ABSTRACT

The feature-based spatial verification method named for its three score components: structure, amplitude, and location (SAL) is applied to cloud data, that is, two-dimensional spatial fields of total cloud cover and spectral radiance. Model output is obtained from the German-focused Consortium for Small-Scale Modeling (COSMO-DE) forward operator Synthetic Satellite Simulator (SynSat) and compared with SEVIRI satellite data. The aim of this study is twofold: first, to assess the applicability of SAL to this kind of data and, second, to analyze the role of external object identification algorithms (OIA) and the effects of observational uncertainties on the resulting scores. A comparison of three different OIA shows that the threshold level, which is a fundamental part of all studied algorithms, induces high sensitivity and unstable behavior of object-dependent SAL scores (i.e., even very small changes in parameter values lead to large changes in the resulting scores). An in-depth statistical analysis reveals significant effects on distributional quantities commonly used in the interpretation of SAL, for example, median and interquartile distance. Two sensitivity indicators that are based on the univariate cumulative distribution functions are derived. They make it possible to assess the sensitivity of the SAL scores to threshold-level changes without computationally expensive iterative calculations of SAL for various thresholds. The mathematical structure of these indicators connects the sensitivity of the SAL scores to parameter changes with the effect of observational uncertainties. Last, the discriminating power of SAL is studied. It is shown that—for large-scale cloud data—changes in the parameters may have larger effects on the object-dependent SAL scores (i.e., the $S$ and $L^2$ scores) than does a complete loss of temporal collocation.

1. Introduction

Verification of numerical model output is essential in the development of successful models for numerical weather prediction. Because of an increase in model resolution, new techniques for the evaluation of spatial fields have emerged during the last decade (see, e.g., Casati et al. 2008; Gilleland et al. 2009; Ebert 2008). Feature-based methods are an important part of this toolkit. These methods use score functions that are defined on objects, not on the spatial field itself (i.e., on a subset of the spatial data usually identified by some external algorithm). Most feature-based methods have been designed with a specific field of application in mind. Verification of precipitation fields is the most prominent application, and various methods have been developed for this kind of spatial data, for example, contiguous rain area (Ebert and McBride 2000; Ebert and Gallus 2009); the method for object-based diagnostic evaluation (Davis et al. 2006a,b); and the structure, amplitude, and location (SAL) method (Wernli et al. 2008).

SAL was developed to measure the quality of a forecast using three distinct scores, which have direct physical interpretations to allow for conclusions on potential sources of model errors. It does not require matching individual objects in observations and forecasts but compares the statistical characteristics of those fields. The resulting scores are close to a subjective visual assessment of the accuracy of the forecast for precipitation data. SAL was originally developed for the verification of precipitation fields in a defined area (e.g.,
river catchments), and today it is widely used in the
evaluation of quantitative precipitation forecasts (e.g.,
2011). Recently, efforts to apply SAL to different kinds
of data have been made: Shi et al. (2014) used SAL for
the evaluation of a soil moisture model and Crocker
and Mittermaier (2013) applied SAL on binary
cloud masks.

The aim of this work is twofold: first, to assess the
benefits and drawbacks of SAL applied on cloud data;
second, to systematically study the role of the object
identification algorithms (OIA) and their parameters.
Spatial fields that describe cloud and precipitation
(CP) processes such as total cloud cover or spectral
radiance may contain large-scale structures. A focus
of this study is thus to investigate how well SAL is able
to deal with large features, and to quantitatively an-
alyze the effect of different OIA parameter settings.
Wernli et al. (2008) investigated the so-called camel
analyze the effect of different OIA parameter settings.
Wernli et al. (2008) investigated the so-called camel
cases on a qualitative level and showed that even small
changes in the threshold level of the OIA can lead to
very different SAL scores. We follow this line of
thought and conduct an extensive statistical analysis
of large sets of spatial cloud data to quantify the
sensitivity of SAL toward three parameters: threshold
ration, smoothing radius, and minimal object size.
Since substantially different threshold levels corre-
respond to different physical situations, we expect the
resulting scores to be different as well. This is true not
only for SAL but virtually any threshold-based veri-
fication method, such as the fractions skill score
(Roberts and Lean 2008) or the intensity-scale skill
score (Casati et al. 2004; Casati 2010). The interesting
question is how the scores react to very small changes
in parameter values, that is, whether the verification
score is numerically stable with respect to its OIA
parameters.

The effect of small perturbations in parameter
values is closely linked to the effect of observational
uncertainties, that is, small perturbations in the data
itself. Observational uncertainties are generally ig-
nored in spatial verification methods (Ebert et al.
2013 and references therein), which might be justi-
fied if observational errors are small relative to model
errors. However, this assumption is not true for re-
motely sensed estimates of variables (e.g., estimates
derived from radar or satellite observations), particu-
larly those related to CP processes. While the in-
strument errors of direct satellite measurements such
as spectral radiance or brightness temperature are
small, this is not true for derived quantities such as
cloud fraction or cloud masks (Zinner et al. 2005; Crocker
and Mittermaier 2013). Additional uncertainties enter
the verification process in the form of spatial inter-
polation due to the discrepancy between the model
grid and the (usually irregular) observational grid.
The evaluation of CP processes in high-resolution
model simulations strongly relies on this kind of re-
motely sensed observations (Evaristo et al. 2014;
Steinke et al. 2015; Hammann et al. 2015; Nam et al.
2014; Eggert et al. 2015). Therefore, it is important to
understand the behavior of SAL with respect to ob-
servational uncertainties.

The article is structured as follows: We first provide
the mathematical definitions of SAL in section 2. Three
different OIA and conceptual scenarios, which allow us
to identify focal points for the analysis of SAL’s pa-
parameter sensitivity, are discussed in section 3. These
points are explored with exemplary cases and an in-
depth statistical analysis using spatial data of total cloud
cover and spectral radiance in section 4. The threshold
parameter is of particular importance, since it is the basis
of all three OIA, is closely connected to observational
uncertainties, and impacts not only SAL but other
threshold-based verification techniques. A priori and a
posteriori indicators, which provide a computationally
effective way to asses SAL’s sensitivity to varying
thresholds, are discussed in section 5. The insights
earned from the mathematical formulation of these in-
dicators are used to establish the link between param-
eter sensitivity and observational uncertainties. Section
6 investigates the ability of the object-dependent SAL
scores to distinguish between two different sets of
data.

