Improving Observations of Precipitation Type at the Surface: A 5-Year Verification of a Radar-Derived Product from the United Kingdom's Met Office

BEN S. PICKERING,[a] STEVEN BEST,[b] DAVID DUFTON,[c] MARYNA LUKACH,[c] DARREN LYTH,[b] AND RYAN R. NEELY III[c]

[a] *Institute for Climate and Atmospheric Science, Leeds, United Kingdom*
[b] *Met Office, Exeter, United Kingdom*
[c] *National Centre for Atmospheric Science, Leeds, United Kingdom*

(Manuscript received 21 June 2020, in final form 25 February 2021)

ABSTRACT: This study aims to verify the skill of a radar-based surface precipitation type (SPT) product with observations on the ground. Social and economic impacts can occur from SPT because it is not well forecast or observed. Observations from the Met Office's weather radar network are combined with postprocessed numerical weather prediction (NWP) freezing-level heights in a Boolean logic algorithm to create a 1-km resolution Cartesian-gridded map of SPT. Here 5 years of discrete nonprobabilistic outputs of rain, mixed-phase, and snow are compared against surface observations made by trained observers, automatic weather stations, and laser disdrometers. The novel skill verification method developed as part of this study employs several tolerances of space and time from the SPT product, indicating the precision of the product for a desired accuracy. In general the results indicate that the tolerance verification method works well and produces reasonable statistical score ranges grounded in physical constraints. Using this method, we find that the mixed precipitation class is the least well diagnosed, which is due to a negative bias in the input temperature height field, resulting in rain events frequently being classified as mixed. Snow is captured well by the product, which is entirely reliant upon a postprocessed NWP temperature field, although a single period of anomalously cold temperatures positively skewed snow scores with low-skill events. Furthermore, we conclude that more verification consistency is needed among studies to help identify successful approaches and thus improve SPT forecasts.

KEYWORDS: Snow; Precipitation; Mixed precipitation; In situ atmospheric observations; Radars/Radar observations; Surface observations

1. Introduction

The type of hydrometeors reaching the surface, known as the surface precipitation type (SPT), can severely impact human activities. In regions where solid precipitation types are common and expected occurrences, long-term adaptations are cost effective, but where solid precipitation types are infrequent and uncommon (midlatitudinal, certain mountainous regions) these adaptations are not cost effective and (as in the case of the United Kingdom) events can significantly disrupt daily life (Kay 2016; Curtis et al. 2017). In the winter of 2009/10, the cost to the U.K. National Health Service from falls on snow and surface ice was £42 million (Beynon et al. 2011). Mitigative actions such as clearing roads, covering exposed crops, and redirecting aircraft are cost associated and require sufficient lead time and confidence (Cornford and Thornes 1996; Rasmussen et al. 2001; Handa et al. 2006; Clark et al. 2009).

Real-time observations are often used by forecasters directly or in nowcasting systems to issue precipitation type guidance, valid for time scales of 0–6 h (Rasmussen et al. 2001; Schmid and Mathis 2004; Haiden et al. 2011). SPT is accurately reported by trained observers but their observations are infrequent, whereas automated ground instruments record continuously but with less accuracy (Bloemink 2005; Landolt et al. 2019). The Met Office operates a network of both station types across the United Kingdom, but these do not provide complete spatial coverage at a high enough temporal resolution sufficient for animated, gridded map products that are essential for SPT nowcasting and public understanding. An ideal measurement system for SPT nowcasting is weather radar because it possesses a high spatiotemporal resolution. Additionally, the U.K. weather radar network has (at most) a 10-min turnaround from measurement to dissemination (Harrison et al. 2000) so it is useful for real-time decision-makers.

This study aims to assess the skill of a U.K. radar-derived SPT product over a 5-yr period. Since the product is deterministic and precipitation type is discrete nonprobabilistic data, there are a limited number of statistical techniques suitable for performing verification. Furthermore, snow and mixed-phase precipitation are an order of magnitude less

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

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-20-0172.s1.

Corresponding author: Ben S. Pickering, b.s.pickering@leeds.ac.uk

DOI: 10.1175/JHM-D-20-0172.1

frequent than rain (Kay 2016; Brown 2019). This discrepancy in the abundance of the classes can deceptively skew some statistical scores (Wilks 2011), further reducing the number of applicable verification techniques.

Added difficulty is introduced with the comparison of a radar-derived spatial product with point surface instruments, since the representative volumes differ. Weather radars measure distribution-weighted three-dimensional volumes of the atmosphere. The verification ''truth'' on the ground (often many hundreds of meters below the peak-weighting of the radar voxel) is a pinpoint measurement, typically a fraction of a cubic meter for automated instruments. Human reporters are capable of broader visual assessment of the precipitation type, but their observation volume is still an order of magnitude less than weather radars.

In this study, a new approach is applied to determine the skill range of radar-based surface precipitation type products against several surface observation datasets, by varying the temporal and spatial tolerance of the product. The verification techniques developed here are further useful for assessment of NWP forecasts of precipitation type (or any discrete nonprobabilistic variable) and thus facilitate more accurate diagnoses of precipitation type in atmospheric science. The ability for the Met Office SPT product to diagnose rain, mixed-phase, and snow precipitation types is examined here. Weaknesses and opportunities for improvement of the radar-based SPT product are also presented. Hail is the fourth SPT class in this product, which uses a separate criterion for diagnosis. However, hail cannot be rigorously verified due to the lack of a reference dataset, primarily due to the rarity of hail in the United Kingdom (Punge and Kunz 2016; Webb et al. 2009). For example, the European Severe Weather Database (ESWD; Dotzek et al. 2009) contains only 32 hail reports in the United Kingdom during the 5-yr study period under examination. The hail class is therefore neglected in this study.

The boundary between rain, mixed-phase, and snow (R–M–S) is important because the presence of mixed-phase precipitation typically indicates that the hydrometeors are melting before they reach the ground and will therefore not accumulate. This is important for several industries—if wet precipitation meets a cold surface (or if it occurs with diurnal cooling), then ice is the primary risk. If the surface is warm (or if it is associated with diurnal heating) then the runoff water will drain away into rivers and lakes, potentially contributing to flood events.

The R–M–S boundaries in the United Kingdom (and similar geographies) are difficult to diagnose and forecast. Cases are often borderline since surface temperatures are nonextreme and fluctuate diurnally between −5° and +10°C in winter (Parker et al. 1992; Brabson and Palutikof 2002), and many factors can influence the change of precipitation phase. The influence of the Northern Hemisphere midlatitude jet stream and the enclosure of the North Atlantic warmed by the Gulf Stream create fluctuating synoptic patterns and coastal micrometeorology. Small changes in the vertical temperature structure of the atmosphere can also shift the R–M–S boundary by hundreds of kilometers horizontally.

a. Met Office SPT product

To overcome the disparity between the radar-observed voxel and the surface precipitation type diagnosis, the Met Office created an SPT product which uses NWP output as input to a parameterized translational process below the lowest-usable radar beam. Since late 2013 the SPT product has been operational with the same spatiotemporal resolution as the Met Office precipitation rate product (1 km^2, 5-min frequency). Figure 1 shows an example of the product at a single point in time. The product has four classes: hail (not examined here due to lack of a suitable reference dataset), snow, mixed-phase, and rain. Note that the term ''mixed-phase'' refers to the mixture of snow and rain and does not include partially melted graupel or hail. These types are determined with a Boolean logic decision tree described in Table 1. The algorithm inputs are radar-derived surface precipitation rate (Harrison et al. 2000), 0°C wet-bulb isotherm altitude (above local surface), and radar reflectivity. The isotherm height is derived from the U.K. postprocessed (UKPP) dataset which uses the Met Office Unified Model run in a Euro4 configuration.

Lumb's critical rate is used for the mixed-phase diagnosis and is defined as

$$R_c = 0.2909 \exp\left[0.004\frac{\mathrm{FZL}}{f(\upsilon)}\right], \tag{1}$$

where R_c is the critical rate (mm h^{-1}), FZL is the 0°C wet-bulb isotherm height above the local surface in meters, and $f(\upsilon)$ is a function of wind speed but is set equal to 1 in the Met Office implementation and is therefore neglected. The notion is that for a given 0°C wet-bulb isotherm height, precipitation will be observed at the ground as still containing a proportion of solid hydrometeors if the critical rate is met, due to evaporative cooling (Lumb 1963).

This process is applied initially to each pixel from all 18 radars (15 Met Office, 2 Met Éireann, and 1 Channel Islands Meteorological Department). All data are then composited onto a Cartesian 1-km^2 grid using the modal value of all contributing pixels since a single location in the United Kingdom is typically observed by many radar sites simultaneously.

b. Verification data

Data which are used to verify the performance of the SPT product are described here. The known capabilities and limitations of the ground instruments are critical to aid the discussion of the results. Table 2 summarizes each dataset and Fig. 2 shows the locations of all surface stations as well as the locations of all radar sites which contribute to the SPT product.

1) Automatic SYNOP

The Met Office operates a network of surface weather stations called SYNOP stations which report observations for the 10-min period leading up to every hour. At the automatic stations, precipitation type is reported using the World Meteorological Organization (WMO) ''present weather'' (PW) code from Table 4680 (WMO 1988, 2019).

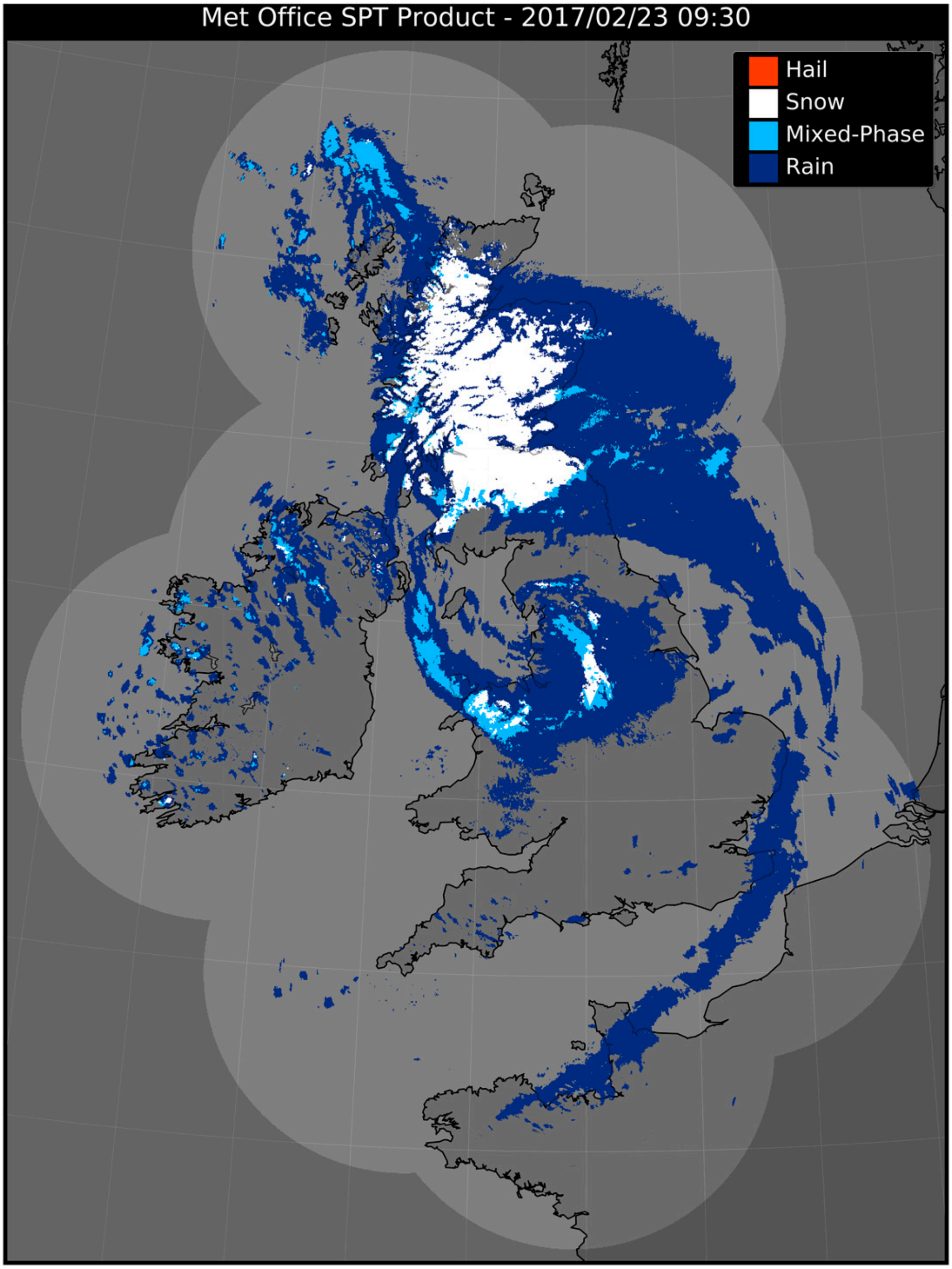


FIG. 1. An example of the Met Office SPT product, during named winter storm Doris at 0930 UTC 23 Feb 2017. An animated video of the whole day is supplied in the supplemental material.

