AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2016JD025199

Key Points:
- SOM analysis of ISCCP data produces a subset of the regimes derived in previous studies using *k*-means clustering with finer detail in areas
- ISCCP flux data and ERA-Interim vertical velocity and lower tropospheric stability demonstrate that the clusters are physically meaningful
- The objective organization of the SOM allows a quantitative analysis of cloud regimes daily transition frequency and persistence

Correspondence to:
A. J. McDonald,
adrian.mcdonald@canterbury.ac.nz

Citation:
McDonald, A. J., J. J. Cassano, B. Jolly, S. Parsons, and A. Schuddeboom (2016), An automated satellite cloud classification scheme using self-organizing maps: Alternative ISCCP weather states, *J. Geophys. Res. Atmos.*, *121*, 13,009–13,030, doi:10.1002/2016JD025199.

Received 17 APR 2016
Accepted 20 OCT 2016
Accepted article online 25 OCT 2016
Published online 14 NOV 2016

An automated satellite cloud classification scheme using self-organizing maps: Alternative ISCCP weather states

Adrian J. McDonald[1], John J. Cassano[2,3], Ben Jolly[1,4], Simon Parsons[1], and Alex Schuddeboom[1]

[1]Department of Physics and Astronomy, University of Canterbury, Christchurch, New Zealand, [2]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado, USA, [3]Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, Colorado, USA, [4]Landcare Research, Palmerston North, New Zealand

Abstract This study explores the application of the self-organizing map (SOM) methodology to cloud classification. In particular, the SOM is applied to the joint frequency distribution of the cloud top pressure and optical depth from the International Satellite Cloud Climatology Project (ISCCP) D1 data set. We demonstrate that this scheme produces clusters which have geographical and seasonal patterns similar to those produced in previous studies using the *k*-means clustering technique but potentially provides complementary information. For example, this study identifies a wider range of clusters representative of low cloud cover states with distinct geographic patterns. We also demonstrate that two rather similar clusters, which might be considered the same cloud regime in other classifications, are distinct based on the seasonal variation of their geographic distributions and their cloud radiative effect in the shortwave. Examination of the transitions between regimes at particular geographic positions between one day and the next also shows that the SOM produces an objective organization of the various cloud regimes that can aid in their interpretation. This is also supported by examination of the SOM's Sammon map and correlations between neighboring nodes geographic distributions. Ancillary ERA-Interim reanalysis output also allows us to demonstrate that the clusters, identified based on the joint histograms, are related to an ordered continuum of vertical velocity profiles and two-dimensional vertical velocity versus lower tropospheric stability histograms which have a clear structure within the SOM. The different nodes can also be separated by their longwave and shortwave cloud radiative effect at the top of the atmosphere.

1. Introduction

A wide range of possibilities exist for evaluating the quality of the cloud representation in climate models and understanding the properties of clouds from satellite observations. Many previous studies have assessed the representation of clouds and their radiative effects in general circulation models (GCMs) with most evaluations being separated into tests of the model climate or case studies. *Jakob* [2003] identified that the evaluation of cloud parameterizations with either highly averaged climatological information or information from individual case studies has serious drawbacks. In particular, the climatological averaging process can hide issues because of compensating errors and the identification of representative case studies can be difficult. For example, the global cloud radiative forcing may perform well in a model because of compensating errors in the temporal frequency and radiative properties of different cloud types [*Webb et al.*, 2001; *Williams and Webb*, 2009]. Thus, climatological analysis can identify first-order problems but may miss subtler errors.

Mixtures of these methodologies and additional analysis approaches have therefore been discussed. In particular, *Jakob* [2003] advocates the use of "intelligent" ways of averaging data, so that the general characteristics of certain cloud systems remain intact even when a large number of cases are averaged. The three most common methods used in this context are the derivation of weather states [*Jakob and Tselioudis*, 2003], compositing cloud properties based on specific ranges of dyanamical parameters [*Tselioudis et al.*, 2000; *Bony et al.*, 2004] and cyclone compositing [*Lau and Crane*, 1995; *Field and Wood*, 2007]. The weather states (WS) or regimes methodology relies on clustering satellite cloud data, most commonly using the *k*-means clustering technique and then classifying the model output in a similar way. The relationship between the WS radiative properties, geographic and seasonal occurrence derived from observations, and the model output