2. Definition of SAL

Let us consider a two-dimensional domain
\( D \subseteq \mathbb{R}^2 \)
composed of \( N \in \mathbb{N} \) grid points with a maximal diameter
\[
d := \sup_{(x,y) \in D} |x - y| > 0.
\]

We now want to evaluate one set of spatial data \( R_i \) on the
domain \( D \) with respect to a second set of data \( R_j \). To this
end, for each field \( R_i \), we define \( n_i \in \mathbb{N} \) objects \( O_{i,k} \subseteq D \)
with \( k \in \{1, \ldots, n_i\}, i \in \{1, 2\} \) using some OIA. The OIAs
are discussed later in section 3. Based on the defined
objects \( O_{i,k} \) the three components of SAL are defined.
Amplitude is
\[
A = \frac{\langle R_1 \rangle_D - \langle R_2 \rangle_D}{0.5(\langle R_1 \rangle_D + \langle R_2 \rangle_D)} \in [-2, 2],
\]
where \( \langle \cdot \rangle_D \) denotes the average over the domain \( D \). A
perfect score of \( A = 0 \) indicates that \( R_1 \) is unbiased with
respect to \( R_2 \). In the case of \( A = 1 \), the spatial-data set \( R_1 \)
is overestimated by a factor of 3, whereas $A = -1$ means that $R_1$ is underestimated by a factor of 3.

For the definition of location, let us denote the center of total mass for the field $R_i$ by $x_i \in \mathbb{R}^2$, for $i \in \{1, 2\}$, the center of mass of each object $O_{i,k}$ by $x_{i,k}$ and the mass of each object by $M_{i,k} \in \mathbb{R}$, for $k \in \{1, \ldots, n_i\}$ $i \in \{1, 2\}$. The $L$ scores are defined as

$$L_1 = \frac{|x_1 - x_2|}{d} \in [0, 1],$$

$$L_2 = 2 \frac{|V_1 - V_2|}{d} \in [0, 1], \quad \text{and}$$

$$L = L_1 + L_2 \in [0, 2].$$

The first $L$ score $L_1$ describes the relative distance between the centers of total mass $x_1$ and $x_2$. The second $L$ score $L_2$ is a measure for the scattering of the identified objects. Since both $L$ scores are fully defined by the centers of total mass and centers of object’s mass, $L$ is rotation invariant, that is, rotating the whole field or an object around its center of mass does not change the $L$ score. For $L = 0$ we have a perfect location match of all centers of mass.

To define structure,

$$V_{i,k} = \frac{M_{i,k}}{\max_{x \in O_{i,k}} \overline{R}_i(x)} \quad \text{scaled mass of object } O_{i,n},$$

$$V_i = \frac{\sum_{k=1}^{n_i} M_{i,k} V_{i,k}}{\sum_{k=1}^{n_i} M_{i,k}} \quad \text{scaled, weighted total mass},$$

and

$$S = \frac{V_1 - V_2}{0.5(V_1 + V_2)} \in [-2, 2],$$

where $V_{i,k}$ is the mass of the object $O_{i,k}$ after its maximal height has been rescaled to 1. The scaled total mass $V_i$ is the weighted and normalized sum over the rescaled masses. The intent of the rescaling is to remove, or at least dampen, the influence of total mass and concentrate on the structure of the objects. An $S$ score of $S = 0$ is obtained for a perfect match for the structures of all objects in both datasets. If $S < 0$, the objects of $R_1$ are too peaked in comparison with those of $R_2$, whereas a positive $S$ score, $S > 0$, implies they are too flat. For visualizations of the SAL properties the reader is referred to Wernli et al. (2008).

3. Object identification algorithms

One central component of SAL is the identification of objects. While $A$ and $L1$ scores are independent of the object identification, since they are directly defined on the $R_i$ fields, $L2$ and $S$ are defined on the sets of objects $O_{i,k}$. As there exist a variety of OIA, which in turn require the specification of parameters, it is imperative to understand the sensitivity of the SAL scores to the choice of the OIA. We start this study with the discussion of some conceptual cases. These theoretical considerations show potential issues, which may lead to very unstable responses of SAL with respect to small changes in parameter values. We will study the parameter sensitivity of three different OIA, which are described in the following section, by looking at the behavior of the object-dependent scores $S$ and $L2$.

To identify cohesive objects in spatial data, OIA typically use a threshold level and define a continuous set of points of threshold exceedances as one object. To filter small-scale noise, many methods apply a smoothing filter prior to object identification or ignore objects smaller than a predefined number of points. From the multitude of existing OIA we apply three methods implemented in the R package SpatialVx (Gilleland 2014).

The OIA “threshfac” is the algorithm originally used with SAL (Wernli et al. 2009). It defines an object as a cohesive set of threshold exceedances. The threshold is defined as $R^f_i = f \times R^{95}_i$, where $R^{95}_i$ is the 95% quantile of the field $R_i$, $i \in \{1, 2\}$, and $f > 0$ a threshold ratio with a default value of $f = 1/15$. This simplistic approach has the advantage that its only parameter, the threshold level, has a direct physical interpretation. The lack of smoothing or filtering makes this method susceptible to the effect of small-scale noise, which might lead to an unrepresentative dominance of very small, scattered objects. This issue is addressed by the following OIA.

The convolution threshold algorithm “convthresh” (Davis et al. 2006a,b) identifies objects in two steps. First, the data fields are convolved with a smoothing process, that is, the value at each grid point is replaced by the mean value over a disc with a radius given by the parameter smoothing. Second, the convolved data is thresholded yielding a binary mask, which in turn is applied to the original fields. The resulting objects thus have the original values at each grid point but smoothed boundaries. The advantage is that the borders of the objects are smooth similar to those a human would draw manually. The method filters out small-scale noise (i.e., small scattered objects), which is either isolated or located at the borders of large objects. The drawback is the introduction of an additional parameter, the smoothing

Unauthenticated | Downloaded 04/01/24 02:40 PM UTC
radius, which has no direct physical interpretation. Therefore it is not obvious how to choose this parameter for a given set of data.

The algorithm “threshsizer” (Nachamkin 2009) defines objects as cohesive threshold exceedances, where objects consisting of less than NContig grid points are omitted. This method is used to filter out small isolated objects. The interpretation of the NContig parameter is more straightforward than the smoothing radius in the previous OIA, but has no effect on the shape of large objects. Hence, it is easier to foresee the consequences of a particular choice of parameter value, but the objects do not look as natural as the ones provided by convthresh.

