

## Object-Based Metrics for Forecast Verification of Convective Development with Geostationary Satellite Data

MARTIN REMPEL,<sup>a</sup> FABIAN SENF, AND HARTWIG DENEKE

*Leibniz Institute for Tropospheric Research, Leipzig, Germany*

(Manuscript received 21 December 2016, in final form 22 April 2017)

### ABSTRACT

Object-based metrics are adapted and applied to geostationary satellite observations with the evaluation of cloud forecasts in convective situations as the goal. Forecasts of the convection-permitting German-focused Consortium for Small-Scale Modeling (COSMO-DE) numerical model are transformed into synthetic observations using the RTTOV radiative transfer model, and contrasted with the corresponding real observations. Threshold-based segmentation techniques are applied to the fields for object identification. The statistical properties of the traditional measures cold cloud cover and average brightness temperature amplitude are contrasted to object-based metrics of spatial aggregation and object structure. Based on 59 case days from the summer half-years between 2012 and 2014, a variance decomposition technique is applied to the time series of the metrics to identify deficits in day-to-day, diurnal, and weather-regime-related variability of cold cloud characteristics in the forecasts. Furthermore, sensitivities of the considered metrics are discussed, which result from uncertainties in the satellite forward operator and from the choice of parameters in the object identification techniques.


### 1. Introduction

The process of evaluating the quality of a forecast is called forecast verification, and is an indispensable part of model-based forecast development in general and numerical weather prediction (NWP) model development in particular. [Murphy \(1993\)](#) defines quality as the correspondence between forecasts and observations. The purpose of forecast verification is to confirm that a model indeed has skill, and to ascertain that recent model changes result in improved forecasting capabilities compared to previous model versions. Moreover, it enables users of the forecasts to assess their quality and to identify typical model deficiencies and systematic errors.

Over the past few decades most of the verification efforts targeting NWP models have focused on scores based on a point-to-point comparison. These verification

measures, referred to here as traditional scores, are insufficient for the evaluation of cloud and precipitation processes in high-resolution NWP forecasts since they do not take into account information about the spatial structure and irregular morphology of these fields [see [Baldwin and Kain \(2006\)](#), [Casati et al. \(2008\)](#), [Rossa et al. \(2008\)](#), and [Wilks \(2011\)](#) for a detailed description of traditional scores]. For example, the traditional scores show counterintuitive behavior if a localized forecast feature of correct size and structure is displaced in time or space. In this case, poor verification scores are obtained because the displacement is penalized twice: first, as the forecast misses the feature identified in the observations and, second, as the forecast feature produces a false alarm.

To avoid this so-called double-penalty problem, and to get scores that correspond more closely to our subjective visual notion of forecast quality, new techniques have been developed for the evaluation of spatial fields over the last decade (see, e.g., [Ebert 2008](#); [Rossa et al. 2008](#); [Casati et al. 2008](#); [Gilleland et al. 2009](#)). One important subset of these spatial approaches features object-based methods, which are concerned with certain structural properties of the meteorological forecast fields. Object-based methods transform a continuous forecast into a categorical field, usually relying on

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

<sup>a</sup> Current affiliation: Deutscher Wetterdienst, Offenbach, Germany.

Corresponding author: Martin Rempel, [martin.rempel@dwd.de](mailto:martin.rempel@dwd.de)

threshold-based segmentation techniques. Metrics are thereafter constructed either from the difference in the individual object properties or from the distributional characteristics of objects. Most of the previous development effort of object-based methods was devoted to the evaluation of precipitation forecasts, for instance, the contiguous rain area (CRA; [Ebert and McBride 2000](#); [Ebert and Gallus 2009](#)), the Method for Object-Based Diagnostic Evaluation (MODE; [Davis et al. 2006a,b, 2009](#)), and the structure, amplitude, and location (SAL; [Wernli et al. 2008](#)) measure. So far, considerably less research has focused on the object-based evaluation of cloud processes. Some more recent studies investigated the application of SAL for the evaluation of binary cloud masks ([Crocker and Mittermaier 2013](#)), the distribution of tropical convective cloud sizes and lifetimes ([Negri et al. 2014](#); [Machado and Chaboureaud 2015](#)), and the use of SAL for upper-tropospheric water vapor fields ([Weniger and Friederichs 2016](#)). The latter study furthermore noted unsatisfying sensitivities of object-based metrics on the parameters of the object identification methods. Based on these investigations, the aim of our present study is to develop and analyze object-based metrics for the evaluation of cloud forecasts using infrared observations from a geostationary satellite. We especially focus on the adaptation of the SAL method for the assessment of cold cloud characteristics and investigate the statistical properties of its individual components.

SAL was proposed to measure the quality of precipitation forecasts in a statistical sense. It does not demand a one-to-one correspondence between individual objects. SAL has three components that are related to the (i) structure, (ii) amplitude, and (iii) location of objects. Over the years, SAL has become a workhorse for the quantitative evaluation of precipitation forecasts from high-resolution forecasts (e.g., [Hanley et al. 2013](#); [Schneider et al. 2014](#); [Kann et al. 2015](#); [Lindstedt et al. 2015](#)). A further measure that we have included in our analysis is the so-called simple convective aggregation index (SCAI) that is related to the arrangement of clouds within the domain. It was developed by [Tobin et al. \(2012\)](#) to investigate the relationships between self-aggregation of tropical deep moist convection and large-scale environmental properties (e.g., middle- and upper-tropospheric humidity, precipitable water, and turbulent surface fluxes). SCAI describes the ratio between the actual degree of convective disaggregation and a potential maximum degree of disaggregation.

In general, the verification of cloud forecasts with geostationary satellite observations can be either based on satellite products (e.g., [Kidd et al. 2013](#)) or utilize an observational forward operator comparing simulated

and real observations (e.g., [Eikenberg et al. 2015](#)). For the latter approach, which is adopted in our study, the model forecasts need to be transferred into observation space using a radiative transfer model. We apply a revised Synthetic Satellite imagery (SynSat) scheme ([Senf and Deneke 2017](#)) here using the computationally very efficient RTTOV model for simulating infrared satellite radiances (e.g., [Saunders et al. 1999](#)). The resulting synthetic satellite images depict the spatial distribution of the top-of-the-atmosphere outgoing radiation for the spectral response of a chosen satellite sensor and a perfect model forecast, and thereby enable a direct and easy comparison with real observations. Synthetic satellite images have been used for model verification for more than 20 years. One of the first studies using synthetic infrared Meteosat images was performed by [Morcrette \(1991\)](#), who evaluated the diurnal cycles of surface temperature and cloudiness of the global European Centre for Medium-Range Weather Forecasts (ECMWF) model. [Jankov et al. \(2011\)](#) investigated several microphysical schemes of the Weather Research and Forecasting (WRF) Model with respect to the accumulated precipitation and the infrared brightness temperatures of the GOES 10.7- $\mu\text{m}$  channel. It was shown that the synthetic satellite images revealed an overestimation of clouds, while the precipitation showed only slight differences. [Bikos et al. \(2012\)](#) depicted the importance of synthetic satellite imagery in operational forecasting, since it allows us to visualize atmospheric processes and monitor cloud development from an integrated perspective instead of providing an analysis of individual model output fields, and therefore a very quick assessment of the accuracy of model forecasted features is possible.

We illustrate the object-based evaluation process on the basis of operational cloud forecasts from the convection-permitting German-focused Consortium for Small-Scale Modeling (COSMO-DE) model. Several earlier studies have identified a distinct bias in the frequency of simulated cold brightness temperatures ([Pfeifer et al. 2010](#); [Böhme et al. 2011](#); [Eikenberg et al. 2015](#)). They report that COSMO-DE significantly overestimates the occurrence frequency of low brightness temperatures in the 10.8- $\mu\text{m}$  channel of the Spinning Enhanced Visible and Infrared Imager (SEVIRI) aboard Meteosat Second Generation (MSG) satellites at around 230 K compared to observations. Recently, [Senf and Deneke \(2017\)](#) discussed how a significant portion of this cold bias can likely be attributed to the uncertainties and inconsistencies in the representation of cirrus-radiative properties. In addition, [Keller et al. \(2016\)](#) reported that biases in the diurnal variation of COSMO-DE's cold cloud cover are reduced when the

model-internal ice-microphysical formulation is switched from the operational one-moment to a two-moment scheme.

The present study is structured as follows. The observational data from the geostationary Meteosat satellite as well as the simulated data from the forecast model COSMO-DE are introduced in [section 2](#). This is followed in [section 3](#) by a description of the case-day selection, the object identification algorithm, and the metrics utilized. In [sections 4](#) and [5](#) we present the results of our study. First, we discuss the distributional characteristics of the considered object-based metrics. Second, we analyze the temporal behavior of the metrics, perform a decomposition of variance into day-to-day and intraday components, and discuss their typical diurnal cycles for different weather regimes. A discussion of the sensitivities of the object-based metrics is presented in [section 6](#). Finally, conclusions are drawn in [section 7](#).

## 2. Data

### a. Observational data

We utilize observations of only one narrowband infrared channel with a central wavelength of  $10.8\mu\text{m}$  of the imaging radiometer SEVIRI aboard the geostationary MSG satellites operated by EUMETSAT ([Schmetz et al. 2002](#)). Besides this channel, SEVIRI has a total of 11 narrowband and 1 broadband high-resolution visible channel. For this study, data from the primary scan service are used, which has an image update cycle of 15 min and an orbital position at  $0^\circ$  longitude. The targeted domain corresponds to that of the COSMO-DE NWP model, which is described more precisely in the next section. Over this domain, the considered SEVIRI channel has an approximate resolution of  $4 \times 6\text{km}^2$ , which is coarser than the COSMO-DE grid size. Before a comparison, we apply nearest-neighbor interpolation to map SEVIRI observations onto the COSMO-DE grid.

The selected SEVIRI  $10.8\text{-}\mu\text{m}$  channel is a window channel, which means that it is only slightly influenced by atmospheric gases and mostly shows the radiative signature of the surface or clouds. Thus, the observed  $10.8\text{-}\mu\text{m}$  brightness temperatures (BT10.8) correspond to the cloud-top temperature for optically thick clouds, but are significantly warmer for semitransparent cirrus as a result of semitransparency and surface contributions.

### b. Simulated data

The operational short-range weather forecast model of the German Weather Service, COSMO-DE ([Baldauf et al. 2011](#)), is a convection-resolving nonhydrostatic NWP

model with a horizontal grid spacing of 2.8 km initialized each 3 h running 21 h ahead. The COSMO-DE domain covers Germany, Switzerland, Austria, the Netherlands, Belgium, and parts of the neighboring European countries.

