A scheme for verifying the spatial structure of extremes in numerical weather prediction: Exemplified for precipitation

Bent H. Sass

Correspondence
Bent H. Sass, Danish Meteorological Institute, Lyngbyvej 100, 2100, Copenhagen, Denmark.
Email: bhs@dmi.dk

Abstract
This paper describes a novel scheme for assessing the ability of a numerical weather prediction (NWP) model to forecast the horizontal spatial structure of local extremes, for example, when comparing accumulated precipitation to a corresponding observed field. Both local maxima and minima are considered. The quality is measured by a score function between 0 and 1, which could potentially be adjusted for specific applications of NWP. The effect of neighbourhood size is included in order to consider the impact of resolvable scales. The scheme complements existing spatial verification schemes with added value and has a potential to be used both in operational verification and as a tool for revealing model deficiencies linked to individual forecast cases. The properties of the scheme are illustrated from various idealized cases. Verification results are presented for a real convective precipitation case and for a frontal precipitation case over several days. The scheme is discussed in the context of operational conditions, using synthetic fields of forecasts and analyses based on statistics of an operational NWP model at the Danish Meteorological Institute (DMI) and related observed data from a 5-year period. Different future options are outlined, including the potential application of the scheme for ensemble prediction systems.

KEYWORDS
neighbourhood methods in NWP, NWP verification, verification of extreme precipitation

1 | INTRODUCTION

From the outset of numerical weather prediction (NWP) modelling, there has been a need to verify the quality of the model predictions. Many aspects of model performance have been studied in the past and have led to a large variety of verifications schemes. This is reflected in detailed documents from international meetings, for example, from the World Weather Research Programme (WWRP) Joint Working Group on Forecast Verification Research (http://www.cawcr.gov.au/projects/verification/).

There is an increasing focus on using NWP models to warn the public about localized precipitation events, for example, cloud bursts. It has become feasible to operate high-resolution cloud-resolving NWP models providing...
detailed precipitation forecasts of extreme events. However, these operational high-resolution systems are not perfect, and value of the increased spatial details need to be assessed by appropriate verification methods. The challenges related with NWP high-resolution model systems are summarized by, for example, Yano et al. (2018).

Providing optimal information and clear communication on precipitation to NWP users is demanding because of the large possible variations of precipitation across regions or municipalities of a country. Each user of rain forecasts would like to know if heavy precipitation or dry weather occurs at a given location. Unexpected large precipitation amounts such as cloudbursts in large cities may lead to significant property damage and related economic losses, as a result of flooding. Knowledge about expected dry conditions in a given area is also very important, especially if such area can be clearly distinguished from neighbouring areas with high probabilities of rain. Activities depending on dry conditions comprise numerous outdoor activities, for example, construction work, sport events, concerts, and so forth. As a consequence, ideal information from NWP to users would comprise a detailed geographical display of predicted rain and dry areas, and there is a need to verify this geographical distribution (structure) of areas with high and low precipitation amounts, in particular the highest and lowest values (dry spots). This requires a special type of verification, which is not easy to provide by traditional verification schemes. This paper describes a novel verification scheme named SLX (Structure of Local Extremes), dedicated to this particular aspect of verification considering both maxima and minima of predicted and analysed precipitation.

Precipitation is challenging to verify because of its large variability in space and time as documented, for example, by Eggert et al. (2015) in their investigation of temporal and spatial scaling impacts on extreme precipitation over Germany. Fiener and Austerwald (2009) documented a variability of observed precipitation down to sub-km scales. Lovejoy and Mandelbrot (1985) pointed to fractal properties of rain. This highlights the need to compare model predictions with observations at different spatial and temporal scales. In recent years, high-resolution precipitation analyses have become available for NWP model verification. They are based on networks of weather radars and rain gauge measurements. An increasing number of precipitation observations are becoming available as crowd-sourced data for NWP (e.g., Hintz et al., 2019). While useful, these data require careful quality control to contribute to improve the quality of precipitation analysis.

The increasing horizontal resolution of operational NWP models down to cloud-resolving scale has led to another verification issue, the so-called ‘double penalty’ problem, for example, Bougeault (2003) among others. First, the model is penalized for not predicting an observed high value at the observed location, and then is further penalized for predicting a high value at a slightly incorrect location. Therefore, a new generation of verification schemes has been developed to address the double penalty problem. Ebert (2008, 2009) summarizes the concepts of spatial verification, and Gilleland et al. (2010) provide an instructive summary of four dominating types of spatial verification: neighbourhood, scale separation, features-based and field deformation, respectively. The neighbourhood methods include information in a neighbourhood around the grid point, which may be observation or forecast point, to assess the quality. This acts as a kind of filter. An example of this type of scheme, which is widely used in limited area NWP, is the Fractions Skill Score (FSS) (Roberts & Lean, 2008). FSS compares binary fields (0 and 1) derived by examining if some threshold is exceeded or not in forecast field and analysis field. Related fractions of 1 in neighbourhoods are compared as input to the score computation over the entire limited area.

Scale separation methods assess quality by a process of decomposing the fields into different scales. An example of this type is Harris et al. (2001) who examine multi-scale properties of a precipitation forecast. Features-based methods identify specific structures, objects or characteristics linked to the observed and forecasted fields and compose a score from that. A popular scheme of this type is the SAL scheme for limited area NWP (Wernli & Paulat, 2008). SAL stands for ‘Structure, Amplitude and Location’.

Recently, the spatial verification methods have been extended to include distance measures. This category measures the spatial distance between certain meteorological events, forecasted and observed (Dorninger et al., 2018). In addition, a comprehensive set of data has been defined for international collaboration on spatial verification methods (Gilleland et al., 2020).

Details of the methodology related with the new scheme (denoted SLX) are described in Section 2. The data used as a basis for verifying the scheme in Danish Meteorological Institute (DMI) are summarized in Section 3. An analysis of idealized examples is included in Section 4 while Section 5 is looking at the scheme in the context of real-world examples. Section 6 is a discussion section, and conclusions are provided in Section 7.

2 | DEFINITION OF THE NEW SCORE

The intention of the novel verification scheme SLX verification (Structure of Local Extremes) is to design a score,
with four sub-components measuring how well diagnosed local extremes (minimum or maximum) of a forecast or analysis field are reproduced in an absolute sense by the other field. Two scores identify, respectively, how forecasts match analysed maxima and minima in the model domain, and two other scores identify, respectively, how analysed values match identified forecasted maxima and minima. All components of the score assign values in the interval [0, 1], with 0 being poorest performance and 1 perfect agreement.

This means that local maxima and minima are identified in the forecast- and analysis fields. When scores related with these values are computed on a daily basis, they do not represent model performance for extremes in a traditional sense since weeks or months may be required to experience precipitation events above a certain threshold. However, post-processing of accumulated daily results over sufficiently long periods will eliminate this issue since results associated with thresholds representing extremes may be extracted for further examination.

