Paper IV

Christina A. Pedersen, Fred Godtliebsen and Andreas C. Roesch, "A Scale-Space Approach for Detecting Significant Differences between Models and Observations Using Global Albedo Distributions", Accepted, *Journal of Geophysical Research*, 2007.

A Scale-Space Approach for Detecting Significant Differences between Models and Observations Using Global Albedo Distributions

Christina A. Pedersen

Norwegian Polar Institute and University of Tromsø, Tromsø, Norway

Fred Godtliebsen

University of Tromsø, Tromsø, Norway

Andreas C. Roesch

Swiss Federal Institute of Technology, Zurich, Switzerland

Abstract. This paper describes how a statistical scale-space technique can be used for evaluating climate models. A difference image between model and validation data is used as input. Hypothesis testing is performed at each difference pixel for a broad range of image resolutions (or scales). This approach circumvents some of the classical problems of hypothesis testing. An area, at a particular scale, is claimed to be significant if it is sufficiently different from zero in the difference image. such differences are called features. As the scale gradually increases from fine to coarse, features are created, they grow and merge and may finally annihilate. The scale-space algorithm produces maps for statistical inference and the degree of significance at different locations. The adapted scale-space technique was applied for validation of ECHAM5 Global Circulation Model surface albedo against a remote sensing surface albedo climatology. Overall, the largest discrepancies were detected over snow and ice-covered areas, and ECHAM5 was found to overestimate the albedo compared to the albedo climatology for all scales in March. Successively coarser spatial scales resulted in more and larger significant areas in the difference image. At the finest scales (280 km) very few areas of significant albedo differences were detected because of relatively high interannual variability for the areas of largest difference. At 1100 km, significant albedo differences were found in the southern part of the Arctic Ocean adjacent to the ice edge, probably because of the different positions of the ice edge in the two datasets. A scale of 2500 km was found to be reasonable for validating albedo as the statistical significance agrees well with differences meaningful from a climatologist's point of view. At this scale most of the snow-covered regions in Northern Eurasia with high positive differences and relatively low interannual variability were found to be significant.

1. Introduction

General circulation models (GCMs) are used for simulating past, present and future climates. Model results are usually evaluated against ground observations. For past climates, proxy data (e.g., information on sediment- and/or ice cores) are sometimes converted into physical parameters and used for evaluation of model outputs; for present climate, both ground-based and remotely sensed information is used for evaluation. GCMs can also be used for sensitivity studies, e.g. by changing the boundary conditions or a physical parameterization, and the experimental run is typically compared against the control run. Even if modifications in the experimental run have no effect on the climate, the difference field between the control and experimental run will deviate from zero and reflect random variations [von Storch and Zwiers, 1999]. Similarly, the difference field between simulated and observed climate parameters can exhibit features (possibly large scale features), even if the model is "correct". Therefore it is necessary to apply statistical techniques to distinguish between deterministic model error and internal model noise in the validation of climate models.

Statistical techniques can identify areas in an image where the difference field between a model and the validation data are higher than the noise level, that is, the difference is statistically significant [Chervin and Schneider, 1976]. However, for a climatologist, "the practical significance" may be of greater importance. To clarify: statistical significance can be interpreted by asking if the means of two datasets are the same or different, which is typical in hypothesis testing. However, when assessing the practical significance, one needs to quantify the magnitude of the difference. This is because a very small, subtle difference can be found to be statistically significant given a large enough sample size, even though the difference is of little practical importance (Wikipedia, 2007).

Albedo is the ratio of reflected to incoming solar radiation, and is crucial for the Earth's heat exchange. Snow and sea-ice albedo is an important parameter in GCMs at high latitudes because of its strong positive feedback properties [Curry et al., 1995]. However, current snow and sea-ice albedo parameterizations are generally still oversimplified [Pedersen and Winther, 2005].

In this paper, we use an objective statistical method for spatial comparison of GCM simulated albedo against remote

0148-0227/07/$9.00

sensing albedo climatology. We advance the comparison from purely qualitative to more quantitative by investigating whether the difference field between the model and validation data is sufficiently different from zero, and if so, in what way and where.

A scale-space methodology, Significance in Scale-Space (S^3), has previously been developed for answering questions like these [Godtliebsen et al., 2004]. S^3 finds features in a noisy image which are strong enough to be distinguished from background noise. Significant features in the image (defined in S^3 as first and second derivatives) are identified through a multi-scale procedure, and changes in image features at different resolutions are detected. At finer scales there is usually a large amount of noise, while successively coarser scales smooth the data and reduce the noise. However at very coarse scales interesting features can be smoothed away entirely. This paper proposes an adapted version of S^3 for finding features in a noisy image which are strong enough to be distinguished from background noise based on the difference image itself. The interesting features are the differences between image intensities, or more specifically, the differences between model and validation data. The adapted version is specifically tuned towards users such as climate scientists; however, it may be used in other contexts as well.

The adapted version differs from the original version in several ways, most importantly it concerns features and makes inference about the signal itself, not the first or second derivative of the signal. This is because estimates of scale-space derivatives are more susceptible to sampling variation than estimates of the scale-space signal itself, hence derivative-based inference will have less statistical power, meaning that fewer (real) features will be flagged as significant [Godtliebsen et al., 2004]. In addition the adapted version gives plots that are easier to interpret when used for validation purposes. This can be illustrated by the following: suppose there is a small peak in the difference image. The adapted version will typically detect a connected area of pixels with a significant difference for a given scale. If derivative based inference (and the original S^3) is used, the image will show a circle of significant pixels around the peak. Hence, there will be areas where the derivative feature plot does not show significant difference although the two images clearly are significantly different. By careful interpretation, this can be understood from the derivative feature plots, but the adapted version shows this in a clearer and more concise way.

The task of producing a quantitative measure of similarity between two images is encountered in many disciplines, including image processing, pattern recognition, landscape ecology, hydrology and multimedia [Haralick and Shapiro, 1992]. Image segmentation [Wealands et al., 2005; Pal and Pal, 1993] is a common approach for detecting structural features, which includes thresholding [Haralick and Shapiro, 1992], clustering [Pauwels and Frederix, 1999] and region growing and -merging.

This paper is broadly divided in two parts, where the methodology behind the adapted S^3 is described in the first part (Section 2). Section 3 describes the surface albedo simulations from ECHAM5 General Circulation Model and the remote sensing PINKER albedo climatology [Pinker, 1985] stated to represent ground truth validation data. In Section 4, the adapted S^3 is used for validating ECHAM5 surface albedo against PINKER albedo climatology. The variability of the method is assessed with various choices of algorithm parameters (scale, variance and significance level). Section 5 discusses the results and the scale-space method, as well as discussing and comparing statistical significance against practical significance. Conclusions are presented in Section 6. The adapted S^3 algorithm, documentation and examples are available at http://www3.npolar.no/∼xtina for the readers convenience. The algorithms can be downloaded and run on a personal computer using Matlab.