The PW code is determined using an arbiter which combines multiple measurements: a Vaisala FD12P present weather sensor, a precipitation detector, a visiometer, a ceilometer, and an air temperature thermometer (Green 2010). Known limitations of the arbiter are insensitivities to weak precipitation rates, poor detection of ''sleet'' (U.K. nomenclature for mixed precipitation), no quantitative uncertainty, and difficulties calibrating or tracing errors since the arbiter ''has many assumptions'' (Lyth and Molyneux 2006; Lyth 2008). A total of 172 automatic SYNOP station locations were available for inclusion during this study.

TABLE 1. The Boolean logic algorithm steps used for the Met Office surface precipitation type product. Note that the term ''mixed-phase'' refers to the mixture of snow and rain and does not include partially melted graupel or hail.

Precipitation type	Criterion
Hail	If a radar reflectivity of ≥45 dBZ occurs ≥1.4 km above the 0°C isotherm height (Waldvogel et al. 1979)
Snow	If the NWP model freezing-level height (0°C wet-bulb isotherm) is negative (i.e., below the surface)
Mixed-phase	If the surface rain rate is higher than Lumb's critical rate (Lumb 1963)
Rain	If none of the previous criteria are satisfied

2) Manual SYNOP

Met Office manual SYNOP stations are those where a qualified employee has physically observed meteorological conditions for the 10-min period leading up to every hour. WMO Table 4677 is used to record PW observations (WMO 1988, 2019). Manual reports are considered to be the highest quality standard of PW observation and observers are well trained with handbooks to minimize inconsistencies between sites. The range of PW codes available cover more obscure weather conditions and many do not refer to precipitation at all. The main limitation of the manual stations is that there are few locations; 38 manual SYNOP station locations were available for inclusion during this study.

3) DiVeN disdrometers

With the support of the Met Office and the National Centre for Atmospheric Science (NCAS), the Disdrometer Verification Network (DiVeN) was installed in the United Kingdom in early 2017 (Pickering et al. 2019). The Thies laser disdrometers (Adolf Thies GmbH and Co. KG 2011) measure the diameter and fall velocity of hydrometeors and use empirical relationships (such as those developed by Gunn and Kinzer (1949) and Locatelli and Hobbs (1974)) to estimate WMO Table 4680 PW codes (WMO 1988, 2019). Prior studies have shown that the Thies laser disdrometers have a good ability to distinguish between solid and liquid precipitation types but less skill in the mixed-phase or during light precipitation (Bloemink 2005; Lyth 2008; Pickering et al. 2019, 2021). Hail detection from the Thies laser disdrometer is possible but is less well studied, so the instruments are not used here for verification of the SPT hail class. Data are openly available (NERC et al. 2019) from February 2017 (18–23 months depending on the site install date) at a 5-min frequency and 14 locations exist.

2. Study period characteristics

In this study, the Met Office SPT product is verified over a 5-yr period of 2014–18 inclusive (60 months total). Before verifying the product an overview of the data characteristics throughout the study period is provided here.

a. Frequency maps

SPT-product classes from the 5-yr study period are summed in time to create total radar-diagnosed frequencies of precipitation, and then each precipitation type as a percentage of total precipitation observed. High-resolution zoomable PDF maps are provided in the online supplemental material. Figure 3 shows the percentage of the 5-yr period where a pixel prescribed precipitation of any kind. The spatial distribution of precipitation frequency in Fig. 3 shows higher precipitation frequency in the north and western areas, and over higher terrain. The radar network covers the whole of the United Kingdom (except the Shetland Islands) but some artifacts are visible. Note that the western and southern edges of the product are constrained by the extent of the UKPP 0°C wet-bulb isotherm field. The furthest extent of the radar network detects precipitation less frequently because the beam is less sensitive with range and may overshoot precipitation.

In a similar fashion, azimuths that experience long-term partial or total beam blockage (by terrain, buildings, or trees) exhibit radial streaks of decreased percentages. The edges of some radar maximum-range boundaries are visible, notably in northern Scotland, and this is due to dual-polarization upgrade downtime at individual sites (see supplemental material). The patches of decreased precipitation frequency are likely due to the removal of ground or sea clutter (reflective human or

TABLE 2. Summary of the three ground verification datasets used in this study. Includes the different measurement techniques, the format of the data when received, the frequency of data available, the number of locations available, and the availability over the duration of this study period.

	Automatic	Manual	DiVeN
Measurement technique	A Vaisala FD12P present weather sensor, precipitation detector, visiometer, ceilometer, and air temperature thermometer combined into an arbiter	Trained meteorological observer	A laser disdrometer measures particle diameter and fall velocity and uses empirical relationships to determine precipitation type
Format	PW Code (WMO Table 4680), 83 codes reported	PW Code (WMO Table 4677), 91 codes reported	PW Code (WMO Table 4680), 21 codes reported
Frequency	Hourly	Hourly	5-min
Locations	172	38	14
Availability	2014–18 (5 years)	2014–18 (5 years)	2017–18 (18–23 months, depending on the install date)

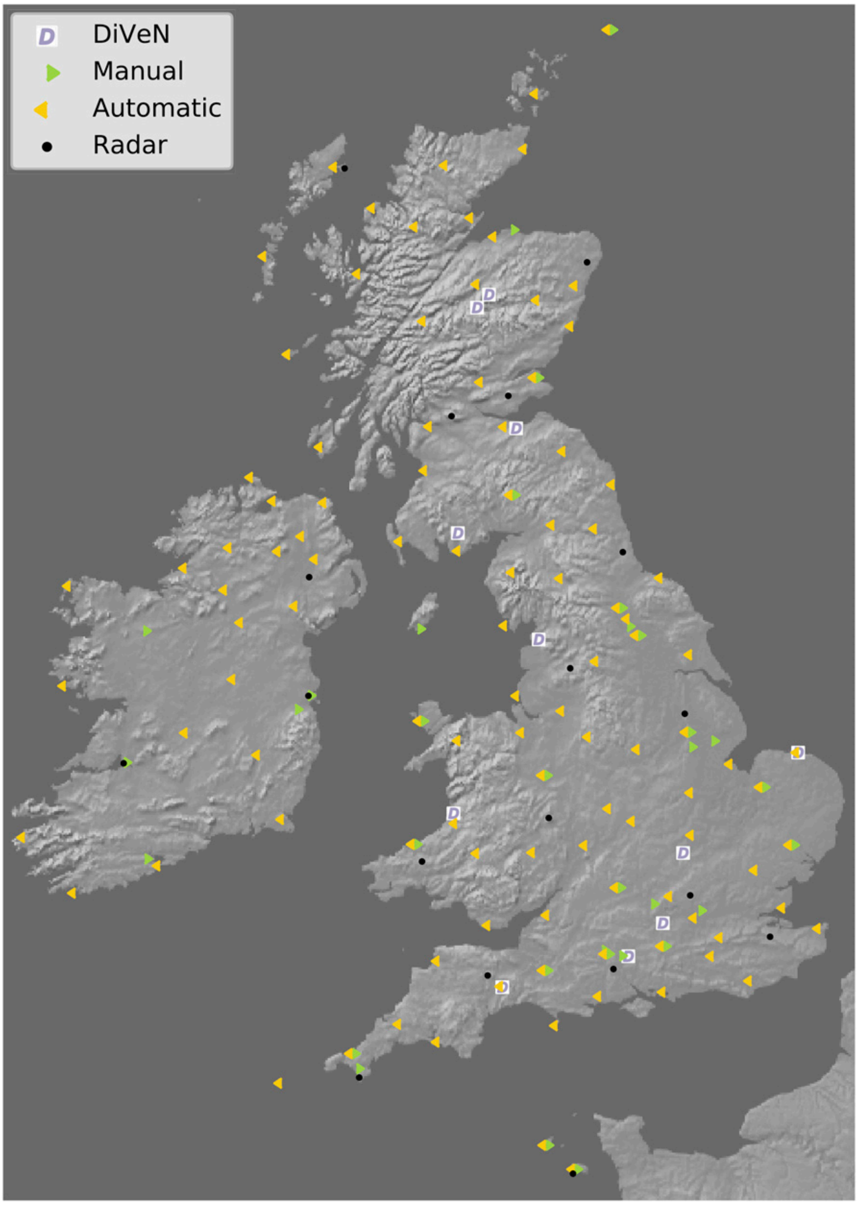


FIG. 2. A map of the United Kingdom showing all surface station sites (automatic, manual, and DiVeN) used in the verification in this study, as well as the locations of all radar sites used in the Met Office SPT product. Some stations are a hybrid (denoted with adjacent yellow left-pointing and green right-pointing triangles), where the observations are mostly automatic but are sometimes overridden with manual observations if an observer is present and disagrees with the automated diagnosis.

natural structures) which also removed some weaker precipitation events. Annual and monthly plots (see supplemental material) show that the Channel Islands (most southern radar) sea clutter has been almost entirely eradicated by the dual-polarization upgrade—a well-documented ability of the technology (Hubbert et al. 2009; Dufton and Collier 2015).

For the precipitation classes, the total occurrences are normalized against occurrences of any precipitation type, e.g., for each pixel, the total number of snow reports as a percentage of the total number of precipitation reports from Fig. 3. Since rain is overwhelmingly common in the United Kingdom (greater than 90% in most areas), the rain frequency map is dominated by the signals shown in Fig. 3 and is therefore not shown here (see supplemental material). Maps for mixed-phase and snow are shown in Figs. 4 and 5 .

Orography is clearly resolved in the SPT product, which can be attributed to the 0°C wet-bulb local height for the mixed and snow classes. The mixed-phase class is also influenced by the