An early version of a scheme addressing the occurrence of local extremes has been operational in DMI for several years (Sass & Yang, 2012). The forecasted precipitation field is compared against three highest and three lowest observed 12-h accumulated precipitation measured at rain gauges in Denmark. The SLX scheme represents a substantial extension of the scheme used in DMI so far, since it is based on complete horizontal fields of analysed and forecasted precipitation to be compared. It has been a priority to design the SLX method such that an output of the scores is produced daily for any input fields of analysis and forecast, including zero-valued fields. Figure 1 illustrates the SLX method. The scheme is based on the availability of a forecast field \( \phi(i,j) \) and an analysis field \( \Psi(i,j) \) of accumulated precipitation over some period being typically 1 or more hours, and \( i, j \) represents the points in the two-dimensional field of the model domain to be verified. The two fields are assumed to have the same horizontal resolution.

Figure 1 illustrates the geometrical principles of the scheme. Neighbourhoods are defined around local extreme points identified for both forecast and analysis. For simplicity, squared neighbourhoods are defined, for example, similar to the neighbourhood treatment defined by Roberts and Lean (2008). The grid point neighbourhood width \( L \) used here is measured in number of grid points. The points up to a distance of \( L \) from the identified extreme point, measured in each of the two horizontal directions, are included in the neighbourhood. \( L = 0 \) means that comparisons are on the level of single grid points. The interval \( [-L, +L] \) defines the number of grid points of the neighbourhood in each of the two directions. The total number of grid points in a squared neighbourhood is then \( (2L + 1)^2 \).

In Figure 1, \( L_{max} \) is the largest neighbourhood width of the chosen set-up. It is equal to the width of a boundary zone that defines a transition between internal points of the verification domain and external points. If there is only one verification domain of a limited area model, the external points will belong to the host model. The boundary zone is represented by the area between the two large black squares of the figure. The algorithm is designed such that the total domain may be divided into a given number of sub-domains defined by a number of grid points in the two horizontal directions. In this case, some boundary zones are between sub-domains and do not represent a transition to an external model. For large model areas of, say 1000 or more grid points in both directions, it is reasonable to partition the model into sub-domains in order to distinguish regional differences.
of the computations in a final output. Alternatively, one may choose to compute averages over sub-domains as a final output. Testing different values of \( L \) \((0 \leq L \leq L_{\text{max}})\) is normally part of using SLX for a given weather situation or an idealized case. It is recommended to compute the score based on internal points only, in order to be able to include the same grid points in the computations and complete neighbourhoods of the points considered. This means that the identified local extremes are internal points.

Selecting extreme points of the analysis and forecast fields requires a decision on a tolerance limit \( \delta \) \((\text{kg/m}^2)\), which reflects the level of uncertainty of the analysis. Choosing a tolerance significantly larger than zero for non-zero precipitation extremes implies increased likelihood that several points will be selected as extremes. A value of \( \delta \) close to zero, for example, at 0.1 \text{kg/m}^2, could be chosen as default.

Zero-valued dry areas will often exist as part of both observation- and forecast fields. In this situation, multiple points of zero value will be automatically selected as minima for the verification process. All the selected points will be treated with the same neighbourhood width \( L \) and contribute with equal weight to the score, as will be treated with the same neighbourhood width \( L \) for the verification process. All the selected points will be automatically selected as minima.

Figure 1 displays the names of local extreme points: \( \text{obmax}(K_1) \), \( \text{obmin}(K_2) \), \( \text{fcmax}(K_3) \) and \( \text{fcmin}(K_4) \) respectively. \( K_1 \) , \( K_2 \) , \( K_3 \) and \( K_4 \) are indexes of the extremes, \( 1 \leq K_1 \leq M_1 \) , \( 1 \leq K_2 \leq M_2 \) , \( 1 \leq K_3 \leq M_3 \) , \( 1 \leq K_4 \leq M_4 \) , with the number of extremes of each type being, respectively: \( M_1 \) , \( M_2 \) , \( M_3 \) and \( M_4 \).

\( \text{obmax}(K_1) \) identifies an observed local maximum point. The value \( \text{obmax}(K_1) \) is compared with \( \phi_{\text{max}}(L, K_1) \), which is the maximum of the forecast field in the neighbourhood of the central point representing \( \text{obmax}(K_1) \) (Equation 1a)

\[
\phi_{\text{max}}(L, K_1) = \text{Max}\{\phi(i,j)\}, i \in [i_{K1} - L,i_{K1} + L], j \in [j_{K1} - L,j_{K1} + L]
\]

(1a)

\( i_{K1}, j_{K1} \) are coordinates of the related forecasted values in central point.

\( \text{obmin}(K_2) \) represents an observed local minimum point. When computing the score, the value \( \phi_{\text{min}}(L, K_2) \) is compared with \( \phi_{\text{min}}(L, K_2) \), which is the minimum of the forecast field in the neighbourhood of the point representing \( \text{obmin}(K_2) \) (Equation 1b)

\[
\phi_{\text{min}}(L, K_2) = \text{Min}\{\phi(i,j)\}, i \in [i_{K2} - L,i_{K2} + L], j \in [j_{K2} - L,j_{K2} + L]
\]

(1b)

\( i_{K2}, j_{K2} \) are coordinates of the related forecasted values in central point.

\( \text{fcmax}(K_3) \) represents a forecasted local maximum point. When computing the local score, the value \( \text{fcmax}(K_3) \) is compared with \( \Psi_{\text{max}}(L, K_3) \), which is the maximum of the analysis field in the neighbourhood of the point representing \( \text{fcmax}(K_3) \) (Equation 1c)

\[
\Psi_{\text{max}}(L, K_3) = \text{Max}\{\Psi(i,j)\}, i \in [i_{K3} - L,i_{K3} + L], j \in [j_{K3} - L,j_{K3} + L]
\]

(1c)

\( i_{K3}, j_{K3} \) are coordinates of the related forecasted values in central point.

\( \text{fcmin}(K_4) \) represents a forecasted local minimum point. When computing the local score, the value \( \text{fcmin}(K_4) \) is compared with \( \Psi_{\text{min}}(L, K_4) \), which is the minimum of the analysis field in the neighbourhood of the point representing \( \text{fcmin}(K_4) \) (Equation 1d)

\[
\Psi_{\text{min}}(L, K_4) = \text{Min}\{\Psi(i,j)\}, i \in [i_{K4} - L,i_{K4} + L], j \in [j_{K4} - L,j_{K4} + L]
\]

(1d)

\( i_{K4}, j_{K4} \) are coordinates of the related forecasted values in central point.

Regarding, as an example, the score computation related with an observed maximum \( \text{obmax}(K_1) \) to be compared with the maximum of the forecast field \( \phi_{\text{max}}(L, K_1) \) in the neighbourhood: If these values are in very close agreement, the contribution of the point to the related total score named \( \text{SLX}_{\text{ob max}} \) is close to 1 (maximum contribution). If the forecast field is very different from \( \text{obmax}(K_1) \) in the neighbourhood, the contribution to the score will be smaller and possibly zero, if the difference is sufficiently large. Correspondingly, the fit of a forecasted extreme maximum point to the extreme of the analysis points in the neighbourhood leads to a contribution to the score component \( \text{SLX}_{\text{fc max}} \). Finally, the scores related with minima are named \( \text{SLX}_{\text{ob min}} \) and \( \text{SLX}_{\text{fc min}} \) respectively: The former determines how well the observed (analysed) minimum point agrees with the minimum of the forecast field in the neighbourhood, and the latter measures the degree of match of a forecasted minimum value to the analysis minimum in the neighbourhood of the forecasted minimum.

