Adapting the SAL method to evaluate reflectivity forecasts of summer precipitation in the central United States

John R. Lawson*† and William A. Gallus Jr

Abstract

The Structure Amplitude Location (SAL) method was originally developed to evaluate forecast accumulated-precipitation fields through identification and comparison of objects in both the forecast and the observed fields. This study describes a small modification for use with instantaneous composite-reflectivity forecasts, where objects’ minimum size and reflectivity thresholds are prescribed. Both the original and modified SAL methods are used to evaluate daily 0000 UTC 12-km North American Model (NAM) forecasts, against NCEP/EMC 4-km Stage IV accumulated-precipitation estimates, during the summer of 2015 for a central US domain. Results show substantial sensitivity to the reflectivity threshold. This is likely related to sampling more signal from convective cell cores, and progressively ignoring stratiform rain areas, as threshold increases. Setting the threshold too high (40 dBZ) yields only 7% of time periods on which error scores can be computed, as opposed to 94% using a low threshold (5 dBZ). The primary difference between the two methods is a larger structural error in SAL using reflectivity, likely related to the unresolved convective peaks in the 12-km NAM forecasts; this error is smoothed out when accumulated precipitation is evaluated. SAL using reflectivity also reveals a diurnal cycle of skill, with minimum skill occurring around 1800–2200 UTC (early to late afternoon local time, before average convective activity reaches its maximum) and maximum skill occurring around 0600 UTC (just before sunrise). We conclude that both methods yield useful results, but results presented herein may not be generalisable to other verification domains or SAL formulations.

Keywords: model evaluation; mesoscale; modelling; predictability; convection; meteorology

1. Introduction

The Structure Amplitude Location (SAL) method (Wernli et al., 2008; hereby W08) evaluates accumulated-precipitation fields by identifying objects in both a forecast and an observed field at a given time, and decomposing differences (i.e. error) into three components. This procedure avoids a double penalisation for timing and locational errors inherent in verification methods such as root mean square error. The errors are normalised by the size of the domain and domain-wide accumulation such that many cases using the same grid can be compared.

The power of SAL also lies in its ability to evaluate the type of error. The structural component S (between −2 and 2) considers the gradient around the object and its size. For instance, a negative S component may indicate, e.g. too high a radial gradient of the forecasted objects, such as a forecast of convective cells when a stratiform area is observed. The amplitude component A (between −2 and 2) considers domain-wide accumulation. Finally, the location component L (between 0 and 2) consists of two parts. One part (L1) measures location differences in centres of mass for the domain-wide observed and forecast fields; the other part (L2) accounts for location differences of all objects weighted by their integrated precipitation. However, as with all so-called objective schemes, there is a subjective element. SAL scores may on occasion be highly sensitive to the choice of minimum threshold, which occurs when a bimodal distribution of precipitation may or may not be split into two objects rather than one (termed the ‘camel effect’ in W08).

SAL has shown its flexibility in previous studies, such as where a potential vorticity anomaly component replaced S (Madonna et al., 2015), or where SAL was used to evaluate interpolated tracer-plume data sets (Dacre, 2011). Given the connection between radar reflectivity and precipitation accumulation, the authors have refactored SAL for use with composite reflectivity for evaluating forecasts of moist convection in the Great Plains of the United States. This study addresses the differences between the original SAL formulation (hereby SAL_{acpc}) and its slight modification for evaluating instantaneous composite reflectivity (hereby SAL_{crfp}) in Section 2. To compare the two SAL methods over an extended period, North American Model (NAM) forecasts of composite reflectivity and 24-h accumulated precipitation were verified over the central United States with radar observations and multisensor...
estimates of precipitation, respectively. The method and data sources are detailed in Section 3, and results are presented and analysed in Section 4. We discuss interpretation of SAL$_{\text{cref}}$, along with concluding statements, in Section 5.

2. SAL modification

The SAL$_{\text{cref}}$ method has three modified aspects: (1) instantaneous composite reflectivity is used instead of accumulated precipitation; (2) the minimum area of the object (herein termed its footprint) is specified; and (3) the minimum object threshold is explicitly set in dBZ. The latter two changes, while small, were motivated by challenges presented by the noisy nature of the instantaneous reflectivity field, and the nonlinear relationship between precipitation and reflectivity that precludes simple extrapolation of SAL$_{\text{acpc}}$ to a reflectivity field.