are then compared to identify sources of errors. The second method attempts to explicitly link clouds and the large-scale atmospheric circulation. *Bony et al.* [2004] link cloud types and cloud radiative forcing in the tropics to the large-scale vertical motion of the atmosphere, identified by the vertical velocity at 500 hPa (ω_{500} expressed in hPa/d). The large-scale tropical circulation is then categorized into a series of dynamical regimes corresponding to different values of ω_{500}. A number of studies have followed this approach, such as *Su et al.* [2008] and *Medeiros and Stevens* [2011]. Classification of cloud properties using pressure anomalies has also been used. For example, *Tselioudis et al.* [2000] examined the quantitative relationship between midlatitude atmospheric dynamics and the properties of the midlatitude clouds in the Northern Hemisphere. The final methodology uses dynamical information to composite data relative to particular structures, cyclone compositing which places satellite data and model output into a coordinate system defined relative to low-pressure centers being particularly popular.

In this study, joint histograms of cloud top pressure versus optical depth (joint histograms from this point forward) derived by the International Satellite Cloud Climatology Project (ISCCP) [*Rossow and Schiffer*, 1991, 1999] are utilized in a self-organizing map (SOM) clustering scheme to test the utility of this clustering technique. The joint histograms provide a simple statistical representation of a satellite scene and have been widely used to evaluate GCMs [*Williams and Webb*, 2009; *Mason et al.*, 2015] and to understand the makeup of clouds globally and regionally [*Tselioudis et al.*, 2000; *Jakob and Tselioudis*, 2003; *Marchand et al.*, 2010; *Tselioudis et al.*, 2013]. An important point to note in later analysis is that the ISCCP output has been suggested to underestimate low-level cloud. This issue likely occurs because the retrieval infers cloud top pressure from cloud top temperature, and an atmospheric profile and stratocumulus clouds often exist under temperature inversions which are poorly represented in the atmospheric profiles [*Marchand et al.*, 2010]. However, all satellite instruments and their retrievals have their strengths and weaknesses [*Stubenrauch et al.*, 2013] and the intent of this work is to test the usefulness of the SOM scheme for future model evaluation efforts.

The usage of cluster analysis to form cloud regimes or weather states to determine the frequency and geographic position of different cloud types has become widespread with the *k*-means clustering methodology being applied to ISCCP [*Jakob and Tselioudis*, 2003; *Rossow et al.*, 2005; *Williams and Webb*, 2009; *Marchand et al.*, 2010; *Oreopoulos and Rossow*, 2011; *Bodas-Salcedo et al.*, 2012; *Zelinka et al.*, 2012; *Klein et al.*, 2013; *Tselioudis et al.*, 2013; *Mason et al.*, 2015; *Rossow et al.*, 2016], Multiangle Imaging Spectroradiometer (MISR) [*Marchand et al.*, 2010], Moderate Resolution Imaging Spectroradiometer (MODIS) [*Williams and Webb*, 2009; *Marchand et al.*, 2010; *Oreopoulos et al.*, 2014; *Bankert and Solbrig*, 2015; *Oreopoulos et al.*, 2016], and CloudSat data [*Zhang et al.*, 2007; *Sassen and Wang*, 2008]. We will compare the results of the application of the SOM analysis in this paper to these previous results with particular reference to geographic structure, seasonal variation, and meteorological context.

Recent work by *Muhlbauer et al.* [2014] has applied an artificial neural network classification scheme to MODIS cloud observations to characterize the properties of marine low-level clouds on a global scale. This expands on previous work in *Wood and Hartmann* [2006] which examined specific regions in the northeast and southeast Pacific. However, *Muhlbauer et al.* [2014] is the only paper to the authors' knowledge that has previously used a neural network classification of cloud data on a global scale. That work classifies a specific cloud type (stratocumulus) based on a more complex range of parameters than those used in the present study. In particular, the neural network uses 32 metrics linked to the power spectra and 40 values linked to the probability distribution functions of the liquid water path scenes to distinguish different types of stratocumulus. Preprocessing based on cloud top temperature over scenes identified as wholly oceanic is required [*Wood and Hartmann*, 2006] and the three-layer backpropagation scheme requires user identification in the training scheme. Our methodology, therefore, has some similarities with this previous work, but we use only the ISCCP joint histogram and apply the SOM algorithm globally. We leave a discussion of the results from *Muhlbauer et al.* [2014] till later in this study to aid comparison between their work and results from the present study.