Let us now consider conceptual cases on a $4 \times 4$ grid with values of different intensity. Figure 1a shows the effect of a varying threshold level. Depending on the threshold level, objects of lower intensity are either identified or ignored. The presence or absence of the lower-intensity objects influences $L_2$ and $S$. On the one hand, this is a deliberate effect, since different threshold levels concentrate the analysis to different physical situations. However, this effect might become problematic when the threshold is close to the lower-intensity value, since an arbitrarily small change in the threshold level may cause the whole object to vanish and lead to a potentially large change in the $L_2$ and $S$ scores.

The effect of the smoothing radius is illustrated in Fig. 1b. Smoothing may cause a low-intensity bridge to fall below the threshold. In our example, smoothing is achieved by averaging over a $3 \times 3$ window. The resulting sets of objects differ in average spread and structure, which in turn leads to changes in the $L_2$ and $S$ scores. If such a bridge is very narrow or its value is close to the threshold level, then even small changes in the strength of the smoothing may have large effects on the $L_2$ and $S$ scores.

In contrast, the effect of the minimal object size parameter (i.e., the Ncontig of the threshsizer algorithm) only affects small objects, regardless of the intensity of the values (Fig. 1c). Since all object-dependent SAL scores are weighted with the mass of the objects, the effect on the scores should be small for small changes in the parameter value. Extreme cases, where for example, only two small objects are present and one of them vanishes because of a slightly raised Ncontig parameter, may be thought of but are very unlikely to occur for actual data with hundreds or thousands of grid points.

Let us consider a conceptual setting (Fig. 2), which allows for an easy calculation of the $L_2$ score. In this setting, the middle grid point has a value equal to the threshold. An arbitrarily small change in the threshold level yields a change in the $L_2$ score from the optimal to the worst possible score. In scenario A the $L_2$ score amounts to 1, whereas in scenario B forecast and observation are identical with a perfect score of $L_2 = 0$. This example demonstrates that there exist situations in which the $L_2$ score is unstable, that is, it changes from the best to the worst $L_2$ value for arbitrarily small parameter changes in the OIA.

4. Analysis of parameter sensitivity

To investigate the sensitivity of SAL regarding cloud processes, we use model data from the Synthetic Satellite Simulator (SynSat; Keil et al. 2006) implemented in the German-focused operational regional weather prediction model of the Consortium for Small-Scale Modeling (COSMO-DE) at the Deutscher Wetterdienst (DWD; German weather service), which computes synthetic spectral radiances and brightness temperatures for eight channels of the Meteosat Second Generation satellite (MSG; Crewell et al. 2008). The model
has a horizontal resolution of 2.8 km and covers a domain with 421 $\times$ 461 grid points containing Germany, Switzerland, and Austria. For each day, the forecast is initiated at 0000 UTC and yields synthetic satellite data with a temporal resolution of 15 min. As observations we use data from the Spinning Enhanced Visible and Infrared Imager (SEVIRI) instrument of the MSG satellite (Crewell et al. 2008; Reuter et al. 2009) for a domain of 302 $\times$ 202 grid points with a maximal horizontal resolution of 3 km.

Nine different variables are studied: total cloud cover (TCIC) and eight channels of spectral radiance. TCIC is derived for the observational data as the fraction of cloudy pixels in a grid box using the NWC SAF MSG v2010 algorithm, which has been developed by the Satellite Application Facility for supporting nowcasting and very short-range forecasting (SAFNWC). The algorithm is based on a multispectral thresholding technique (Derrien and Le Gléau 2005, 2010). The COSMO model uses a parameterization based on relative humidity in its radiation scheme and a statistical cloud scheme (Sommeria and Deardorff 1977) within the turbulence model to parameterize boundary layer clouds (Schättler et al. 2013). We have chosen spectral radiance over brightness temperature to study parameter sensitivity, since it allows us to use SAL in its original formulation (implemented in the R package SpatialVx). For brightness temperature, the definition of threshold levels based on the 95% quantile is problematic because the minimal value of the fields is far greater than zero. Therefore, only a very small range of threshold ratios around $f = 1$ would yield sensible thresholds. Thus, by using spectral radiances, we avoid additional choices how to normalize the data or change the threshold routine and can concentrate on the effects of different OIA and their parameters. If one is primarily interested in the direct verification results (e.g., to evaluate a specific model setup) and not in a technical analysis of the verification method itself, this decision should be revisited. In this case it would be interesting to compare verification scores derived from spectral radiance and brightness temperature, which essentially describe the same physical quantity. Because of the strictly monotone increasing relation between brightness temperature and spectral radiance based on Planck’s law, a one to one conversion of all data points and thresholds would not change the results of the sensitivity study. We refer to the technical reports for SEVIRI for more details (EUMETSAT 2012a,b).

The results shown concentrate on the spectral radiance at IR6.2, that is, the water vapor band at a wavelength of 6.2 $\mu$m. For each variable, we compare observed and synthetic IR6.2 radiance values every 3 h (starting at 0000 UTC) between 1 January 2012 and 19 February 2012, resulting in 400 pairs of spatial fields. Note that each set of 400 spatial fields includes forecasts of eight different lead times, which have a large impact on the verification scores. However, we are not interested in absolute SAL values, but rather in the difference of two SAL values calculated for the same fields with different OIA parameter settings. Examples in form of two case studies are stated below. The consideration of different lead times allows us to cover the whole range of small and large SAL values for the study of parameter sensitivity. Since SAL requires all the data to be on the same grid, the model output was interpolated onto the shared area (Fig. 3) of the coarser observational grid using a straightforward nearest-neighbor method. While the effect of different interpolation methods—for example, bilinear, weighted, spline-based, or via kriging (Li and Heap 2008)—may have very significant effects on verification scores, a systematic statistical analysis is out of the scope of this work. To focus on the study of different OIA parameter settings, we use the computationally least expensive interpolation method.

FIG. 2. Conceptual case demonstrating a potential effect of small changes in parameter value. Black squares indicate points with high intensity, while gray squares denote an intensity value near the threshold level. Obs. A yields the worst possible $L_2 = 1$ score, while obs. B is a perfect match with $L_2 = 0$. 

Unauthenticated | Downloaded 04/01/24 02:40 PM UTC
Before we explore the statistical consequences different OIA and parameter settings have on SAL, let us consider two exemplary cases in which the conceptual processes from section 3 can be observed for meteorological data. Both cases exhibit large changes in $S$ and $L_2$ scores due to small changes in the parameters of the OIA. Following the line of thoughts established in section 3, these object decomposition processes may occur for the convthresh and thresfac algorithm. We distinguish between two different types of object decomposition: one where the (spatial) shape of a large object is the deciding factor and one where the intensity structure of an object is the most important criterion.