\( \text{SLX}_{\text{ob max}}, \text{SLX}_{\text{fc max}}, \text{SLX}_{\text{ob min}} \) and \( \text{SLX}_{\text{fc min}} \) are defined in (4)–(7) utilizing values of the score defined below.

The assessment of the local value of the score is defined by a score function \( S \) (Figure 2), and formulas (2a)–(2c), (3a)–(3b), detailed below, are used for both the local maxima and minima.
The SLX score function used in this study. The score is defined between zero (poorest quality) and 1 (perfect forecast). The function is piecewise linear in the intervals \([0,\text{OB-k}], [\text{OB-k}, \text{OB}], [\text{OB}, 5\times\text{OB}]\). ‘ob’ means observed extreme, ‘\(S\)’ means forecasted value. \(k/\text{(kg/m}^2\text{)}\) represents a value below which a forecast should be assigned a perfect score of 1. It should be equal to or closely related to observation uncertainty at small precipitation amounts. Currently, \(k = 0.1\) (kg/m\(^2\)): \(A\) is a dimensionless constant representing how fast the score is decreasing towards zero when forecast is larger than observation.

If \(\text{ob} > k\):

\[
\begin{align*}
\text{If } \phi < \text{ob} - k : & \quad S = \phi / (\text{ob} - k) \quad (2a) \\
\text{If } \text{ob} - k \leq \phi \leq \text{ob} : & \quad S = 1 \quad (2b) \\
\text{If } \phi > \text{ob} : & \quad S = \text{Max}\{1 - (\phi - \text{ob}) / (A \times \text{ob}), 0\}, A = 4 \quad (2c)
\end{align*}
\]

If \(\text{ob} \leq k\):

\[
\begin{align*}
\text{If } \phi \leq k : & \quad S = 1 \quad (3a) \\
\text{If } \phi > k : & \quad S = \text{Max}\{1 - (\phi - k) / (A \times k), 0\}, A = 4 \quad (3b)
\end{align*}
\]

It is seen that the present choice of score function \(S\) is piecewise linear and very simple. The most characteristic feature is that \(S\) is asymmetric due to the application of \(A = 4\) in \((2c)\) and \((3b)\). In DMI, we find this asymmetry useful because of the risk of under-forecasting warning conditions for different user groups who depend on critical thresholds. Over-forecasting will trigger a warning and thus avoid or reduce economic losses. However, too much over-forecasting is not useful and may cause significant and unnecessary expenses to the users. Over-prediction relative to observation by a factor of 5 is set as a limit for useful forecasts. The score function should be thought of as a construct between model developers and NWP users, for example, forecasters. Other specific user groups may define the \(S\) function somewhat differently depending on their practical application. The challenge of defining score function for precipitation is discussed further in Section 6.

\(S\) decreases linearly from 1 to zero as the fraction \((\phi - k)/(A \times k)\) increases from zero to 1 (Equation 3b). Using \(k = 0.1\) (kg/m\(^2\)), \(S\) decreases to zero at 0.5 kg/m\(^2\). A gradual decrease towards zero for \(\phi > 0.1\) kg/m\(^2\) is used in order to avoid a large sensitivity to small changes of forecast precipitation close to \(k\) (kg/m\(^2\)). With these definitions, \((3a)\) and \((3b)\) define some asymmetry of the score function near zero. For simplicity, the same coupling as expressed by \((3a)\) and \((3b)\) has been adopted in \((2b)\) and \((2c)\). This implies also small asymmetry in score function (Figure 2) near the observation ‘ob’, since from \((2b)\)

\[
\text{S} = 1 \quad \text{if } \text{ob} - k \leq \phi \leq \text{ob} \quad \text{while } S < 1 \quad \text{for } \phi > \text{ob}.
\]

However, this asymmetry close to the observed value for \(\text{ob} > k\) may be adjusted for by imposing \(S = 1\) in the interval \([\text{ob} - k, \text{ob} + k]\) and an associated small adjustment of \(S\) above \(\text{ob} + k\) by computing \(S = \text{Max}\{1 - (\phi - \text{ob} - k) / (A \times \text{ob} - k), 0\}\). This will retain the formulation of a linear reduction of \(S\) to zero at 5\(\times\)ob. The impact of this type of asymmetry may be considered small when using \(k = 0.1\) kg/m\(^2\).

The steps of the verification process for the score component \(\text{SLX}_{\text{ob max}}\) may be summarized as follows, assuming that a decision has been made on a possible partitioning of the verification in sub-domains, as part of the verification set-up:

I. Choose the set of different neighbourhood widths to be used in the experiments.
II. Determine the value (and positions in grid) of the observed maxima \(\text{ob max}(\text{K}_1)\).
III. Compute the forecasted maxima \(\phi_{\text{max}}(\text{L}, \text{K}_1)\) in the neighbourhood(s) of the extreme point(s).
IV. Insert the values \(\text{ob max}(\text{K}_1)\) as value = OB and \(\phi_{\text{max}}(\text{L}, \text{K}_1)\) as a forecast to the score function \(S\) of Figure 2. The values and distance between the two determine the value of the score \(S_{\text{ob max}}(\text{K}_1)\).
V. The procedure includes consideration of multiple extreme points, \(1 \leq \text{K}_1 \leq M_1\) (Equation 4), and the average of all computations is computed as the final value of the score \((\text{SLX}_{\text{ob max}})\) linked to the chosen neighbourhood width.
VI. Repeat (III)-(V) for all chosen neighbourhood widths.

The values of individual scores, taking into account multiple occurrences of extremes, are given by \((4)-(7)\), which describe the averages for each type of score.
The total combined SLX score is defined as an average of the four individual components according to (8).

\[ \text{SLX} = \frac{1}{4} \left( S_{\text{ob,max}} + S_{\text{ob,min}} + S_{\text{fc,max}} + S_{\text{fc,min}} \right) \]  

\[ (8) \]

4 | IDEALIZED EXAMPLES

The properties of the SLX scheme are first illustrated using idealized examples. The response of the scheme is examined for different precipitation structures displaced between analysis and forecast, for example, coherent precipitation structures and single grid points representing extremes. Also, an idealized precipitation band is studied. The forecasted precipitation band is parallel to the analysed one, but displaced by a number of grid points. Finally, the response of the scheme is studied with regard to field perturbations in forecast and analysis. The domain size used in the computations is 200 × 200 grid points.