SAL$_{\text{acpc}}$ deals with a smoothed field (for instance, precipitation accumulated where frontal systems have traversed), whereas SAL$_{\text{cref}}$ is an instantaneous snapshot of the reflectivity field. The noisier nature of a reflectivity field hence necessitates a minimum footprint. Smaller footprint and threshold parameters yield more, smaller objects in SAL$_{\text{cref}}$ than in SAL$_{\text{acpc}}$. Note the increased likelihood of multiple objects in at least one data set means the L component is more likely to be larger (due to a non-zero $L_L$ component), and increases the potential frequency of the camel effect. As $S$ is computed with the average of all objects’ scaled volume, a large structural error will occur if, e.g. observations have a quasi-linear convective system represented by many strong convective cores joined by weak stratiform rain (e.g. in the trailing stratiform mode) appearing as one object, but if the simulated field has less stratiform precipitation and appears as numerous cell objects.

In addition, it is desirable that the footprint be large enough to ignore radar clutter and clear-air backscatter, but not large enough that growing convective cells suddenly appear or disappear as objects between forecast hours and cause a large step increase or decrease in SAL error. Preliminary testing with object identification on a 4-km grid found a footprint of ~200 gridpoints (3200 km$^2$) subjectively related best to the field, with little variation when the footprint was varied between 100 and 500 gridpoints (1600 and 8000 km$^2$). When a footprint value is not applied, the resultant large number of objects in the composite-reflectivity field degrades the signal-to-noise ratio (i.e. the camel effect occurs more often). Hence, all SAL$_{\text{cref}}$ computations in this study use a footprint of 200 gridpoints (3200 km$^2$).

A constant reflectivity threshold is set to (1) focus further on moist convection as the threshold increases and (2) ignore clear-air backscatter and/or radar clutter. An increase in threshold should not detract from the component’s rationale. Note the $A$ component is not sensitive to the threshold. However, as mentioned before, the camel effect is compounded as the number of objects increases, and $L$ is more likely to be non-zero and/or large. If the threshold becomes too large, SAL$_{\text{cref}}$ ignores an increasingly large number of cases in which the minimum threshold is not reached in either the model or observations, as noted in W08, yielding results biased towards well-forecast cases.

3. Data and methods

We evaluated the 12-km NAM model using 0000 UTC initializations between 1 April 2015 and 31 August 2015 inclusive, verifying 24-h precipitation accumulation between 12 and 36 forecast hours, and composite reflectivity forecasts valid hourly between 12 and 36 forecast hours exclusive (The 36th hour was not included to ensure the first and last hours were not accounted for twice in statistics).

A 4-km grid was arbitrarily defined inside the continental United States (Figure 1) as a common grid to which observations and forecast data were interpolated linearly. (Stage IV and NAM data were interpolated using the NumPy method `scipy.interpolate.griddata`, because of the non-rectangular grid. Composite reflectivity observations, due to the large data set, were interpolated using the more efficient NumPy method `scipy.interpolate.RectBivariateSpline`, with no smoothing (i.e. interpolation only) allowed by the regular latitude–longitude grid of the data set). This is a similar method to W08. As in Barrett et al. (2015), our preliminary testing found negligible sensitivity of the reflectivity and precipitation fields to our re-projection and interpolation methods. We may expect the NAM forecast model – in which convection is parameterised, and afterwards is interpolated to a finer grid – to develop objects too flat ($S > 0$). This may be perceived as a limitation of our methodology, but served as a check for our tests.

Verification of forecasted reflectivity was performed with composite NEXRAD Level III radar reflectivity from archives at the Iowa State University (https://mesonet.agron.iastate.edu/docs/nexrad_composites/, accessed 1 February 2016). Base reflectivity product data are composited through the GEMPAK program `nex2img`, after which suspected false echoes are removed through comparison with the Net Echo Top product. Gridded accumulated-precipitation data sets (NCEP/EMC 4-km Stage IV), created using rain gauge and radar observations, were obtained from the Earth Observing Laboratory (http://data.eol.ucar.edu/).