The focus in this study is on convective situations, since knowledge about the NWP model's performance as well as the rapid use of observational data is essential in such situations. Therefore, forecasts are selected for 59 case days for the years 2012 (20 days), 2013 (20 days), and 2014 (19 days), containing deep moist convection in the domain of COSMO-DE. A further description of the case selection is found below. Three-dimensional forecast fields including thermodynamic and hydrometeor properties have been retrieved from the data archive for four initialization times (0300, 0600, 0900, and 1200 UTC). The output has a temporal frequency of 1 h, and only the time period between 0600 and 1800 UTC was chosen for further analysis. Therefore, for each day we consider 13 scenes from the COSMO-DE initializations at 0300 and 0600 UTC and 10 (7) scenes from COSMO-DE 0900 UTC (1200 UTC) and, thus, altogether 2537 scenes were included. A revised SynSat scheme was applied to the COSMO-DE dataset for the calculation of synthetic brightness temperatures [see [Senf and Deneke \(2017\)](#) for further details on the method]. Operationally available SynSat schemes are also examined to assess the sensitivity of the object-based metrics to changes in the satellite forward operator.

As an illustrative example, the scene at 1400 UTC 6 August 2013 is depicted in [Fig. 1](#). [Figure 1a](#) shows the observed MSG SEVIRI BT10.8 field, and [Fig. 1b](#) the synthetic BT field for the 0600 UTC initialization of COSMO-DE (COSMO-DE 0600 UTC). The observations show three convective systems over Germany and Switzerland, as well as smaller convective clouds over France, the Czech Republic, Poland, and the Dinaric Alps. Additional cirrus clouds are also visible over parts of Belgium, the Netherlands, and Germany. Compared to this, COSMO-DE 0600 UTC shows only two of these convective systems. The convective clouds over the eastern part of Germany, Poland, and the Czech Republic are not present in the model forecast, but numerous developing cells are visible from southern Germany to the Dinaric Alps. Further, COSMO-DE 0600 UTC shows a significantly higher amount of cirrus cloud cover across the whole domain.

## 3. Method

### a. Case days

For the selection of suitable case days, the archive of convective forecasts by the European Storm Forecast

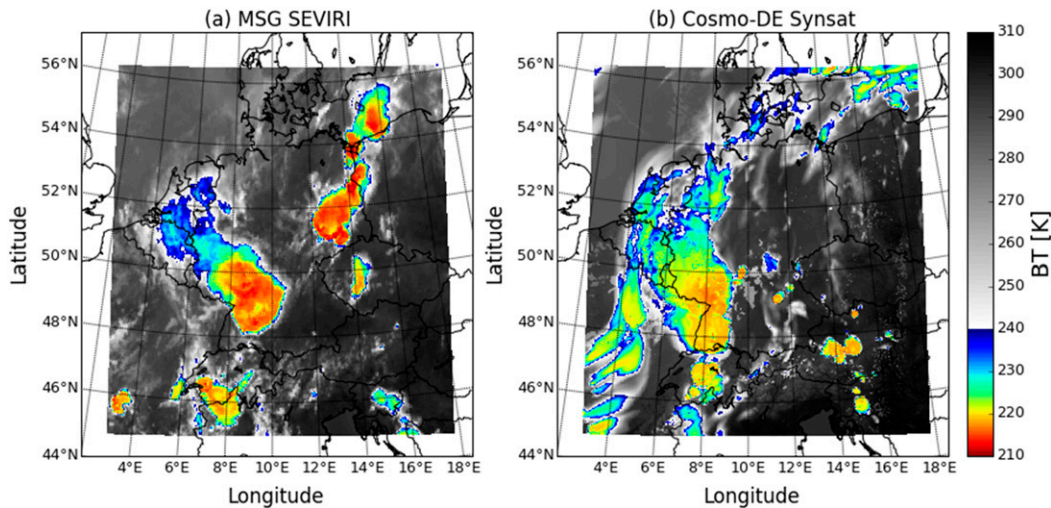


FIG. 1. Satellite imagery of a convective scene at 1400 UTC (1500 LT) 6 Aug 2013 over central Europe. The (a) observed and (b) synthetic BT10.8 fields for COSMO-DE initiated at 0600 UTC. For plotting, standard color enhancement is used to improve the perception of cold cloud tops. Thus, BTs warmer than 240 K are shown in gray-shaded colors and colder BTs up to 210 K are highlighted in colors from blue to red.

Experiment (ESTOFEX) network ([www.estofex.org](http://www.estofex.org)) was consulted as the first criterion in identifying potential convective situations. Here, the focus was placed on the period from April to September for the years 2012–14. Days with such a forecast were further analyzed by querying the European Severe Weather Database (ESWD; [Dotzek et al. 2009](#)). Within this database, entries localized in Germany and especially those with associated phenomena of severe thunderstorms (e.g., hail, strong winds, heavy precipitation, or tornadoes) were considered. Afterward, days that had both a convective forecast and at least one entry in ESWD were subjectively reviewed, considering different instability parameters based on the Global Forecast System (GFS) reanalysis, with consideration of the KO index, lifted index, convective available potential energy (CAPE), and the vertical motion at 500 hPa. The aim was to also obtain an overview of the synoptic situation of each day. In the end, 59 case days were found, which will be used for the statistical analysis.

[Akkermans et al. \(2012\)](#) and [Böhme et al. \(2011\)](#) showed that the forecast quality differs for various large-scale synoptic situations. Therefore, the large-scale synoptic situation must also be considered for the investigation of a diurnal cycle of deep convective clouds. In our study, the operational circulation pattern classification of the German Weather Service was utilized, which applies the subjective method of [Baur \(1963\)](#). The frequency of the obtained weather regimes is depicted in [Fig. 2](#). Here, the dark gray bars indicate the three most

frequent patterns, which are considered for the investigation of the diurnal cycle of metrics.

#### b. Object identification

The computation of object-based metrics requires the identification of individual cloud objects within the considered domain both in the observed and synthetic satellite images. Here, the aim is to identify cold cloud cover that is associated with the development of deep convective clouds. For the distinction between such cloud cover and the environment, a BT10.8 threshold of 240 K is used, which is consistent with numerous previous studies ([Roca and Ramanathan 2000](#); [Tobin et al. 2012](#); [Feidas and Giannakos 2012](#)). However, this means that cirrus clouds are also included in the investigation.

The individual components of SAL are unfortunately sensitive to the considered value range (i.e., are not invariant to linear transformations). We therefore transform the BT10.8 field into a first analysis step. For this, the threshold value of 240 K is subtracted from the original BT10.8 field, and negative values are set to zero afterward. The resulting BT10.8 amplitude is termed  $\Delta T_i$ , where  $i$  is the index of subsequently numbered grid points in the domain. In a second step, connected structures are identified with a standard segmentation method that labels contiguous nonzero grid boxes with a unique index using eight connectivity (i.e., two grid boxes are combined into the same cluster if they are adjacent in the horizontal, vertical, or diagonal directions). Each compound of connected grid boxes with the same index is called an object in the following.

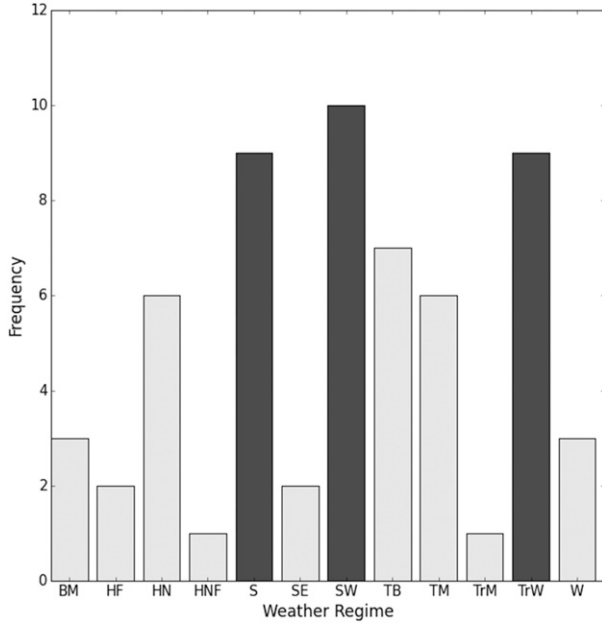


FIG. 2. Number of weather regimes occurring within the considered case days. The dark gray bars indicate the three most frequent weather regimes, which are used for the investigation of the diurnal cycle. The following abbreviations are used: BM, zonal ridge across central Europe; HF, Scandinavian high; HN, Icelandic high; HNF, high over Scandinavia–Iceland; S, southerly; SE, southeasterly; SW, southwesterly; TB, low over the British Isles; TM, low (cut off) over central Europe; TrM, trough over central Europe; TrW, trough over western Europe; and W, westerly.

Furthermore, we denote a set of individual objects as a cluster. In the last step, objects with an area-equivalent diameter smaller than 20 km are excluded because of large fluctuations in their numbers [see, e.g., [Weniger and Friederichs \(2016\)](#) for a discussion of sensitivities due to this parameter].

### c. Definition of object properties

Each grid box  $i$  carries the information about the field value  $\Delta T_i$ , the position  $\mathbf{x}_i$ , and the gridbox area  $A_i$ . With these parameters, specific object characteristics can be constructed as follows. An object  $m$  is composed of individual grid boxes whose indices are collected in the index set  $\mathcal{T}_m$ , which is a subset of the full domain index set. The object  $m$  can be characterized by its maximum BT ( $\Delta T_{m,\max}$ ) and a mean BT:

$$\langle \Delta T \rangle_m = \frac{1}{N_m} \sum_{i \in \mathcal{T}_m} \Delta T_i, \quad (1)$$

where  $N_m$  denotes the number of grid boxes in the  $m$ th object and the brackets are a shortcut for the arithmetic average over all object points.

For the object position  $\mathbf{X}_m$ , we define a second type of average, an amplitude-weighted average, to take the typical distribution of  $\Delta T$  within the object into account. For this, the weights

$$w_i = \frac{\Delta T_i}{N_m \langle \Delta T \rangle_m} \quad (2)$$

are defined as the ratio between individual values  $\Delta T_i$  and the object-average  $\langle \Delta T \rangle_m$ . The center of mass for each object is hence given by

$$\mathbf{X}_m = \sum_{i \in \mathcal{T}_m} w_i \mathbf{x}_i. \quad (3)$$

We furthermore define the object area  $A_m$  as the sum of individual gridbox areas, as well as the object volume  $V_m$  and object shape  $S_m$ . The volume represents a metric that aims to capture the overall structure of the objects. This method was introduced by [Wernli et al. \(2008\)](#) and provides combined information about the object size and shape. For this, individual field values  $\Delta T_i$  are normalized by the object-maximum value  $\Delta T_{m,\max}$  and finally summed over the object

$$V_m = \sum_{i \in \mathcal{T}_m} \frac{\Delta T_i}{\Delta T_{m,\max}} = \frac{N_m \langle \Delta T \rangle_m}{\Delta T_{m,\max}}. \quad (4)$$

This step is illustrated in [Fig. 3b](#). Here, the gray-shaded area corresponds to  $V_m$ . The volume is high (low) when the objects are large (small) and/or flat (more peaked).