4.1 | Impact of neighbourhood and scales

We first demonstrate how neighbourhoods and horizontal distribution of extremes determine the score components. The influence of increasing neighbourhood size is studied in the context of situations with coherent groups of extremes or single points separated at some distance. These extreme points may be thought of as mesoscale precipitation areas that are not precisely forecasted with regard to size and location. Single points of maxima separated in space may represent maxima related with (local) showers. A small default tolerance value of \( \delta \) [0 kg/m²] for defining extremes favours selection of single points as maxima. Hence, coherent areas of maxima would normally require larger values of \( \delta \) Regarding minima, the focus is on zero-valued coherent areas, which may have any size without increasing \( \delta \). In Figure 3, the white circular areas represent observed and forecasted areas of zero precipitation, respectively. The yellowish areas represent observed and forecasted maxima, respectively, that are larger than the minimum \( k \) (kg/m²) defined in the score function (formulas 2a–2c and 3a–3b). The grey areas represent intermediate precipitation amounts also

\[ \text{SLX}_{\text{ob,max}} = \frac{1}{M_1} \sum_{K_1=1}^{M_1} S_{\text{ob,max}}(K_1) \]  

\[ (4) \]

\[ S_{\text{ob,max}}(K_1) \] is obtained by inserting for ‘ob’ and ‘\( \Psi \)’ the respective values, \( \text{ob}_{\text{max}}(K_1) \) and \( \Psi_{\text{max}}(L, K_1) \), in Equations (2a)–(2c), (3a) and (3b).

\[ \text{SLX}_{\text{ob,min}} = \frac{1}{M_2} \sum_{K_2=1}^{M_2} S_{\text{ob,min}}(K_2) \]  

\[ (5) \]

\[ S_{\text{ob,min}}(K_2) \] requires respectively for ‘ob’ and ‘\( \Psi \)’ (in Equations 2a–2c, 3a and 3b) the values of \( \text{ob}_{\text{min}}(K_2) \) and \( \Psi_{\text{min}}(L, K_2) \)

\[ \text{SLX}_{\text{fc,max}} = \frac{1}{M_3} \sum_{K_3=1}^{M_3} S_{\text{fc,max}}(K_3) \]  

\[ (6) \]

\[ S_{\text{fc,max}}(K_3) \] requires respectively for ‘\( \Psi \)’ and ‘ob’ (in Equations 2a–2c, 3a and 3b) the values of \( \text{fc}_{\text{max}}(K_3) \) and \( \Psi_{\text{max}}(L, K_3) \)

\[ \text{SLX}_{\text{fc,min}} = \frac{1}{M_4} \sum_{K_4=1}^{M_4} S_{\text{fc,min}}(K_4) \]  

\[ (7) \]

\[ S_{\text{fc,min}}(K_4) \] requires respectively for ‘\( \Psi \)’ and ‘ob’ (in Equations 2a–2c, 3a and 3b) the values of \( \text{fc}_{\text{min}}(K_4) \) and \( \Psi_{\text{min}}(L, K_4) \)

\[ M_1, M_2, M_3 \text{ and } M_4 \] are the number of extreme points identified in each category.

Methods; (i) Kriging with external drift (KED), and (ii) Kriging with radar-based error corrections (KRE), which are described in the open literature (Goudenhoofdt & Delobbe, 2009; Sinclair & Pegram, 2005; Wackernagel, 2003). Further, both parametric and non-parametric variograms, describing the spatial statistical variation of precipitation, have been used (Schiemann et al., 2011; Velasco-Forero et al., 2009). The final product evaluated is the variance weighted mean of these four products (KED and KRE with both parametric and non-parametric variograms).

3 | DATA

The scheme is tested using idealized set-ups and forecasts from the HARMONIE-AROME cloud-resolving model system (Bengtsson et al., 2017) as used operationally at DMI. The verification is performed at a horizontal grid point resolution down to 750 m.

A quantitative precipitation analysis system has been developed at DMI, combining in situ observations with a detailed analysis of data from meteorological radars. The radar products, at a horizontal grid size down to 500 m and accumulation periods varying from 10 min to 24 h, have operational status and are being used routinely by forecasters for monitoring of weather and by NWP model developers for model validation and verification.

The theoretical techniques used to merge rain gauge- and radar data in near real time are based on Kriging methods; (i) Kriging with external drift (KED), and (ii) Kriging with radar-based error corrections (KRE), which are described in the open literature (Goudenhoofdt & Delobbe, 2009; Sinclair & Pegram, 2005; Wackernagel, 2003). Further, both parametric and non-parametric variograms, describing the spatial statistical variation of precipitation, have been used (Schiemann et al., 2011; Velasco-Forero et al., 2009). The final product evaluated is the variance weighted mean of these four products (KED and KRE with both parametric and non-parametric variograms).
Figure 3  Neighbourhood and scales: \( \text{o}_{\text{b, min}} = \text{f}_{\text{c, min}} = 0 \) in white areas. \( \text{o}_{\text{b, max}} \) and \( \text{f}_{\text{c, max}} \) represent areas with maxima observed and forecasted, respectively. \( d_{\text{min}} \) is the distance in grid points between areas with minima. \( d_{\text{max}} \) is the distance between areas with maxima. Grey areas outside the minima and maxima represent intermediate precipitation amounts. \( d_{\text{o, b-min}} \) represents smallest necessary neighbourhood width to make \( \text{SLX}_{\text{o, b-min}} = 1 \). \( d_{\text{f, c-min}} \) represents the smallest necessary neighbourhood width to make \( \text{SLX}_{\text{f, c-min}} = 1 \). For details, see text

Regarding the scores of maxima, \( \text{SLX}_{\text{o, b, max}} \) and \( \text{SLX}_{\text{f, c, max}} \), the values will again be determined by a weighted average of the scores involving the accepted maximum points. A critical neighbourhood width for comparing maxima is then \( d_{\text{max}} \). If \( \text{f}_{\text{c, max}} \) exceeds \( \text{o}_{\text{b, max}} \) by more than a factor of 5, both \( \text{SLX}_{\text{f, c, max}} \) and \( \text{SLX}_{\text{o, b, max}} \) will be zero for comparisons involving sufficiently large neighbourhoods that will include values from both areas of maxima. In all other situations, a non-zero score will be obtained since it is assumed that the values to compare with are above \( k \) (kg/m²), but values may get close to zero, for example, for large observed extremes \( \text{o}_{\text{b, max}} \), if \( \text{f}_{\text{c, max}} \) is small.

As a summary, the set-up of Figure 3 demonstrates that the combined \( \text{SLX} \) score may become zero for \( L < d_{\text{min}} \) for separated dry areas in combination with areas of maxima involving sufficiently different extremes. The width of the neighbourhood sometimes plays a critical role when trying to quantify at which scales the minima and maxima are predicted with reasonably high scores.

4.2  Precipitation band displaced

Precipitation along meteorological fronts is a common feature in the atmosphere and is highly relevant in the context of verification. We may consider the idealized situation where forecasted accumulated precipitation versus analysed precipitation occurs in frontal structures separated by a number of grid points (D). This is illustrated in Figure 4. The idealized precipitation bands (observed and forecasted) are both 10 grid points across, with a gradual increase of precipitation in steps of 5 (kg/m²) and a maximum of 15 (kg/m²) accumulation in the middle. Each step is thus 2 grid points wide. The observed and forecasted precipitation are zero outside the precipitation bands. The main characteristics of this set-up, with respect to resulting values of scores, are the number of grid points separating the precipitation bands, the neighbourhood width chosen and the fraction of grid points in the precipitation bands compared with total number of points in the area verified. Figure 5 shows the \( \text{SLX} \) score for two displacements between the precipitation bands equal to 15 and 25 grid points, respectively. The common decreasing scores below 0.50 at decreasing neighbourhood widths below 6 points occur for the following reason: Some of the forecasted minima (= 0) chosen during the verification process occur under the non-zero precipitation values of the analysed precipitation band. Computing \( \text{SLX}_{\text{f, c, min}} \) related with these points requires a certain neighbourhood width > 0 to contain analysed point values with zero precipitation, needed to
achieve a positive contribution to $\text{SLX}_{\text{fc, min}}$. A similar argument may be used regarding zero analysed precipitation points under the non-zero forecasted precipitation band. Increasing the areal fraction of the precipitation bands compared with total domain size would decrease the score further due to this argument.