The first type, which is shown on a conceptual level in Fig. 1b, is responsible for most of the large deviations when using the convthresh OIA. Figure 4 shows a case for IR6.2 using the convthresh algorithm with a smoothing radius of 0 and 1, respectively. Without smoothing the OIA identifies one dominating object in both observation and forecast. Although these do not match perfectly, they are very similar, which leads to small scores of $S = 0.12$ and $L_2 = 0.04$. Using the smallest possible smoothing radius of 1 grid point causes the small interconnecting bridge in the center of the object in the observations to vanish. This leads to the decomposition of the dominant object into two large

![Fig. 4](image_url)
ones. Since the dominant object in the forecast is unaffected, $S$ and $L^2$ scores exhibit large changes: the object in the forecast is too large, which results in a large positive structure score ($S = 0.72$). The spread of objects is too small in the forecast resulting in a large $L^2$ score ($L^2 = 0.44$). In this case, the shape of the bridge is crucial; that is, it has to be thin to vanish because of smoothing, while its intensity values are only of secondary importance.

The second type of object decomposition is shown on a conceptual level in Fig. 1a. In real meteorological data the situation is not as palpable most of the time, but the defining aspect that a large part of an object’s mass falls below a varying threshold level can be clearly observed. Figure 5 shows a case for IR6.2 where the threshfac algorithm is applied with threshold ratios of 0.9 and 1, respectively. For the lower threshold, the forecast is dominated by a large elongated object. Raising the parameter causes most of this object to fall below the now higher threshold level. The remaining mass is then identified as a cluster of smaller objects. The objects in the observations become smaller but are otherwise unaffected. For the lower threshold the dominant object in the forecast is too large and thus responsible for a large positive $S$ score ($S = 1.4$). The situation is reversed for the higher threshold: the structure of the clustered small objects in the forecast is too small, which leads to a large negative $S$ score ($S = -1.11$). While the effect on the $L^2$ score is small in this example ($L^2_{\text{fac}=0.9} = 0.05$ and $L^2_{\text{fac}=1} = 0.18$), we have observed other cases where it exhibits large changes. The unpredictable behavior of the $L^2$ score is one reason for the low correlation between absolute changes in $S$ and $L^2$ scores for the threshfac OIA, which will be discussed at the end of section 4.

The exemplary cases show that object decomposition and the resulting large changes in SAL scores can be caused by small changes in parameter values of the OIA. Let us now take closer look at the statistical effects on the distribution of SAL scores over large datasets of $N = 400$ pairs of spatial fields $R_{ij}$, where $i \in \{1, 2\}$ denotes forecast and observation, and $j \in \{1, \ldots, 400\}$ denotes the

![Fig. 5. Case study: object decomposition for the threshfac algorithm for IR6.2 (1200 UTC 21 Jan 2012). For a threshold ratio of 0.9, one large object dominates the forecast. Raising the threshold ratio to 1 causes most of the object’s mass to fall below the threshold level. Effectively the object decomposes into many small ones, which leads to vastly different $S$ scores but nearly constant $L^2$ scores.](image-url)
temporal index. We denote SAL as maximum stable with respect to a parameter $p$ of an OIA, if small changes in the value of this parameter ($\Delta p = p_2 - p_1$) can only cause small changes in the resulting SAL scores ($\Delta SAL$), that is

$$\max_{j \in \{1, \ldots, N\}} |\Delta SAL(R_{1,j}, R_{2,j}, p_1, p_2)| \leq C|\Delta p|,$$

for a constant $C > 0$. SAL is mean stable with respect to a parameter $p$, if there exists a constant $C > 0$ with

$$\frac{1}{N} \sum_{j=1}^{N} |\Delta SAL(R_{1,j}, R_{2,j}, p_1, p_2)| \leq C|\Delta p|.$$

While maximal changes in SAL scores represent worst-case scenarios, the mean value of score changes presents a starting point to a distributional analysis of parameter sensitivity. We denote SAL as maximum unstable or mean unstable with respect to a parameter of an OIA, if no bounding constants exist; that is, small changes in the parameter value can lead to large changes, or even unbounded responses, in the resulting SAL scores.

Figures 6–8 show the responses of $L2$ and $S$ scores to parameter changes of the OIA for IR6.2. For stable parameters we expect a linear decrease in (maximal and mean) absolute score differences for decreasing differences in parameter value.

Figure 6 shows that the threshold ratio, which varies for the threshfac algorithm, clearly induces maximum-unstable and mean-unstable SAL scores. The maximum of absolute differences in $L2$ score is as high as 1.0. As discussed for the conceptual example this corresponds to the difference between the best and the worst score possible. This holds true for the $S$ score as well, with a maximum difference of 2.5. Naturally, the effect on the mean value is smaller but still very significant at about 0.2 for the $L2$ score and 0.4 for the $S$ score.

The convthresh algorithm with a varying smoothing radius and a fixed threshold ratio is studied in Fig. 7. Here, the results are more complex: first, we would like to point out the change in scale on the vertical axis relative to Fig. 6. And second, while the maxima of absolute score differences indicate a maximum-unstable behavior, the mean values indicate mean-stable SAL scores. Largest $L2$ and $S$ score differences are of about 0.6. These are smaller than in Fig. 6 by a factor of 1.7 for the $L2$ score and 4 for the $S$ score. The differences in the mean values are smaller by a factor of 20 for both scores. Conclusively, the worst cases for varying smoothing radii are less severe and occur less frequently than for the threshold ratio of the threshfac algorithm. Note, however, that a different behavior might be obtained with another dataset. Whether SAL can be regarded as stable with respect to changes in the smoothing radius depends on two aspects: first, the data, and second, the question one wishes to address with the SAL verification. If latter includes the interpretation of quantiles other than the median (e.g., the interquartile range), a more elaborate statistical analysis is necessary.

Figure 8 shows the same stability analysis for the threshsizer algorithm, where the NContig parameter (i.e., the minimal size of objects) varies. Here, both
maxima and mean values exhibit a stable behavior of both scores. The differences are much smaller than for the previous cases, which is indicated by the much smaller scale on the vertical axis relative to Figs. 6 and 7. Interestingly, even large changes in the parameter values lead to small changes in $S$ and $L_2$ scores.

In summary, we have an unstable behavior of SAL with respect to the threshold parameter and a stable behavior with respect to the minimum object size. Varying smoothing radii seem to induce mean-stable behavior but may result in unstable SAL scores for some cases, which calls for a closer look at the distributional properties of score changes.