2. Scale-space methodology and the adapted S^3 algorithm

The ideas behind scale-space were first introduced by Lindeberg [1994] in the field of computer vision, where successive smoothing was applied to represent coarser scales. Chaudhuri and Marron [1999] first developed the methodology for detecting significant zero-crossings of the derivative in a one-dimensional signal (SiZer). Significance in scale-space methodology also exists for bivariate density estimation [Godtliebsen et al., 2002], random design [Ganguli and Wand, 2004], dependent time series [Park et al., 2004] and in a Bayesian framework [Godtliebsen and Øigård, 2005]. The one dimensional scale-space methodology has been used by Godtliebsen et al. [2003] and Karløf et al. [2005] for detecting significant peaks in climatological time series. Other approaches are also used for multi-scale image processing. Hay et al. [2002] reviewed the recent development of a scale-space technique via a non-mathematical primer, and present a multi-scale analysis method, where scale-space-blob features are detected from a stack of scale-space smooths. Itti et al. [1998] present a visual attention system, where multi-scale image features are combined into a single topographical saliency map and a neural network selects attended locations. For more in-depth treatment of scale see e.g. Marceau [1999] and Marceau and Hay [1999].

The basic idea behind scale-space is to develop a system for determining the scale of a feature, and how to search for it, before knowing what kind of feature being studied and where it is located [Lindeberg, 1994]. Scale is said to represent the filter or measuring tool with which the data are viewed or quantified, and a feature can only exist over a specific range of scales [Levin, 1992]. This means that the type of information obtained is largely determined by the relationship between the actual size of the feature and the resolution of the filters used to extract information [Hay et al., 2002]. When the scale is unknown, the most reasonable approach is to investigate the data at all or at least at multiple scales.

The basic quantity in the adapted S^3 is the difference image (model data minus validation data), and we define a "feature" as areas where pixels in the difference image are significantly different from zero. A feature can be positive or negative, and a positive (negative) difference means the model response is too high (low). These features, or areas of significantly different pixels, will follow the path of scale-space events [Hay et al., 2002]; (i) creation, where a new feature appears, (ii) growing, where a feature grows in size, (iii) merging, where two features merge into one and (iv) annihilation, where a feature disappears. The Gaussian kernel (discussed later) does not allow splitting, the last scale-space event where one feature splits into two [Hay et al., 2002].

2.1. The adapted S^3

The content below is parallel to Section 3 in Godtliebsen et al. (2004), but full details are given here to make the underpinnings of the adapted S^3 clear. The statistical model for the difference image is

$$Y(i,j) = f(i,j) + \epsilon(i,j), \tag{1}$$

where $i = 1,\ldots,n$ and $j = 1,\ldots,m$ index the pixel locations, $f(i,j)$ is the underlying unknown deterministic difference image, and $\epsilon(i,j)$ is the stochastic noise signal, assumed to be independent random variables. A smoothed scale-space version of f, f_h, is estimated by

$$\hat{f}_h(i,j) = \sum_{i'=1}^{n}\sum_{j'=1}^{m} Y(i',j')K_h(i-i',j-j'), \tag{2}$$

where K_h is a smoothing kernel function. Choosing K_h as the spherical, symmetric Gaussian density has many advantages over using other kernels [Chaudhuri and Marron, 1999; Hay et al., 2002], e.g. the number of zero crossings of the smooth is always a decreasing function of the scale (which is not true for any other kernel). As Hay et al. [2002] state: "the Gaussian kernel's use in scale-space theory is not by chance, but instead reflects strict purpose, design and evaluation". The bandwidth parameter h in the Gaussian kernel controls the degree of smoothness (i.e. the scale), and in practice the user defines the range of scales.

2.2. Variance estimation and hypothesis testing

The variance of the smoothed image plays an important role in the hypothesis test underlying the decision, and is estimated from

$$\begin{aligned} &\mathrm{Var}\left[\hat{f}_h(i,j)\right] = \\ &\mathrm{Var}\left[\sum_{i'=1}^{n}\sum_{j'=1}^{m} Y(i',j')K_h(i-i',j-j')\right] = \\ &\sum_{i'=1}^{n}\sum_{j'=1}^{m} K_h(i-i',j-j')^2\mathrm{Var}[Y(i',j')], \end{aligned} \tag{3}$$

where the variance of $Y(i,j)$ must be specified. Since the signal f is assumed to be deterministic, the variance of Y reflects the noise variance (cf. equation (1)). The variance can be known apriori or it can be estimated from the images, and furthermore, it can be assumed constant over the image, or it may be spatially varying. Summarized, four approaches for variance estimation are presented and it is left to the user to choose the approach that suits the particular problem best:

1. Apriori knowledge of constant variance, i.e. $\mathrm{Var}[Y(i,j)] = \sigma^2$, where σ^2 is known.

2. Apriori knowledge of spatially varying variance, i.e. $\mathrm{Var}[Y(i,j)] = \sigma^2(i,j)$, where $\sigma^2(i,j)$ is known.

3. Assumes constant variance, i.e. $\widehat{\mathrm{Var}[Y(i,j)]} = s^2$, is estimated globally from the image. The standard variance estimator or other estimators can be used.

4. Assumes spatially varying variance, i.e. $\widehat{\mathrm{Var}[Y(i,j)]} = s^2(i,j)$, is estimated locally from the image. The window size for estimating $s^2(i,j)$ should depend on the smoothing level. The standard variance estimator or other estimators can be used.

Approach 1 and 3 above are normally required for in-situ sampled data on a spatially regular grid under equal weather, light and atmospheric conditions. In other words, if the surface reflectance is sampled on a regular grid, each point measurement can be assumed to represent one pixel in a scene, and the variance associated with each measurement can be assumed constant, presupposing equal conditions. If the accuracy of the measuring device is known, approach 1 is selected; if not, approach 3 is used. When simulating a geophysical surface parameter by a climate model, the variance is often assumed to be spatially varying depending on the area and surface type, i.e. when modeling the surface albedo with a GCM, the variance is assumed to be larger over snow-covered areas than over open water. Depending on whether the uncertainty associated with the model is known or not, GCM output is assumed to meet the 2nd or 4th approach. The same argument holds for a remote sensing scene, where the product is more accurate at nadir than at extreme angles.

Returning to the hypothesis test for the scale-space smooth, pixels in the difference image are flagged as significantly different from zero when $\hat{f}_h(i,j)$ is higher than the noise level by testing

$$\begin{aligned} &H_0 : f_h(i,j) = 0 \text{ against} \\ &H_1 : f_h(i,j) \neq 0. \end{aligned} \tag{4}$$

The test statistic, $T(i,j)$, is calculated from

$$T(i,j) = \frac{\hat{f}_h(i,j)}{\sqrt{\mathrm{Var}[\hat{f}_h(i,j)]}}, \tag{5}$$

and the null hypothesis is rejected for those pixels where the absolute value of the test statistic is larger than the appropriate quantile q, i.e. when $|T(i,j)| > q(\alpha')$, where α' is the significance level for each test (defined later). An important component of the statistical inference is the number of sample points inside the kernel window, called effective sample size (ESS). ESS is used to highlight regions where the data are too sparse for doing inference and partly prevents the problem caused by large sample size (discussed in Section 5). By using the standard binomial rule of thumb, these are regions where $ESS < 5$. ESS is defined for each (i,j,h) as

$$\begin{aligned} &ESS(i,j,h) = \\ &\frac{\sum_{i'=1}^{n}\sum_{j'=1}^{m} K_h(i-i',j-j')}{K_h(0,0)}. \end{aligned} \tag{6}$$