When computing the scores of $\text{SLX}_{\text{ob, max}}$ and $\text{SLX}_{\text{fc, max}}$ involving points of $15 \text{ kg/m}^2$, the distance between the precipitation bands determines the scores, which become equal to 1 when the neighbourhood width reaches the distance between the precipitation bands, as indicated in Figure 5. The fact that forecasted and observed minima agree over large parts of the domain implies that the combined $\text{SLX}$ score is close to 0.50 for neighbourhood sizes less than the separation distance between the precipitation bands. All scores will be 1 if the two precipitation bands are not separated. Then all values will agree on the grid scale.

### 4.3 | Field perturbations

Figure 6 illustrates situations when the range of forecast values gets increasingly separated from the range of analysed values. In this situation, the definition of the score function implies that $\text{SLX}$ will decrease.

Figures 7 and 8 illustrate this in a specific set-up of perturbed forecasts related to an analysis field. Figure 7a, b illustrate observed and forecasted fields, respectively, with perturbed $10 \times 10$ squares of magnitude $c = 5$ displaced differently between analysis and forecast. The ‘background’ values of observed- and forecasted fields are $b$ and $d$, respectively. The results of $\text{SLX}$ obtained with the forecast background values increasing from 5 to 40 are shown in Figure 8. When $b = d = 5$, the extreme areas of analysis and forecast are just displaced and the score will increase for sufficiently large neighbourhoods (dashed blue line). However, as expected, in these experimental conditions, the results show that the score will decrease at high forecast background values towards a low value. This is consistent with Figure 6.

Next, the impact of perturbing the background field $b$ with the same amount $c$ is shown, regarding both analysis and forecast (Figure 9a,b). Figure 10 shows the results applicable to different background fields up to 40. It is seen that improved scores are obtained as the general
background level is increased while the perturbation ±c is fixed. This result is a consequence of the score function since the ratio between forecast and analysis gets closer to 1. In view of the limited distance between the squares with perturbed values, it is seen that increased size of neighbourhood leads to higher scores.

5 | REAL-WORLD EXAMPLES

The purpose of the SLX score is the verification of operational NWP, with output on a daily basis. To demonstrate what can be expected in an operational context, results are presented from real-world weather situations. The first is a summer convection event where forecasts from different origin times are compared at the same valid time. At first sight, these different forecasts are quite accurate, but the SLX verification is able to distinguish quality of the individual forecasts. For the second case, some forecasts in December 2020 are examined, as part of a pre-operational set-up. The period studied includes consecutive days linked to changing weather conditions, as a high pressure ridge affecting Denmark is replaced by frontal rain. Finally, in order to simulate the characteristics of some longer time statistics, forecast- and observation data for a 5-year period, valid for summer conditions, are used as input to simulate properties of the SLX scheme under different assumptions, for example, forecasts simulating operational conditions and forecasts based on estimated ‘climatological’ precipitation frequencies of five precipitation categories. Also forecasts with bias are examined, and forecasts based on a uniform and random value applied to the entire grid.

5.1 | A case of summer convection

The case studied is from 27 August 2019 with convection developing during the day over Denmark in a warm air mass being convectively unstable. The analysis system and operational forecast model are used at a grid size of 2.5 km (see Section 3). The results of the forecasts are shown in Figure 11a–d. These forecasts of accumulated precipitation are all valid for 3 h, from 9 UTC to 12 UTC 27 August. The initial time of the individual forecasts are shown in the figures representing forecast lengths of 3, 6,
9 and 12 h respectively. The corresponding analysis is shown in Figure 11e. The black crosses indicate the location of the maximum accumulation values of the domain excluding the dark area outside the frames. The individual SLX components are shown in Figure 12a–d. The combined score is shown in Figure 12e (see figure text). The best scores (SLX and local maximum components) are obtained for the 6-h forecast. This agrees well with a visual inspection of Figure 11a–d showing that this forecast has a rather accurate position and magnitude of the precipitation maximum. Lower scores for the 3-h forecast are sometimes seen in operational forecasts and may be linked to model spin-up. This concept is mentioned by Ulmer and Balss (2016). Significantly lower scores are seen for the earliest forecast started at 00 UTC 27 August, especially when considering forecasting of extremes on the grid scale. A quite general feature is that scores become higher at a neighbourhood size of 5–10 grid points. This interval represents the smallest scales that the model can resolve. Another feature is that scores tend to be higher when considering scores related with the minima as compared with the maxima. This could be interpreted as larger areas with minima, for example, with zero-valued precipitation, occurring in both forecast and analysis.

5.2 | Moving weather systems

Changing weather related to moving weather systems across Europe play an important role in daily verification context. The SLX scores are illustrated for weather conditions occurring over Denmark from 24 December to 28 December 2020. On 24 December, a low pressure system and its associated precipitation affect the southernmost Denmark. This system moves towards the east and is followed by rising pressure and a flow of dry air from the north, with only few scattered areas of precipitation. 25 December is dominated by dry weather, but precipitation from a frontal system reaches Denmark on the following day. This is linked to a deep cyclone (central pressure < 960 hPa) developing north of Scotland. Precipitation covering most of Denmark develops during the afternoon of 26 December and continues over 27–28 December.

Figure 13a illustrates precipitation over Denmark, occurring early in the almost dry period. The 3-h accumulation is shown for 04–07 UTC on 24 December. It represents precipitation on 24 December early in the day. Later, only small areas of scattered precipitation occur.
Figure 13b shows accumulation from 16–19 UTC on 26 December 2020. It represents frontal precipitation over Denmark dominating from the afternoon of 26 December and on the following 2 days.

The different precipitation characteristics for this changing weather pattern are reflected in the SLX scores of Figure 14. The model area verified is covering Denmark in a latitude band of 54.5°N–57.8°N, and a longitude band of 7.5°E–12.7°E. The operational model HARMONIE-AROME, see Section 3, is tested in a pre-operational verification domain using a grid size of 750 m. For simplicity, the results of Figure 14 are for one neighbourhood size only ($L = 6$ points). The verification is split into nine equally sized sub-domains with separate SLX computations that are averaged. Figure 14a represents forecast verification of the mainly dry period while 14B applies to the period with frontal type of precipitation.
All operational runs are verified up to a forecast range of 48 h. They are executed every 3 h. The verification is based on 3-h accumulated precipitation. Figure 14a shows the averaged verification of all individual forecasts executed on 24 December. The latest forecast up to 48 h ends on 26 December 21 UTC. Similarly, Figure 14b shows the averaged verification of all 48-h forecasts executed on 26 December.