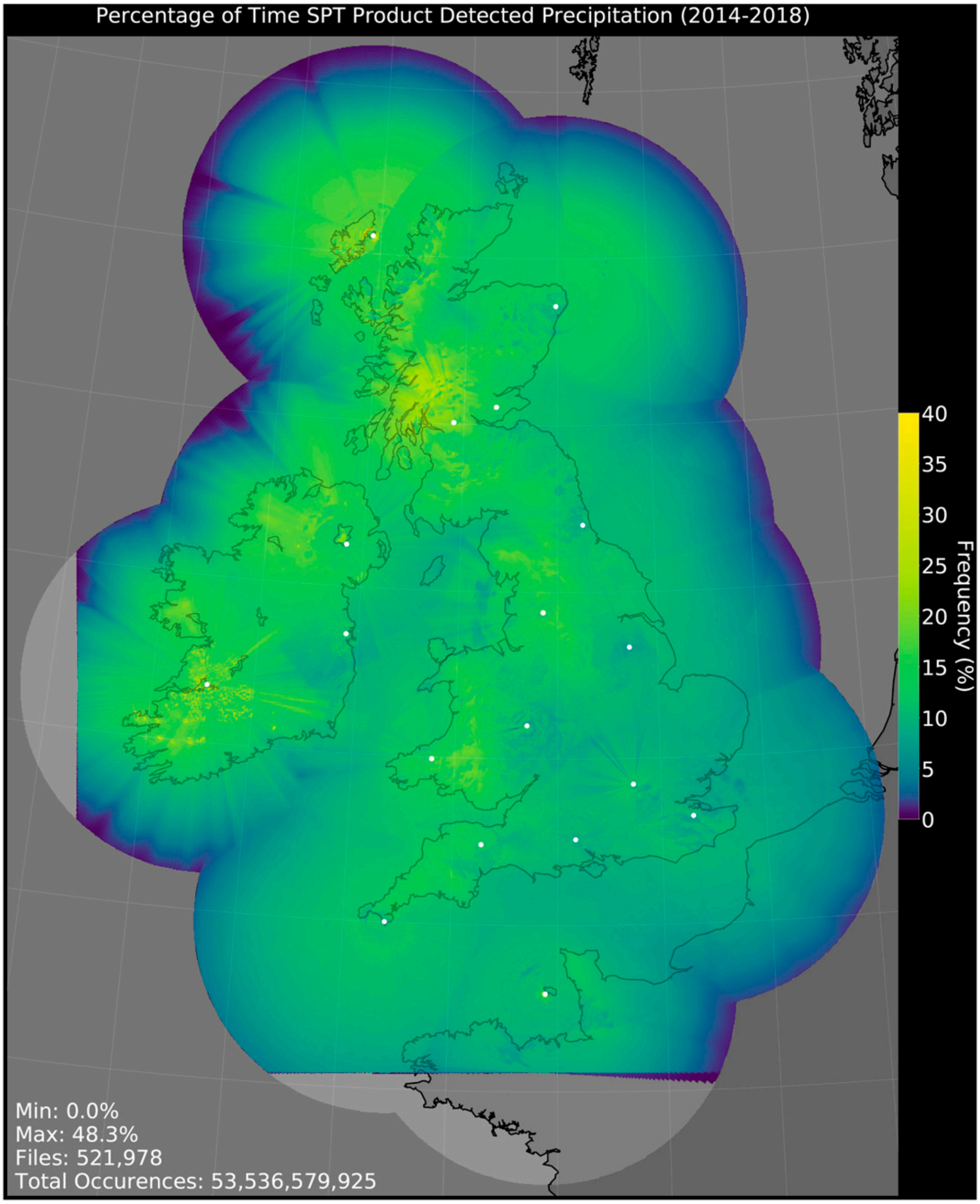

Percentage of Time SPT Product Detected Precipitation (2014-2018)
40
35
30
25
20
15
10
5
0
Frequency (%)
Min: 0.0%
Max: 48.3%
Files: 521,978
Total Occurences: 53,536,579,925

FIG. 3. Percentage of time that precipitation of any class is detected by the Met Office radar network from the start of 2014 to the end of 2018 (5 years). The Met Office, Met Éireann, and the Channel Islands Meteorological Department radar locations are marked as white dots.

enhancement of precipitation rate over orography applied by the Met Office (Harrison et al. 2000) due to Lumb's critical rate. The highest snow frequency is over the Scottish mountains where 45.2% of the precipitation detected receives a snow classification. Between 2014 and 2018, every square-kilometer pixel of U.K. land is diagnosed as experiencing snow at least once. Lowland areas of England typically experience ~0.5%–1.0% of precipitation as mixed-phase and ~3%–4% of precipitation as snow. The mixed-phase class occurs more frequently over the western-facing coasts of Scotland and the Republic of Ireland, which experience heavier precipitation more often due to exposure to westerly dominated synoptic weather and thus meet Lumb's critical rate more frequently.

In Figs. 4 and 5, offshore wind farms are visible east of London and the Thames Estuary. Wind turbines are reflective so the precipitation rate will be falsely higher and thus Lumb's critical rate will be met more often. Mixed-phase frequency also decreases in both plots where a reflectivity correction is

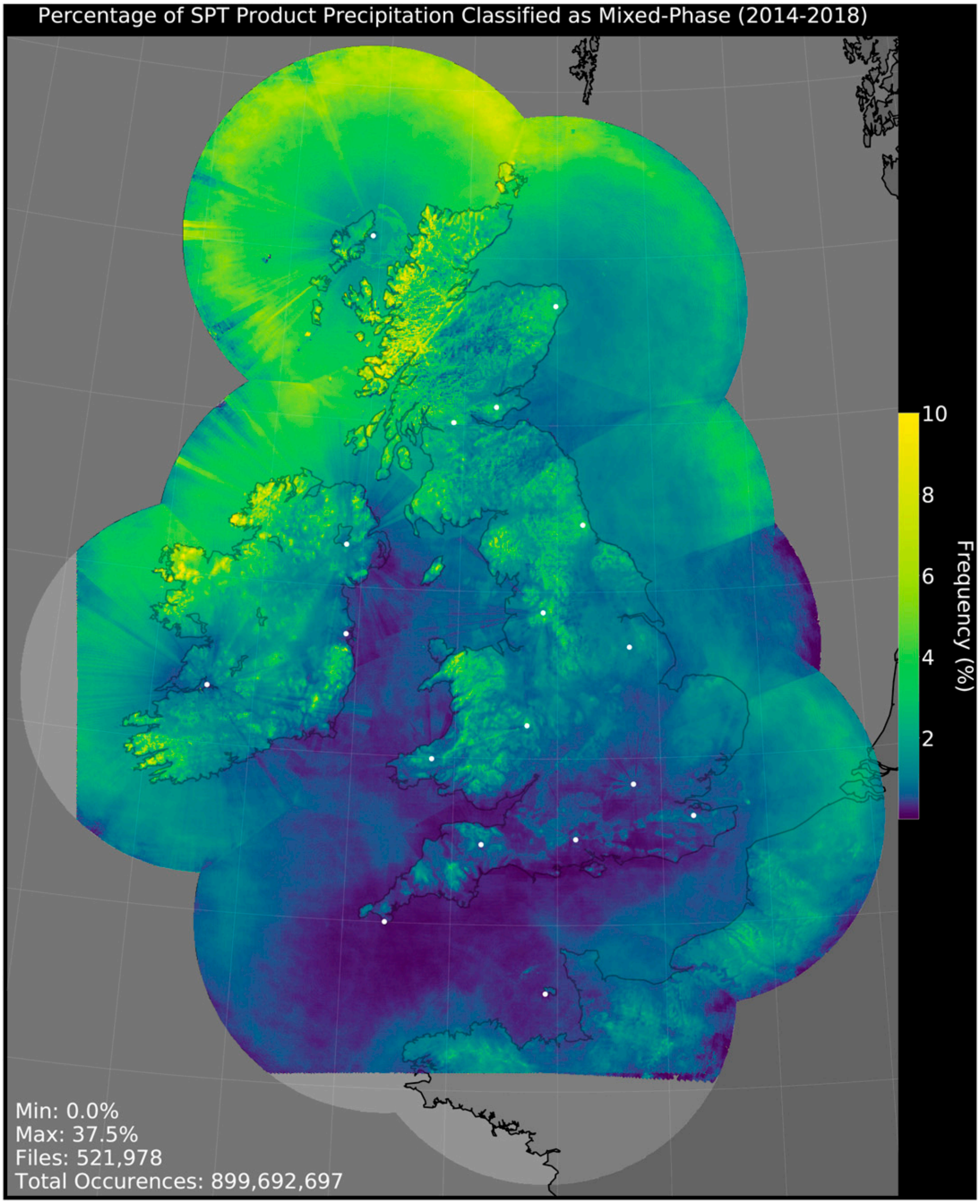

Percentage of SPT Product Precipitation Classified as Mixed-Phase (2014-2018)
Min: 0.0%
Max: 37.5%
Files: 521,978
Total Occurences: 899,692,697
Frequency (%)
10
8
6
4
2

FIG. 4. Percentage of precipitation detected by the Met Office radar network that the SPT product diagnosed as the precipitation type mixed-phase, between 2014 and 2018 inclusive. The Met Office, Met Éireann, and the Channel Islands Meteorological Department radar locations are marked as white dots. The scale is set from 0% to 10%.

made for known wind farms; for snow, this means the minimum reflectivity for precipitation diagnosis is met less often. These plots show that the correction is too strong and that the polygon is not large enough since a halo effect is seen around these locations, even after the dual-polarization upgrade. A feathered-edge polygon would give improved results.

The Ingham radar [Lincolnshire, see Fig. 1 in Harrison et al. (2015)] has fewer mixed-phase precipitation events at maximum range from the radar, caused by lower reflectivity such that Lumb's critical rate is met less frequently. Borders between preferred radars during the compositing process are visible but mainly over the ocean (with the exception of East Anglia). Banding occurs in the mixed and snow plots particularly around the edge of the network; the insensitivity to weaker precipitation at long ranges (because the radar is less sensitive generally and the beam is at a high altitude)

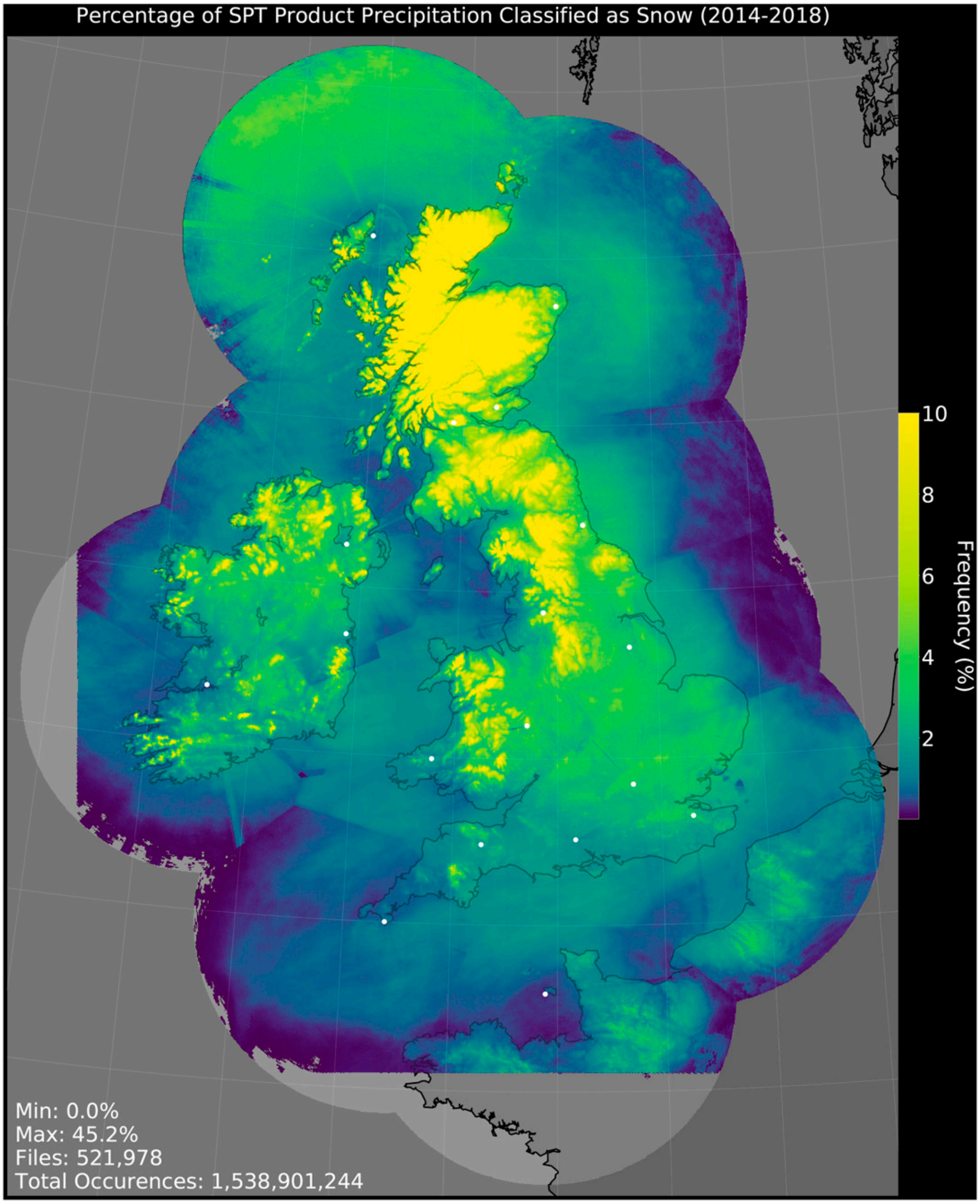

Percentage of SPT Product Precipitation Classified as Snow (2014-2018)
10
8
6
4
2
Frequency (%)
Min: 0.0%
Max: 45.2%
Files: 521,978
Total Occurences: 1,538,901,244

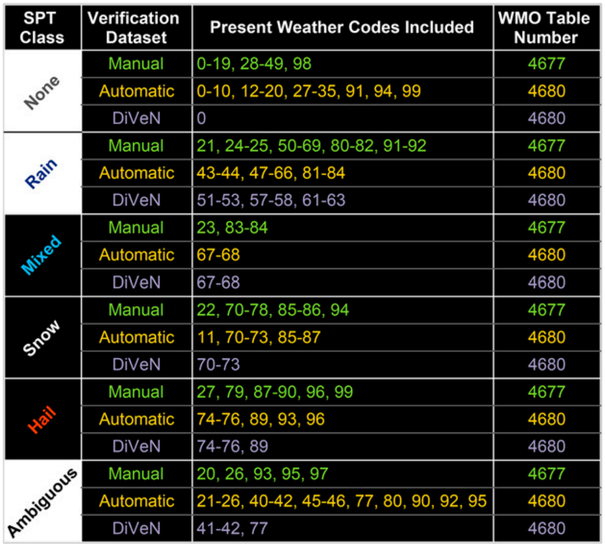
FIG. 5. Percentage of precipitation detected by the Met Office radar network diagnosed as snow by the SPT product, between 2014 and 2018 inclusive. The Met Office, Met Éireann, and the Channel Islands Meteorological Department radar locations are marked as white dots. The scale is set from 0% to 10% to highlight features. The maximum percentage is 45.2%, which occurs over the Scottish Grampians.

means that the percentage of events detected that are heavy (and are therefore more likely to meet Lumb's critical ratio) is higher.