Note that if K_h is a uniform kernel, $ESS(i,j,h)$ is the number of data points in the kernel window centered at (i,j). The test statistic in equation (5) can be assumed to be normally distributed if ESS is greater than 5 [Chaudhuri and Marron, 1999], leading to a quantile $q(\alpha')$ from the normal distribution. In Chaudhuri and Marron [1999] different candidates for calculating an appropriate quantile $q(\alpha')$ for multiple testing were investigated, and based on their results, the quantile is chosen to be approximately simultaneous over (i,j) Gaussian quantiles based on the number of independent averages,

$$\begin{aligned} q(\alpha') &= \Phi^{-1}\left(1-\frac{\alpha'}{2}\right) \\ &= \Phi^{-1}\left(1+\frac{(1-\alpha)^{1/l}}{2}\right), \end{aligned} \tag{7}$$

where l is the number of independent averages,

$$l(h) = \frac{nm}{\overline{ESS}(h)}, \tag{8}$$

and $\overline{ESS}(h)$ is the average of $ESS(i,j,h)$ over i and j. The overall significance level, α, is usually set at $\alpha = 0.05$.

2.3. The adapted S^3 standard output

To summarize, the adapted S^3 algorithm provides three images at each scale (for an example see Figures 2-3):

1. To give the user an idea of the resolution, the first panel shows the smoothing of the image, $\hat{f}_h$, performed with the 2D Gaussian kernel using scale parameter h.

2. The second panel shows the absolute value of the test statistic for the hypothesis test (cf. equation (5)). This is a measure of significance probability, where darker pixels correspond to a stronger significance. Another way of interpreting this is that darker pixels lead to significance even for more conservative significance levels. This measure of significance probability weakens the effect of a fixed significance level.

3. The decision plot is displayed in the third panel, with pixels significantly above (below) zero are marked in blue (red).

Three smoothing levels have been selected and presented here, but a broader range of smoothing levels can be displayed (perhaps as a movie). The significance test was also applied on the difference image with no smoothing, that is, at the scale of the pixel resolution.

In addition, an overall decision plot based on many scales simultaneously was included to sharpen the statistical inference. The overall decision plot shows the number of times a pixel is detected as significant for all the investigated scales without taking into account if the significance is positive or negative, indicating the creation and growing of features through scale-space towards merging and annihilation.

3. Model and Data

ECHAM5 GCM surface albedo simulations were intercompared and validated against a remote sensed surface albedo climatology taken to represent ground truth albedo.

3.1. ECHAM5 surface albedo

Thirty-years present-day simulations were performed with the ECHAM5 GCM from the Max Planck Institute for Meteorology [Roeckner et al., 2003]. The simulations were performed at 1.1° resolution, and used observed sea surface temperature and sea ice distribution for the period 1961-1990. The snow free surface albedo was estimated from three blended data sets [Claussen et al., 1994]. The snow albedo is a linear function of temperature between -5° and freezing, and is fixed above and below these limits. The total albedo is a weighted mean of the model-prescribed background albedo and the snow albedo according to the calculated snow cover fraction. Over sea-ice, the albedo follows a similar linear approach, whereas the sea-ice fraction is fixed based on monthly observations [Roeckner et al., 2003].

ECHAM5 captures the main features of the surface albedo pattern in March (Figure 1a), with high reflectivities in the snow-covered areas and other bright surfaces such as deserts (e.g. the Sahara and the Arabian desert). The surface albedo for snow free land and open water regions are prescribed, leading to negligible interannual variability for these areas (Figure 1b). The largest interannual variability is found for areas where the sea-ice cover fluctuates widely from year to year, e.g. the Sea of Okhotsk north of Japan and Baffin Bay. Previous work on ECHAM4 albedo shows that the ECHAM4 gives a high positive bias over snow-covered boreal forests and the Himalayas. In contrast, the model is probably too low compared with observations over extended parts of the Sahara and the adjacent steppes [Roesch et al., 2004]

3.2. Surface albedo climatology

The remote sensed surface albedo climatology, compiled from version 2.1 of the surface albedo algorithm developed at the University of Maryland for the period July 1983 to December 1998 [Pinker, 1985; Pinker and Laszlo, 1992] is taken to be ground truth surface albedo. This dataset (hereinafter called PINKER) is based on an algorithm that uses input data from the International Satellite Cloud Climatology Project (ISCCP) data D1 [Schiffer and Rossow, 1985] at 2.5° resolution provided by the Goddard Institute for Space Studies. The surface albedo was derived from satellite observations by constructing a look-up-table based on the delta-Eddington radiative transfer approximation under a wide range of atmospheric conditions, surface types and angle dependencies [Roesch et al., 2002]. The PINKER surface albedo has previously been compared against other surface albedo climatologies [Roesch et al., 2002], and coupled global climate models [Roesch, 2006], and therefore provides a good basis for validation. PINKER has been shown to generally underestimate the surface albedo in snow-covered regions [Roesch et al., 2002].

PINKER also captures the main features of the surface albedo pattern, with high reflectivities in the snow-covered areas and bright deserts (Figure 1c). The variance of PINKER (Figure 1d) shows greater spatial variability compared to ECHAM5, but has similar characteristics, with highest variance over land and snow-covered areas. The highest albedo variances in March are found in and around Kazakhstan, due to this being a sparsely forested area which often contains the snow line in March.

4. Validation of model simulated surface albedos by the adapted S^3 for March

Both the simulated surface albedo and satellite estimates provide the upward and downward surface radiation fluxes, and the surface albedos were derived from the ratio of these. The ECHAM5 time-slice simulation runs from 1961-1990, while PINKER estimates are available from 1983-1998, meaning that the two time series only partly overlap. A different reference period may thus lead to certain differences mainly in snow-covered regions, as snow cover extent exhibits considerable interannual and decadal variations. The dataset was re-sampled on the same T42 grid, providing a pixel resolution of 280 km. Surface albedo is spatially very heterogeneous; however, GCMs only have grid average surface albedos [Roesch et al., 2004]. The intracell variability for the T42 grid, that is, the spatial heterogeneity within a grid cell, was determined from MODIS albedo [Schaaf et al., 2002] and found to be largest in snow-covered areas, particularly mountainous regions, and forested areas [Roesch et al., 2004].

As indicated previously, the variance is taken to be apriori unknown and spatially varying (approach 4). The interannual variabilities of ECHAM5 and PINKER albedo are taken to be an estimate of the noise variance. ECHAM5 (model albedo) and PINKER (satellite albedo) are assumed to be independent and the variance of the albedo difference (ECHAM5 minus PINKER) is simply the sum of the individual variances (calculated as the sample variance derived from the standard estimator). All months were studied; however, this paper focuses on March as it is characterized by both an extended snow cover and sufficient global radiation allowing accurate albedo estimates over Northern Eurasia . The northernmost latitudes in the images, where there are Polar Nights in March (i.e. missing data), were removed from the image maps.