It turns out that the dry conditions are well forecasted. As a consequence, the scores related with minima are very close to 1.0 in Figure 14a. However, long forecasts starting late on 24 December do not predict...
well the observed precipitation increase on 26 December. This is clearly reflected in the scores of maxima, in particular the score of forecast maxima reaching values below 0.50.

For the wet period, all scores are much more alike (Figure 14b). This may be due to the dominating stratiform precipitation, which is more uniformly distributed horizontally than convective precipitation. Therefore, the scores related with minima are similar to the scores of maxima. The typical levels of SLX are smaller than the corresponding ones of the earlier dryer period. This may be explained by the difficulties in forecasting horizontal variations of non-zero precipitation. Towards the end of the second period, in the evening of 28 December, the scores tend to increase. This is related to more dry conditions starting to appear (not shown).

5.3 Simulating operational conditions

It is a goal to carry out detailed tests of the SLX scheme using NWP data for Denmark. However, there is not yet a multi-year high-quality observed precipitation analysis data set available. To simulate how the SLX scheme would operate in an operational context, a simple method has been designed and tested. Archived NWP data and rain gauge data are used for this. The NWP model is the HARMONIE-AROME model (Bengtsson et al., 2017) operated at a grid size of 2.5 km. The data used to generate synthetic tests are from a summer month (June) covering 5 years from 2014 to 2018. The simulated verification results will only represent average performance of the model over the 5-year period since the model version has changed somewhat over the 5-year period. The forecasts and observations consist of data partitioned into five precipitation intervals (categories) and are provided as a contingency table, see Table 1. These data are valid for 12-h periods. The observations are based on 45 Danish quality controlled stations. The forecast data apply to 6–18 h forecasts. The data of Table 1 comprising $5 \times 5 = 25$ precipitation classes represent 6545 data values. The class of dry or almost dry conditions ($\leq 0.2$ kg/m$^2$) dominates (4514 values).

We assume that the observation data can be used to approximate observed precipitation in a grid ($2.5 \times 2.5$ km) to be compared with the NWP model data produced and stored at this resolution. A feature of R programming is used for selecting an arbitrary number between 1 and the total number of 6545. The selected number is converted into a specific precipitation class out of the 25, utilizing that each class is associated with an interval directly proportional to the relative frequency out of the 6545 values. Having chosen one of the 25 classes, specific values of forecast and observation are selected. A precipitation interval linked to the precipitation class is divided into 100 equidistant sub-intervals of equal probability. A random number selected between 1 and 100 is then used to choose a specific sub-interval and its corresponding precipitation value. Separate random numbers are selected for observation and forecast to fix the precise values used in the precipitation class. The forecasted and observed precipitation fields of the model domain are simulated based on this. The upper bound of the highest class is set to 40 (kg/m$^2$). This value is estimated from regional extreme precipitation analysis over Denmark (Madsen et al., 2017), indicating that 12-h
accumulations of 40 (kg/m²) on a regional scale occur on a time scale of ~5 years.

However, it is not realistic to execute a procedure on the grid scale for all grid points independently since it will be almost impossible to generate quasi-constant large-scale precipitation accumulations, which occur in reality for domain sizes considered. Therefore, an approximation is made by selecting a scale to compose the total field, choosing among squares of different sizes by random selection, for example, $5 \times 5$ squares of grid points.
or 10 × 10 squares, and so forth up to a maximum size representing the entire domain. Constant values of forecast and analysis respectively are used inside each square. In this way, a spectrum of possible sizes is considered, from the smallest resolvable scale (approximately 5 × 5 grid points) up to a scale representing the entire domain. As a result, the relative occurrence of a precipitation class as well as different horizontal scales is taken into account. Figure 15 presents the results generated with this approach. In this case, a small model domain of 100 × 100 grid points has been chosen, with horizontal grid boxes of 5 × 5, 10 × 10, 20 × 20, 50 × 50 or 100 × 100, selected as random options. A total of 720 grids of model- and observed states respectively have been generated to do an experiment. This would, for example, correspond to verification of each hour over a month (24 × 30 = 720) in an imagined operational situation.

Figure 15 shows a comparison between SLX scores of different simulations as a function of neighbourhood width. Simulated operational forecast data are represented by ‘fc-stat’, based on the 25 classes of Table 1. This is compared with forecast verification ‘fc-clim’ based on ‘climate’, which is not a constant value. It is based on the observed ‘climatic’ frequency of the individual precipitation categories. A modified contingency table has been constructed for this, using a forecast probability of each category based on the observed relative frequency of that category derived from the 5-year period. Also, the impact of adding a constant value (0.5 kg/m²) to a standard forecast has been studied (‘fc-bias-min’). Finally, a uniform random value forecast is verified (‘fc-unif-random’). The field is generated by assuming equally likely choices in the possible interval of [0, 40] (kg/m²).

As expected, Figure 15 shows that the forecast model clearly produces superior scores compared with other alternatives. ‘fc-clim’ produces a reasonably high forecast score. It benefits from a high statistical weight of the smallest quasi-dry precipitation class. The combined SLX score studied here is significantly affected by adding a constant bias of 0.5 kg/m² to all forecasts. The current choice of score function penalizes biases when dry conditions are observed. The constant-bias score is much lower than the ‘climatic’ forecast. Since all observed dry conditions will be over-forecasted by at least 0.5 kg/m², these contribute with zero values to the total score. This result is reasonable if one accepts the definition of the score function for dry observed conditions, penalizing significantly small precipitation amounts. It emphasizes that a full understanding of the SLX verification requires that attention is paid to individual components related with minima and maxima. However, the averaged score provides a general overview of the quality. This subject is discussed further in Section 6. As expected, ‘fc-unif-random’ produces a very poor score. It is worthwhile noting that this type of forecast is associated with a large bias.

6 | DISCUSSION AND FUTURE PROSPECTS

In this section, several aspects of the SLX verification are addressed: the definition of an adequate score function, the impact of defining tolerance for selecting extreme points, use of the SLX scheme operationally, and finally some future development prospects.

6.1 | Score function

Verifying precipitation using a score function is challenging for several reasons. First, the fact that precipitation is a quantity with a lower bound = 0, while no limit applies upward, creates a special situation that is not favourable for defining symmetry of the score function. For another parameter such as temperature, it would be a
straightforward possibility to choose a symmetric score function centred around the observation. The score function used in the present study is very simple, with only three linear parts. A natural goal for an improved formulation could be to make a more advanced function allowing flexibility, with the possibility of several parameter settings in accordance with demands from users who might want a dependency on particular thresholds. Symmetry around observation in some interval would be a natural choice for non-dry conditions.

6.2 Tolerance for selecting extreme points

Choosing tolerance $\delta = 0$ (kg/m$^2$) for defining an analysed extreme means that the uncertainty is not taken into account. However, this uncertainty $\delta_{ob}$ (kg/m$^2$) might be included in some way. It implies that points deviating up to $\delta_{ob}$ (kg/m$^2$) from the analysed extreme may be selected for the score computation.