To distinguish between object size and shape, a metric is constructed to identify only the shape of the objects

$$S_m = \frac{V_m}{N_m}. \quad (5)$$

From [Eq. \(4\)](#), the object shape is simply the ratio between the object-average and object-maximum  $\Delta T$  values. In [Fig. 3c](#), the gray-shaded area represents again the cross section through the center of a sine-shaped circular cone while the reddish rectangle shows a cylinder with similar base and height. The object shape  $S_m$  is now the ratio between both. Values of  $S_m$  can range between 0 and 1, whereby 0 depicts a strongly peaked object with a high maximum in only one grid box and very small values in the other ones within the object. An  $S_m$  value of 1 implies a plane object with the same features but an arbitrary BT within the whole object. Conically or hemispherically shaped objects attain respective values of  $1/2$  and  $2/3$  [see [Wernli et al. \(2008\)](#) for illustrative examples].



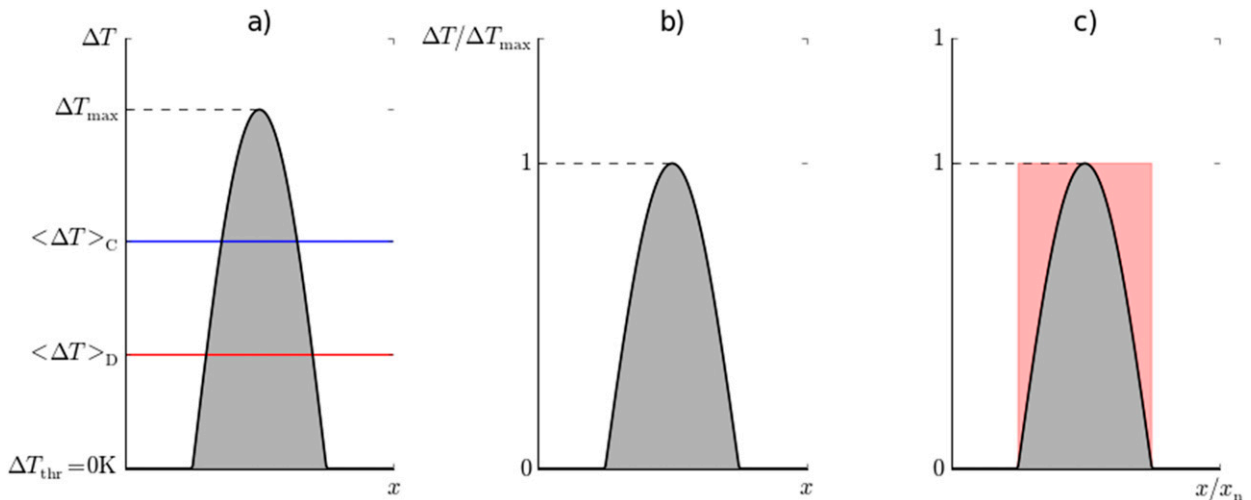


FIG. 3. Section across the center of a sine-shaped circular cone (gray shaded area) as an idealized object to illustrate metrics with information about the intensity of an object. (a) Here,  $\langle \Delta T \rangle_C$  (blue line) and  $\langle \Delta T \rangle_D$  (red line) denote the cluster or domain average of the transformed BT field and  $\Delta T_{\max}$  is the maximum value of the object. (b) Cross section scaled with  $\Delta T_{\max}$ . The gray-shaded area now corresponds to a scaled volume  $V_m$  of the object. (c) As in (b), but after the scaled volume is divided by the number of grid boxes  $N_m$  of the object. The individual shape  $S_m$  then corresponds to the ratio between the gray-shaded area and a cylinder with the same base and height. The latter is depicted with the red-shaded rectangle.

#### d. Definition of cluster properties

As stated above, we define a cluster as a set of individual objects, in our case all objects that are located within the interior of the domain (i.e., which do not touch the domain edges) and fulfill the minimum size criterion. The cluster averages are defined similarly to those defined on objects. Following Wernli et al. (2008), we subdivide the cluster characteristics into properties related to (i) amplitude or amount, (ii) location or spatial arrangement, and (iii) structure. The final set of derived object-based metrics is listed in Table 1.

##### 1) AMPLITUDE

Most basic cluster characteristics related to the number of objects are the total number of objects  $M$  and the total areal coverage of objects  $cc$ , which is calculated from the ratio of the total cluster area  $A_C$  and domain area  $A_D$ . For a cluster amplitude estimate, we define the cluster-average BT as an arithmetic mean:

$$\langle \Delta T \rangle_C = \frac{1}{M} \sum_{m=1}^M \langle \Delta T \rangle_m. \quad (6)$$

It describes a mean cloud-top temperature excess over the considered threshold of 240 K, only accounting for selected cloud objects and assuming that the clouds are opaque. The domain-mean  $\langle \Delta T \rangle_D$  is calculated similarly; however, it includes all  $\Delta T_i$  values that have been set to zero. It is therefore strongly coupled to the total areal coverage:

$$\langle \Delta T \rangle_D \approx cc \times \langle \Delta T \rangle_C. \quad (7)$$

The latter is an adaptation of the amplitude measure used in Wernli et al. (2008). A higher (lower) domain-mean BT corresponds to a large (small) cold cloud cover and/or lower (higher) cloud-top temperatures. An example for both mean values is shown in Fig. 3a.

##### 2) LOCATION

The center of mass of the cluster  $\mathbf{X}_C$  is calculated from the individual object centers. We apply a similar strategy as in Eq. (3) and define a weighted average as

$$\mathbf{X}_C = \sum_{m=1}^M W_m \mathbf{X}_m, \quad (8)$$

TABLE 1. Overview of all metrics used.

Variable	Metric
Amplitude metrics	
$M$	No. of objects
$cc$	Cold cloud cover
$\langle \Delta T \rangle_C$	Cluster-mean BT
$\langle \Delta T \rangle_D$	Domain-mean BT
Location metrics	
$R_D$	Mass distance
$R_C$	Compactness radius
SCAI	Simple convective aggregation index
Structure metrics	
$V_C$	Volume
$S_C$	Shape

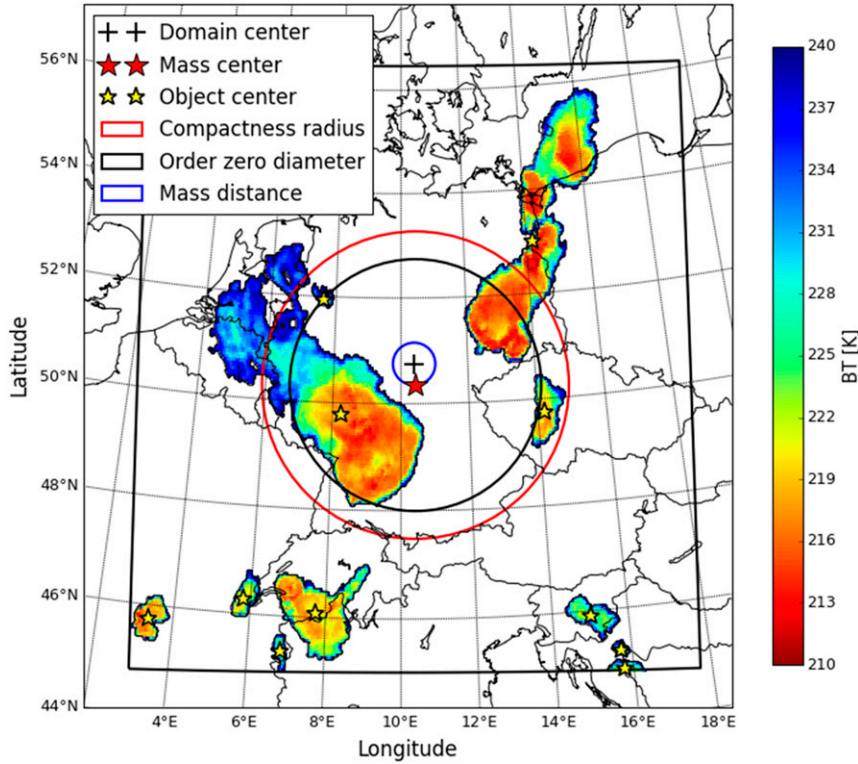


FIG. 4. Observed example scene from Fig. 1 presented to illustrate the considered metrics that provide information about the spatial distribution of the field. The black border shows the considered domain whose center is marked with a black cross. The colored areas are the identified objects. Each yellow star depicts the center of mass of the respective object, whereas the red star marks the center of mass of the whole cluster. The radii of the shown circles indicate different metrics for the spatial distribution: blue, mass distance; black, half of order-zero diameter; and red, compactness radius.

where now the weights  $W_m$  for the different objects are derived from the ratio between object-average and cluster-average BT amplitude; that is,

$$W_m = \frac{N_m \langle \Delta T \rangle_m}{N_C \langle \Delta T \rangle_C}, \quad (9)$$

where  $N_C$  denotes the total number of grid boxes in the cluster. With this, two spatial measures can be defined. The first, the mass distance  $R_D$ , is related to the average placement of the cloud cluster and is computed from the distance between the cluster center  $\mathbf{X}_C$  and domain center  $\mathbf{X}_D$  by

$$R_D = |\mathbf{X}_C - \mathbf{X}_D|. \quad (10)$$

The second spatial measure, the compactness radius  $R_C$ , is an aggregation measure that originates from the second part of the location component of SAL (see Wernli et al. 2008). Here, the distances between the object centers  $\mathbf{X}_m$  and the total center of mass  $\mathbf{X}_C$  are

computed. Afterward, these distances are weighted with  $W_m$  to obtain

$$R_C = \sum_{m=1}^M W_m |\mathbf{X}_m - \mathbf{X}_C|. \quad (11)$$

Thus,  $R_C$  depicts the radius of a circle around  $\mathbf{X}_C$ , which includes most of the mass (see Fig. 4). A scattered (more compact) spatial distribution of the objects is given when  $R_C$  is large (small).