The difference map (ECHAM5-PINKER) reveals large differences, primarily over regions covered with snow and ice (Figure 1e). The albedo differences over snow-covered land and ice-covered areas frequently exceeds 0.2, while corresponding deviations over ice-free sea are rarely higher than 0.02. Positive deviations are generally found over snow and sea-ice covered areas, deserts such as the Sahara and the Arabian desert, and also in dry regions of South America and Australia. In contrast, negative deviations are found over tropical rain forests such as the Amazonian forest and African rain forest as well as tundra areas in West Siberian Plain. The total interannual variability of the surface albedo difference (Figure 1f) reveals that the variance is highest in areas adjacent to the snow and ice line, as also seen in the individual variances. Low variances are generally found over snow-free areas, ice-free oceans, and regions covered by thick snow pack or a closed ice cover.

4.1. Practical significance

The problem with statistical significance is that it may not be particularly meaningful for a given application because of the natural variability of the phenomena or the arbitrariness of a defined statistical threshold (as discussed in

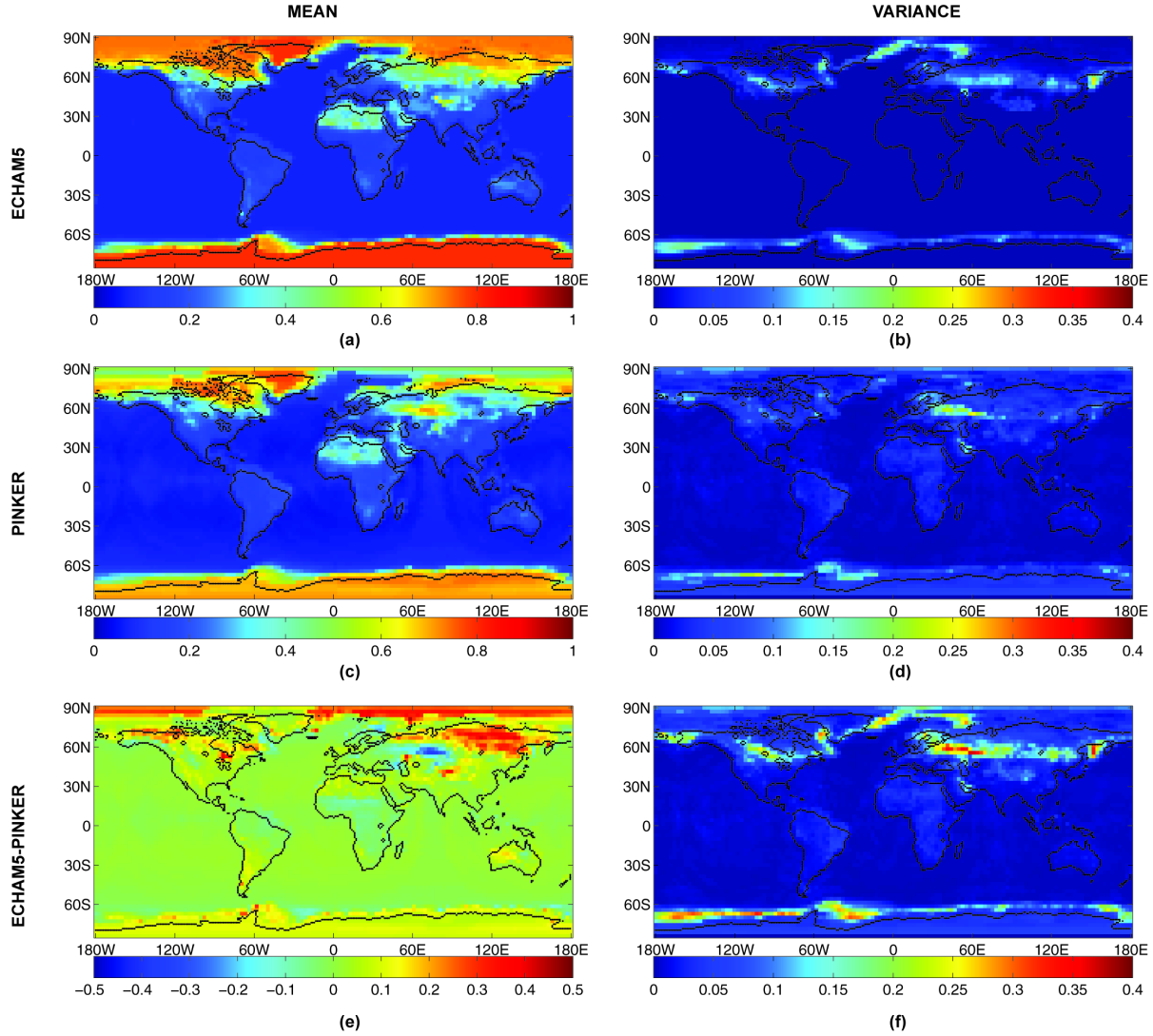

Figure 1. Mean simulated surface albedo (a) and interannual variability (variance, b) for ECHAM5 in March, and mean PINKER surface albedo climatology (c) and interannual variability (variance, d) in March. The albedo difference (ECHAM5-PINKER) in (e) and the total interannual variability of the difference in (f). Note the distinction in the scales.

the introduction). Experienced climatologists, on the other hand, can decipher whether the difference between, e.g., two albedo images is climatologically meaningful (herein called 'practical significance'), but this process is subjective and greatly depends on the skill of the observer. Also, there is no exact definition of practical significance. Often the normalized albedo difference (difference divided by the standard deviation) is used as a measure of practical significance, and values above one are considered to be practically significant. Calculating statistical significance across multiple scales helps to overcome the problems of both statistical and practical significance and provides a more objective procedure to determine the practical significance of areas of change across an image. In the next sections, areas where large discrepancies are identified by means of the adapted S^3 will be highlighted as statistically significant, and compared to the practical significance.

4.2. Statistical significance - assessing the range of scales

Both for defining the scale and choosing the range of scales, we used an approach similar to that suggested by Chaudhuri and Marron [1999]. The scale is defined as four times the bandwidth parameter h. This definition is based on the fact that approximately 95% of the probability mass of the 2D Gaussian kernel is inside the circle of diameter four times the standard deviation (here h corresponds to the standard deviation). The widest possible range of the scale has the lower boundary limited by the binned implementation and the upper boundary is the range of the data. However, a smaller range of scales is investigated here.