One may try to consider typical observation uncertainty related with rain gauges. According to Yang et al. (1998), such uncertainty varies substantially, for example, depending on meteorological conditions, and may amount to at least 20%. In the case of dry weather, the observation uncertainty is in practice zero. The currently used analysis should be considered an advanced combination of radar data and data from rain gauges, with quality control. Hence, the resulting uncertainty of the analysis field is difficult to estimate and is beyond the scope of the present paper.

Regarding forecasts, a concept of non-zero tolerance $\delta_{fc}$ (kg/m$^2$) implies that more extreme forecast points tend to be selected to form coherent groups. For maxima, this means that points in an interval of $\delta_{fc}$ lower than the extreme maximum value will be selected as points to be verified. For minima, larger values with deviations up to a magnitude of $\delta_{fc}$ will be used. It will impact on the scores of SLX$_{fc_{\text{max}}}$ and SLX$_{fc_{\text{min}}}$. This might be desirable, but it implies that a strict definition of an extreme value is relaxed.

Specifying both $\delta_{ob}$ (kg/m$^2$) and $\delta_{fc}$ (kg/m$^2$) as small fractions, for example, 5%–10% of the extreme value observed and forecasted, respectively, might be an attractive way of allowing non-zero tolerance of extremes. It tends to keep sensitivity of score computations at a low level since small fractions were involved. On the contrary, a high sensitivity related with minima could result if a constant and relatively high tolerance of $\delta$ were chosen. For maxima, for example, 10% of a large value increases the likelihood that coherent groups of points would be realized.

6.3 SLX in an operational context

Prediction of accurate precipitation extremes is frequently requested by, for example, duty forecasters in an operational context. The SLX scores contribute in this respect with complementary information not available from, for example, FSS and SAL verification schemes mentioned in the introduction.

Forecasters normally represent the last step in a forecast chain. Sometimes it is verified if the forecaster improves NWP forecasts in this last step, when measuring quality by a verification score. It is then relevant to ask if the SLX score(s) are easy to improve by a forecaster, for example, by ‘hedging’, which implies some sort of degradation of the original forecast. Considering the combined score, it will be difficult to hedge for the following reasons:

First, regarding minima, a forecast bias tends to reduce the scores, as shown in Section 5.3, investigating the impact of bias in the case of small precipitation amounts. Therefore, forecasters need to avoid biases of small precipitation amounts as much as possible.

Second, regarding maxima, it is relevant to mention another feature of hedging: According to Murphy (1993, 1978), hedged forecast fields are smooth, and they lack sharpness. Hence, a horizontal smoothing of the predicted precipitation field is an act of hedging. This would be attractive to forecasters in order to avoid large errors near the observed maxima, for example, measured in an absolute sense by mean absolute error or SLX. However, a horizontal smoothing of high forecast values might lead to incorrect non-zero forecasts related with minima, for example, overlapping the areas with observed minima equal to zero. As a consequence, the associated score SLX$_{\text{min}}$ would be reduced. In general, achieving high scores related with maxima is challenging, as this requires predicting both location and amount with sufficient accuracy. Introducing a significant positive bias tends to provide a lower score, as shown in the field perturbation experiment of Section 4.

When operating SLX on a daily basis, one may produce not only output of the score components but also the associated high and low values of the extremes. This provides an opportunity for computing verification results as a function of specific thresholds, for example, observed or forecasted maxima above high thresholds. In this way, SLX may be used for studying model behaviour for high extreme precipitation values selectively, as a post-processing step. Other options for post-processing are relevant, for example, to compute SLX and its separate components as distributions, showing the number of score values occurring...
in separate equidistant intervals between 0 and 1. When accumulating results from daily runs, statistics may be carried out for longer periods, for example, months or years.

For large model domains, division into sub-domains is recommended. This could imply that large limited area NWP models of dimensions, say 1000 or 2000 grid points in each direction, would be divided into, for example, ~100 sub-domains. Output from sub-domains could be averaged or stored separately. The latter provides an opportunity to diagnose problematic model output in specific sub-domains.

The examples of Sections 4 and 5 illustrate the important role of neighbourhood width for computations inside a given model domain or sub-domain, showing at which scales various local extremes are captured. It is recommended to compute results for a number of neighbourhood widths from grid scale to a maximum width determined by a prioritized scale for predicting extremes, for example, 50 km.

In operational verification, it is often relevant to investigate how different models perform, for example, as a function of model resolution. From the users' point of view, the skill of a forecast should be considered on an absolute spatial scale, for example, squares of 50 × 50 km. Then, the number of grid points for this scale will depend on the model resolution. The high-resolution model will then have more grid points in such squares and should be able to benefit from that in a comparison between the two models. Alternatively, the coarse mesh model could be interpolated to the fine mesh grid. It is assumed that the analysed precipitation is to be applied at a resolution of the high-resolution model. Otherwise possible details predicted at high resolution would not be rewarded.

The SLX concept can, in principle, be applied to a global model system. The maxima and minima could then be computed for any selection of neighbourhood sizes without considering the influence of lateral boundaries. A challenge would be to produce a sufficiently accurate precipitation analysis.

The current implementation is coded in R (https://www.R-project.org/). For modest domain sizes with dimensions of few hundred grid points in each direction, the execution time on, for example, a single CPU, GNU 4.8.4 compiler, Intel processor COREi5 1.70 GHz is within few minutes. Code optimizations may be useful in the case of large model domains that are not tested in this investigation. In the case of many extreme points, the current code will run faster if the computations are based only on a subset of extreme points. However, such subset should be sufficiently large, in order not to affect the result significantly.

6.4 | Future development prospects

The potential of the scheme to be used in the context of ensemble prediction systems of limited area models, for example, the study by Froger et al. (2019), is obvious. One may, for example, compute SLX in the context of percentiles as an alternative to the median. In the future, the convection resolving model and the precipitation analysis may be run at sub-km grid sizes, for example, down to 500 m. The general strategy of the international collaboration developing this model system (https://hirlam.org) is to develop ensemble systems down to sub-km grid size with frequent updates and short forecast lengths, for example, shorter than 24 h. This would mean using SLX at short scales in space and time, with precipitation accumulation periods down to 1 h or less in the future.

If model developers are to compare experiments for the same atmospheric case, it will be important to agree on common settings for the comparisons. First of all the score function (Figure 2) should be the same. In addition, the size and location of model domain and accumulation periods should be agreed on.

The general concept of verifying local maxima and minima implies that the scheme could also be adapted to other variables than precipitation. For each parameter to be considered, an appropriate score function needs to be defined, and an appropriate analysis needs to be available for comparison with the forecast.

7 | CONCLUSIONS

The evolution of high-resolution numerical weather prediction (NWP) in recent decades has led to much more detailed forecasts in space and time. However, there is a need to verify these more detailed forecasts, both in order to improve models further and in order to communicate forecast skill to the NWP users. Precipitation is a challenging parameter to verify due to its very fine scale structures. A new generation of high-resolution field verification of precipitation has been developed to meet the requirements. This verification is dependent on a detailed precipitation analysis. Based on experiences with the precipitation analysis developed in Danish Meteorological Institute (DMI), it is concluded that quality control of observational data is a very important part of the system in order to secure quality.