Six of the 153 days in our period had missing archived NAM forecast data, and one other day had missing accumulated-precipitation data. These days were removed, leaving 146 days for SAL$_{\text{acpc}}$ and 147 days for SAL$_{\text{cref}}$. We ran SAL$_{\text{cref}}$ for four thresholds: 5, 15, 30, and 40 dBZ. As the SAL methodology requires...
identification of at least one object in both the observed and the forecast fields, times or periods that resulted in spurious SAL-component scores (i.e. exactly 0 or ±2) were removed. This did not affect the number of SAL_acpc days, but reduced the 3672 instances of composite reflectivity to 3458, 3408, 2560, and 254 times, for 5-, 15-, 30-, and 40-dBZ thresholds, respectively. The number of ignored instances of composite reflectivity increases with threshold because higher thresholds eliminate more areas of reflectivity, and is more likely to remove an object in that field. This is particularly drastic for the 40-dBZ threshold, where only 7% of times contained objects in both forecast and observational data sets.

4. Results

4.1. Accumulated precipitation (SAL_acpc)

We found that NAM forecasts only weakly overestimate accumulated precipitation on the majority of days in the data set (Figure 2). Objects were too flat, which is likely related to the 12-km horizontal resolution of the NAM forecasts. The positive correlation between $S$ and $A$ components is unsurprising, as discussed in W08, due to the physical relationship between larger objects and larger domain-wide accumulation. There is no obvious relationship between these two components and $L$-component error.

4.2. Composite reflectivity (SAL_cref)

Due to the high volume of data, we focus on the 30-h NAM reflectivity forecast (i.e. 0600 UTC on day 2) as an example here. This is around the time of maximum thunder occurrence for summer months in the central United States (Easterling and Robinson, 1985). At lower thresholds (5 and 15 dBZ; Figure 3(a) and (b)), the positive correlation between $S$ and $A$ components is similar to that for SAL_acpc (Figure 2). However, at 30 dBZ (Figure 3(c)), the line of best fit (not shown) is more parallel with the $x$-axis: the $A$-component error remains positive regardless of $S$ error for almost all points. This suggests the stronger signal ratio from convective cells over stratiform precipitation results in positive $A$ error not simply because of the size of objects, but from radial reflectivity gradients in the objects. As discussed in W08, positive $S$ and negative $A$ (bottom right quadrant) can occur when a forecast misses an observed convective cell. We note the $S–A$ relationships shown here, valid at forecast-hour 30 for all four thresholds, are consistent throughout the whole 24-h period. We also find the variation in all three components, represented by the ‘spread’ of points and their colours in Figure 3, is smallest around 24 forecast hours (0000 UTC) and largest between 30 and 36 forecast hours (0600 and 1200 UTC; not shown). In other words, systematic errors in simulated composite reflectivity dominate SAL statistics in the early night period, while random errors dominate towards sunrise.

Median $S$-component error, denoted by the black vertical broken line in Figures 2 and 3, is similar (0.6–0.7) in both SAL_acpc (Figure 2) and SAL_cref (Figure 3), but only when the threshold of the latter is set at 15 dBZ or higher. However, despite the larger $S$ component when using 5-dBZ threshold (Figure 3(a)), the $A$ component is around the same as at other thresholds (~0.5; Figure 3(b)–(d)). This suggests that weak stratiform (<15 dBZ) precipitation areas are forecasted too large in area coverage and too weak in magnitude.

When median $S$ and $A$ components are plotted for each hour over the whole data set, we find a diurnal loop as shown in Figure 4. Note 40 dBZ is not discussed here.
Adapting SAL to evaluate reflectivity forecasts in the United States

2.0
1.5
1.0
0.5
0.0
–0.5
–1.0
–1.5
–2.0
0.0
0.25
0.5
0.75
1.0
1.5
2.0

Figure 2. SAL scores of 24-h accumulated precipitation, valid at 36 forecast hours (i.e. 1200 UTC on day 2), for all days in summer 2015. Scatter points involving spurious scores (0 or ±2), resulting from insufficient precipitation during the period, are not shown. Each scatter point is coloured by its location component (see inset). The white box spans the middle two quartiles of structural (x-axis) and amplitude (y-axis) components. Dotted line denotes the median structure- and amplitude-component scores. 

| Location component |
|---------------------|
| 0.0 | 0.25 | 0.5 | 0.75 | 1.0 |
| Structural component |
| –0.5 | 0.0 | 0.5 | 1.0 | 1.5 | 2.0 |

because of its small sample size and noisy nature. For all three thresholds (5, 15, and 30 dBZ), the S and A components increase from 1200 UTC; the trajectories reach their maximum A error at 1800 UTC and maximum S error at 2200 UTC (2100 UTC for 30 dBZ). While diurnal differences in S and A components are O(0.5), L differences are an order of magnitude smaller (not shown). The L component varies similarly to the other two: all thresholds have their largest L-component error at 1800 UTC, and this error reduces through the course of the night. As with S and A, values of L also retreat towards the origin (i.e. better forecasts) with increasing threshold; the 24-h minimum occurs at 1000 UTC. Note the three trajectories terminate slightly further from the origin than their initial points (i.e. 23 h previously), representing the decrease of forecast skill with time if extrapolated to the 24th hour.