The adapted S^3 provides information on the statistical significance of the difference between the simulated and

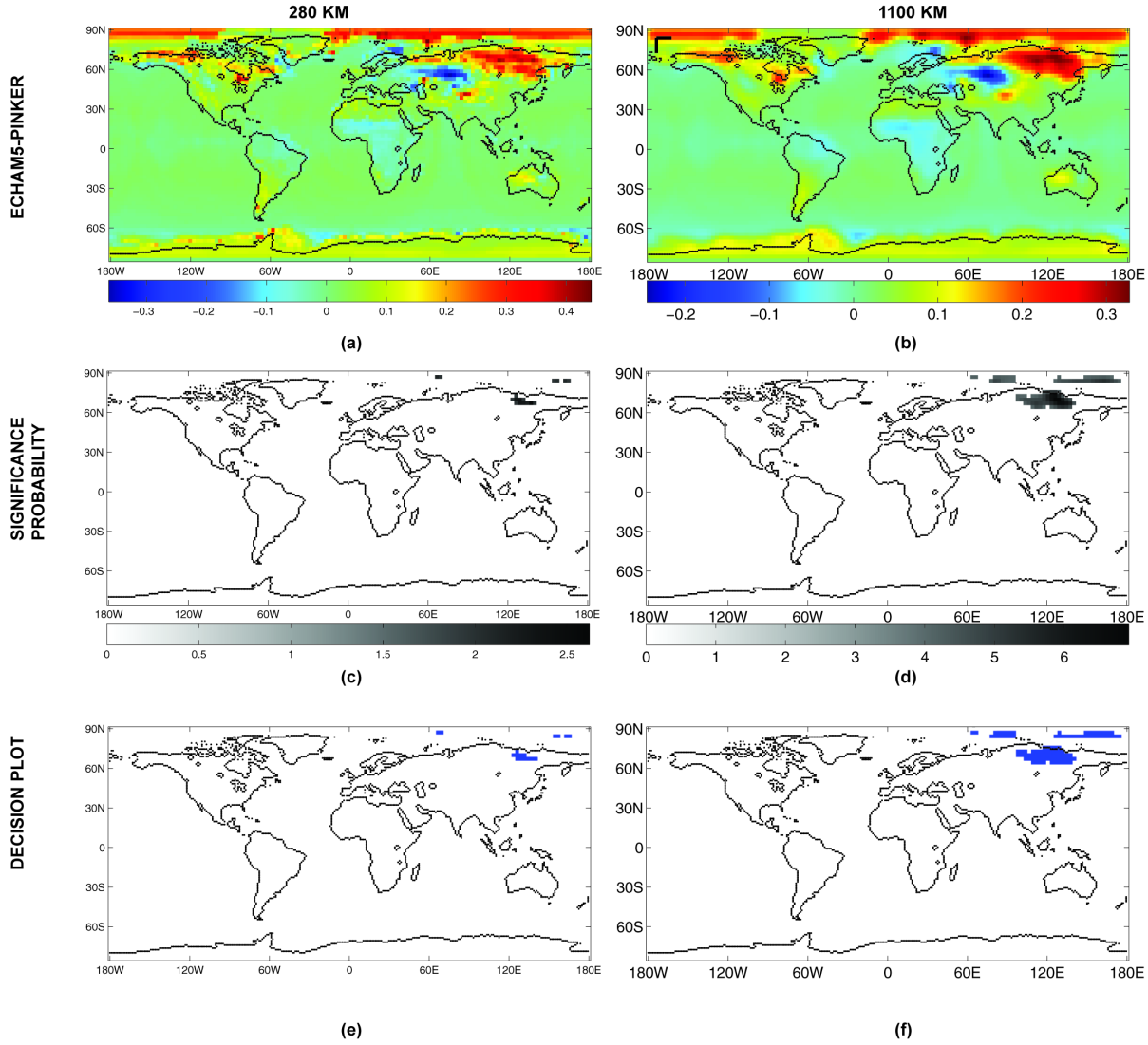

Figure 2. The difference image (ECHAM5-PINKER), $\hat{f}_h$, at 280 km resolution (no smoothing) in (a) and smoothed at 1100 km resolution in (b). The scales are marked in black in the upper left corner of (b). The measure of significance probability for 280 km resolution in (c) and 1100 km resolution in (d). The decision plot at 280 km in (e) and at 1100 km in (f), identifying the pixels in the smoothed difference image significantly different from zero at a 5% significance level, with blue for significant positive pixel, and red for significant negative pixel (only significant positive pixels are detected here). The variance is spatially varying and estimated from the data.

observed surface albedo for four scales from 280 km (no smoothing) to 4000 km (Figures 2 and 3). The finest scale equals the pixel resolution of 280 km and produces very few pixels (0.21%) where the albedo in the two images is significantly different (Figure 2e, significantly different pixels will hereafter be abbreviated SDP). This may be somewhat surprising to a climatologist as the normalized albedo difference (not shown here) exceeds unity in extended regions over both land and sea such as major parts of snow-covered Northern Eurasia and North America as well as the Southern Ocean. As the interannual variability of the above mentioned regions differs considerably, the albedo difference over these areas is not statistically significant. Only for a few pixels in north-eastern Siberia and some areas in the Arctic Ocean are the differences significant. A few more SDP (1.2%) are found at 1100 km resolution (Figure 2f). The critical areas include the northern and eastern boreal forests and some regions close to the ice edge. The positive differences over snow-covered boreal forests are primarily attributed to a poor representation of the sky view factor in ECHAM5, which plays a major role in the parameterization of the forest albedo under snow-covered conditions [Roesch and Roeckner, 2006]. Significant differences are also found in the southern part of the Arctic Ocean adjacent to the ice edge, which is attributed to the different positions of the ice edge in ECHAM5 and PINKER.

The statistical SDP at 2500 km resolution (Figure 3e) coincide well with the practical significance from a climatologist's view (8.2% SDP). Most of the snow-covered regions in northern Eurasia and northern Canada/Alaska with positive differences above 0.2, and relatively low interannual variability show SDP. The regions further to the south and close to the spring snow line tend to show differences of the same magnitude, but as the interannual variation is more pronounced in these regions with thin snow cover, the albedo