This paper documents an example of a novel field verification scheme: The Structure of Local EXtremes (SLX) scheme. It includes verification of both local maxima and minima of observed and forecasted precipitation and has properties of both features-based and neighbourhood-
based verification schemes. The properties of the scheme have been illustrated for both idealized precipitation cases, and for real-world forecast examples. The results indicate that the scheme is potentially very useful for investigating model weaknesses in individual cases and more generally for routine verification.

The scheme provides flexibility since the verification output can be produced for separate sub-domains. Moreover, linked with the four separate scores of the SLX scheme, a variety of diagnostic output can be presented for different aspects of forecast quality. Statistics for high extreme thresholds may be generated as part of a post-processing step. The scheme will produce results for any analysis- and forecast field including zero-valued fields. The fact that the scheme verifies local extremes fits well with an increasing focus on using ensemble prediction systems in the NWP community. The score function can be adapted to special needs of the user community. SLX complements well other spatial verification schemes such as FSS and SAL.

ACKNOWLEDGEMENTS

The author would like to thank Kasper Hintz and Henrik Feddersen (DMI) for useful discussions and for collaboration on the real-world cases included. Rashpal Gill (DMI) is acknowledged for supplying detailed information on the concepts of the precipitation analysis developed in DMI. Finally, Michael Butts (DMI) is acknowledged for contributing with editorial improvements of the manuscript.

AUTHOR CONTRIBUTIONS

Bent Sass: Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); software (lead); writing – original draft (lead); writing – review and editing (lead).

ORCID

Bent H. Sass https://orcid.org/0000-0002-9724-2529

REFERENCES

Bengtsson, L., Andrae, U., Aspelien, T., Batrak, Y., Calvo, J., de Rooy, W. et al. (2017) The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. Monthly Weather Review, 145, 1919–1935. https://doi.org/10.1175/MWR-D-16-0417.1

Bougeault, P. (2003) WMO WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. Appendix C of report of the 18th session of the CAS/JSC WGNE. Available at: http://www.wmo.ch/web/wcrp/wgnepublications.htm [Accessed July 14, 2021].

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M.P., Ebert, E. E., Brown, B.G. et al. (2018) The setup of the MesoVICT project. Bulletin of the American Meteorological Society, 99, 1887–1906. https://doi.org/10.1175/BAMS-D-17-0164.1

Ebert, E.E. (2008) Fuzzy verification of gridded high resolution forecasts: a review and proposed framework. Meteorological Applications, 15, 51–64.

Ebert, E.E. (2009) Neighborhood verification: a strategy for rewarding close forecasts. Weather and Forecasting, 24, 1498–1510.

Eggert, B., Berg, P., Haerter, J.O., Jacob, D. & Moseley, C. (2015) Temporal and spatial scaling impacts on extreme precipitation. Atmospheric Chemistry and Physics, 15, 5957–5971.

Fiener, P. & Austerwald, K. (2009) Spatial variability of rainfall on a sub-kilometre scale. Earth Surface Processes and Landforms, 34, 848–859.

Frogner, I., Andrae, U., Bojarova, J., Callado, A., Escribá, P., Feddersen, H. et al. (2019) HarmonEPS – the HARMONIE ensemble prediction system. Weather and Forecasting, 34, 1909–1937. https://doi.org/10.1175/WAF-D-19-0030.1

Gilleland, E., Ahijevych, D.A., Brown, B.G. & Ebert, E.E. (2010) Verifying forecasts spatially. Bulletin of the American Meteorological Society, 91, 1365–1375.

Gilleland, E., Skok, G., Brown, G.B., Casati, B., Dorninger, M., Mittermaier, M.P. et al. (2020) A novel set of geometric verification test fields with application to distance measures. Monthly Weather Review, 148, 1653–1673.

Goudenhoofdt, E. & Delobbe, L. (2009) Evaluation of radar-gauge merging methods for quantitative precipitation estimates. Hydrology and Earth System Sciences, 13, 195–203.

Harris, D., Foufoula-Georgiou, E., Drogemeier, K.K. & Levit, J.J. (2001) Multiscale statistical properties of a high resolution precipitation forecast. Journal of Hydrometeorology, 2, 406–418.

Hintz, K.S., O’Boyle, K., Dance, S.L., al-Ali, S., Ansper, I., Lovejoy, S. & Mandelbrot, B.B. (1985) Fractal properties of rain and a fractal model. Tellus, 37A, 209–232.

Madsen, H., Gregersen, I.B., Rosbjerg, D. & Arnbjerg-Nielsen, K. (2017) Regional frequency of short duration rainfall extremes using gridded daily rainfall data as a co-variate. Water Science and Technology, 75, 1971–1981. https://doi.org/10.1016/j.wst.2017.089

Murphy, A.H. (1978) Hedging and the mode of expression of weather forecasts. Bulletin of the American Meteorological Society, 59, 71–373.

Murphy, A.H. (1993) What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and Forecasting, 8, 281–293.

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

Sass, B.H. & Yang, X. (2012) A verification score for high resolution NWP: idealized and preoperational tests. HIRLAM Technical Report Number: 69, pp. 28 (December 2012). Available at: http://www.hirlam.org

Schiemann, R., Erdin, R., Willi, M., Frei, C., Berenguer, M. & Sempere-Torres, D. (2011) Geostatistical radar-rain gauge combination with non-parametric correlograms: methodological
considerations and application in Switzerland. *Hydrology and Earth System Sciences*, 15, 22.

Sinclair, S. & Pegram, G. (2005) Combining radar and rain gauge rainfall estimates using conditional merging. *Atmospheric Science Letters*, 6, 19–22.

Ulmer, F.-G. & Balss, U. (2016) Spin-up time research on the weather research and forecasting model for atmospheric delay mitigations of electromagnetic waves. *Journal of Applied Remote Sensing*, 10(1), 016027. https://doi.org/10.1117/1.JRS.10.016027

Velasco-Forero, C.A., Sempere-Torres, D., Cassiraga, E. & Gomez-Hernandez, J.J. (2009) A non-parametric automatic blending methodology to estimates of rainfall fields from rain gauge and radar data. *Advances in Water Resources*, 32, 17.

Wackernagel, H. (2003) *Multivariate geostatistics, an introduction with applications*, 3rd edition. Heidelberg: Springer.

Wernli, H. & Paulat, M. (2008) SAL – a novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, 136, 4470–4486.

Yang, D., Goodison, B.E. & Metcalfe, J.R. (1998) Accuracy of “NWS8” standard nonrecording precipitation gorges, results and application of WMO intercomparison. *Journal of Atmospheric and Oceanic Technology*, 15, 54–68.

Yano, J.-I., Ziemiański, M.Z., Cullen, M., Termonia, P., Onvlee, J., Bengtsson, L. et al. (2018) Scientific challenges of convective-scale numerical weather prediction. *Bulletin of the American Meteorological Society*, 99(4), 699–710. https://doi.org/10.1175/BAMS-D-17-0125.1

**How to cite this article:** Sass, B. H. (2021). A scheme for verifying the spatial structure of extremes in numerical weather prediction: Exemplified for precipitation. *Meteorological Applications*, 28(4), e2015. https://doi.org/10.1002/met.2015