To assess the sensitivity of the SAL scores to changes in the OIA parameters on a distributional level, we study the null hypothesis of equal distributions of the $S$ and $L_2$ scores. The following OIA parameters are considered: threshold ratio $f \in \{1/15, 0.2, 0.5, 0.75, 0.85, 0.9, 0.95, 1\}$, smoothing radius smoothpar $\in \{0, 1, 2, 5, 10\}$, and minimum object size Ncontig $\in \{5, 25, 50, 100, 250, 500, 1000\}$.

The null hypothesis is tested using three complementary hypothesis tests, notably the Kolmogorov–Smirnov (K;
Table 1. Significant differences in score distributions for different values of the threshold ratio \( f \) in the threshfac algorithm for 400 fields of IR6.2. The capital letters denote hypothesis tests (defined in section 4) detecting differences at a 5% level of significance. The upper-right triangular section shows results for the comparison of two \( S \) score distributions (italic font), while the lower-left triangular section shows the results for \( L2 \) (boldface font).

<table>
<thead>
<tr>
<th>( f )</th>
<th>0.20</th>
<th>0.50</th>
<th>0.75</th>
<th>0.85</th>
<th>0.90</th>
<th>0.95</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/15</td>
<td>—</td>
<td>( K )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
</tr>
<tr>
<td>0.20</td>
<td>—</td>
<td>( K )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
</tr>
<tr>
<td>0.50</td>
<td>( K )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
</tr>
<tr>
<td>0.75</td>
<td>( K )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
</tr>
<tr>
<td>0.85</td>
<td>( K )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
</tr>
<tr>
<td>0.90</td>
<td>( K )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
</tr>
<tr>
<td>0.95</td>
<td>( K )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
</tr>
<tr>
<td>1</td>
<td>( K )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
<td>( M )</td>
</tr>
</tbody>
</table>

Kolmogorov 1933; Smirnov 1939), median (M), and quantile (Q) test, where the test statistic is the interquartile distance (i.e., the distance between the 25% and 75% quantiles). The M and Q tests are permutation tests (Good 2000) with 10 000 iterations and are particularly interesting for the study of SAL, since both median and interquartile distance are important quantities for the interpretation of SAL scores (Wernli et al. 2008).

For large variations of the threshold ratio in the threshfac algorithm, one potentially looks at different physical situations. Thus, significant differences in the distributions of the \( S \) and \( L2 \) scores are expected for large differences \( \Delta f > 0.1 \). Table 1 summarizes the results of all hypothesis tests and all combinations of threshold ratios for 400 fields of IR6.2 data. The upper-right triangular section shows results for the comparison of two \( S \) score distributions (italic font), while the lower-left triangular section shows the results for \( L2 \) (boldface font), for example, for the two \( S \) distributions derived with \( f = 0.75 \) and \( f = 0.85 \) only the quantile test (Q) indicates significant differences, while the \( L2 \) distributions with the same parameter settings differ in all three test statistics (KMQ). Table 1 confirms our expectations for IR6.2 for all but the largest threshold ratios of the threshfac algorithm. These results are consistent with the stability analysis for the threshold ratio in the previous section. The Q test is tailored to detect changes in spread and therefore well suited for a two-sided score. Accordingly, only the Q test is able to distinguish between \( S \) distributions for threshold ratios 0.75 and 0.85 (Table 1). Note that changes in \( S \) scores due to object decomposition are symmetric, since the decomposition can happen in both observations and/or forecasts.

Both the convthresh and threshsizer algorithms show significant differences only in the distribution of the \( L2 \) but not the \( S \) scores (Tables 2 and 3). The reason for this is twofold: first, \( S \) is a two-sided score and changes in the score may cancel out in accumulated statistics like mean or median. Therefore, the M test is less able to detect changes in the \( S \) score distribution. Second, while—in principle—the K test is able to detect changes in spread, it has issues when only the outer tails of the distributions are affected (e.g., Mason and Schuenemeyer 1983). In summary, we can identify symmetric changes in distributions for the \( S \) score only if they affect the interquartile distance. By definition, the interquartile distance is largely unaffected by small variation of individual values. Hence, changes in the \( S \) score are expected to be small for the less critical parameters, which is consistent with the results observed in Tables 2 and 3.

The distributional analysis has confirmed the unstable behavior of the SAL scores with respect to the threshold ratio in the threshfac algorithm. Whether smoothing radius and minimum object size can be considered as uncritical parameters depends on the interpretation of the SAL scores: if one is interested in the statistical quantities mean, median and interquartile distance both can be considered to give fairly stable SAL scores for IR6.2. However, for a varying smoothing radius this highly depends on the data. Table 4 shows the results for TCIC, where both median and interquartile distance change significantly for many parameter pairings.

It is often stated that the components of SAL are independent (e.g., Früh et al. 2007; Zimmer et al. 2009). However, this is not true in the sense of the mathematical definition of statistical independence. Quite contrarily, Fig. 9 shows significant Pearson correlation coefficients between the absolute \( L2 \) and \( S \) differences of each of the 400 fields for all algorithms and standardized parameter changes (i.e., the difference of two

Table 2. As in Table 1, but for different values of the smoothing radius smoothpar in the convthresh algorithm.

<table>
<thead>
<tr>
<th>smoothpar</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>( K )</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>10</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 3. As in Table 1, but for different values of the minimum object size Ncontig in the threshsizer algorithm.

<table>
<thead>
<tr>
<th>NContig</th>
<th>5</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>250</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>25</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>50</td>
<td>( K )</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>100</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>250</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>500</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1000</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>( K )</td>
<td>—</td>
</tr>
</tbody>
</table>

For large variations of the threshold ratio in the threshfac algorithm, one potentially looks at different physical situations. Thus, significant differences in the distributions of the \( S \) and \( L2 \) scores are expected for large differences \( \Delta f > 0.1 \). Table 1 summarizes the results of all hypothesis tests and all combinations of threshold ratios for 400 fields of IR6.2 data. The upper-right triangular section shows results for the comparison of two \( S \) score distributions (italic font), while the lower-left triangular section shows the results for \( L2 \) (boldface font), for example, for the two \( S \) distributions derived with \( f = 0.75 \) and \( f = 0.85 \) only the quantile test (Q) indicates significant differences, while the \( L2 \) distributions with the same parameter settings differ in all three test statistics (KMQ). Table 1 confirms our expectations for IR6.2 for all but the largest threshold ratios of the threshfac algorithm. These results are consistent with the stability analysis for the threshold ratio in the previous section. The Q test is tailored to detect changes in spread and therefore well suited for a two-sided score. Accordingly, only the Q test is able to distinguish between \( S \) distributions for threshold ratios 0.75 and 0.85 (Table 1). Note that changes in \( S \) scores due to object decomposition are symmetric, since the decomposition can happen in both observations and/or forecasts.