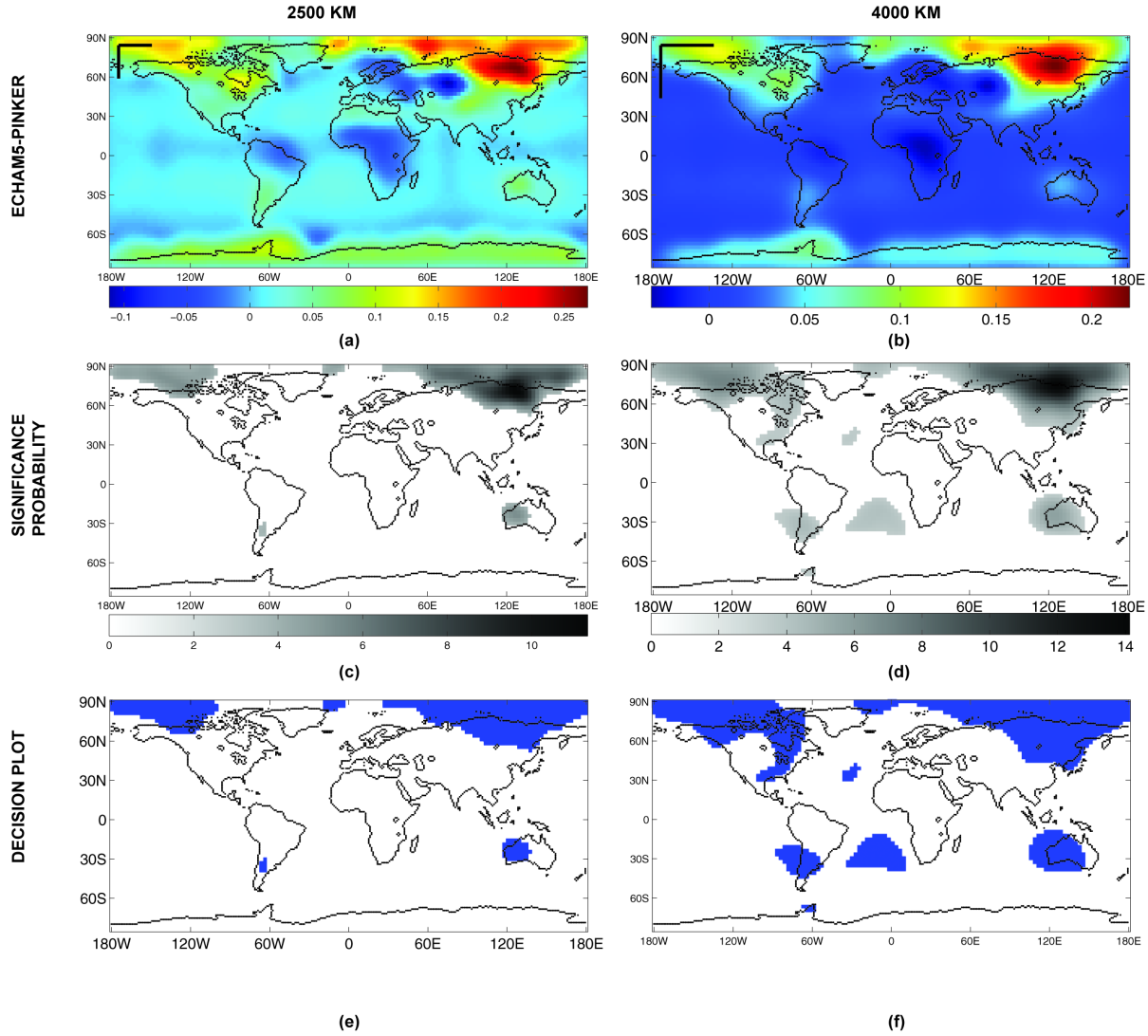

Figure 3. Same as for Figure 2, only for the scales 2500 km in (a,c,e) and 4000 km in (b,d,f).

difference is not statistically significant. The quite low interannual variability in western Australia gives SDP despite the small albedo difference. South America shows differences of the same magnitude, but has slightly higher interannual variation, and therefore very few pixels are significantly different. SDP are also found in the snow and ice covered Arctic Ocean and Beaufort Sea. The rest of the sea-ice covered Arctic does not show SDP, probably because of changing ice edge over the years and high interannual variability. Also these northernmost areas are largely cloud-covered with low global radiation during spring. At 4000 km resolution (Figure 3f) the snow-covered regions of SDP continue to grow and merge further south (17.6% SDP). More areas of SDP are detected in both the North and South Atlantic Ocean. All these plots prove that ECHAM5 overall overestimates the albedo, and this is confirmed when investigating larger scales (indicated in Figure 4). No areas of negative SDP are found at any scale, meaning that ECHAM5 albedo is never significantly lower than PINKER albedo. However, it should be borne in mind that previous studies [Roesch et al., 2002] showed that PINKER albedo was underestimated in snow-covered regions.

The overall decision plot in Figure 4 is a composite of individual decision plots at several scales ranging from 1120 km to 6950 km with steps of 224 km (and also including the un-smoothed scale of 280 km). The plot indicates which regions contain SDP at different scales. For increasing scales, the SDP areas merge and grow until one reaches scales comparable to the Earth size (not shown here) where all pixels are SDP. Inference at scales equal to the range of the data should be avoided since the statistical significance at these very coarse scales is not of any practical significance.

4.3. Statistical significance - assessing the choice of variance

The use of the interannual albedo variability as a measure of the noise variance can be debated, as a high variance in a window surrounding a pixel in one area of an image might arise through natural and true variability, whereas a part of the image with lower variance can in fact be noisier. However, for spatially varying variance, the variance in the window of surrounding pixels is a reasonable estimate of the noise variance if the difference image can be assumed to be reasonably smooth. If the image contains abrupt changes, the variance will fail, and a more robust estimate e.g. of the nearest neighbor differences is required. This idea was further exploited in Godtliebsen and Øigård (2005), and with the additional knowledge about the noise distribution, it

works very well for a broad range of signals. The method can be used to obtain variance estimates both locally and globally (and is included in the adapted S^3 package on the web). For this albedo validation, we believe the interannual albedo variability is the best estimate of the noise variance, and the estimate that will provide the most realistic statistically significant areas (Figure 2 and 3).

As the variance estimate is often associated with errors, it is crucial that the adapted S^3 is somewhat stable against small variance biases (say plus/minus 10%). The adapted S^3 was tested for stability by applying it to the difference map (ECHAM5-PINKER) and changing the variance in turn to be 10% higher and 10% lower than the original spatially varying variance used for Figure 2-3. It is obvious that areas of SDP will be reduced by larger variance, and increased by smaller variance, as a larger variance allows more variability before the difference becomes significant (cf. equation (5)). A variance 10% higher than the original variance gave 11-62% fewer pixels detected for the four scales, with largest percent deviations for the finest scale. However, the absolute difference was only 0.21% for the same scale. A variance 10% lower than the original variance gave 12-24% more pixels detected for the four scales, again with largest deviations for smallest scales. The same areas of SDP were found after increasing/decreasing the variance: the only change was the size of the areas of SDP. The only two exceptions were an area outside the east coast of Australia which was detected as SDP after a 10% variance reduction at 4000 km, and areas in the Arctic Ocean which were not detected as SDP after a 10% variance increase at 280 km.

4.4. Statistical significance - assessing the level of significance

The significance level α is important for the outcome of the statistical test, and a social norm of 0.05 is common in most scientific fields [Germano, 1999]. The adapted S^3 reduces the problems inherent in using a fixed α, as the measure of significance probability gives a clear indication about the strength of the significance. The measure of significance probability gives the absolute value of the test statistics for the SDP pixels. The value of the quantile decreases for successively larger scales because of more smoothing, as is evident from equation (7). The quantile is calculated based on the significance level α, but α cannot be interpreted directly from the measure of significance probabilities. It is, however, clear that the SDP in the darkest areas of the measure of significance probability, will be SDP also for a more conservative significance level. Another interesting observation is that areas detected as significant at fine scales, have a stronger significance at coarser scales (cf. Figure 2-3 c-d).

Figures 2-3 refer to SDP at 5% significance level, and a decrease or increase of α will affect the number of SDP. The adapted S^3 was applied on the difference image (ECHAM5-PINKER) with a conservative 1% significance level and was found to detect 22-100% fewer SDP compared to the 5% level, again with largest deviations for the smallest scale and again with small absolute decreases. The 100% decrease was for 280 km resolution, where no SDP were detected. A less conservative 10% significance level, resulted in 14-252% more SDP being detected. Again, the greatest increase was for the finest scale, and again the absolute difference was small. Despite the variable number of SDP, the overall pattern of SDP was unchanged. These findings are in agreement with the original S^3 [Godtliebsen et al., 2004], where S^3 had little sensitivity to the significance level.