These times of maximum A/L- and S-component errors are around 1 and 5 pm, respectively, for local time in most of the domain (Central Daylight Time; UTC-5). The earlier peak in A may represent forecasted cell initiation that grows too quickly, while the later peak in S may be related to upscale growth (forecast reflectivity objects are too stratiform). The progression of the trajectories towards the origin suggests an increase in forecast skill towards and after sunset, as diurnal convection decays and nocturnal systems develop. This signal that nocturnal systems are better forecast may be due to the propensity of mesoscale convective systems to occur at night (Markowski and Richardson, 2010 and references therein), whose length scales are larger than (typically daytime) single-cell convection, and whose predictability is therefore theoretically larger (Lorenz, 1969; Clark et al., 2007). Yan and Gallus (submitted to Monthly Weather Review) found NAM forecasts of precipitation to be more skillful between midnight and early morning, and least skillful near noon. While this corroborates results presented here, we note that our location error (L) is an order of magnitude smaller than structural (S) and amplitudinal (A) error, whereas displacement error was the main source of low forecast skill in Yan and Gallus. However, as SAL components are normalised, we expect L to be small because of our large domain size.

As each SAL component is normalised (i.e. by how wrong a forecast could possibly be in each component), the authors propose that the total of absolute SAL component values (taSAL) at each time

© 2016 The Authors. Atmospheric Science Letters published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.
or time period allows an estimate of forecast skill (\(taSAL = |S| + |A| + |L|\)). Ultimately subjective in nature because of its formulation, a skill score may be considered an objective method to reflect how the human eye subjectively perceives numerous error components: quantifying how the total error limits the utility of a forecast. For example, the Model Evaluation Tool (MET: http://www.dtcenter.org/met/users/) software package Method for Object-Based Diagnostic Evaluation (MODE) weights its error components subjectively and empirically, each potentially with different values. Herein, we propose a simple combination of equal weighting to all three components as a starting point. A different weighting combination should be tested in subsequent studies. The median taSAL values at each forecast time are shown in Figure 5(a) for 5-, 15-, and 30-dBZ reflectivity thresholds. The decrease of error with increased dBZ threshold may be related to a smaller area of variation in scaled volume to occur within each object, and objects restricted to convective cores, both of which lower the ‘area of freedom’ for potential dBZ values. Additionally, as noted in Section 2 and W08, the smaller error could simply be due to the increasing neglect of missed events and false alarms as the threshold is raised. Note, given that stratiform precipitation is increasingly ignored with larger thresholds associated with lower error, more power in the SAL signal is given to less predictable convective precipitation (Zhang et al., 2003). Hence, one may have instead expected reduced skill of convective-object forecasts.

We also present in Figure 5(b) the average percentage of pixels in both the observed (darker shades) and forecast (paler shades) fields that are active, i.e. above the thresholds denoted by the same colour scheme.

Figure 3. As Figure 2, but for hourly composite reflectivity SAL scores at 30 forecast hours (0600 UTC), for four thresholds: (a) 5 dBZ, (b) 15 dBZ, (c) 30 dBZ, and (d) 40 dBZ. As in Figure 2, scatter points are omitted if they contain a spurious value.
Adapting SAL to evaluate reflectivity forecasts in the United States

Figure 4. Trajectories of SAL through structure–amplitude component space for the full 24-h period (12–36 forecast hours) for three thresholds: 5 dBZ (orange), 15 dBZ (light blue), and 30 dBZ (green). The coloured square denotes the start of the trajectory (i.e. 12 forecast hours; 1200 UTC on day 1). The line decreases in transparency as time progresses; each line’s terminus denotes 36 forecast hours (i.e. 1200 UTC the next day). Note the 40-dBZ threshold was omitted for clarity because of its noisy nature, and the plot axes are zoomed into the top right quadrant from Figures 2 and 3 to show detail.