Both the convthresh and threshsizer algorithms show significant differences only in the distribution of the \( L2 \) but not the \( S \) scores (Tables 2 and 3). The reason for this is twofold: first, \( S \) is a two-sided score and changes in the score may cancel out in accumulated statistics like mean or median. Therefore, the M test is less able to detect changes in the \( S \) score distribution. Second, while—in principle—the K test is able to detect changes in spread, it has issues when only the outer tails of the distributions are affected (e.g., Mason and Schuenemeyer 1983). In summary, we can identify symmetric changes in distributions for the \( S \) score only if they affect the interquartile distance. By definition, the interquartile distance is largely unaffected by small variation of individual values. Hence, changes in the \( S \) score are expected to be small for the less critical parameters, which is consistent with the results observed in Tables 2 and 3.

The distributional analysis has confirmed the unstable behavior of the SAL scores with respect to the threshold ratio in the threshfac algorithm. Whether smoothing radius and minimum object size can be considered as uncritical parameters depends on the interpretation of the SAL scores: if one is interested in the statistical quantities mean, median and interquartile distance both can be considered to give fairly stable SAL scores for IR6.2. However, for a varying smoothing radius this highly depends on the data. Table 4 shows the results for TCIC, where both median and interquartile distance change significantly for many parameter pairings.

It is often stated that the components of SAL are independent (e.g., Früh et al. 2007; Zimmer et al. 2009). However, this is not true in the sense of the mathematical definition of statistical independence. Quite contrarily, Fig. 9 shows significant Pearson correlation coefficients between the absolute \( L2 \) and \( S \) differences of each of the 400 fields for all algorithms and standardized parameter changes (i.e., the difference of two
parameter values is divided by the maximum difference we investigated for this parameter). The correlation coefficients between the absolute $L_2$ and $S$ differences are even close to one for the convthresh algorithm, where most significant changes are due to object decomposition (see section 3). The decomposition of a large object into two, or the emergence of more small objects impacts the structure ($S$) and the spread of objects ($L_2$), simultaneously.

The correlation coefficients for the threshsizer algorithm are slightly lower with values between 0.7 and 0.9. Here, an increasing $N_{Contig}$ parameter implies that objects of increasing size are omitted. Since only the small-scale objects are affected, the structure is shifted toward larger objects with more mass. At the same time the spread is reduced. For threshfac we observe a large spread in the correlation coefficients varying between 0.4 and 0.8. This underlines the unstable behavior with respect to a varying threshold level, since many different effects can occur: small changes can lead to object decomposition, while large changes can cause objects of arbitrary size to vanish. Latter can lead to vastly different $S$ scores, but nearly unchanging $L_2$ scores, which in turn leads to lower correlation coefficients. This behavior can be observed for the exemplary case in Fig. 5.

5. Indicators for parameter sensitivity and observational uncertainty

Of the three investigated OIA parameters the threshold ratio is most important one for two reasons. First, all three OIA depend on the threshold level. Second, the SAL scores show the largest sensitivity to changes in the threshold ratio. Therefore, we concentrate solely on the threshfac algorithm in this section. Since the calculation of SAL scores for a multitude of different threshold ratios rapidly becomes computationally expensive, it is useful for practical applications to find a quantity that indicates whether or not a given set of data exhibits a high sensitivity toward small variations in threshold ratios, that is, whether SAL is mean stable with respect to threshold ratio. To derive such a sensitivity indicator, we concentrate on changes in the $L_2$ score, which is mathematically more accessible. Section 4 shows that the results hold true also for the $S$ score.

![Figure 9](image)

**Fig. 9.** The Pearson correlation coefficients between absolute changes in $S$ and $L_2$ scores for IR6.2 are plotted against the normalized difference in parameter values. For the convthresh algorithm we vary only the smoothing radius (smoothpar), for the threshsizer algorithm we vary only the minimal object size ($N_{Contig}$), and for the threshfac algorithm we vary only the threshold ratio (fac).
Since we are interested in the response to small parameter changes, we vary each of the eight threshold ratios \( f \in \{1/15, 0.2, 0.5, 0.75, 0.85, 0.9, 0.95, 1\} \) additionally by \( \pm 0.05 \), and calculate \( L2 \) scores for all 24 resulting parameter values. Let us denote the original threshold level by \( f_i \), the perturbed levels by \( f_i^- \) and \( f_i^+ \), and the resulting \( L2 \) scores by \( L2_i^- \), \( L2_i^+ \), and \( L2_i^* \), for \( i \in \{1, \ldots, 8\} \).

Largest sensitivity is expected in cases where a slight increase in the threshold ratio causes a large number of grid points to fall below the threshold. We are therefore interested in the ratio of grid points that vanish for a given field because of an increase in threshold ratio. This quantity can be approximated via the univariate empirical cumulative distribution function (ECDF) of the set of forecasts fields (see Fig. 10a), which describes the probability that IR6.2 is below a threshold. The ECDF of observational data is defined as

\[
eCDF_i(t) = \frac{1}{N \#(D)} \sum_{i=1}^{N} \#\{x \in D \mid R_i^{\text{obs}}(x) \leq t\},
\]

where \( \#() \) denotes the number of elements in a set. The ECDF of the set of forecasts fields is denoted by \( \text{ecdf}_2 \) and defined analogously. The ECDF ratios, which are functions of the parameter value \( f_i \), are given by

\[
p_i^k = \frac{\text{ecdf}_k(f_i^+) - \text{ecdf}_k(f_i^-)}{1 - \text{ecdf}_k(f_i^+)}, \quad k \in \{1, 2\}, i \in \{1, \ldots, 8\}.
\]

The lowest threshold \( f_i^- \) is used in the denominator to ensure that the ratio has an upper bound equal to one. This is necessary to allow the interpretation of \( p_i^k \) as a first-order approximation of a “decomposition probability,” that is, the probability for the event that a large object decomposes into two or more smaller objects. This interpretation is intuitive as seen for the following two extreme cases. First, if no point vanishes when raising the threshold, the probability that a large object decomposes is zero and \( p_i^k = 0 \). Second, if all points vanish, the probability that a large object vanishes is one and \( p_i^k = 1 \). Therefore, \( p_i^k \) approximates the decomposition probability while ignoring any effects of spatial correlation. This helps us to estimate the sensitivity of \( L2 \), since we can now quantify the probability that object decomposition occurs and hence sensitivity is high. It may seem overly simplified to use a single ECDF for a whole set of 400 spatial fields instead of 400 independent ECDFs. However, we are not interested in single worst-case scenarios but in an indicator for mean stability, which is a statistical quantity depending on the average score.
deviations in the dataset. This justifies the above approach as a reasonable and computational effective first guess.1