A spatial metric that includes information about the amount and the average spatial extent of the cluster is SCAI. It was introduced by Tobin et al. (2012). SCAI depends on the total number of objects  $M$  and the possible distances between all objects. This dependence might be compared to the compactness radius  $R_C$  that depends on the distances between each object and the cluster center as well as the object size and amplitude distribution due to the weighting function  $W_m$ . SCAI is defined as

$$\text{SCAI} = \frac{M}{M_{\max}} \times \frac{R_0}{L} \times 1000. \quad (12)$$

Here,  $R_0$  denotes the so-called order-zero diameter, the geometric mean of all possible distances between the objects. In Fig. 4,  $R_0$  is illustrated as a black circle. Furthermore, SCAI depends on the characteristic domain length  $L$  and potential maximum number of objects  $M_{\max}$ . We define  $L$  as the area of our fixed domain divided by the longest possible distance within the domain. These domain-specific values are  $N_D \approx 2 \times 10^5$ ,  $L = 871.4$  km, and  $M_{\max} = 1898$ , where the latter value takes the minimum object size of 20 km into account. SCAI is large (small) for a scattered (more aggregated) spatial distribution of the objects.

### 3) STRUCTURE

We define the cluster volume  $V_C$  and the cluster shape  $S_C$  as weighted averages over the individual object volume and shape values. The calculation is done similar to Eq. (8) and is not stated separately. Please keep in mind that this averaging strategy gives a higher weight to the volume and shape of the larger objects.

## 4. Distributional characteristics

In the following, the overall distribution of the proposed object-based metrics is analyzed. We especially show how a direct comparison of observed and simulated distributions helps to isolate specific deficits of cloud forecasts.

### a. Comparison between observations and simulations

Figure 5 shows box plots of the frequency distributions for the set of metrics. For clarity, and because the forecast distributions are very similar, only one model run (COSMO-DE 0600 UTC) is depicted. It can be seen that the forecasts exhibit a systematic and significant overestimation in six out of nine metrics. We applied the so-called two-sample Kolmogorov–Smirnov test (see, e.g., Wilks 2011) to determine if the observations and forecasts share the same empirical cumulative distribution function. As the time series of the metrics show significant serial correlation, we incorporated the effective degrees of freedom in the distribution tests. The estimates are based on the autocorrelation behavior for each metric (see section 5a for more information on signal decomposition). The effective sample size is typically found to be a factor 3–6 smaller than the actual sample size.

For the object number and cold cloud cover (Figs. 5a,b), the observed distributions show a median of 19 objects and extrema between 1 and 60 objects. These

values cover a domain fraction from 0.2% up to 36% with a median of 11%. For both metrics, the model exhibits not only higher values but also a wider range. The interquartile range (IQR) of the difference between forecasts and observations lies between 2 and 15 objects for the number of objects, and from 1% to 10% for cloud cover. This overestimation may correspond to more convective cells and/or to a more peaked structure of high clouds. As indicated by the higher values in the cloud fraction, the COSMO-DE forecasts considered strongly overestimate the presence of high clouds, which is in agreement with earlier studies (Böhme et al. 2011; Eikenberg et al. 2015).

The relations between the frequency distributions of domain-mean BT amplitude (Fig. 5c) for observations and forecasts are similar to those of the cold cloud cover. In the observations, the IQR lies between 0.6 and 1.9 K, whereas in the model, this range extends from 1.0 to 2.6 K. Extrema reach up to 5.1 K in both distributions. Typical amplitude differences between COSMO-DE and SEVIRI are in the range of 0.0–1.1 K. To distinguish the effects of cloud cover and average cloud-top temperature excess on the amplitude metric, we consider the cluster-mean BT amplitude in Fig. 5d. It can be seen that both distributions are centered around a median  $\langle \Delta T \rangle_C$  of 9.6 and 9.9 K for the observations and model forecasts, respectively. However, the modeled distribution is narrower, which shows that the forecasts fail to reproduce the observational variability in cloud-top temperature.

To address the question of how the objects are spatially distributed in the simulations and observations, three metrics are investigated. First, the compactness radius is shown in Fig. 5e. The typical radius  $R_C$  that encloses the majority of the mass is around 300 km for the observations, as well as for the forecasts. The forecasted distribution of the compactness radius is slightly shifted toward higher values, probably due to the overestimate in cloud cover. Therefore, a larger amount of mass has to be covered. Differences in  $R_C$  between the COSMO-DE and SEVIRI observations range from –38 to 111 km and can also be caused by a more scattered spatial object distribution within the model. Negative distances herein represent situations in which the forecasted objects are spatially more aggregated than the objects within the observations. For the mass distance (Fig. 5f), the forecasts show a slight underestimation, with a median of 245 km instead of 266 km. Moreover, the IQRs of the differences between both distributions range from –80 to 61 km. The lower values indicate a more widespread object cluster in the model. Forecast extreme values reach up to a distance of 643 km, whereas for the observations, they go up to 790 km.



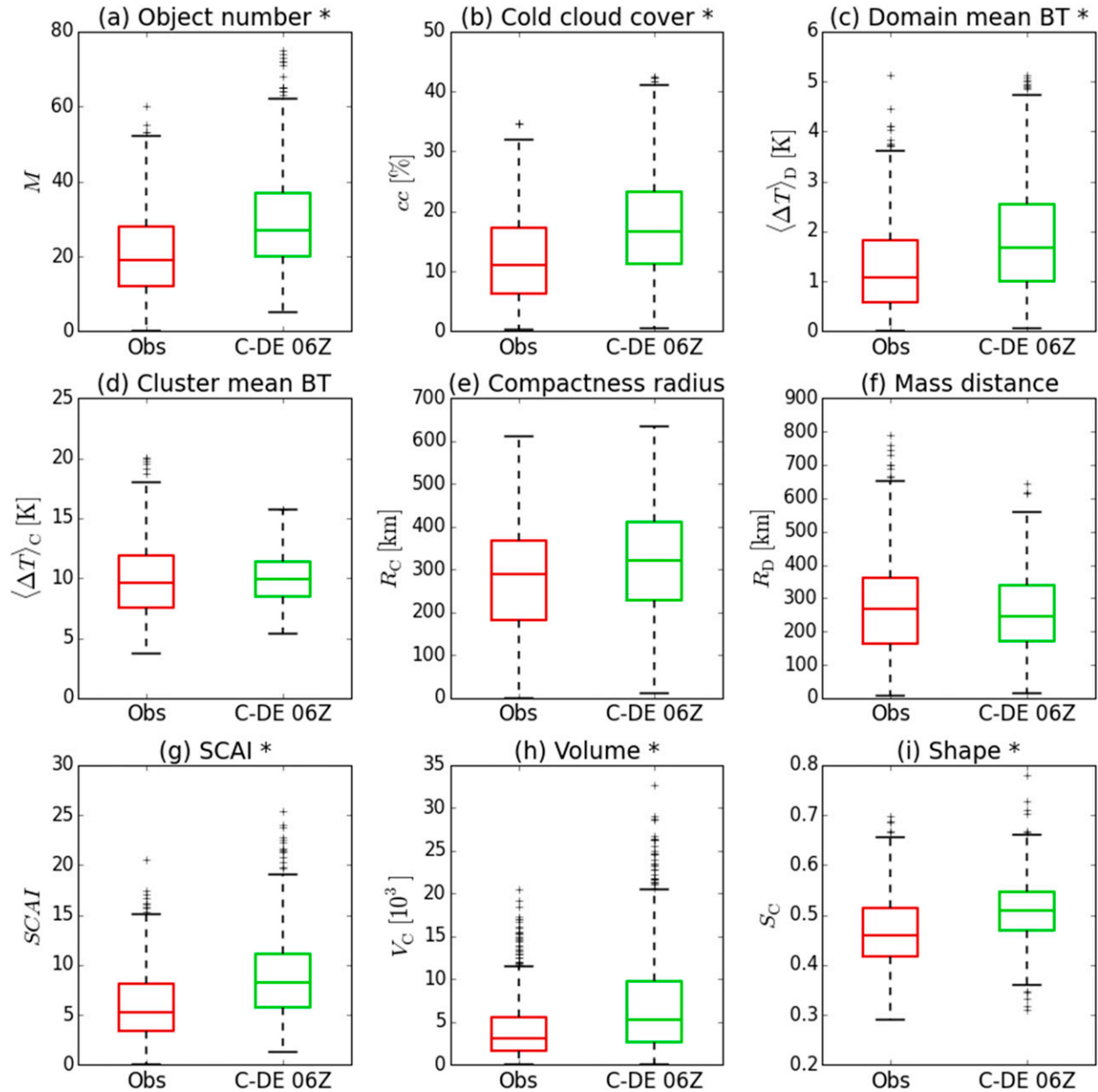


FIG. 5. Boxplots of frequency distributions of (a) object number, (b) cold cloud cover, (c) domain-average BT amplitude, (d) cluster-mean BT amplitude, (e) compactness radius, (f) mass distance, (g) SCAI, (h) volume, and (i) shape. Red boxes depict the observed frequency distribution whereas the green boxes show the predicted distribution of COSMO-DE 0600 UTC. The bottom (top) lines represent the 25th (75th) percentiles, and the box-center lines the median. Whiskers span 2.5 times the interquartile range and black crosses indicate outliers. A star behind the panel title marks distributions that are different at the 5% significance level.

These extremes represent clusters of a few cells near the edges or corners of the domain, since the longest distance between the domain center and a corner is 862 km. The last spatial metric is SCAI (Fig. 5g). The SCAI distribution of the model forecasts and the observations shows a similar pattern for this metric as for the object number. The median is shifted from 5.27 to 8.24, and in

addition the IQR is wider in the forecasts. If SCAI is, on the one hand, interpreted as an aggregation measure, the analysis suggests that the objects are more scattered in COSMO-DE compared to the SEVIRI observations. On the other hand, we already saw that the overestimation of the object number seems to be coupled to the bias in cold cloud cover, which is problematic for the

assessment of the simulated aggregation state. Furthermore, a higher value of SCAI can arise for different reasons, either as a result of a smaller number of objects that are far apart, or a larger number of objects that are clustered together. Moreover, the order-zero diameter  $R_0$  approaches a constant value for a sufficiently large number of objects.

The last two metrics consider the average structure of the objects. For this, first, the volume  $V_C$  is shown in Fig. 5h, which is a combination of object sizes and cloud shapes. We find a large overestimation of forecast  $V_C$ . The median is shifted to larger values, and also the IQR is wider than in the observations. The overestimation of volume is caused by two effects. First, because of the overestimation of cloud cover, the average object size is too large. This effect is not compensated by the simultaneous increase in object number. Second, the forecasts underestimate the maximum  $\Delta T_i$  values of each object (i.e., the intensity of the convective cores). This effect is identified by the consideration of the average cluster shape  $S_C$ , which is the average ratio between the mean and maximum  $\Delta T_i$  values within each object. In Fig. 5i, the IQR extends from 0.42 to 0.51 in the observations. In contrast, the forecasts exhibit an IQR that ranges from 0.47 to 0.55. Both distributions show extreme values above 0.68, which indicate very flat objects. However, the model reveals only a few values below 0.37, which indicate nearly conically shaped objects.