In general, long-term frequency plots are useful for exposing artifacts, events, and trends within the radar and SPT product data. The sensitivity of the SPT product to changes in reflectivity and radar scan geometry are well highlighted here. A limitation of using this method to find radar artifacts is that many years of observations are needed if seasonal changes are to be observed.

b. *Verification data statistics*

The SYNOP (automatic and manual) reports are hourly and cover the full 5-yr study period. DiVeN began in February 2017 and therefore contributes 18–23 months of data (depending on the site install date), but every 5 min. The automatic stations



SPT Class	Verification Dataset	Present Weather Codes Included	WMO Table Number
None	Manual	0-19, 28-49, 98	4677
	Automatic	0-10, 12-20, 27-35, 91, 94, 99	4680
	DiVeN	0	4680
Rain	Manual	21, 24-25, 50-69, 80-82, 91-92	4677
	Automatic	43-44, 47-66, 81-84	4680
	DiVeN	51-53, 57-58, 61-63	4680
Mixed	Manual	23, 83-84	4677
	Automatic	67-68	4680
	DiVeN	67-68	4680
Snow	Manual	22, 70-78, 85-86, 94	4677
	Automatic	11, 70-73, 85-87	4680
	DiVeN	70-73	4680
Hail	Manual	27, 79, 87-90, 96, 99	4677
	Automatic	74-76, 89, 93, 96	4680
	DiVeN	74-76, 89	4680
Ambiguous	Manual	20, 26, 93, 95, 97	4677
	Automatic	21-26, 40-42, 45-46, 77, 80, 90, 92, 95	4680
	DiVeN	41-42, 77	4680

FIG. 6. Conversion lookup table (LuT) for converting ground observations from WMO present weather code into the SPT product classes for this study to verify. Also shown are the ranges of PW codes supported by each instrument and the specific table used, since autonomous and human observations use different WMO tables. Many of the codes available in the WMO tables are ambiguous (contain multiple SPT product classes) and are shown in the last row. All supported PW codes from each surface dataset are assigned an "SPT class" in the table. Note that the term "mixed-phase" refers to the mixture of snow and rain and does not include partially melted graupel or hail.

contributed a total of 330 369 precipitating PW code reports, of which 321 111 (97.20%) were rain, 2408 (0.73%) were mixed, and 6850 (2.07%) were snow. Manual sites are less common and contributed 75 647 precipitation reports, consisting of 73 609 (97.31%) rain, 716 (0.95%) mixed-phase, and 1322 (1.75%) snow. DiVeN disdrometer instruments contributed 148 441 precipitation reports, of which 135 083 (91.00%) were rain, 2787 (1.88%) were mixed-phase, and 10 571 (7.12%) were snow. DiVeN sites observe higher frequencies of mixed and snow cases because several of the sites are at high elevation (5 sites > 250 m MSL out of 14 total). The Met Office SYNOP sites are more commonly at lower elevations on flat terrain (~10% > 250 m MSL).

3. Methodology

The aim of this study is to verify the skill of the Met Office SPT product over a 5-yr period. To achieve this, several ground-based datasets are used to increase the volume of data available and to have multiple perspectives since all ground-based data have their own artifacts and biases. The sections below outline the steps taken to verify the skill of the SPT product.

a. Data handling and quality control

A limitation of the ground-based data is that all are coded using the PW system; many codes contain multiple precipitation types or are ambiguous (i.e., multiple conditions are described). To facilitate comparison to the SPT product, the WMO Table 4680 and 4677 codes are translated into the Met Office SPT product classes (none, rain, mixed-phase, snow, hail) or "ambiguous" as shown in Fig. 6. The number of ambiguous (containing more than one SPT product class) reports were as follows: manual 16 961 (1.8%), automatic 489 481 (7.7%), DiVeN 12 888 (0.5%).

In this study, an event constitutes one surface observation paired with a collocated SPT product diagnosis. There are 9 894 007 events in total available to this study from combined automatic, manual, DiVeN sites. The purpose of this study is to examine the SPT-classification skill of the product, not whether the radar correctly detects precipitation. Therefore, events that contain no precipitation (from either or both data sources), events that are erroneous (SPT data missing, codes outside of the PW coding scheme) or are ambiguous, are removed (562 590 events remain). The SPT product should also be functioning nominally in the wider vicinity; if the SPT product has any erroneous flags in the 5 km × 5 km ± 15-min SPT pixel region around the ground report location, then the event-pair is discarded (555 993 events remain). Additionally, events where either of the event-pair report hail are removed. After filtering, 554 457 events remain from which the analysis is performed.

Ground-based observations are paired with the next available SPT file because output files are labeled with the end time of a 5-min period. Note that the Met Office operates a 10-min radar scan strategy with three elevation descents containing both high- and low-elevation angles.

b. Confusion matrices and contingency table metrics

Discrete nonprobabilistic datasets are typically verified by confusion matrices where events are allocated a position in the matrix based on the ground-truth dataset (the class-designated column) and the dataset under examination (the class designated row). Table 3 shows the confusion matrix that will be employed in this analysis. The top-left to bottom-right diagonal entries are therefore instances where the dataset under examination is in agreement with the truth and a "hit" occurs. The remaining entries reveal where the scrutinized dataset (the SPT product) is misdiagnosing.

Furthermore, the confusion matrix ($n \times n$) is reformulated into dichotomous (yes/no) contingency tables (2×2, shown in Table 4) for each of the SPT product precipitation classes (Wilks 2011). Three metrics are then applied to each table: frequency bias (B), probability of detection (POD), false alarm ratio (FAR):

TABLE 3. The structure of the 3 × 3 confusion matrix applied in this study.

		Surface			
		Rain	Mixed	Snow	
SPT	Rain	r	s	t	y_1
	Mixed	u	v	w	y_2
	Snow	x	y	z	y_3
		x_1	x_2	x_1	Total, n

TABLE 4. The layout of the 2 × 2 contingency table used in this study.

		Surface		
		Yes	No	
SPT	Yes	Hit, a	False alarm, b	y_1
	No	Miss, c	Correct null, d	y_2
		x_1	x_2	Total, n

$$B = \frac{(a+b)}{(a+c)}, \tag{2}$$

$$\mathrm{POD} = \frac{a}{(a+c)}, \tag{3}$$

$$\mathrm{FAR} = \frac{b}{(a+b)}, \tag{4}$$

where a = hit, b = false alarm, and c = miss. Bias shows whether the class is being under or overdiagnosed by the SPT product, which can range from 0 (underdiagnosis) to ∞ (overdiagnosis); 1 is the perfect score. POD is the chance of a correct diagnosis when the precipitation type does occur and thus ranges from 0 (the event is never detected) to 1 (the event is always detected). FAR is the chance of a false diagnosis when the event is diagnosed and ranges from 0 (no false alarms) to 1 (all diagnoses are false alarms).

c. *Heidke skill score and bootstrapping*

An overall score is sought for the SPT product, before narrowing in to identify the strengths and weaknesses of the product on a per-precipitation-class basis. Generally, a skill score (SS) takes the form:

$$\mathrm{SS} = \frac{V - V_{\mathrm{ref}}}{V_{\mathrm{perf}} - V_{\mathrm{ref}}}, \tag{5}$$

where V is the verification metric, V_{ref} is the verification metric for a reference diagnosis, and V_{perf} is the verification metric for a perfect diagnosis. Several scores exist and each come with strengths and limitations. Since the SPT data are discrete nonprobabilistic (rain, mixed-phase, or snow) as opposed to dichotomous (yes or no), two appropriate higher-dimension generalized skill scores are considered: the Heidke skill score (HSS) and the Peirce skill score (PSS). The n-dimension HSS is defined following the structure of Eq. (6) as

$$\mathrm{HSS} = \frac{\sum_{i=1}^{I} p(y_i, x_i) - \sum_{i=1}^{I} p(y_i)p(x_i)}{1 - \sum_{i=1}^{I} p(y_i)p(x_i)}, \tag{6}$$

where $\sum_{i=1}^{I} p(y_i, x_i)$ is the proportion correct (the normalized sum of all diagonal confusion matrix terms),$\sum_{i=1}^{I} p(y_i)p(x_i)$ is the random proportion correct (the product of diagnosed and observed normalized probabilities summed over each class), 1 is the perfect score, I is the length of the confusion matrix, y_i is the ith row, and x_i is the ith column (Doolittle 1888; Heidke 1926). The HSS indicates the fractional improvement in diagnosis over the probability of a correct diagnosis by chance, which would score zero. The highest score (V_{perf}) is 1, and the lowest possible score is −∞; negative values therefore indicate that a random guess would have been more skillful. For a dichotomous 2 × 2 contingency table the HSS collapses to

$$\mathrm{HSS} = \frac{2(a \times d - b \times c)}{(a+c)(c+d) + (a+b)(b+d)}, \tag{7}$$

where d = correct nulls. Applying the HSS to both the higher-dimension classifier (all classes simultaneously) and the individual classes allows the contributions from each precipitation type to be quantified.

The PSS is a modification on the HSS where the denominator V_{ref} term is the unbiased random proportion $\sum_{j=1}^{J}[p(x_j)]^2$, defined by the climatology of the observation dataset. If the climatology of the verification region differs substantially, or if seasonal changes occur during a verification period, the score must be recalculated for each subset of the events (Wilks 2011). This adds computational expense and obscures the analysis as the subsets of events have no rigorous boundaries for climatology or seasonality. Therefore, this study uses the HSS as an overall SPT product metric, which is applied to each ground-based dataset (automatic, manual, and DiVeN) separately.

To show the stability of the overall skill score, a bootstrapping technique is employed (Efron and Tibshirani 1994; Chernick 2011). A similar approach for SPT verification is taken by Wandishin et al. (2005) and Elmore et al. (2015). Events are extracted at random with replacement (an event can be extracted multiple times) to form a new subset of data. Bootstrapping is repeated 100 times to create many new subsets of randomized events which give an indication of the sensitivity of the HSS to rare events.