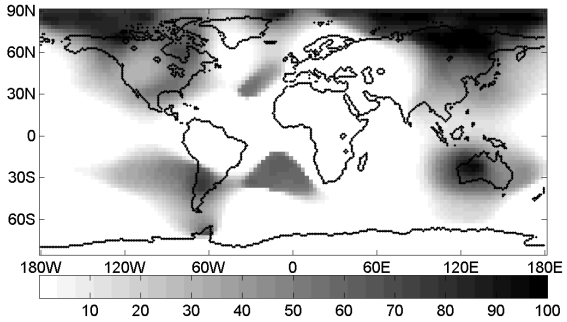

Figure 4. The overall decision plot as a composite of individual decision plots for scales ranging from 1120 to 6950 km with steps of 224 km (also including the no smoothing at 280 km), altogether 28 scales. The colorbar indicates the number of times (in percent) a pixel is significant for the chosen range of scales.

5. Discussion

The use of scale-space clearly adds great new information to statistical testing in climatology, as a simple t-test applied to the original data (without any smoothing), will detect very few SDP as shown in Figure 2e. These few statistical SDP do not coincide with practical significance; however, when the data are smoothed and inference made at coarser scales the statistical significance agrees well with the practical significance from a climatologist's point of view.

Areas of large discrepancies over boreal forests and deserts will strongly influence the temperature pattern and possibly the Monsoon regime, and are indeed of practical significance. Also, areas close to the ice edge and snow line are of practical significance, as the discrepancies are probably due to cloud screening problems and interannual variations in ice coverage.

As stated previously the information obtained is a relationship between the actual size of the feature and the resolution of the filter used to extract information. From a statistical point of view it is not reasonable to compute averages over a single or very few GCM grid boxes, as climate models are inappropriate for estimating small-scale regional averages. It is appropriate, however, to provide averages over a larger number of grid cells. What contributes as a reasonable number of cells depends, among other things, on the parameter in question and the topography. In view of the significant albedo variations both in time and space, it might be reasonable to compare observed and simulated albedos only at scales of 1000 km and above. A scale of 2500 km proved to be appropriate for albedo validation, as the statistical significance seen at this scale coincided well with the practical significance. However, we want to underline that the strength of the adapted S^3 methodology is that all scales are investigated: looking only at the scale that supports the preconceived notions and putting most emphasis on that one should be avoided.

The difficulties connected with hypothesis testing and inference based on hypothesis testing should not be ignored. One of the main objections to hypothesis testing is that the zero hypothesis is false in nature [Germano, 1999; Anderson et al., 2000]. E.g. two datasets will never be quite identical, hence tests similar to equation (4) are false in nature. Large samples enhance this weakness, because a small difference will become statistically significant given a large enough sample size [Katz, 1992]. S^3 partly avoids the problems of large sample size by defining the effective sample size (cf. equation (6)). For this paper the effective sample size was below 80 for all individual tests for the albedo validation, meaning that reasonable inference can be obtained for reasonable scales [von Storch and Zwiers, 1999]. A common mistake is to take the p-value (the probability of getting the observed value or something more extreme) as the probability of the research hypothesis H_1 being true. However, it should be kept in mind that the only conclusion to be made from hypothesis testing is to reject or retain H_0. Indeed the significance level α is the probability of rejecting a true H_0. The adapted S^3 reduces the effect of a fixed significance level as the measure of significance gives a clear indication about the strength of a significant feature. In addition, the adapted S^3 proved to be reasonably stable for variations in significance level. For a more thorough and general description of advantages and disadvantages of hypothesis testing, see Germano [1999] and Anderson et al. [2000].

6. Conclusions

Numerous applications would benefit from quantifying the difference image (difference between model and validation data or difference between two model runs) by using statistical methods for making statistical inference. Such applications include comparisons of model parameterizations for GCMs with control runs [Marshall and Oglesby, 1994; Douville et al., 1995] and validation of model results [Roesch et al., 2002; Wei et al., 2001; Liston, 2004]. This paper presents an adapted version of Significance in Scale-Space (S^3) for detecting significant features (defined as areas where pixels in the difference image are sufficiently different from zero) at different scales specifically adapted for validation purposes in climatology. The scale issue is introduced for making inference at different levels of resolution. As the scale gradually goes from fine to coarse, features are created, they grow and merge and may finally annihilate. The adapted S^3 algorithm produces maps for statistical inference as well as showing the strength of the significance.

The adapted S^3 was applied to the difference map of ECHAM5 simulated surface albedo and PINKER surface albedo climatology to validate and intercompare the two datasets from a statistical point of view. Overall, snow and ice covered areas had the largest discrepancies. Only positive statistically significant areas of pixels were detected, which means that the ECHAM5 model overestimates the albedo compared to the PINKER climatology for all scales - at least in March. At the finest scales (280 km) very few areas of significant albedo differences were detected because of relatively high interannual variability for the areas of largest difference, such as major parts of snow-covered Northern Eurasia and North America as well as the Southern Ocean. At 1100 km, significant albedo differences were found in the southern part of the Arctic Ocean adjacent to the ice edge, probably because of the different positions of the ice edge in ECHAM5 and PINKER. At 2500 km the statistical significance coincided well with the practical significance, as most of the snow-covered regions in Northern Eurasia with positive differences above 0.2, and relative low interannual variability were marked as significant. It is not possible to put a threshold on the albedo bias to judge if a certain bias is large enough to substantially change the large-scale circulation. The threshold would depend on both the spatial extent of the bias and the magnitude of the bias, as well as the location. Previous studies show that surface albedo biases should be within +/-0.02 to +/-0.05 in order to allow reliable climate simulations [Henderson-Sellers and Wilson, 1983]; however, these limits are widely exceeded in this validation.

The adapted S^3 proved to be both a powerful and an easy approach for detecting significantly different pixels taking into account the observed interannual variability. It will be a helpful tool for climatologists attempting to make objective inference from large amounts of climate data that need to be validated and intercompared (cf. IPCC).

Acknowledgments. We would like to thank J.-G. Winther, D. K. Hall, E. Roeckner, B. Ivanov and M. Køltzow for comments during the initial phases of this work. J. Holmén and J. Kohler is acknowledged for their review of the manuscript. We would also like to thank the anonymous reviewers for many valuable comments. The work is supported by the Research Council of Norway, the Norwegian Polar Institute and the University of Tromsø.

References

Anderson, D. R., Burnham, K. P. and Thompson, W. L. (2000). Null Hypothesis Testing: Problems, Prevalence, and an Alternative. *Journal of Wildlife Management*, 64(4):912–923.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association, Theory and Methods*, 94(447):807–823.

Chervin, R. M. and Schneider, S. H. (1976). On Determining the Statistical Significance of Climate Experiments with General Circulation Models. *Journal of Atmospheric Science*, 33(3):405–412.

Claussen, M., Lohmann, U., Roeckner, E. and Schulzweida, U. (1994). A Global Data Set of Land-Surface Parameters. Technical Report 135, Max Planck Institute for Meteorology.

Curry, J., Schramm, J. L. and Ebert, E. E. (1995). Sea Ice-Albedo Climate Feedback Mechanism. *Journal of Climate*, 8:240–247.

Douville, H., Royer, J.-F. and Mahfouf, J.-F. (1995). A New Snow Parameterization for the Météo France Climate Model. *Climate Dynamics*, 12:21–35.