In general, this analysis shows that the two amplitude and structure metrics—domain-average BT amplitude  $\langle \Delta T \rangle_D$  and volume  $V_C$ —which are components of the SAL scoring method, are both sensitive to biases in the amount of cloud cover. The overestimation of forecasted cc is hence penalized twice in SAL, which is an undesirable behavior for the evaluation of cloud forecasts.

#### *b. Interdependencies between metrics*

There are several simple interdependencies between the considered metrics that arise from the construction of composite metrics as products and ratios of base metrics. Obviously, the term composite or base metric is somewhat arbitrary and depends on the viewpoint. Ignoring higher-order terms, the variance of a product can be decomposed into the individual variance contributions of its components and the covariance between them. This decomposition is now applied to the domain-average BT amplitude, which is approximately the product of the cloud cover and the cluster-mean BT amplitude. For observations and forecasts, respectively,  $\langle \Delta T \rangle_D$  has normalized variances of 0.5 and 0.34. Around 70% of these values are found to arise from variability in cloud cover, which is consistent in forecasts and

observations. In the observations, the remaining 30% of the  $\langle \Delta T \rangle_D$  variability results equally from the variability in  $\langle \Delta T \rangle_C$  and the covariability between  $\langle \Delta T \rangle_C$  and cc. In the forecasts, the  $\langle \Delta T \rangle_C$  variability is however a factor of 2 smaller than the covariability between  $\langle \Delta T \rangle_C$  and cc. Another composite metric is SCAI. The aggregation index achieves normalized variances of 0.33 and 0.22 for the observations and forecasts, respectively. Eighty (70) percent of the variability in SCAI can be attributed to the variability in the object number  $M$  for the observations (forecasts). The remainder is again approximately equally distributed over the variability in geometric object distance  $R_0$  and the covariability between  $M$  and  $R_0$ . Hence, if the forecasts, as observed here, produce too many objects, the forecasted larger values of SCAI do not have to be indicative of a less aggregated model state.

Figure 6 presents the joint occurrence frequencies of the metrics cc,  $\langle \Delta T \rangle_C$ ,  $R_C$ , SCAI, and  $S_C$ . In general, the joint distributions show a large scatter. Larger values of SCAI are obtained for larger values of cc and  $R_C$ . In addition, there is a strong dependence between  $\langle \Delta T \rangle_C$  and  $S_C$ , with clusters of larger BT amplitude leading to a flatter shape. Besides the previously noted biases, there is no obvious mismatch between the observations and the forecasts with regard to the interdependency of metrics.

## **5. Temporal behavior**

### *a. Variance decomposition*

Cold clouds and their associated characteristics can exhibit variations on diurnal to day-to-day time scales depending on the external synoptic forcing. Hence, evaluating the temporal behavior of cold cloud metrics can help to identify deficits in internal transformation processes, for instance microphysical conversion rates, in response to these external forcings. In the following, we study the variability of the proposed object-based metrics on different time scales and decompose them into daily mean, average diurnal, and residual intraday signals. The daily mean values are obtained by averaging the hourly time series in the interval between 0600 and 1800 UTC, thus representing typical daytime values. In total, 59 daily average values result for the chosen set of case days. The distributions of daily mean values of cc,  $\langle \Delta T \rangle_C$ ,  $R_C$ , SCAI, and  $S_C$  are shown in Fig. 6. The distributions of daily mean values present similar characteristics compared to the full distributions. The most probable daily mean values are consistently overestimated by the forecasts. In general, the missing extreme values lead to higher peaks in the central parts of

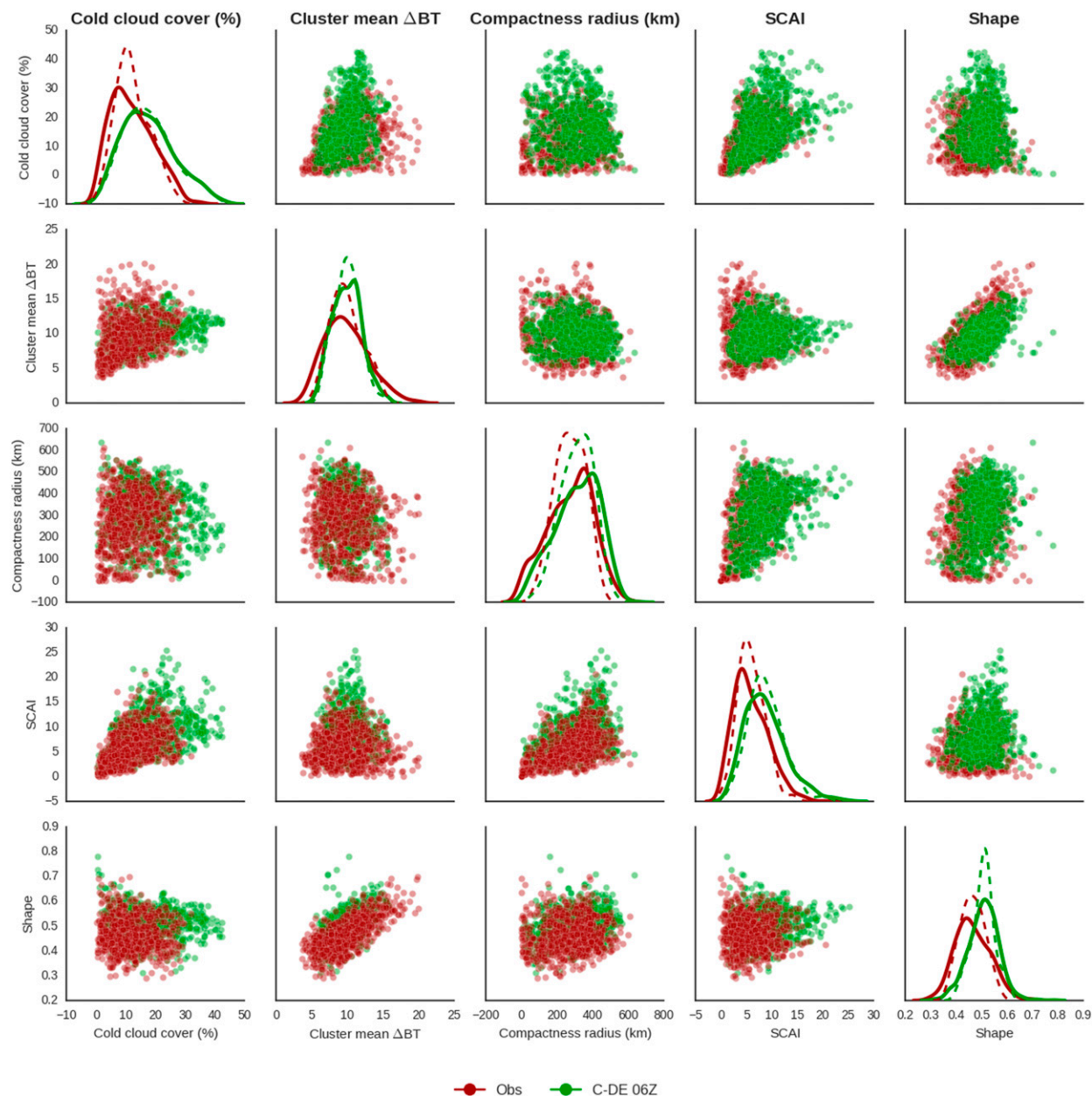


FIG. 6. Scatterplots of the joint distribution of two metrics. From left to right and from top to bottom the following variables are shown: cloud cover, cluster-mean BT amplitude, compactness radius, SCAI, and shape (red for observation and green for COSMO-DE 0600 UTC simulations). The diagonal shows the probability density functions of total (solid lines) and daily mean values (dashed lines) for the different parameters in arbitrary units. The scatterplots above the diagonal have their counterparts at transposed positions below the diagonal, however, in reversed plotting order, i.e., observations have been plotted on top of the simulations.

the distributions. One special case is the forecasted cc distribution, for which the daily mean density function already matches its full counterpart quite closely.

We furthermore consider the different contributions to the intraday variability. Therefore, the daily average values are subtracted from the time series in a first step. Thereafter, an average diurnal signal is calculated and

subtracted by aggregating and averaging values at the same time of day. A residual time series results that includes all the variability that is not explained by day-to-day or typical diurnal variations. In the following, we compare the forecast to the observed variance for different components of the time series of the metrics. All variance values have been normalized by the full

TABLE 2. Relative variance contributions (in %) for different metrics. The total observed variance was chosen to normalize the variance components related to day-to-day, average diurnal, and residual variability. The simulated variances have been tested against the observed variance using the nonparametric Brown–Forsythe test. The residual time series was subsampled in 6-h intervals before the variance test to remove effects from serial correlation. Boldface variance numbers indicate simulated variance contributions that are significantly different from the observation at the 5% level.

Statistic	Cold cloud cover		Cluster-mean BT		Compactness radius		SCAI		Shape	
	C-DE 0600		C-DE 0600		C-DE 0600		C-DE 0600		C-DE 0600	
	Obs	UTC	Obs	UTC	Obs	UTC	Obs	UTC	Obs	UTC
Relative day-to-day variance	57.8	<b>128.2</b>	47.6	32.3	39.5	45.8	54.1	94.7	44.1	39.4
Relative diurnal-mean variance	21.1	<b>2.8</b>	20.2	<b>1.8</b>	2.7	1.7	19.5	17.7	10.1	<b>4.6</b>
Relative residual variance	23.6	21.5	31.5	<b>9.8</b>	59.2	46.8	25.2	<b>34.8</b>	45.7	32.0

observed variance and are listed in Table 2. The variance components have been tested for the significance of their difference using the so-called Brown–Forsythe test, which scores the absolute deviations of the different time series (Brown and Forsythe 1974). In addition, the residual intraday time series have been examined for serial correlations. We have found decorrelation lengths between 3 and 6 h depending on the metric considered and thereafter decided to subsample the residual time series in 6-h intervals before the application of the significance test. The observed variance of cloud cover is to 58% attributable to day-to-day fluctuations. The remaining fractions are nearly equally governed by the average diurnal behavior and the residual intraday variability. The forecasted cloud cover, however, possesses more than twice the observed day-to-day variability and furthermore significantly underestimates the concurrent diurnal signal. Other significant deficits of forecast metrics are the missing intradaily variability of  $\langle \Delta T \rangle_C$ , the overestimation of residual intraday variance of SCAI, and a too small diurnal cycle signal in the forecasts of shape properties. From the above analysis, it becomes apparent that especially the correct representation of the diurnal cycle component is challenging for the model.