The spread of HSS for the subset of data produced by bootstrapping is heavily dependent on the number of random samples taken in each bootstrap and must, therefore, be chosen with physical justification. The more data that are ingested, the less variability the HSS exhibits with a random subset. The full 5-yr dataset will have a narrow spread when bootstrapped, whereas a single event could have any HSS value and therefore the maximum possible spread. This study aims to show the realistic range of HSS values possible with a single month and a single year of the SPT product. Two bootstrap sample sizes are chosen to represent the number of events typically reported (after the quality control procedures described in section 3a) in one month (5506, 1261, 7069) and in one year (66 074, 15 129, 84 823), from each ground observation dataset, respectively (automatic, manual, and DiVeN).

d. *Tolerance*

Due to the disparity of the lowest-usable radar beam height and the surface, precipitation observed by radar is often not vertically collocated with the surface. Sandford (2015) showed that the uncertainty in radar drift estimates can vary from 1 km below the melting layer to 10 km at the extreme distance of the maximum range of a radar. The terminal fall velocity of different SPTs differs (Langleben 1954; Zikmunda 1972; Locatelli and Hobbs 1974; Matson and Huggins 1980; Böhm 1989), so

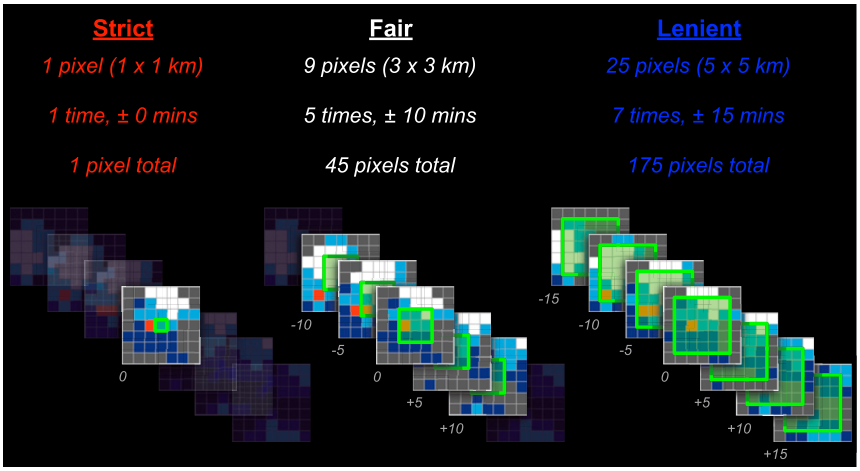


FIG. 7. An example of a time series of the SPT product stacked to represent time (5-min frequency). The green-outlined area is the sample used for verification in three tolerances. The strict tolerance uses only the pixel collocated with the ground report. The fair tolerance uses a 3 km × 3 km region around the ground report and ±10 min product outputs for a total of 45 pixels. The lenient tolerance uses a 5 km × 5 km region around the ground report and ± 15 min product outputs for a total of 175 pixels. If any of the green-shaded pixels are in agreement with the ground observation, then the SPT product is correct and a "hit" is recorded.

the descent time varies between precipitation types. Furthermore, the horizontal wind advects precipitation as it falls and, therefore, the amount of horizontal displacement during descent will also differ between precipitation types.

There are several factors determining the trajectory of hydrometeors as they fall to Earth's surface, which makes verification difficult. Here a general solution is applied which increases the spatial and temporal tolerance for the SPT product to inform how the product skill is impacted. This informs a user of what spatiotemporal specificity corresponds with a desired accuracy. Three tolerances of the SPT product are used; strict: only the 1 km × 1 km area and 5-min period collocated with the surface report; fair: a 3 km × 3 km area and ± 10 min around the surface observation will be considered; lenient: a 5 km × 5 km area and ± 15 min around the surface observation will be considered. Figure 7 shows the three tolerances diagrammatically.

If any of the SPT product pixels in the fair or lenient tolerances agrees with the surface, then it is considered a hit. Note that new false alarms can be introduced when moving from a strict to a more lenient tolerance, since the SPT class under examination may appear in the larger tolerance window. For example, if the SPT under examination is "snow," the ground instrument does not record snow and neither does the central radar pixel (strict tolerance), then the outcome is a correct null label of the event. However, if within the larger tolerance there is a snow detection, this event becomes a false alarm. The lenient tolerance is approximately the maximum reasonable displacement (~2.5-km radius) and fall time (15 min) a hydrometeor could experience from the lowest usable beam height given the Met Office radar network coverage. To apply this verification technique to other products, the choice in tolerance may differ. There must exist a physical meaning to the minimum (strict) and maximum (lenient) possible extent of the gridded product under examination, which is dependent upon the specific variable being examined and also the measurement technique.

4. Results

a. Heidke skill score and bootstrapping

First, the higher-dimension generalized HSS is examined to give an overall value to the SPT product, before examining each precipitation class. Note that only the SPT product pixel which directly encapsulates the location and time of the ground-based observation is used here (i.e., strict tolerance). While the hit and correct null quadrants are simple, the higher-dimension thresholds for false alarm or miss criterion from multiple SPT pixels would be subjective.

Figure 8 shows the higher-dimension HSS for all classes of the SPT product. Overall, the SPT product has absolute HSS values (using the full dataset without bootstrapping, indicated by black dots on Fig. 8) from 0.48 for automatic, 0.60 for manual, and 0.73 for DiVeN. If all surface-based observations are combined, the HSS of the SPT product is 0.61.

The spread of HSSs represents the possible scores if a random month or random year of data were considered. HSS distributions are markedly different between yearly and monthly bootstrap representations, with a much narrower spread for the yearly than monthly. Between verification datasets there are also differences. The manual station verification has the largest spread with a standard deviation (2σ) of 0.147 monthly and 0.038 yearly. Automatic stations give the second largest spread but the lowest overall score, with a standard deviation (2σ) of 0.058 monthly and 0.018 yearly (approximately half compared to manual sites). The DiVeN dataset has the highest scores and a standard deviation (2σ) of 0.024 monthly and 0.008

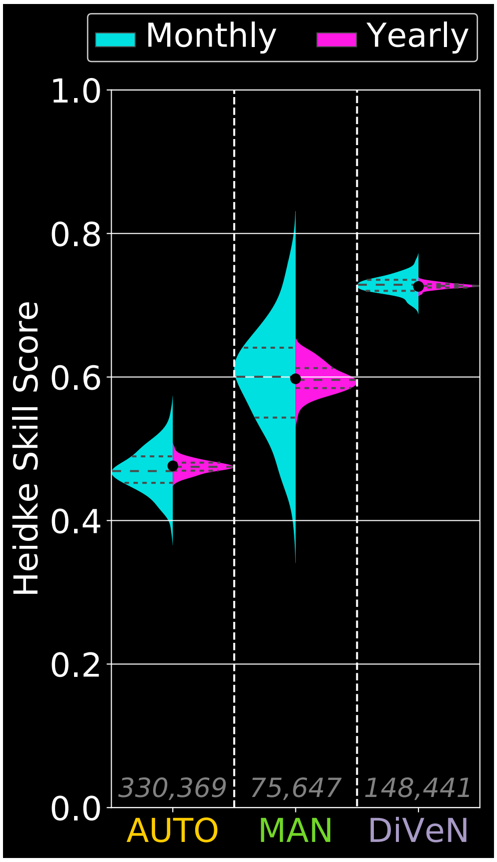


FIG. 8. Higher-dimension HSS (rain, mixed-phase, and snow simultaneously) with probability distributions produced by a bootstrapping technique. Note that each distribution is scaled to fit half the width of the column for ease of viewing. Each ground dataset is shown (automatic, manual, DiVeN) and each has monthly and yearly representative distributions. The black dot indicates the HSS for the full dataset.

yearly (approximately half compared to automatic sites). Ultimately the differences in HSS spread tell us more about the ground-based dataset than the SPT product, but taking into consideration all three ground-based datasets gives a broader picture of the variability of the skill of the SPT product on different time scales, from approximately 0.4 to 0.8.

The HSS is recalculated with adjustments to some of the SPT product classes. Including the hail class of the SPT product into the calculation makes little difference because the HSS gives proportional weighting to rare events, and the ground-based datasets rarely report hail; automatic stations never report hail. If the mixed precipitation class is removed, the score (for all ground-based datasets) improves significantly from 0.61 to 0.77. This is unhelpful as the SPT product would in this scenario have an ‘‘unknown’’ class for these events. If all mixed-phase diagnoses are reclassified as rain the HSS increases to 0.73 and if all mixed-phase diagnoses are reclassified as snow then the HSS decreases slightly to 0.59. This indicates that mixed diagnoses are more likely to be rain than either mixed-phase or snow.

b. *Confusion matrices*

Confusion matrices are useful for showing where each class is being misdiagnosed. Figure 9 shows the results for the rain, mixed-phase, and snow classes for each of the three ground observation sets available. Note again that the tolerance approach cannot be applied (see previous section), so the values shown are using only the encapsulating SPT product pixel area and time.

First for the overall frequency of diagnoses, the rain type is underdiagnosed by the SPT product for automatic stations (−1.94%) but is close to the observed occurrences by manual (+0.18%) and DiVeN (+0.08%) sites. For mixed precipitation, the SPT product diagnoses this class twice as often compared with automatic sites, around the same compared with manual sites, and half as often compared with DiVeN sites. Finally, snow is diagnosed 50% more by the SPT product compared with automatic stations, around the same for manual stations, and 12% more for DiVeN sites.

Next, the rows of the confusion matrices are examined so that for a given SPT product diagnosis, the true observed precipitation type can be discussed. For example, given that the rain class is diagnosed, it is correct most often, but there are some miss events where the ground station observed mixed-phase or snow and in all ground datasets the mixed-phase class is the missed truth more often. The mixed class is poorly diagnosed, and rain is the observed ground event 23.8, 4.7, and 7.1 times more often (automatic, manual, and DiVeN). Finally, the snow diagnosis is correct 52.5%, 78.4%, and 77.7% of the time (automatic, manual, and DiVeN). The miss events differ between ground datasets. For automatic, the majority of miss events are rain (41.3% of all snow diagnoses), with 6.3% miss events being mixed. For manual, miss events are more evenly split over rain (10.3%) and mixed (11.4%). For DiVeN, rain is the missed event for 14.1% of the snow diagnoses and mixed is the missed event for 8.3% of the snow diagnoses.

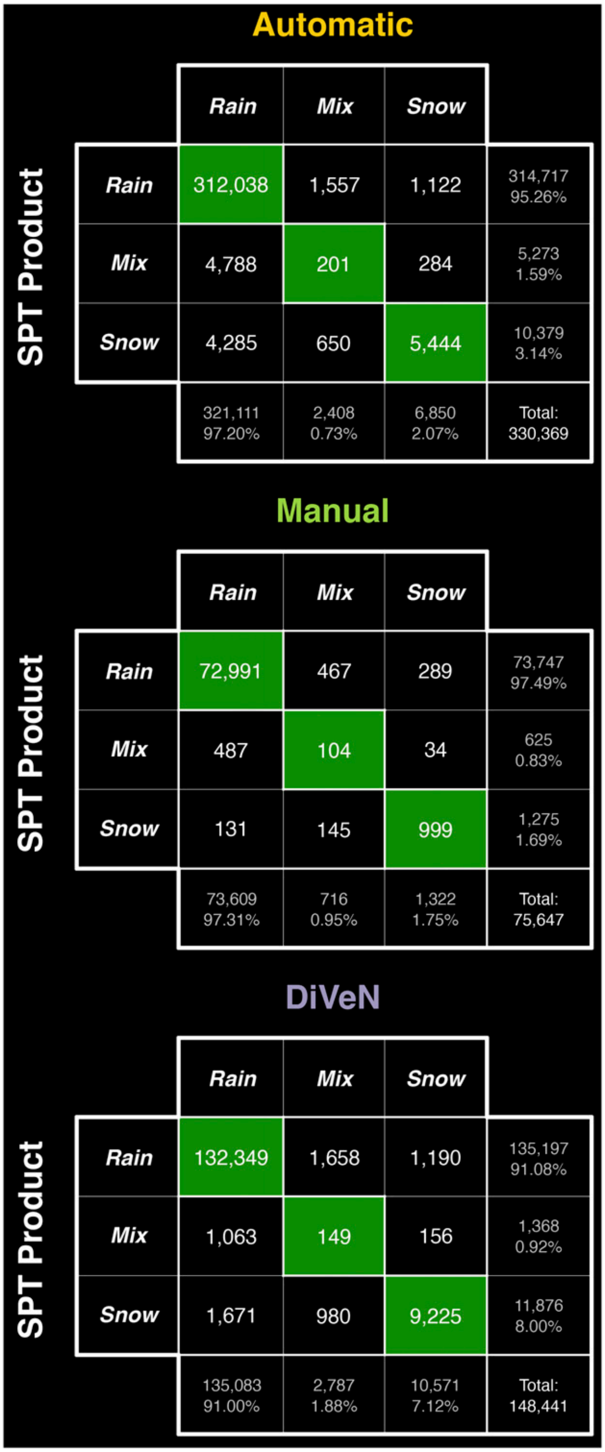
c. *Contingency table metrics with tolerance*

Next, skill scores are examined for each precipitation class where a contingency table has been produced from the confusion matrices. Three realistic tolerances based on the maximum horizontal displacement during descent from the lowest-usable radar beam have been applied to the SPT product as described in section 3d. All of the results are composed into Fig. 10.