Ganguli, B. and Wand, M. P. (2004). Feature Significance in Geostatistics. *Journal of Computational and Graphical Statistics*, 13(4):954–973.

Germano, J. D. (1999). Ecology, Statistics, and the Art of Misdiagnosis: The Need for a Paradigm Shift. *Environmental Review*, 7:167–190.

Godtliebsen, F., Marron, J. S., and Chaudhuri, P. (2002). Significance in Scale-Space for Bivariate Density Estimation. *Journal of Computational and Graphical Statistics*, 10:1–21.

Godtliebsen, F., Marron, J. S., and Chaudhuri, P. (2004). Statistical Significance of Features in Digital Images. *Image and Vision Computing*, 22:1093–1104.

Godtliebsen, F., Olsen, L. R., and Winther, J.-G. (2003). Recent Developments in Statistical Time Series Analysis: Examples of Use in Climate Research. *Geophysical Research Letters*, 30(12):1654–1657.

Godtliebsen, F. and Øigård, T.-A. (2005). A Visual Display Device for Significant Features in Complicated Signals. *Computational Statistics and Data Analysis*, 48:317–343.

Haralick, R. M. and Shapiro, L. G. (1992). Computer and Robot Vision. *Reading MA: Addison-Wesley.*

Hay, G. J., Dubé, P., Bouchard, A. and Marceau, D. J. (2002). A Scale-Space Primer for Exploring and Quantifying Complex Landscapes. *Ecological Modeling*, 153:27–49.

Henderson-Sellers, A. and Wilson, M. F. (1983). Surface Albedo Data for Climate Modeling. *Reviews of Geophysics*, 21(8):1743–1778.

Itti, L., Koch, C. and Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Karløf, L., Øigård, T.-A., Godtliebsen, F., Kaczmarska, M., and Fischer, H. (2005). Statistical Techniques to Select Detection Thresholds for Peak Signals in Ice-Core Data. *Journal of Glaciology*, 51(175):655–622.

Katz, R. W. (1992). Role of Statistics in the Validation of General Circulation Models. *Climate Research*, 2:35–45.

Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology. *Ecology*, 73:1943–1967.

Lindeberg, T. (1994). Scale-Space Theory in Computer Vision. *Kluwer Academic Publisher, Dordrecht, The Netherlands*.

Liston, G. E. (2004). Representing Subgrid Snow Cover Heterogeneities in Regional and Global Models. *Journal of Climate*, 17(6):1381–1397.

Marceau, D. J. (1999). The Scale Issue in the Social and Natural Sciences. *Canadian Journal of Remote Sensing*, 25:347–356.

Marceau, D. J. and Hay, G. J. (1999). Contributions of Remote Sensing to the Scale Issues. *Canadian Journal of Remote Sensing*, 25:357–366.

Marshall, S. and Oglesby, R. J. (1994). An Improved Snow Hydrology for GCMs. Part 1: Snow Cover Fraction, Albedo, Grain Size and Age. *Climate Dynamics*, 10:21–37.

Pal, N. R. and Pal, S. K. (1993). A Review on Image Segmentation Techniques. *Pattern Recognition*, 26(9):1277–1294.

Park, C., Marron, J. S. and Rondonotti, V. (2004). Dependent Sizer: Goodness-of-Fit Tests for Time Series Models. *Journal of Applied Statistics*, 31(8):999–1017.

Pauwels, E. J. and Frederix, G. (1999). Finding Salient Regions in Images - Nonparametric Clustering for Image Segmentation and Grouping. *Computer Vision and Image Understanding*, 75:73–85.

Pedersen, C. A. and Winther, J.-G. (2005). Intercomparison and Validation of Snow Albedo Parameterisation Schemes in Climate Models. *Climate Dynamics*, 25:351–362.

Pinker, R. (1985). Determination of Surface Albedos from Satellites. *Advanced Space Research*, 5:333–343.

Pinker, R. and Laszlo, I. (1992). Modeling of Surface Solar Irradiance for Satellite Application on a Global Scale. *Journal of Applied Meteorology*, 31:194–211.

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kirchner, I., Kornblueh, L., Manzini, E., Rhodin, A., Schlese, U., Schulzweida, U., and Tompkins, A. (2003). The Atmospheric General Circulation Model ECHAM5 - Part 1. Technical Report 349, Max Planck Institute for Meteorology.

Roesch, A. (2006) Evaluation of Surface Albedo and Snow Cover in AR4 Coupled Climate Models. *Journal of Geophysical Research*, 111(D15).

Roesch, A. and Roeckner, E. (2006). Assessment of Snow Cover and Surface Albedo in the ECHAM5 General Circulation Model. *Journal of Climate*, 19(16):3828–3843.

Roesch, A., Schaaf, C. and Gao, F. (2004). Use of Moderate-Resolution Imaging Spectroradiometer Bidirectional Reflectance Distributing Function Products to Enhance Simulated Surface Albedos. *Journal of Geophysical Research*, 109(D12105).

Roesch, A., Wild, M., Pinker, R., and Ohmura, A. (2002). Comparison of Spectral Surface Albedos and Their Impact on the General Circulation Model Simulated Surface Climate. *Journal of Geophysical Research*, 107(D14).

Sandve, I. L. (2006). Comparison of Noise Variance Estimators. Statistical Report from University of Tromsø, Norway.

Schaaf, C. B., Gao, F., Strahler, A. H., Lucht, W., Li, X., Tsang, T., Strugnell, N. C., Zhang, X., Jin, Y., Muller, J.-P., Lewis, P., Barnsley, M., Hobson, P., Disney, M., Roberts, G., Dunderdale, M., Doll, C., d Éntremont, R. P., Hu, B., Liang, S., Privette, J. L. and Roy, D. (2002). First Operational BRDF, Albedo Nadir Reflectance Products from MODIS. *Remote Sensing of the Environment*, 83:135–148.

Schiffer, R. and Rossow, W. (1985). ISCCP Global Radiance Data Set: A New Resource for Climate Research. *Bulletin of the American Meteorological Society*, 66:1498–1505.

von Storch, H. and Zwiers, F. W. (1999). Statistical Analysis in Climate Research. Cambridge University Press.

Wealands, S. R., Grayson, R. B. and Walker, J. P. (2005). Quantitative Comparison of Spatial Fields for Hydrological Model Assessment - Some Promising Approaches. *Advances in Water Resources*, 28:15–32.

Wei, X., Hahmann, A. N., Dickinson, R. E., Yang, Z. L., Zeng, X. B., Schaudt, K. J., Schaaf, C. B. and Strugnell, N. (2001). Comparison of Albedos Computed by Land Surface Models and Evaluation Against Remotely Sensed Data. *Journal of Geophysical Research*, 106(D18):20687–20701, September.

Christina A. Pedersen, Norwegian Polar Institute and University of Tromsø, Tromsø, Norway. (xtina@npolar.no)

Fred Godtliebsen, University of Tromsø, Tromsø, Norway. (fred@math.uit.no)

Andreas C. Roesch, Swiss Federal Institute of Technology, Zurich, Switzerland. (andreas.roesch@env.ethz.ch)