We further need to quantify the effect that a process of object decomposition has on $L_2$. Recall that the $L_2$ score at threshold level $f_i$ is defined as

$$L_2_i = 2 \frac{|r_1(f_i) - r_2(f_i)|}{d}, \quad i \in \{1, \ldots, 8\},$$

where $r_1$ and $r_2$ describe the scattering of objects (see section 2). $L_2$ as well as $r_1(f_1)$ and $r_2(f_2)$ are statistical estimators. The variance—or standard deviation—of a statistical estimator is closely related to its robustness. We therefore use the empirical standard deviation $\sigma$ over the whole set of spatial fields of the three following quantities as a measure of the effect an object decomposition would have on the $L_2$ score. The effect is quantified as

- $2r_1/d$, if object decomposition occurs only in the first spatial field;
- $2r_2/d$, if object decomposition occurs only in the second spatial field; or
- $L_2$, if object decomposition occurs in both spatial fields simultaneously.

Note that all standard deviations are calculated based only on SAL values for the threshold $f_i$ and do not use any SAL values or calculations for the perturbed thresholds $f_i^-$ and $f_i^+$.

Using the ECDF ratios $p^k_i$ as first-order approximation for the “decomposition probabilities” we can estimate the expected $L_2$ sensitivity for threshold $f_i$ as follows:

$$SI_i = p^1_i (1 - p^2_i) \sigma \left[ \frac{2r_1 (f_i)}{d} \right] + p^2_i (1 - p^1_i) \sigma \left[ \frac{2r_2 (f_i)}{d} \right] + p^1_i p^2_i \sigma (L_2),$$

for $i \in \{1, \ldots, 8\}$. Figure 10b shows that $SI$ is indeed a good indicator for $L_2$ sensitivity: if $SI < 0.05$, the change of $L_2$ is bounded by 0.1. Note that once $L_2$ scores have been calculated for a given threshold, $SI$ can be computed with very little computational cost, since all its components but the univariate ECDF ratios have already been calculated for $L_2$. For an a priori indicator that uses only the ECDF information and no SAL values, we set $2r_1/d$, $2r_2/d$, and $L_2$ equal to 1 and obtain

$$\bar{SI}_i = p^1_i (1 - p^2_i) + p^2_i (1 - p^1_i) + p^1_i p^2_i, \quad i \in \{1, \ldots, 8\}.$$

Figure 10c shows that the correlation between the $L_2$ sensitivity and $SI$ is still strong. This indicates that a large number of the stability issues are founded in the univariate field distributions, that is, the slopes of the univariate ECDFs, while spatial correlations only play a minor role.

There exists a close link between the sensitivity of SAL to varying threshold levels and the effects of observational uncertainties: if we look at the thresholded field, it makes no difference whether we raise the threshold level by a certain amount or lower the intensity of the field by the same constant amount at each grid point. The latter is equivalent to the effect of observational uncertainties with infinite spatial correlation length, that is, an additive constant. Therefore, the results for the sensitivity of SAL to varying threshold levels carry over to its sensitivity to large-scale uncertainties.

Do the results also contain valuable information about the effect of small-scale uncertainties? The previous section shows that $SI$ is a good indicator for the $L_2$ sensitivity with infinite spatial correlation length. However, $SI$ employs only univariate ECDF information of the fields, that is, it ignores any information regarding spatial correlations. Therefore, the correlation length of the (observational) uncertainties can only play a minor role for the sensitivity of SAL. This strongly suggests that the sensitivity of SAL to uncertainties of arbitrary correlation length is close to SAL’s sensitivity to varying threshold levels. Consequently, $SI$ and $SI$ are good indicators not only for the sensitivity of SAL toward varying threshold levels, but also for the sensitivity of SAL toward observational uncertainties.

6. Discrimination power of SAL

The sensitivity of the SAL parameters is closely linked to the ability of SAL to discriminate between good and bad forecasts. We have shown that the SAL parameters are sensitive to changes in the OIA, and as argued in section 5, a similar effect is expected for observational uncertainties. It is thus of interest to investigate the sensitivity of SAL scores toward artificial changes in the data themselves. A somewhat savage way to produce an artificially bad set of forecasts is to destroy the temporal collocation of the 400 pairs of forecasted and observed fields by a random permutation of the

---

1To be on the safe side, we have also calculated indicators analogous to $SI$ and $SI$ defined in (1) and (2) but based on single-field ECDFs. However, no significant improvements could be observed.
fields in time, that is, for each observation a random forecast is drawn (from the set of 400 available forecast fields). We then ask whether SAL is able to distinguish the quality of the original forecast and the randomly permuted forecast. This is achieved by testing the null hypothesis that the SAL scores over the 400 pairs of fields in the original and permuted dataset follow the same distribution.

Table 5 provides results of the different hypothesis tests on the SAL values using the threshfac OIA with different threshold ratios for the two different variables TCIC and IR6.2. We included TCIC in this analysis to demonstrate that the results are strongly dependent on the type of data. For TCIC and IR6.2 the default threshold ratio of $\frac{1}{15}$ exhibits almost no differences between both distributions (see Fig. 11). Only the quantile permutation test is able to significantly detect differences in the $S$ score distribution of the IR6.2 data.

Figure 12 shows the density of the $L_2$ and $S$ score distributions for IR6.2 for a higher threshold ratio of 0.85. Although significant statistical differences can be observed for this threshold, the distributions do not look as different as one would expect for such a hyperbolic case.

The traditional mean-square error (MSE) is able to clearly discriminate between the original and the permuted set of IR6.2 data, as shown in Fig. 13. In the original dataset the MSE is significantly smaller than in the permuted dataset.

For TCIC and threshold ratios of $\frac{1}{15}$ or 0.2, none of the two object-dependent scores are able to discriminate between the original and the perturbed dataset (Table 5). This changes for higher thresholds between 0.5 and 0.95, which suggests that the loss of discriminating power is due to very large objects at low thresholds. However, the situation is not as clear cut: the highest studied threshold ratio of 1, which identified the smallest objects of all OIA settings, again yields indistinguishable score distributions.