The hierarchy of the next section is as follows: each verification metric is discussed individually, going through the precipitation types (as some scores have interdependencies between the precipitation classes) and commenting on differences between the ground datasets and tolerances throughout.

1) BIAS

The frequency bias indicates the scale to which precipitation classes are being under or overdiagnosed. Generally speaking, the mixed-phase and snow classes are overdiagnosed at the expense of rain. The high frequency of rain events makes the bias close to 1 but a slight underdiagnosis is occurring. Bias changes with increased tolerance are also small. The mixed-phase has the largest positive biases of any class, with the highest being 8.87 (automatic, lenient tolerance), whereas



Automatic

SPT Product		Rain	Mix	Snow	
	Rain	312,038	1,557	1,122	314,717 95.26%
	Mix	4,788	201	284	5,273 1.59%
	Snow	4,285	650	5,444	10,379 3.14%
		321,111 97.20%	2,408 0.73%	6,850 2.07%	Total: 330,369

Manual

SPT Product		Rain	Mix	Snow	
	Rain	72,991	467	289	73,747 97.49%
	Mix	487	104	34	625 0.83%
	Snow	131	145	999	1,275 1.69%
		73,609 97.31%	716 0.95%	1,322 1.75%	Total: 75,647

DiVeN

SPT Product		Rain	Mix	Snow	
	Rain	132,349	1,658	1,190	135,197 91.08%
	Mix	1,063	149	156	1,368 0.92%
	Snow	1,671	980	9,225	11,876 8.00%
		135,083 91.00%	2,787 1.88%	10,571 7.12%	Total: 148,441

FIG. 9. Confusion matrices of SPT product against ground observations, for each ground observation type. (a) Automatic SYNOP, (b) manual SYNOP, and (c) Disdrometer Verification Network (DiVeN).

some are close to an ideal bias (0.87, manual, strict tolerance). The strict DiVeN result shows an underdiagnosis of mixed-phase (0.49) but increased tolerance shows an overdiagnosis (1.67 and 2.40). For snow, biases are overall smaller than the mixed-phase class but still show a positive tendency. With strict tolerance, biases against the manual and DiVeN data are 0.96 and 1.12, whereas bias against automatic is 1.52.

2) Probability of detection

The POD tells us the probability of the SPT product being correct given that the precipitation class is occurring. Again the rain class is weighted by the frequency of occurrence (91%–97% of precipitation) in the study period and has values close to a perfect score of 1. The lowest rain class POD score is in the automatic dataset (0.97, strict) due to underdiagnosis. For the mixed-phase class, POD is low, ranging from 0.08, 0.15, and from 0.05 to 0.24, 0.44, and 0.19 (automatic, manual, DiVeN). The snow class has POD values similar to rain, with lenient/fair tolerances consistently 0.91–0.94 for all ground datasets. The strict tolerance varies: 0.79 (automatic), 0.76 (manual), and 0.87 (DiVeN). Given that an SPT is occurring, increasing tolerance makes a correct diagnosis more likely.

3) False alarm ratio

The FAR indicates the probability of a false alarm when the SPT product diagnoses a precipitation type. The rain FAR is consistently low due to its high occurrence frequency. The DiVeN dataset gives a slightly higher rain FAR of 0.04 (lenient tolerance), which is indicative of the lower occurrence frequency from DiVeN (91% versus 97% of precipitation for the other datasets). The mixed-phase class has high FAR (from 0.83 to 0.97) for all verification datasets consistent with a positive bias. The snow class has different FAR depending on the verification dataset: against manual and DiVeN, FAR values are around 0.22–0.39 but against automatic, FAR values are 0.48–0.66. Increasing the SPT product tolerance increases the chance of a false alarm.

4) Heidke skill score

The HSS indicates the fractional improvement of the SPT product diagnoses over random diagnoses, where a value of 0 is no skill and a value of 1 is a perfect diagnosis every time. The decimal value can be described as a percentage improvement over random chance. The HSS values for rain take into account the high frequency of occurrence and range between 0.51 and 0.64 for automatic and strict, but are higher (0.70–0.77) for DiVeN (lower rain occurrence frequency). The HSS values are not correlated with increasing or decreasing tolerance as is the case with the other verification metrics; this is explained in the discussion (section 5d). The weaknesses in the mixed-phase class are highlighted by the HSS, with low values across the ground datasets and tolerances. Automatic observations give the lowest scores (~0.04), DiVeN the middle scores (0.06–0.09), and manual the highest scores (0.15–0.19), but all indicate poor skill. Snow has skill on par or better than the rain class, with values ranging between 0.73 and 0.81 for manual and DiVeN datasets, while the automatic dataset gives scores slightly lower with a wider range from 0.49 to 0.62.

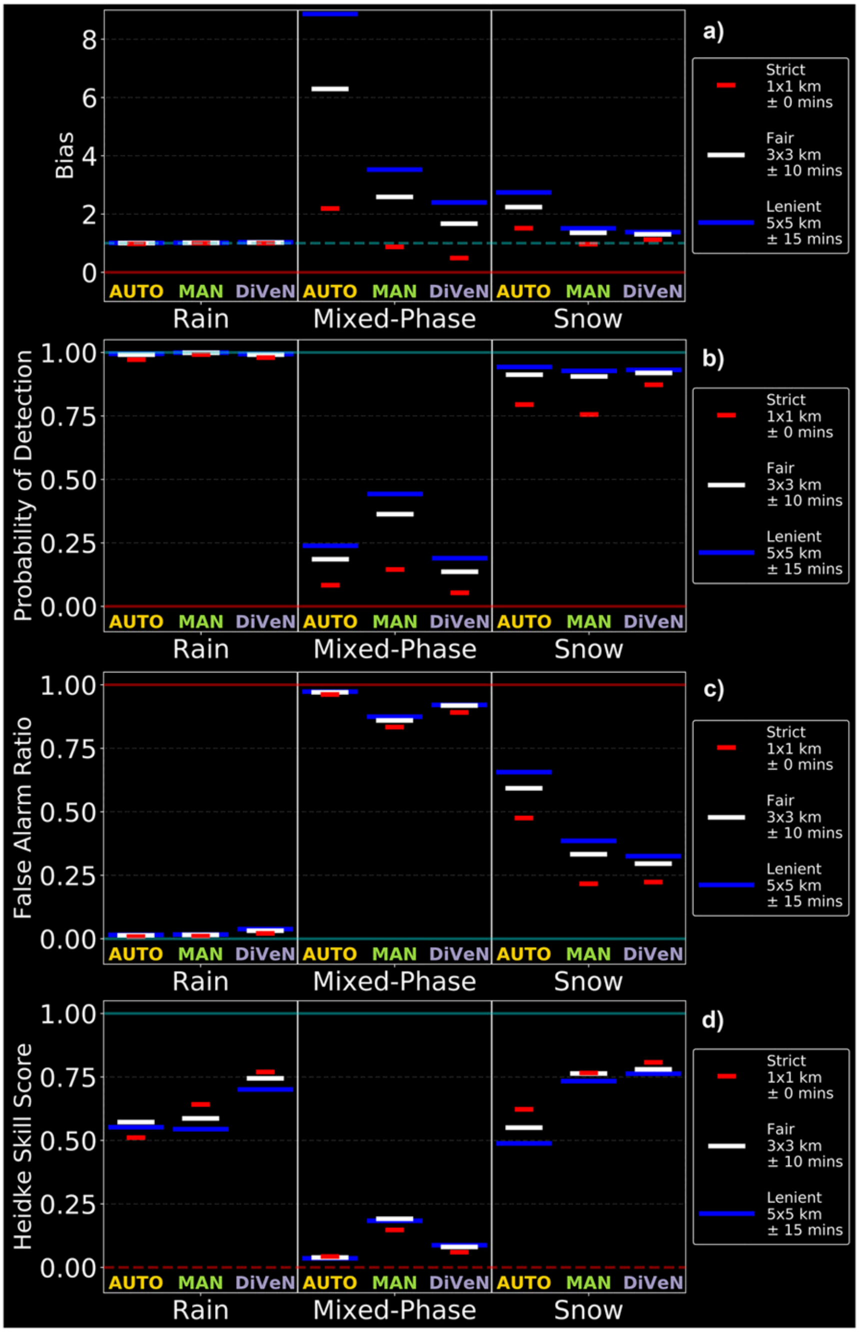


FIG. 10. Skill scores for each precipitation class and ground dataset. (a) Bias, (b) POD, (c) FAR, and (d) HSS. Cyan horizontal lines indicate a perfect score, and red horizontal lines indicate a ''no skill'' score. Solid cyan or red lines are fixed value limits, dashed are surpassable (bias and HSS).

5. Discussion

a. Rain

Since rain is the dominant class with >90% frequency, most skill scores for this precipitation type are skewed. The bias appears close to 1 but is underdiagnosed, POD is deceptively high and, similarly, FAR is deceptively low. The HSS takes the frequency into account and shows a 50%–65% improvement over random chance diagnoses which are caused by the mixed-phase class diagnosing rain events. A fairer verification should not include low-skill rain cases; product users would not look for snow during heatwaves, for example. Events could be limited to the Met Office snow warnings, or periods of 0°C wet-bulb isotherm below 500 m MSL, criteria that operational forecasters use (S. Lee, MeteoGroup, 2019, personal communication). Alternatively, the occasions when the SPT product is opened could be recorded to build up cases targeted to user activity. The number of events would be reduced but SPT frequencies would be more equitable and the verification more applicable to certain product users, dependent on the criteria used.

b. *Mixed-phase*

POD for the mixed-phase class ranges from 0.08 to 0.24. Combined with a positive bias tendency up to 8.87, this indicates that the mixed-phase class has very little skill. This is reinforced by FAR values ranging from 0.83 to 0.97 and HSS scores between 0.04 and 0.19. Typically overdiagnosis increases the POD, but the mixed-phase class in the Met Office SPT product is the most overdiagnosed and still has the lowest POD of any class.

The HSS reclassification results (section 4a) and the confusion matrices in Fig. 9 show that the mixed-phase class diagnoses are more often rain than mixed-phase or snow. Combining all verification datasets, 87.2% of mixed-phase class diagnoses are rain, 6.2% are correct, and 6.5% are snow. The height of the mixed-phase to rain boundary being too low would be consistent with these results. Assuming Lumb's critical rate to be correct, this bias would be attributable to either a negative bias in the local 0°C wet-bulb isotherm height, a positive bias in precipitation rate diagnosed by the radar, or both. Figures 4 and 5 showed the sensitivity of the SPT product to precipitation rate, as "corrected" artifacts in precipitation rate still show a signal in the mixed-phase frequency map.

Lumb's critical rate uses the work of Langleben (1954), setting the boundary between rain and mixed-phase at 90% of the precipitation as liquid, based on the behavior of the velocity of the particle. Lumb (1963) also assumed spherical aggregates and a saturated atmospheric column. Note that the data used in the derivation of Lumb's critical rate only covered 1–4 mm h^{-1} precipitation rates. These assumptions and limitations of Lumb's critical rate should be revisited and examined with modern measurement techniques to ensure that the SPT product is valid under all atmospheric conditions.

Finally, the effect of topographic representativity must be discussed. The method of calculating the local 0°C wet-bulb isotherm height results in a topographic resolution of 1 km^2. For the majority of the United Kingdom this is an acceptable approach. Where deviations of surface altitude are large such as in mountainous regions, if the station providing verification data is situated in a valley or on a peak in the terrain, then the verification will have systematic errors, since the SPT product is calculating precipitation type for the average topographic altitude within 1 km^2. To combat this, a higher-resolution topography could be used with the existing framework, for higher-resolution product output. Topographic representativity will also affect the snow diagnosis since the local height of the 0°C wet-bulb isotherm is the only criterion, meaning a perfect diagnosis at 1-km^2 resolution is not possible.

c. *Snow*

Overall the snow class has similar HSS to rain diagnoses, but is overdiagnosed and, thus, has a higher FAR than rain. Since the diagnosis is entirely dependent on the height of the UKPP 0°C wet-bulb isotherm being below the ground (i.e., surface temperatures below zero), the results suggest that the height is negatively biased. This conclusion would also agree with the results of the mixed-phase precipitation class.