Figure 5. (a) Median total absolute SAL error for 5-, 15-, and 30-dBZ thresholds (coloured according to legend) and (b) average active pixels in observed (darker shades) and forecast (lighter shades), for each forecast hour in the composite-reflectivity data set. Active pixels are defined as those belonging to an object at the given threshold and footprint at a given time. The vertical black line in panel (a) denotes 0000 UTC on day 2 (forecast hour 24), for reference.
and belonging to an object larger than the prescribed footprint. The average is computed for each hour over the whole summer, and because of the propensity of thunderstorms within our chosen domain, serves as a proxy for convective activity. Figure 5(b) shows that peak activity occurs at 2300 UTC (6 pm local time) in simulated data, and at 0400 UTC (11 pm local time) in the observed data, with both displaying a similar sinusoidal pattern as in taSAL (Figure 5(a)). The 5-h difference in convective maxima between observed and forecast fields suggests that forecast error could well contain a large timing component. Furthermore, the peak taSAL error occurs earlier – while observed and forecast convective activity is growing – suggesting that forecast error may develop within growing late afternoon convection.

5. Conclusions

This study presented modifications to the original SAL methodology (SALacpc) to verify composite reflectivity fields (SALcref), instead of accumulated precipitation. We evaluated NAM forecasts for a summer (April–August inclusive) season in the central United States with both SAL methods to gauge the modifications’ impact. The two methods draw different conclusions from their respective fields. SALcref demonstrated a larger positive S-component error, likely related to the inability of the convection-parameterising forecast model to resolve peak maxima associated with convective cells, and the smoother nature of the accumulated-precipitation field in SALacpc. The positive correlation between S and A components is expected because of the physical relationship between object size and domain-wide composite reflectivity. However, this correlation is not apparent when the minimum threshold of SALcref is raised to 30 dBZ. SALcref reveals a diurnal cycle of skill, with forecasts best in the early morning and worst around noon, with similar patterns in all three SAL components. The lag in convective activity after the largest total SAL error suggests that timing and simulated character of developing convection may be responsible for a substantial proportion of forecast error.

These results show the need to set a threshold and footprint small enough to give a sufficient sample size, but large enough to capture the signal of interest – be it convective or stratiform in nature. Further work may investigate whether SAL scores of hourly maximum composite reflectivity forecasts show less sensitivity to threshold, because of the smoother nature of that field. Our results also reiterate that interpretation of SAL must be isolated to the threshold and field chosen. For instance, use of a large domain reduces the impact of the L component as object displacements are normalised by the diagonal length. Further investigation is needed into the optimal weighting of SAL components to create a skill score. A remaining limitation of the SAL method relating to moist convection is its inability to measure error in orientation of objects. The authors are aware of the MET package MODE, another object-based evaluation system that considers orientation, but which lacks some simplicity and portability of SAL. A fourth component that considers the mode and orientation of convection may improve SAL’s utility for reflectivity fields.

Acknowledgements

All statistical processing was performed with the NumPy Python package, and figures were produced with the matplotlib Python package. We thank the editor and two anonymous reviewers for improving the rigour and clarity of this article.

References

Barrett AJ, Gray SL, Kirshbaum DJ, Roberts NM, Schultz DM, Fairman JG. 2015. Synoptic versus orographic control on stationary convective banding. Quarterly Journal of the Royal Meteorological Society 141: 1101–1113.

Clark AJ, Gallus WA, Chen T-C. 2007. Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. Monthly Weather Review 135: 3456–3473.

Dacre HF. 2011. A new method for evaluating regional air quality forecasts. Atmospheric Environment 45: 993–1002.

Easterling DR, Robinson PJ. 1985. The diurnal variation of thunderstorm activity in the United States. Journal of Applied Meteorology and Climatology 24: 1048–1058.

Lorenz EN. 1969. The predictability of a flow which possesses many scales of motion. Tellus 21: 289–307.

Madonna E, Boettcher M, Grams CM, Joos H, Martius O, Wernli H. 2015. Verification of North Atlantic warm conveyor belt outflows in ECMWF forecasts. Quarterly Journal of the Royal Meteorological Society 141: 1333–1344.

Markowski P, Richardson Y. 2010. Mesoscale Meteorology in Mid-latitudes. Wiley-Blackwell: Chichester, UK.

Wernli H, Paulat M, Hagen M, Frei C. 2008. SAL – a novel quality measure for the verification of quantitative precipitation forecasts. Monthly Weather Review 136: 4470–4487.

Zhang F, Snyder C, Rotunno R. 2003. Effects of moist convection on mesoscale predictability. Journal of the Atmospheric Sciences 60: 1173–1185.