Conclusively, the ability of object-dependent SAL scores to distinguish between the two datasets largely depends on the data and cannot be guaranteed a priori. It is important to note that randomly permuting the

<table>
<thead>
<tr>
<th>$f$</th>
<th>$S$ (TCIC)</th>
<th>$S$ (IR6.2)</th>
<th>$L_2$ (TCIC)</th>
<th>$L_2$ (IR6.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/15</td>
<td>—</td>
<td>$Q$</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>0.20</td>
<td>—</td>
<td>$Q$</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>0.50</td>
<td>—</td>
<td>$Q$</td>
<td>$KQ$</td>
<td>—</td>
</tr>
<tr>
<td>0.75</td>
<td>$Q$</td>
<td>$KQ$</td>
<td>$KM$</td>
<td>$KMQ$</td>
</tr>
<tr>
<td>0.85</td>
<td>$Q$</td>
<td>$KQ$</td>
<td>$KMQ$</td>
<td>$KMQ$</td>
</tr>
<tr>
<td>0.90</td>
<td>$KQ$</td>
<td>$Q$</td>
<td>$KMQ$</td>
<td>$KMQ$</td>
</tr>
<tr>
<td>0.95</td>
<td>$KQ$</td>
<td>—</td>
<td>$KMQ$</td>
<td>$KM$</td>
</tr>
<tr>
<td>1</td>
<td>—</td>
<td>$Q$</td>
<td>—</td>
<td>$KM$</td>
</tr>
</tbody>
</table>

**TABLE 5.** Significant differences between score distributions of 400 fields derived from original observations vs observations randomly permuted in time at various threshold ratios $f$ in the threshfac algorithm. The capital letters denote hypothesis tests defined in section 4 that were able to detect differences with a 5% level of significance. Results are shown for TCIC and IR6.2 for the $S$ score (italic font) and $L_2$ score (boldface font).

Fig. 11. Density of (a) $L_2$ and (b) $S$ score of original and permuted TCIC with the default threshold ratio $f = 1/15$.  

2104 JOURNAL OF APPLIED METEOROLOGY AND CLIMATOLOGY V OLUME 55
observations is synonymous with a complete loss of temporal collocation between forecast and observation, which is expected to lead to significantly worse scores, as is the case for the MSE (Fig. 13).

On its own, the inability to distinguish between the original and permuted dataset is not necessarily a disaster. SAL investigates the statistics of the objects, and one might conclude that this is relatively homogeneous in time and thus insensitive to the loss of temporal collocation. However, in view of the large sensitivity that the SAL scores show with respect to the choice of the OIA parameters, caution is advised when interpreting the results of SAL.

7. Conclusions

The aim of this work is twofold: first, to study the applicability of SAL for cloud processes; second, to identify and understand the importance of OIA and their parameters on feature-based verification methods and the link to observational uncertainties. Three different OIA have been used for the comparison of COSMO-DE SynSat data with SEVIRI satellite observations. In this process varying values have been used for the three parameters threshold ratio, smoothing radius, and minimal size of objects. On a conceptual level we have shown that small changes in threshold levels or smoothing radii can potentially lead to very large score differences because of object decomposition, which is confirmed by two exemplary case studies of IR6.2 data.

To study SAL’s parameter sensitivity on a distributional level, we denote SAL as unstable with respect to a parameter of an OIA if small changes in the value of this parameter can lead to large changes, or even unbounded responses, in the resulting SAL scores. SAL is unstable with respect to threshold ratio and stable with respect to minimal object size. With respect to varying smoothing radii, SAL is mean stable but maximum unstable; that is, there are rare worst-case scenarios in which large score deviations occur. In-depth statistical analysis using three different hypothesis tests confirms these results. For varying threshold ratios the observed large score deviations translate into significant changes in the distribution of $S$ and $L_2$ scores. Consistent with the prior stability assessment the statistical implications for varying smoothing radii are much weaker.

The threshold ratio is of particular interest, not only because it is the most sensitive parameter but because it links the field of parameter sensitivity to observational uncertainties: in cases where the intensity of a spatial field is close to the threshold level, changes in the threshold ratio (parameter sensitivity) lead to similar results as changes in the intensity of the data itself (observational uncertainties). An a posteriori indicator (SI) for the stability of SAL to the threshold ratio parameter shows promising results to assess the sensitivity without the need of expensive computations of multiple threshold levels (section 5). The a priori indicator $SI$ is based solely on univariate ECDF.
information of the spatial fields and can therefore be calculated with very little computational effort. Both quantities can also be employed to assess SAL’s sensitivity to observational uncertainties (see section 5).

Highly sensitive parameters are particularly problematic if the changes in scores due to varying parameters outweigh score deviations caused by actual differences in the data. Such a case is discussed in section 6, where $S$ and $L^2$ scores were unable to reliably detect the complete loss of temporal collocation between forecast and observation.

To summarize, the choice of OIA and its parameters has a significant effect on the resulting SAL scores. Therefore it is essential to explicitly state the algorithm and all parameter settings when using SAL for verification. The use of complementary hypothesis tests has shown that it is advisable to include statistical quantities beside median and interquartile range for the interpretation of SAL scores. The high sensitivity toward the threshold level implies a potentially high impact of observational uncertainties. This is particularly true for SAL’s original field of application, that is, the verification of quantitative precipitation fields against radar observations. By defining sensitivity indicators SI and $SI$, we were able to quantify the connection between parameter sensitivity and the effect of observational uncertainties. The fact that small changes in parameter values have a larger impact than even drastic changes in data implies that SAL is not well equipped to verify this specific kind of data.

On a more technical level, object decomposition in conjunction with noncontinuous operations, for example, thresholding, during object identification has been established as the major cause for unstable behavior. Because of the similarity between threshold sensitivity and observational uncertainties, this study is a step toward the construction of feature-based verification methods that are robust with respect to observational uncertainties. The importance of the univariate ECDFs in the definition of SI and $SI$ suggests, that the normalization of continuous data or the application of thresholds solely based on quantiles could significantly reduce the high sensitivity to parameters and uncertainties.

Acknowledgments. We gratefully acknowledge financial funding by the project High Definition Clouds and Precipitation for advancing Climate Prediction HD(CP)$^2$ funded by the German Ministry for Education and Research (BMBF) under Grant FK 01LK1209B. The authors thank Sonja Reitter (Universität Köln) and the DWD, who provided data from the COPS/GOP project, which was founded by the Deutsche Forschungsgemeinschaft under Grant WU 356/4-2. We further appreciate the help of Jennifer Slobodda and Justus Franke (Institut für Weltraumwissenschaften, Freie Universität Berlin), who prepared SEVIRI data as part of the DFG-ICOS program.

REFERENCES


Gilleland, E., 2014: SpatialVx: Spatial forecast verification, version 0.2-0 R package. [Available online at http://CRAN.R-project.org/package=SpatialVx.]