For the snow class the skill of the Euro4 temperature field is essentially being verified, which itself has many influencing factors. The only other source for misclassification is the previously mentioned 1-km^2 resolution of the local terrain input data. The SPT product might be seen as an attractive candidate for verifying NWP model SPT forecasts against. However, be aware that this would be a closed-loop verification for the snow class since its diagnosis is entirely reliant upon the model.

DiVeN data give higher verification metric values (73%–81% improvement over random chance). The sites contain more snow events (5 sites > 250 m MSL) which are often observed when the 0°C wet-bulb isotherm height is several hundred meters below the surface. Borderline cases are less common in DiVeN compared with the other data. Similar to rain cases being low skill in summer, low-skill winter events make a difference to the snow verification results. In late February and early March 2018, the exceptional snowfall associated with the "Beast from the East" (Galvin et al. 2019; Greening and Hodgson 2019) brought many low-skill snow cases into the verification dataset. If 2018 data are removed then scores using all datasets are reduced dramatically. The SPT product has diminished value in these scenarios since it is clear to users that all precipitation will reach the surface as snow.

d. *Tolerance method*

The tolerance method used in this study demonstrates the sensitivity of the product's skill when adjusting the spatiotemporal inclusion, which a user typically considers when viewing a graphical map. Given the spatiotemporal range (from 1 pixel at one time, to 175 pixels over 30 min) the range of values provided by this method is often quite narrow, and is therefore informative to users. A wide range of score results would add negligible value to a single verification score result with no indication of spread. The tolerance method is therefore applicable to future verification of precipitation type diagnosis from any spatial-coverage product using single-point reference datasets.

When viewing a contingency table, the sum of all events remains constant between strict, fair, and lenient tolerances but events can only move vertically in a contingency table between tolerances. If more events move from "miss" to "hit" compared with the number moving from "correct null" to "false alarm," then the HSS improves, and vice versa. The initial distribution of events differs significantly between precipitation class and ground dataset, hence the HSS values sometimes increase and sometimes decrease (notably rain against automatic observations) between SPT product tolerances in the results of this study.

If a user desires a higher POD then a larger domain should be considered from the SPT product. If a lower FAR is desirable then a smaller domain should be taken around a desired location, which will depend on the specific user and their application of the Met Office SPT product. The results are more complex for the HSS values. If a user wants a higher skill score, then the spatiotemporal sample should be different for each precipitation class and always dependent on which ground dataset is most trusted. For the mixed-phase class, the HSS values reveal that a larger sample increases the skill of the diagnosis (except when considering the automatic data which has the lowest HSS of any ground dataset). Generally for rain and snow, using the specific pixel encapsulating a location area and time increases product skill. Note that this does not take into account the skill of detecting or not detecting precipitation accurately since all events that feed into the verification have precipitation in both sources.

e. *Comparison to other verifications*

Comparing other SPT products in the literature is difficult since there are many variables affecting the verification. Table 5 shows a sample of literature verifying many SPT products based on NWP and various observational inputs. In addition to the different inputs to each algorithm, the location, time period, method and verification scores also differ, influencing the verification results of each study. Here it was noted that even the inclusion of a fifth year onto four existing years dramatically changed the true climatology and therefore the overall results.

Different statistical approaches are applied in different studies. Chen et al. (2016) and Gascón et al. (2018) use critical success index (CSI) as a verification metric, but CSI cannot be applied to the higher-dimension confusion matrices. Wandishin et al. (2005) use Brier skill score (BSS) but this is only applicable for probabilistic data. Elmore et al. (2015) use the PSS for contingency and the Gerrity skill score (GSS; Gerrity 1992) for higher-order which emphasizes weighted ranking to each class based on climatological rarity. The PSS is as justifiable as the HSS as a skill metric and both have higher-order applicability and give similar score values (Wilks 2011). However, using the same score for contingency tables and confusion matrices (as was done here) demonstrates the contributions from each class to the overall score.

6. Summary and further work

Reliable observations of precipitation type are needed both to verify and improve forecast microphysics, and also to operationally force NWP models with more accurate initial conditions through data assimilation. The Met Office surface precipitation type (SPT) product was examined with three datasets of ground-based observations over 5 years (2014–18). The product uses Boolean logic to diagnose hail (not examined here due to lack of a suitable reference dataset), snow, mixed-phase, and rain using an empirical relationship based on radar precipitation rate and the 0°C wet-bulb isotherm from an NWP model. In this paper snow, mixed-phase, and rain were verified. An overall product score was obtained using the higher-order Heidke skill score (HSS) and a bootstrapping technique to infer the monthly and yearly sensitivity to the overall product score. Statistical metrics applied to individual precipitation classes from contingency tables were bias (B), probability of detection (POD), false alarm ratio (FAR), and the HSS. A novel tolerance method was introduced which shows the realistic spatiotemporal spread of scores taking into consideration the fall time and the horizontal displacement precipitation may experience between the lowest-usable radar beam from the Met Office radar network and the ground.

The results show that the 0°C wet-bulb isotherm from the UKPP (interpolation from the Euro4 NWP model) is too low, causing an overdiagnosis of snow ($B > 1$) leading to FAR values of 0.22–0.48 (strict tolerance). The 0°C wet-bulb isotherm height also controls the height at which mixed-phase precipitation is fully melted into rain, and may contribute to the significant overdiagnosis of mixed-phase ($B \gg 1$) with FAR values of 0.83–0.97 and POD values of 0.05–0.44 (all verification datasets and tolerances). Due to the overdiagnosis of snow and mixed-phase, by elimination the rain class is underdiagnosed. Rain has a bias of just under 1 which is skewed by the high frequency of the rain class, 91%–97% between verification datasets. The HSS takes into account high frequency of occurrence, and this gives values of 0.51–0.77, which are similar to snow where HSS values are 0.49–0.81 (all verification datasets and tolerances). The mixed-phase has low HSS values of 0.04–0.19.

Overall the higher-dimension HSS value for all datasets combined is 0.61, which improves to 0.73 if all mixed-phase diagnoses are relabeled as rain. Between verification datasets, the higher-dimension HSS are 0.48 ± 0.058 (automatic), 0.60 ± 0.147 (manual), and 0.73 ± 0.024 (DiVeN), where the uncertainty is representative of a 2σ confidence interval produced through bootstrapping.

Ground-based observations should capture the climatology of the location or target audience of the users of the product. Thus, the representativity of the data used to verify the product at a certain location is important. The automatic and manned SYNOP stations run by the Met Office may not capture the most extreme climatologies of the United Kingdom due to their siting requirements for optimal measurement standards. Similarly, the Disdrometer Verification Network likely does not capture the U.K. climatology since many instruments are located at high elevations.

Improvements to the Met Office radar-based SPT product are ongoing based on the results of this study. The Euro4 model has been marked for deprecation at the end of 2021 and there has been a freeze on scientific upgrades for several years. The implementation of a newer, higher-resolution NWP model temperature field, particularly a model with improved microphysics schemes, should improve the snow class diagnosis in a future Met Office SPT product. Note that to verify the improvement in future SPT products, the current SPT product can be statistically implemented as a baseline. Currently, V_{ref} in Eq. (5) is set here as the random proportion correct but this can be changed to be the proportion correct from this ‘‘baseline’’ SPT product instead. Thus the score can then be used to show the percentage improvement over the current SPT product. The methods employed here may be easily

TABLE 5. Comparative literature on verifications of surface precipitation type products, based on various input data, verification data, study location, time period, and methods. This study is also shown, along with results from all studies for comparison. M-P stands for mixed-phase, T_{dry} and T_{dew} are dry-bulb and dewpoint temperature, PSS is the Peirce skill score, and GSS is the Gerrity skill score (Gerrity 1992).

Paper	Product	Method	Results
Pickering et al. (2021)	Radar + NWP observation rain, M-P, snow	United Kingdom, 5 years, automatic, manual, DiVeN, varied spatiotemporal tolerance	Bias: rain ~1, M-P > > 1, snow > 1 POD: rain ~0.98, M-P 0.05–0.44, snow 0.76–0.94 FAR: rain ~0.02, M-P 0.83–0.97, snow 0.22–0.66 HSS: rain 0.51–0.77, M-P 0.04–0.19, snow 0.49–0.81 Overall HSS: 0.61 (0.4–0.8 w/bootstrapping)
Gascón et al. (2018)	NWP ensemble 0–24-h forecast Rain, snow, others	Europe, 4 months (winter) Manual reports Strict tolerance	POD: rain 0.52, M-P 0.07, snow 0.55 • Probability of any precipitation must exceed 50% • Highest probabilistic type chosen
Reeves (2016)	Sounding Observation Rain, snow, others	United States east of Rockies, 27 months Automatic + crowdsourced Auto: strict tolerance Crowd: closest (max 35 km ± 1 h)	POD: rain 0.68–0.92, snow 0.79–0.90 • SPT must be consistent for >4 h • Lower rain frequency (62%)
Schmid and Mathis (2004)	$T_{\mathrm{dry}} + T_{\mathrm{dew}}$ Observation Rain, snow	Switzerland, 2 months (winter) Automatic reports Strict tolerance	Bias: rain 0.92, snow 1.10 POD: rain 0.81, snow 0.99 • Higher snow frequency (63%)
Chen et al. (2016)	Radar + NWP Observation Rain, snow	8 U.S. cities, 4 months (winter) Crowdsourced reports Strict tolerance	POD: rain 0.93–0.98, snow 0.69–0.82 FAR: rain 0.08–0.29, snow 0.02–0.15
Wandishin et al. (2005)	NWP ensemble 0–48 h forecast Rain, snow, others	United States, 3 months (winter) Manual reports Strict tolerance	POD: Rain 0.45–0.95, Snow 0.70–0.95 • Probability thresholds • Low-skill rain ($T_{\mathrm{dry}} > 5°\mathrm{C}$) removed
Elmore et al. (2015)	NWP 0–24-h forecast Rain, snow, others	United States, 8 months (2 winters) Crowdsourced reports Strict tolerance	Bias: rain 1.10–1.60, snow 0.80–1.05 PSS: rain 0.58–0.76, snow 0.58–0.74 Overall GSS: 0.48 (3 h), 0.34–0.45 (18 h)
Ikeda et al. (2013)	NWP 1–8-h forecast Rain, M-P, snow	Eastern United States, 2 months (winter) Automatic reports 18 km × 18 km ± 6-min tolerance Fractional confusion matrices	POD: rain 0.90, M-P 0.66, snow 0.86 • "No precipitation" class included

implemented for verifying a range of observation-based or model-based classifiers; however, the most important aspect of verification is consistency of score choice between studies to enable comparisons and to identify successful SPT diagnosis techniques.

Acknowledgments. This work is funded by a NERC Industrial CASE studentship (NE/N008359/1). Publication fees were funded by the University of Leeds Library through the UKRI Open Access grant. Ryan R. Neely III, David Dufton, and Maryna Lukach were funded through the NCAS underpinning science programme through UKRI. Thanks goes to the National Centre for Atmospheric Science and Atmospheric Measurement and Observation Facility (AMOF) for site support at CAO.

Data availability statement. The Thies LPM data are available from NERC et al. (2019). The SPT product data and the Met Office SYNOP data are not publicly available.

REFERENCES

Adolf Thies GmbH & Co. KG, 2011: Laser Precipitation Monitor - Instruction for Use. Tech. Rep., Adolf Thies GmbH & Co. KG, 66 pp.

Beynon, C., S. Wyke, I. Jarman, M. Robinson, J. Mason, K. Murphy, M. A. Bellis, and C. Perkins, 2011: The cost of emergency hospital admissions for falls on snow and ice in England during winter 2009/10: A cross sectional analysis. *Environ. Health*, **10**, 60, https://doi.org/10.1186/1476-069X-10-60.

Bloemink, H., 2005: Precipitation type from the Thies disdrometer. *WMO Technical Conf. on Meteorological and Environmental Instruments and Methods of Observation (TECO-2005)*, Bucharest, Romania, WMO, 7 pp., https://www.wmo.int/pages///prog//www/IMOP/publications/IOM-82-TECO_2005/Papers/3(11)_Netherlands_4_Bloemink.pdf.