### b. Diurnal cycles

In the remaining section, we shed more light on the diurnal cycle of the five observed and forecasted metrics: cold cloud cover, cluster-mean BT, compactness radius, SCAI, and shape. The temporal evolution is analyzed and discussed separately for three different weather regimes.

#### 1) SOUTHWESTERLY

The time series in general can be split up into three periods. The morning hours are associated with a

cloud-dissipating phase, which is followed by a period of developing convection. Afterward, anvils of deep convective clouds begin to merge, which lead to a merging phase. The diurnal cycles of cloud properties are depicted in Fig. 7 for the southwesterly flow regime. The observed cluster-mean BT amplitude decreases until 1100 UTC during the cloud-dissipating phase. In contrast, observed cold cloud cover remains constant or even increases lightly. Furthermore, noticeable increases in the compactness radius and SCAI as well as shape can be identified. This implies that residual clouds do not become smaller, but split into more objects as a result of cloud-top warming. The convective growth phase can be detected from 1100 to 1400 UTC, when all observational metrics increase. Afterward, the compactness radius, SCAI, and shape begin to decrease, whereas cold cloud cover and cluster-mean BT continue to increase. The forecasts exhibit good agreement with the observed temporal evolution of the metrics. However, cold cloud cover and cluster-mean BT show a lower diurnal amplitude. The decrease in cluster-mean BT amplitude during the morning as well as the subsequent increase during the afternoon of cluster-mean BT is not well represented in the forecasts, which leads to negative biases of up to  $-3$  K in COSMO-DE 1200 UTC. The compactness radius and SCAI reveal a persistent overestimation due to a higher amount of objects. Further, COSMO-DE shows a more distinct peak in the compactness radius. Together with a small overestimation during the morning hours, the forecasted shape metric peaks around 2 h too early in the afternoon.

#### 2) TROUGH OVER WESTERN EUROPE

In contrast to the previous situation, the observations for the trough over western Europe weather regime show a generally higher cold cloud cover and



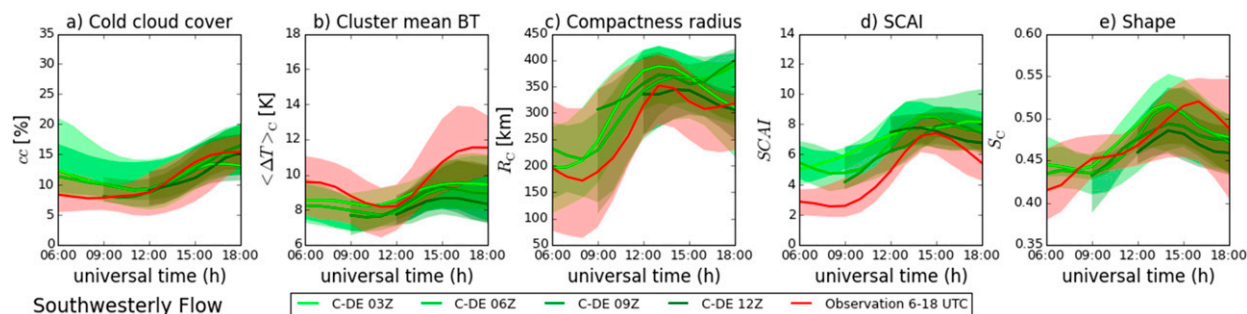


FIG. 7. Temporal evolution of five different metrics for 10 days with a southwesterly regime and for the time period from 0600 to 1800 UTC. (a) Cold cloud cover, (b) cluster-mean BT, (c) compactness radius, (d) SCAI, and (e) shape. The green curves indicate the median for four COSMO-DE runs (0300, 0600, 0900, and 1200 UTC) while the red line shows the median of the observations. The envelopes denote the interquartile range.

colder clouds (see Fig. 8). The observed cloud cover exhibits a small decrease during the morning hours and heavily increases around noon, resulting in a twice as large cold cloud fraction during the evening. Compactness radius and shape exhibit a decrease during the morning, which may be associated with merging effects, albeit, SCAI increases. The development of convection starts from 0900 UTC onward, when all metrics rise. The observed diurnal amplitude is much smaller for this weather regime compared to the southwesterly regime, with cold cloud cover being one notable exception. The forecasts also reveal a large positive bias for cold cloud cover, SCAI, and shape during the morning hours, whereas the cluster-mean BT and compactness radius are in a good agreement with the observations. The model forecasts more and smaller objects, which cover a larger area, however, with a similar spatial distribution and cloud-top temperature to the observations. From 0900 UTC onward, compactness radius and SCAI increase. This indicates new objects with a more spread out spatial distribution, which implies developing convection. However, the increase in compactness radius varies among each COSMO-DE run. The forecasts reach the peak in compactness radius and SCAI during

the afternoon, which is not in accordance with the observations.

### 3) SOUTHERLY

For case days with a southerly flow pattern, Fig. 9 depicts the diurnal cycle of the metrics. It can be seen that the observations show a distinct diurnal cycle. The cloud dissipation period can be detected between 0600 and 0900 UTC. The median cold cloud cover decreases as well as the cluster-mean BT amplitude. SCAI shows a slight increase during the morning hours. Before 1200 UTC, observed cold cloud cover, BT amplitude, and SCAI start to rise reaching maximal values during the late afternoon or evening. In contrast, compactness radius and shape exhibit a broad minimum between 0900 and 1500 UTC and strongly increase thereafter. Hence, all observation-based metrics show a pronounced diurnal cycle. In contrast, the forecasts are not able to capture this general temporal behavior. The forecasted cloud cover is heavily overestimated, and remains relatively constant throughout the day (similar for the simulated BT amplitude). Forecasted SCAI time series reach a maximum around 1500 UTC, but also have a much too small amplitude. The temporal evolution of the forecasted compactness radius and shape does not match

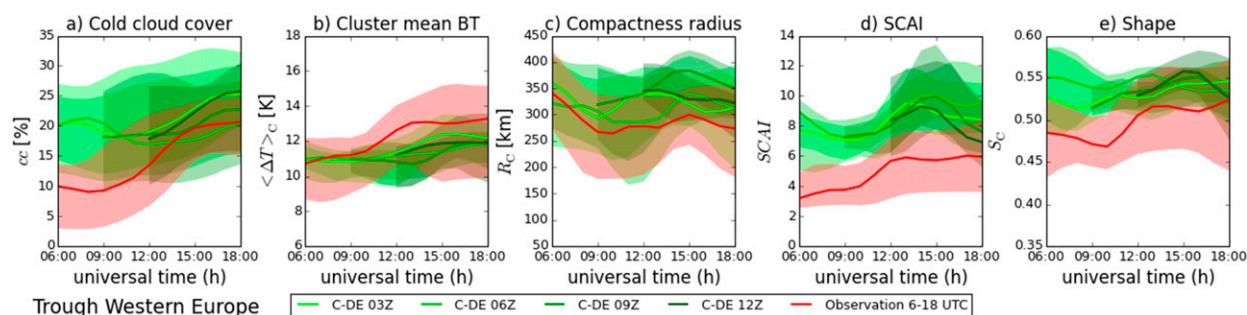


FIG. 8. Temporal evolution for 9 days during the trough over western Europe regime. Metrics and plot conventions are as in Fig. 7.



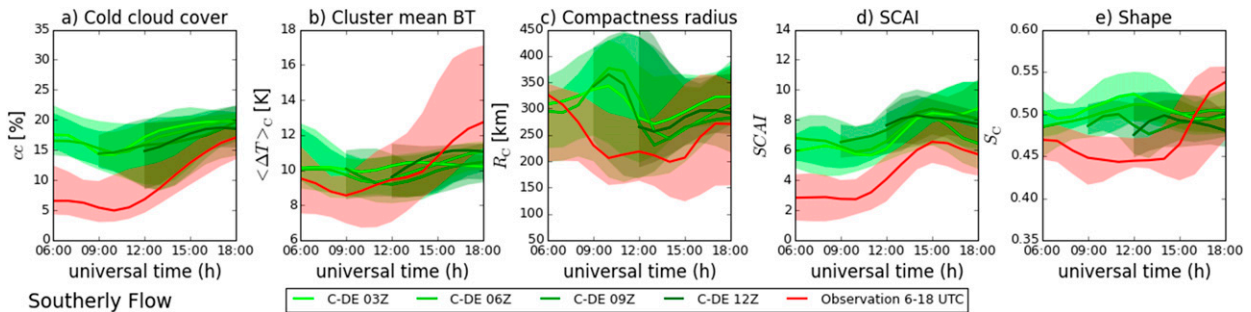


FIG. 9. Temporal evolution for 9 days during a southerly regime. Metrics and plot conventions are as in Fig. 7.

with the observation. Hence, depending on the weather situation, the forecasted diurnal behavior can exhibit large deviations from the observation.

## 6. Sensitivities

In the following, two types of sensitivity experiments that have been performed to assess possible weaknesses of the considered object-based metrics are discussed. First, as discussed by Senf and Deneke (2017), the calculation of synthetic satellite images is affected by some inherent uncertainties, which are caused by unconstrained microphysical and optical properties of ice and mixed-phase clouds. Resulting changes in infrared cloud emissivity lead to variations in window-channel BTs on the order of 2–4 K. To quantify this effect, the complete analysis presented so far has been repeated for the operationally available synthetic satellite imagery that has a less consistent formulation of cirrus particle size and subgrid-fractional cloud coverage. As a result, cold cloud cover remains significantly overestimated, with a median relative bias with respect to the observation increasing from 49% for the revised SynSat scheme to 91% for the operational SynSat scheme. Other parameters that are related to cc like domain-mean BT amplitude  $\langle \Delta T \rangle_D$  and volume  $V_C$  show an even more pronounced increase in the positive bias. The overestimation of the total number of objects and SCAI, however, decreases from 50% to around 30%. The decomposition of variance into different contributions remains similar for the operational SynSat. The main deficits again appear in the day-to-day variance of cloud cover, which is even more strongly overestimated, and the average diurnal component remains too weak. Additional differences are related to the underestimation of subdaily variance of cluster-mean BT amplitude and shape in the cold cloud cluster.

