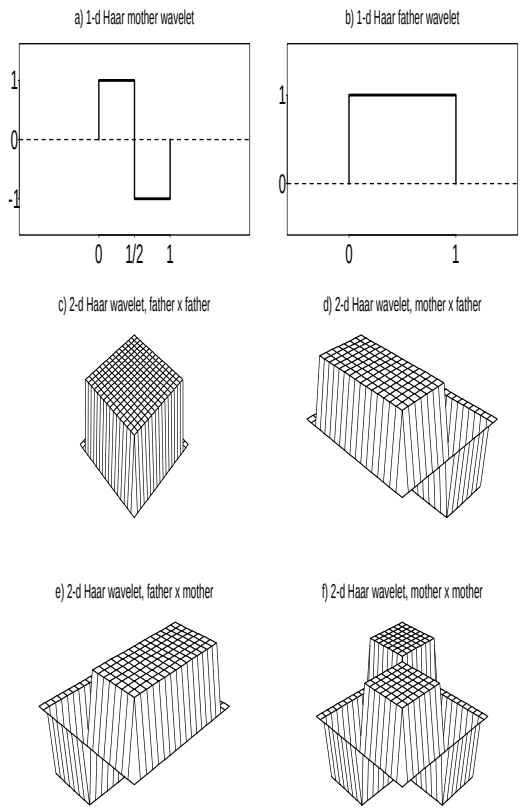

Figure 13: One- and two- dimensional Haar wavelets. In the one-dimensional discrete Haar wavelet decomposition of a real function, the mother wavelet components are generated by the mother wavelet shown in panel (a), the father wavelet components are generated by the father wavelet shown in panel (b). In two dimensions, the father wavelet components are generated from the two-dimensional Haar father wavelet shown in the panel (c). The mother wavelet components are generated from the two-dimensional Haar wavelets shown in the panels (d), (e) and (f).

discrete wavelets of the same family, and therefore as a sum of components on different spatial scales. As for the Fourier series, the coefficient assigned to each wavelet is equal to the integral of the absolute value of the product of the function and the corresponding wavelet (i.e. the L^2 inner product[1] of the function and the corresponding wavelet). Note that discrete wavelets are orthogonal (i.e. their L^2 inner product[1] is zero), therefore the integral over the spatial domain of the product of two different wavelets of the same discrete family is equal to zero The spatial scale components obtained from a discrete wavelet decomposition are also orthogonal. Note also that the discrete wavelet decomposition is a linear operator, i.e. the wavelet decomposition of a linear combination of functions is the linear combination of the wavelet decomposition of each function.

The 2D discrete Haar wavelet filter can be explained by an algorithm based on spatial averaging over $2^j \times 2^j$ pixel domains. In this work we illustrate the two-dimensional Haar wavelet filter with this approach. The two-dimensional Haar wavelet filter is applied to a spatial field Z defined over a spatial domain of $2^J \times 2^J$ pixels.

The Haar wavelet filter at its first step decomposes the spatial field Z into the sum of a coarser *mean* field (the first father wavelet component) and a detail *variation-about-the-mean* field (the first mother wavelet component). The father wavelet component is obtained from the spatial field Z by a spatial averaging over 2×2 pixels. The mother wavelet component is obtained as the difference between the spatial field Z and the father wavelet component.

At its second step the Haar wavelet filter decomposes the father wavelet component obtained from the first step into the sum of a coarser *mean* field (the second father wavelet component) and a detail *variation-about-the-mean* field (the second mother wavelet component). The second father wavelet component is obtained from the spatial field Z by a spatial averaging over 4×4 pixels. The second mother wavelet component is obtained as the difference between the second father wavelet component and the first father wavelet component.

The process is recursive and at each step the Haar wavelet filter decomposes the father wavelet component obtained from the $(j-1)^{th}$ step into the sum of a coarser *mean* field (the j^{th} father wavelet component) and a detail *variation-about-the-mean* field (the j^{th} mother wavelet component). The j^{th} father wavelet component is obtained from the spatial field Z by a spatial averaging over $2^j \times 2^j$ pixels. The j^{th} mother wavelet component is ob-

tained as the difference between j^{th} and $(j-1)^{th}$ father wavelet components. The process stops when the father wavelet component corresponding to the largest scale (J) is found. The spatial field Z is decomposed into the sum of the mother wavelet components on the spatial scales $j = 1, \ldots, J$ and the J^{th} father wavelet component:

$$Z = \sum_{j=1}^{J} Z_{j,m} + Z_{J,f}, \tag{25}$$

where the mother $Z_{j,m}$ wavelet components on the scale j have resolution equals to 2^{j-1} pixels and the father wavelet component $Z_{J,f}$ on the largest scale J has resolution equal to 2^J pixels. Note that the largest (J^{th}) father wavelet component is equal to the mean of Z over the whole $2^J \times 2^J$ pixel spatial domain. Therefore

$$Z = \sum_{j=1}^{J} Z_{j,m} + \overline{Z}, \tag{26}$$

where the overbar indicates averaging over all the pixels in the domain.

References

Brier, G. W. (1950). Verification of forecasts expressed in term of probability. *Monthly Weather Review*, **78**, 1–3.

Briggs, W. M. and Levine, R. A. (1997). Wavelets and field forecast verification. *Monthly Weather Review*, **125**, 1329–1341.

Burrows, W., Price, C., and Wilson, L. (2005). Warm season lightning probability prediction for Canada and the northen United States. *Weather and Forecasting*. in press.

Casati, B., Ross, G., and Stephenson, D. (2004). A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Application*, **11**, 141–154.

Côté, J., Gravél, S., Méthot, A., Patoine, A., Roch, M., and Staniforth, A. (1997). The operational CMC/MRB Global Environment Multiscale (GEM) model: part I - Design consideration and formulation. *Monthly Weather Review*, **126**, 1373–1395.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

De Elia, R., Laprise, R., and Denis, B. (2002). Forecasting skill limits of nested, limited-area models: a perfect model approach. *Monthly Weather Review*, **130**, 2006–2023.

Denis, B., Laprise, R., and Caya, D. (2003). Sensitivity of a regional climate model to the resolution of the lateral boundary conditions. *Climate Dynamics*, **20**, 107–126.

Harris, D. and Foufoula-Georgiou, E. (2001). Multiscale statistical properties of a high-resolution precipitation forecast. *Journal of Hydrometeorology*, **2**, 406–418.

Jolliffe, I. T. and Stephenson, D. B., editors (2003). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on pattern analysis and machine intelligence*, **II**(7), 674–693.

Zepeda-Arce, J. and Foufoula-Georgiou, E. (2000). Space-time rainfall organization and its role in validating Quantitative Precipitation Forecasts. *Journal of Geophysical Research*, **105**, 10,129–10,146.