Böhm, H. P., 1989: A general equation for the terminal fall speed of solid hydrometeors. *J. Atmos. Sci.*, **46**, 2419–2427, https://doi.org/10.1175/1520-0469(1989)046<2419:AGEFTT>2.0.CO;2.

Brabson, B. B., and J. P. Palutikof, 2002: The evolution of extreme temperatures in the central England temperature record. *Geophys. Res. Lett.*, **29**, 2163, https://doi.org/10.1029/2002GL015964.

Brown, I., 2019: Snow cover duration and extent for Great Britain in a changing climate: Altitudinal variations and synoptic-scale influences. *Int. J. Climatol.*, **39**, 4611–4626, https://doi.org/10.1002/joc.6090.

Chen, S., J. J. Gourley, Y. Hong, Q. Cao, N. Carr, P. E. Kirstetter, J. Zhang, and Z. Flamig, 2016: Using citizen science reports to evaluate estimates of surface precipitation type. *Bull. Amer. Meteor. Soc.*, **97**, 187–193, https://doi.org/10.1175/BAMS-D-13-00247.1.

Chernick, M. R., 2011: *Bootstrap Methods: A Guide for Practitioners and Researchers*. John Wiley & Sons, 400 pp.

Clark, J. P. B., S. Solak, Y. H. Chang, L. Ren, and A. E. Vela, 2009: Air traffic flow management in the presence of uncertainty. *Proc. Eighth USA/Europe Air Traffic Management Research and Development Seminar*, Napa, CA, FAA and EUROCONTROL, 571–580.

Cornford, D., and J. E. Thornes, 1996: A comparison between spatial winter indices and expenditure on winter road maintenance. *Int. J. Climatol.*, **16**, 339–357, https://doi.org/10.1002/(SICI)1097-0088(199603)16:3<339::AID-JOC40>3.0.CO;2-U.

Curtis, S., A. Fair, J. Wistow, D. V. Val, and K. Oven, 2017: Impact of extreme weather events and climate change for health and social care systems. *Environ. Health*, **16**, 128, https://doi.org/10.1186/s12940-017-0324-3.

Doolittle, M. H., 1888: Association ratios. *Bull. Philos. Soc. Washington*, **7**, 122–127.

Dotzek, N., P. Groenemeijer, B. Feuerstein, and A. M. Holzer, 2009: Overview of ESSL's severe convective storms research using the European Severe Weather Database ESWD. *Atmos. Res.*, **93**, 575–586, https://doi.org/10.1016/j.atmosres.2008.10.020.

Dufton, D. R. L., and C. G. Collier, 2015: Fuzzy logic filtering of radar reflectivity to remove non-meteorological echoes using dual polarization radar moments. *Atmos. Meas. Tech.*, **8**, 3985–4000, https://doi.org/10.5194/amt-8-3985-2015.

Efron, B., and R. J. Tibshirani, 1994: *An Introduction to the Bootstrap*. CRC Press, 456 pp.

Elmore, K. L., H. M. Grams, D. Apps, and H. D. Reeves, 2015: Verifying forecast precipitation type with mPING. *Wea. Forecasting*, **30**, 656–667, https://doi.org/10.1175/WAF-D-14-00068.1.

Galvin, J., M. Kendon, and M. McCarthy, 2019: Snow cover and low temperatures in February and March 2018. *Weather*, **74**, 104–110, https://doi.org/10.1002/wea.3469.

Gascón, E., T. Hewson, and T. Haiden, 2018: Improving predictions of precipitation type at the surface: Description and verification of two new products from the ECMWF ensemble. *Wea. Forecasting*, **33**, 89–108, https://doi.org/10.1175/WAF-D-17-0114.1.

Gerrity, J. P., Jr., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2709–2712, https://doi.org/10.1175/1520-0493(1992)120<2709:ANOGAM>2.0.CO;2.

Green, A., 2010: From Observations to Forecasts – Part VII. A new meteorological monitoring system for the United Kingdom's Met Office. *Weather*, **65**, 272–277, https://doi.org/10.1002/wea.649.

Greening, K., and A. Hodgson, 2019: Atmospheric analysis of the cold late February and early March 2018 over the UK. *Weather*, **74**, 79–85, https://doi.org/10.1002/wea.3467.

Gunn, R., and G. D. Kinzer, 1949: The terminal velocity of fall for water droplets in stagnant air. *J. Meteor.*, **6**, 243–248, https://doi.org/10.1175/1520-0469(1949)006<0243:TTVOFF>2.0.CO;2.

Haiden, T., A. Kann, C. Wittmann, G. Pistotnik, B. Bica, and C. Gruber, 2011: The Integrated Nowcasting through Comprehensive Analysis (INCA) system and its validation over the eastern Alpine region. *Wea. Forecasting*, **26**, 166–183, https://doi.org/10.1175/2010WAF2222451.1.

Handa, H., L. Chapman, and X. Yao, 2006: Robust route optimization for gritting/salting trucks: A CERCIA experience. *IEEE Comput. Intell. Mag.*, **1**, 6–9, https://doi.org/10.1109/MCI.2006.1597056.

Harrison, D., S. J. Driscoll, and M. Kitchen, 2000: Improving precipitation estimates from weather radar using quality control and correction techniques. *Meteor. Appl.*, **7**, 135–144, https://doi.org/10.1017/S1350482700001468.

——, K. Norman, T. Darlington, D. Adams, N. Husnoo, C. Sandford, and S. Best, 2015: The evolution of the Met Office radar data quality control and product generation system: Radarnet. *37th Conf. on Radar Meteorology*, Norman, OK, Amer. Meteor. Soc., 14B.2, https://ams.confex.com/ams/37RADAR/webprogram/Paper275684.html.

Heidke, P., 1926: Berechnung Des Erfolges Und Der Güte Der Windstärkevorhersagen Im Sturmwarnungsdienst. *Geogr. Ann.*, **8**, 301–349, https://doi.org/10.1080/20014422.1926.11881138.

Hubbert, J. C., M. Dixon, S. M. Ellis, and G. Meymaris, 2009: Weather radar ground clutter. Part I: Identification, modeling, and simulation. *J. Atmos. Oceanic Technol.*, **26**, 1165–1180, https://doi.org/10.1175/2009JTECHA1159.1.

Ikeda, K., M. Steiner, J. Pinto, and C. Alexander, 2013: Evaluation of cold-season precipitation forecasts generated by the hourly updating high-resolution rapid refresh model. *Wea. Forecasting*, **28**, 921–939, https://doi.org/10.1175/WAF-D-12-00085.1.

Kay, A. L., 2016: A review of snow in Britain: The historical picture and future projections. *Prog. Phys. Geogr.*, **40**, 676–698, https://doi.org/10.1177/0309133316650617.

Landolt, S. D., J. S. Lave, D. Jacobson, A. Gaydos, S. Divito, and D. Porter, 2019: The impacts of automation on present weather–type observing capabilities across the conterminous United States. *J. Appl. Meteor. Climatol.*, **58**, 2699–2715, https://doi.org/10.1175/JAMC-D-19-0170.1.

Langleben, M. P., 1954: The terminal velocity of snowflakes. *Quart. J. Roy. Meteor. Soc.*, **80**, 174–181, https://doi.org/10.1002/qj.49708034404.

Locatelli, J. D., and P. V. Hobbs, 1974: Fall speeds and masses of solid precipitation particles. *J. Geophys. Res.*, **79**, 2185–2197, https://doi.org/10.1029/JC079i015p02185.

Lumb, F. E., 1963: Downward penetration of snow in relation to the intensity of precipitation. *Meteor. Mag.*, **92**, 1–14.

Lyth, D., 2008: Results from UK Met Office investigations into new technology present weather sensors. *WMO Technical Conf. on Instruments and Methods of Observation (TECO-2008)*, St. Petersburg, Russia, WMO, 27–29.

——, and M. Molyneux, 2006: Results of using present weather instruments in the United Kingdom. Tech. Rep., Met Office, 13 pp.

Matson, R. J., and A. W. Huggins, 1980: The direct measurement of the sizes, shapes and kinematics of falling hailstones. *J. Atmos. Sci.*, **37**, 1107–1125, https://doi.org/10.1175/1520-0469(1980)037<1107:TDMOTS>2.0.CO;2.

NERC, Met Office, B. S. Pickering, R. R. Neely III, and D. Harrison, 2019: The Disdrometer Verification Network (DiVeN): particle diameter and fall velocity measurements from a network of Thies Laser Precipitation Monitors around the UK (2017-2019). Centre for Environmental Data Analysis

(CEDA), accessed 21 June 2020, https://doi.org/10.5285/602f11d9a2034dae9d0a7356f9aeaf45.

Parker, D. E., T. P. Legg, and C. K. Folland, 1992: A new daily central England temperature series, 1772–1991. *Int. J. Climatol.*, **12**, 317–342, https://doi.org/10.1002/joc.3370120402.

Pickering, B. S., R. R. Neely, and D. Harrison, 2019: The Disdrometer Verification Network (DiVeN): A UK network of laser precipitation instruments. *Atmos. Meas. Tech.*, **12**, 5845–5861, https://doi.org/10.5194/amt-12-5845-2019.

——, R. R. Neely III, J. Jeffery, D. Dufton, and M. Lukach, 2021: Evaluation of multiple precipitation sensor designs for precipitation rate and depth, drop size and velocity distribution, and precipitation type. *J. Hydrometeor.*, **22**, 703–720, https://doi.org/10.1175/JHM-D-20-0094.1.

Punge, H. J., and M. Kunz, 2016: Hail observations and hailstorm characteristics in Europe: A review. *Atmos. Res.*, **176–177**, 159–184, https://doi.org/10.1016/j.atmosres.2016.02.012.

Rasmussen, R., and Coauthors, 2001: Weather Support to Deicing Decision Making (WSDDM): A winter weather nowcasting system. *Bull. Amer. Meteor. Soc.*, **82**, 579–595, https://doi.org/10.1175/1520-0477(2001)082<0579:WSTDDM>2.3.CO;2.

Reeves, H. D., 2016: The uncertainty of precipitation-type observations and its effect on the validation of forecast precipitation type. *Wea. Forecasting*, **31**, 1961–1971, https://doi.org/10.1175/WAF-D-16-0068.1.

Sandford, C., 2015: Correcting for wind drift in high resolution radar rainfall products: A feasibility study. *J. Hydrol.*, **531**, 284–295, https://doi.org/10.1016/j.jhydrol.2015.03.023.

Schmid, W., and A. Mathis, 2004: Validation of methods to detect winter precipitation and retrieve precipitation type. *12th SIRWEC Conf.*, Bingen, Germany, Standing International Road Weather Commission, 8 pp., http://www.sirwec.org/wp-content/uploads/Papers/2004-Bingen/D-27.pdf.

Waldvogel, A., B. Federer, and P. Grimm, 1979: Criteria for the detection of hail cells. *J. Appl. Meteor.*, **18**, 1521–1525, https://doi.org/10.1175/1520-0450(1979)018<1521:CFTDOH>2.0.CO;2.

Wandishin, M. S., M. E. Baldwin, S. L. Mullen, and J. V. Cortinas, 2005: Short-range ensemble forecasts of precipitation type. *Wea. Forecasting*, **20**, 609–626, https://doi.org/10.1175/WAF871.1.

Webb, J. D., D. M. Elsom, and G. T. Meaden, 2009: Severe hailstorms in Britain and Ireland, a climatological survey and hazard assessment. *Atmos. Res.*, **93**, 587–606, https://doi.org/10.1016/j.atmosres.2008.10.034.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

WMO, 1988: Manual on Codes. WMO Publ. 306, 203 pp.

——, 2019: Measurement of precipitation. WMO Guide to Meteorological Instruments and Methods of Observation (the CIMO Guide), Tech. Rep., World Meteorological Organization, 454 pp., https://library.wmo.int/index.php?lvl=notice_display&id=13617#.YKK2eS1Q3ik.

Zikmunda, J., 1972: Fall velocities of spatial crystals and aggregates. *J. Atmos. Sci.*, **29**, 1511–1515, https://doi.org/10.1175/1520-0469(1972)029<1511:FVOSCA>2.0.CO;2.