The second sensitivity experiment assesses the sensitivity of the object identification technique for different threshold choices. It has been intensively discussed in

Weniger and Friederichs (2016) that the SAL location and structure parameters can become unstable for certain parameter values. The chosen threshold for image segmentation is one of the most sensitive parameters. In contrast to our study, Weniger and Friederichs (2016) applied their object identification algorithm to geostationary water vapor imagery during winter conditions, and defined patches of moister upper-level air as targets for their object-based investigations. In general, we argue here that convective clouds occurring during the summertime seen by window-channel observations are closer to the subjective notion of distinct objects, at least in comparison to the more continuous moist air structures. We thus expect that the sensitivity of our object properties to changes in parameters is less pronounced. Nevertheless, we repeatedly derived the time series of object-based metrics for different BT thresholds in the wide range between 210 and 270 K. Each time series was again decomposed into daily average, average diurnal, and residual signals. In a further analysis step, the linear correlation between the time series components at the various BT thresholds with that of the threshold of 240 K was calculated. The correlation decreases monotonically as a function of the difference of the threshold and 240 K. This loss of correlation indicates a decrease in the coherence between the two signals. The faster the correlation decays, the more sensitive the considered metric is to changes in the threshold. We determined the  $e$ -folding values for two threshold ranges above and below 240 K. Figure 10a illustrates that there is a broad corridor between 230 and 250 K in which the daily mean and the higher-frequency residual component remain strongly correlated with the base setup, respectively. In general, setting the threshold to a colder BT by a certain amount has a larger impact than vice versa. This is also expected, as more and more convective cloud tops disappear for colder threshold values. The most sensitive parameters are the compactness radius  $R_C$ , the aggregation index SCAI, and the cluster shape  $S_C$ . We furthermore analyzed the impact of the threshold choice on

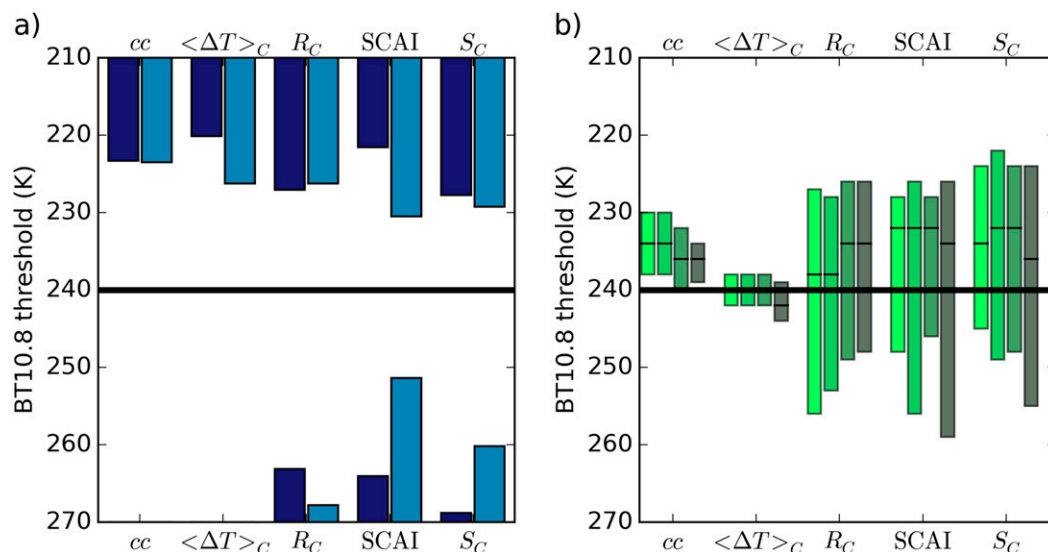


FIG. 10. Visualization of threshold sensitivity of five metrics: cloud cover, cluster-mean BT amplitude, compactness radius, SCAI, and shape. (a) The dark blue bars indicate the range at which the correlation between the daily average time series derived with a certain threshold and the one with 240 K falls below the  $e$ -folding value. Across the  $e$ -folding values derived from observations and the simulations, the one closest to the base threshold of 240 K was chosen. Light blue bars indicate the same, but for the residual time series component. If no blue bars appear, then the  $e$ -folding value is outside the plotting range. (b) The green bars show the interquartile range of optimal threshold values that minimize the daily bias between observations and simulations. From left to right, the COSMO-DE initialization times increase from 0300, 0600, 0900, to 1200 UTC (light to dark green). The median is also indicated by black horizontal lines within the green boxes.

the day-to-day bias between forecasts and observations. Therefore, we calculated the optimal threshold that minimizes the bias between forecast and observed metrics. The threshold adjustment was only performed for the forecasts and was done for each of the case days independently. The resulting distributions of optimal forecast thresholds are shown in Fig. 10b. Biases in  $cc$  and  $\langle \Delta T \rangle_C$  can be compensated by much smaller changes of the threshold compared to the changes for  $R_C$ , SCAI, and  $S_C$ . The significant spread of optimal thresholds for the latter three metrics again suggests that the metrics are more sensitive to and are influenced less systematically by the threshold choice. In addition, if the spread approaches the  $e$ -folding values, no meaningful coherence between forecasts and observations is expected.

## 7. Conclusions

In this study, object-based metrics are investigated as part of an evaluation of forecasts of cold cloud characteristics with infrared geostationary satellite observations. For this purpose, operational forecasts of the German Weather Service's COSMO-DE model as well as Meteosat SEVIRI observations have been examined. The convection-permitting cloud forecasts have been

transferred into synthetic brightness temperatures using radiative transfer calculations with the RTTOV model utilizing a revised SynSat scheme, in order to facilitate a direct comparison with Meteosat observations. The identification of individual cloud objects in the  $10.8\text{-}\mu\text{m}$  brightness temperature fields is based on an image segmentation technique that uses a threshold of 240 K and identifies connected areas of colder brightness temperatures. Several object-based metrics have been defined on the overall cluster, which comprise the total set of cloud objects. Similar to the SAL technique that was developed to evaluate precipitation forecasts in river catchments, the proposed metrics were sorted into three categories related to (i) the amplitude or amount, (ii) the location or spatial arrangement, and (iii) the structure of the cluster. In particular, no direct match between individual objects was required in the statistical analysis. A dataset of 59 case days with active deep moist convection spanning the summer-half years between 2012 and 2014 has been collected, and forecasts initialized at four times (0300, 0600, 0900, and 1200 UTC) were evaluated. We have applied a temporal decomposition of the metric time series into different components related to the daily mean day, average diurnal cycle, and residual variability. Furthermore, diurnal cycles of several metrics between 0600 and 1800 UTC have been

analyzed with regard to their temporal behavior and ability to distinguish different weather regimes.

The distributional characteristics of different object-based metrics have been analyzed for and compared between observations and forecasts. It is shown that the forecasts significantly overestimate six out of nine metrics. The affected metrics are either related to the amplitude and amount of simulated cold clouds, such as cold cloud cover and domain-mean BT amplitude, or to the structure of the cloud cluster, such as volume or shape. Because of its close connection to the total object number, the aggregation measure SCAI is also overestimated. However, we question whether this is in fact due to a less aggregated convection state within the forecasts, but present evidence that this is due to an overestimation of the amount of cloud itself. This is further supported by the analysis of the variability of SCAI, which shows that around 80% of the variability is determined by the variability in the object number, whereas only the remaining smaller part is influenced by the spatial arrangement of the cold clouds. In addition, it is discussed that a general overestimate of cloud cover influences the domain-mean BT as an amplitude metric and the volume as a structure metric in a similar way. Such a model deficit is therefore penalized twice, if both are combined into a composite metric such as SAL, which might not be desirable for the purposes of evaluating cloud forecasts. In addition, we have studied the joint behavior of the metrics cloud cover, cluster-mean BT amplitude, compactness radius, SCAI, and shape. In general, the joint distributions between two metrics show a large scatter, but are in good agreement for observations and forecasts, besides the already mentioned biases of the individual metrics. Systematically larger values of SCAI are found for larger values of cloud cover and compactness radius. In addition, a strong dependence between cluster-mean BT amplitude and cluster shape is identified, leading to the situation where clusters with higher average BT values possess a flatter shape.

The temporal behavior of the observed and forecasted cloud cover, cluster-mean BT amplitude, compactness radius, SCAI, and shape is thoroughly analyzed. The cloud cover variability for daily means is significantly overestimated by the forecasts, whereas the opposite is true for the average diurnal cycle. Further model weaknesses are related to the underrepresentation of intradaily variability in cluster-mean BT amplitude and shape. The average diurnal cycle is additionally presented for the three most frequent weather regimes among the case days. For a southwesterly regime, the forecasts exhibit the most convincing agreement with the observations. However, the observed late afternoon

increases in cloud cover, cluster-mean BT amplitude, and SCAI are underestimated. The agreement deteriorates for the other two flow regimes: trough over western Europe and southerly. For both regimes, the observed cloud cover shows a pronounced diurnal cycle with a doubling of the cloud amount from morning to late afternoon. The forecasts significantly overestimate the cloud cover in the morning hours, therefore leading to a much smaller amplitude of the diurnal cycle in cloud cover. Similarly, deficits appear for the diurnal behavior of the cluster-mean BT amplitude and shape. This illustrates that both aspects, the change in the areal extent of convective clouds as well as the change and magnitude of convective cloud-top temperature, are inherently problematic for the considered convection-permitting forecasts.

The sensitivity of the proposed object-based metrics to changes in the satellite forward operator as well as to changes to the threshold used in the object identification has been investigated. First, we have reexamined the operationally available synthetic imagery that can deviate from our revised SynSat scheme by several degrees kelvin. The analysis shows an even more pronounced overestimation of cold cloud cover. Deficits in the intradaily variability of cloud cover, cluster-mean BT amplitude, and shape remain significant for the operational synthetic imagery. Second, we have assessed the temporal coherence between metric time series components derived with a standard threshold of 240 K, and a threshold shifted to either smaller or larger BT values. In general, we find that threshold changes between 230 and 250 K do not lead to a loss of temporal coherence, indicating a certain level of robustness of our results to threshold changes. Decreasing the threshold toward colder BTs results in higher sensitivity compared to an increase. We furthermore analyzed day-to-day threshold adjustments to minimize the bias between the observed and forecast metrics. For cloud cover and cluster-mean BT amplitude, these bias adjustments show smaller spread compared to compactness radius, SCAI, and shape, which illustrates that the latter measures are more uncertain, and are also harder to constrain by the observations.

The current study utilizes measurements of cloud-affected infrared radiation from only a single spectral channel, the 10.8- $\mu\text{m}$  window channel. An obvious extension of the current approach for object identification is the application of multispectral methods (e.g., [Derrien and Le Gléau 2005](#)) to distinguish transparent or semi-transparent cirrus clouds from deep convective cores, which we will pursue in the future. This would allow for a more specific and targeted evaluation of cloud properties depending on the predefined cloud types. We

additionally emphasize here that a joint evaluation of cloud and precipitation characteristics can have the potential to further increase our understanding of current state-of-the-art model deficits and might illuminate pathways of future improvements of simulating the effects of deep convective clouds for the climate system.

**Acknowledgments.** The study was initialized within the framework of the Hans Ertel Center for Weather Research (Project OASE) by the German Weather Service (DWD). FS's research was partly funded by the abovementioned project, and the HD(CP)<sup>2</sup> initiative funded by the BMBF under Grant 01LK1507C. We thank DWD for providing COSMO-DE data, EUMETSAT for provision of MSG SEVIRI data, and the NWPSAF team for their efforts in the development of RT-TOV. We further thank two anonymous reviewers for their valuable comments and suggestions.

## REFERENCES

- Akkermans, T., and Coauthors, 2012: Regime-dependent evaluation of accumulated precipitation in COSMO. *Theor. Appl. Climatol.*, **108**, 39–52, doi:[10.1007/s00704-011-0502-0](https://doi.org/10.1007/s00704-011-0502-0).
- Baldauf, M., A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer, and T. Reinhardt, 2011: Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Mon. Wea. Rev.*, **139**, 3887–3905, doi:[10.1175/MWR-D-10-05013.1](https://doi.org/10.1175/MWR-D-10-05013.1).
- Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, doi:[10.1175/WAF933.1](https://doi.org/10.1175/WAF933.1).
- Baur, F., 1963: *Grosswetterkunde und langfristige Witterungsvorhersage*. Akademische Verlagsgesellschaft, 91 pp.
- Bikos, D., and Coauthors, 2012: Synthetic satellite imagery for real-time high-resolution model evaluation. *Wea. Forecasting*, **27**, 784–795, doi:[10.1175/WAF-D-11-00130.1](https://doi.org/10.1175/WAF-D-11-00130.1).
- Böhme, T., and Coauthors, 2011: Long-term evaluation of COSMO forecasting using combined observational data of the GOP period. *Meteor. Z.*, **20**, 119–132, doi:[10.1127/0941-2948/2011/0225](https://doi.org/10.1127/0941-2948/2011/0225).
- Brown, M. B., and A. B. Forsythe, 1974: Robust tests for the equality of variances. *J. Amer. Stat. Assoc.*, **69**, 364–367, doi:[10.1080/01621459.1974.10482955](https://doi.org/10.1080/01621459.1974.10482955).
- Casati, B., and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.*, **15**, 3–18, doi:[10.1002/met.52](https://doi.org/10.1002/met.52).
- Crocker, R., and M. Mittermaier, 2013: Exploratory use of a satellite cloud mask to verify NWP models. *Meteor. Appl.*, **20**, 197–205, doi:[10.1002/met.1384](https://doi.org/10.1002/met.1384).
- Davis, C. A., B. G. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, doi:[10.1175/MWR3145.1](https://doi.org/10.1175/MWR3145.1).
- , —, and —, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, doi:[10.1175/MWR3146.1](https://doi.org/10.1175/MWR3146.1).
- , —, —, and J. Halley-Gotway, 2009: The Method for Object-Based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24**, 1252–1267, doi:[10.1175/2009WAF2222241.1](https://doi.org/10.1175/2009WAF2222241.1).
- Derrien, M., and H. Le Gléau, 2005: MSG/SEVIRI cloud mask and type from SAFNWC. *Int. J. Remote Sens.*, **26**, 4707–4732, doi:[10.1080/01431160500166128](https://doi.org/10.1080/01431160500166128).
- 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, doi:[10.1016/j.atmosres.2008.10.020](https://doi.org/10.1016/j.atmosres.2008.10.020).
- Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, doi:[10.1002/met.25](https://doi.org/10.1002/met.25).
- , and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.*, **239**, 179–202, doi:[10.1016/S0022-1694\(00\)00343-7](https://doi.org/10.1016/S0022-1694(00)00343-7).
- , and W. A. Gallus Jr., 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. *Wea. Forecasting*, **24**, 1401–1415, doi:[10.1175/2009WAF2222252.1](https://doi.org/10.1175/2009WAF2222252.1).
- Eikenberg, S., C. Köhler, A. Seifert, and S. Crewell, 2015: How microphysical choices affect simulated infrared brightness temperatures. *Atmos. Res.*, **156**, 67–79, doi:[10.1016/j.atmosres.2014.12.010](https://doi.org/10.1016/j.atmosres.2014.12.010).
- Feidas, H., and A. Giannakos, 2012: Classifying convective and stratiform rain using multispectral infrared Meteosat Second Generation satellite data. *Theor. Appl. Climatol.*, **108**, 613–630, doi:[10.1007/s00704-011-0557-y](https://doi.org/10.1007/s00704-011-0557-y).
- Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, doi:[10.1175/2009WAF2222269.1](https://doi.org/10.1175/2009WAF2222269.1).
- Hanley, K., D. Kirshbaum, N. Roberts, and G. Leoncini, 2013: Sensitivities of a squall line over central Europe in a convective-scale ensemble. *Mon. Wea. Rev.*, **141**, 112–133, doi:[10.1175/MWR-D-12-00013.1](https://doi.org/10.1175/MWR-D-12-00013.1).
- Jankov, I., and Coauthors, 2011: An evaluation of five ARW-WRF microphysics schemes using synthetic GOES imagery for an atmospheric river event affecting the California coast. *J. Hydrometeorol.*, **12**, 618–633, doi:[10.1175/2010JHM1282.1](https://doi.org/10.1175/2010JHM1282.1).
- Kann, A., I. Meirold-Mautner, F. Schmid, G. Kirchengast, J. Fuchsberger, V. Meyer, L. Tüchler, and B. Bica, 2015: Evaluation of high-resolution precipitation analyses using a dense station network. *Hydrol. Earth Syst. Sci.*, **19**, 1547–1559, doi:[10.5194/hess-19-1547-2015](https://doi.org/10.5194/hess-19-1547-2015).
- Keller, M., O. Fuhrer, J. Schmidli, M. Stengel, R. Stöckli, and C. Schär, 2016: Evaluation of convection-resolving models using satellite data: The diurnal cycle of summer convection over the Alps. *Meteor. Z.*, **25**, 165–179, doi:[10.1127/metz/2015/0715](https://doi.org/10.1127/metz/2015/0715).
- Kidd, C., E. Dawkins, and G. Huffman, 2013: Comparison of precipitation derived from the ECMWF operational forecast model and satellite precipitation datasets. *J. Hydrometeorol.*, **14**, 1463–1482, doi:[10.1175/JHM-D-12-0182.1](https://doi.org/10.1175/JHM-D-12-0182.1).
- Lindstedt, D., P. Lind, E. Kjellström, and C. Jones, 2015: A new regional climate model operating at the meso-gamma scale: performance over Europe. *Tellus*, **67A**, 24138, doi:[10.3402/tellusa.v67.24138](https://doi.org/10.3402/tellusa.v67.24138).
- Machado, L. A. T., and J.-P. Chaboureaud, 2015: Effect of turbulence parameterization on assessment of cloud organization. *Mon. Wea. Rev.*, **143**, 3246–3262, doi:[10.1175/MWR-D-14-00393.1](https://doi.org/10.1175/MWR-D-14-00393.1).
- Morcrette, J.-J., 1991: Evaluation of model-generated cloudiness: Satellite-observed and model-generated diurnal variability of

- brightness temperature. *Mon. Wea. Rev.*, **119**, 1205–1224, doi:[10.1175/1520-0493\(1991\)119<1205:EOMGCS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1991)119<1205:EOMGCS>2.0.CO;2).
- Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, doi:[10.1175/1520-0434\(1993\)008<0281:WIAGFA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2).
- Negri, R. G., L. A. T. Machado, S. English, and M. Forsythe, 2014: Combining a cloud-resolving model with satellite for cloud process model simulation validation. *J. Appl. Meteor. Climatol.*, **53**, 521–533, doi:[10.1175/JAMC-D-12-0178.1](https://doi.org/10.1175/JAMC-D-12-0178.1).
- Pfeifer, M., and Coauthors, 2010: Validating precipitation forecasts using remote sensor synergy: A case study approach. *Meteor. Z.*, **19**, 601–617, doi:[10.1127/0941-2948/2010/0487](https://doi.org/10.1127/0941-2948/2010/0487).
- Roca, R., and V. Ramanathan, 2000: Scale dependence of monsoonal convective systems over the Indian Ocean. *J. Climate*, **13**, 1286–1298, doi:[10.1175/1520-0442\(2000\)013<1286:SDOMCS>2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013<1286:SDOMCS>2.0.CO;2).
- Rossa, A., P. Nurmi, and E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts. *Precipitation: Advances in Measurement, Estimation and Prediction*, S. C. Michaelides, Ed., Springer, 419–452.
- Saunders, R., M. Matricardi, and P. Brunel, 1999: An improved fast radiative transfer model for assimilation of satellite radiance observations. *Quart. J. Roy. Meteor. Soc.*, **125**, 1407–1425, doi:[10.1002/qj.1999.49712555615](https://doi.org/10.1002/qj.1999.49712555615).
- Schmetz, J., P. Pili, S. Tjemkes, D. Just, J. Kerkmann, S. Rota, and A. Ratier, 2002: An introduction to Meteosat Second Generation (MSG). *Bull. Amer. Meteor. Soc.*, **83**, 977–992, doi:[10.1175/1520-0477\(2002\)083<0977:AITMSG>2.3.CO;2](https://doi.org/10.1175/1520-0477(2002)083<0977:AITMSG>2.3.CO;2).
- Schneider, S., Y. Wang, W. Wagner, and J.-F. Mahfouf, 2014: Impact of ASCAT soil moisture assimilation on regional precipitation forecasts: A case study for Austria. *Mon. Wea. Rev.*, **142**, 1525–1541, doi:[10.1175/MWR-D-12-00311.1](https://doi.org/10.1175/MWR-D-12-00311.1).
- Senf, F., and H. Deneke, 2017: Uncertainties in synthetic Meteosat SEVIRI infrared brightness temperatures in the presence of cirrus clouds and implications for evaluation of cloud microphysics. *Atmos. Res.*, **183**, 113–129, doi:[10.1016/j.atmosres.2016.08.012](https://doi.org/10.1016/j.atmosres.2016.08.012).
- Tobin, I., S. Bony, and R. Roca, 2012: Observational evidence for relationships between the degree of aggregation of deep convection, water vapor, surface fluxes, and radiation. *J. Climate*, **25**, 6885–6904, doi:[10.1175/JCLI-D-11-00258.1](https://doi.org/10.1175/JCLI-D-11-00258.1).
- Weniger, M., and P. Friederichs, 2016: Using the SAL technique for spatial verification of cloud processes: A sensitivity analysis. *J. Appl. Meteor. Climatol.*, **55**, 2091–2108, doi:[10.1175/JAMC-D-15-0311.1](https://doi.org/10.1175/JAMC-D-15-0311.1).
- Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487, doi:[10.1175/2008MWR2415.1](https://doi.org/10.1175/2008MWR2415.1).
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.