2. Data Set and Methodology

In this study the self-organizing map methodology is applied to the ISCCP D1 data set. Details of the ISCCP data set and its derivation are discussed in *Rossow and Schiffer* [1991, 1999]. The ISCCP D1 data set occurs at 3-hourly temporal resolution on a 280 km by 280 km equal area grid (6596 of these grid points cover the Earth) over the period July 1983 to December 2009. For each grid box, the number of cloudy pixels (each pixel is approximately 5 km by 5 km) that belong to one of seven pressure levels and six optical thickness categories

is identified. Thus, all cloudy pixels in a grid box are placed in one of 42 bins forming a joint histogram of the cloud optical depth-cloud-top pressure. The summation of the cloud fraction in each bin therefore allows the calculation of the total cloud cover. It should be noted that the ISCCP retrieval of optical depth utilizes visible wavelengths, and the histograms are therefore only available during daytime.

The SOM scheme is applied to the joint probability distribution function histograms of the cloud top pressure versus optical depth. In this initial work, we study the data for one calendar year (2000) to demonstrate the potential of this technique. However, analysis of a number of other individual years displays very similar clusters (not shown) and we are principally interested in testing the SOM methodology.

We also use the ISCCP FD data [*Zhang et al.*, 2004] to provide clear and overcast radiative fluxes for the same period at identical spatial and temporal resolution. The radiative fluxes are derived using a broadband radiative transfer code which takes the ISCCP retrievals of cloud properties and surface albedo as inputs. From the clear and overcast radiative fluxes, we derive the cloud radiative effect (CRE) using the formulation defined in *Oreopoulos and Rossow* [2011]:

$$\mathrm{CRE}_{\mathrm{SW/LW}} = \mathrm{CF}[F_{\mathrm{SW/LW}}(\mathrm{clr}) - F_{\mathrm{SW/LW}}(\mathrm{over})] \tag{1}$$

where the fluxes (F) at the top of the atmosphere in the shortwave (SW) and longwave (LW) are differenced between clear (clr) and overcast (over) conditions and multiplied by the cloud fraction (CF).

This study also employs output from the ERA-Interim reanalysis [*Dee et al.*, 2011] on a 0.75° latitude/longitude grid. These data are then resampled to a 2.5° by 2.5° grid to allow direct comparison with the ISCCP output once it has been reprojected from an equal area grid using a nearest neighbor interpolation scheme.

Self-organizing maps (SOMs) are artificial neural networks commonly used to reduce the dimensionality of a data set by clustering [*Kohonen*, 1990]. The SOM scheme is an iterative unsupervised learning process which adjusts a set of reference vectors on the basis of differences between the reference vector and each input record. The initial set of reference vectors, where each reference vector represents the 42 values of cloud fraction in the joint histogram, are initialized linearly along the data's greatest eigenvectors. A learning rate determines how the adjustment is related to the difference between the reference vector and the input data. Training then consists of many iterations of reference vector adjustment until stable values are reached. In each iteration, the best matching reference vector is found for each input record and updated to more closely resemble the input data. Mathematically, individual vectors in the input matrix x are compared to a set of i reference vectors m_i to identify the best matching node c using the Euclidean distance:

$$c = \arg\min\{\|x - m_i\|\} \tag{2}$$

During the learning process, the reference vectors within a defined neighborhood of the input vector are updated, using the algorithm

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \tag{3}$$

where $h_{ci}(t)$ is the neighborhood kernel which can be written as

$$h_{ci}(t) = \alpha(t)\exp\left(\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right) \tag{4}$$

where $\alpha(t)$ is the learning rate, $\sigma(t)$ is the width of the kernel, and r_c and r_i are the radius vectors for reference vectors c and i. The learning rate and width of the kernel are reduced as a function of time such that the SOM evolves more quickly initially. The Euclidean distance term associated with the radius vectors for different node means is used to identify reference vectors within a certain range of the best matching vector. These neighboring reference vectors are then updated to a lesser degree than the best matching node resulting in adjacent nodes having the strongest similarity. This feature of the scheme produces the coherent organization of output which is a useful property of the scheme. This training process ultimately produces reference vectors that each represent a distinct portion of the multidimensional input space. Training of the neural network is unsupervised with the user specifying only the size and shape of the SOM as well as the training parameters (learning rate and width of the kernel); the end result is an objective set of distinct maps (referred to as nodes from this point on) that are representative of the entire data set. This makes SOMs useful tools for cluster

analysis of large, complex data sets, particularly because the algorithm is relatively simple and computationally efficient. However, it is worth clarifying that the SOM technique differs from cluster analysis in that it identifies points in the data space that are representative of the surrounding data rather than simply grouping the data [*Huth et al.*, 2008]. Thus, the nodes within the SOM reflect the entire range of the data, which is clearly a desirable property of a classification, but can make interpretation of the results challenging. While SOMs have been used for a wide array of studies covering many disciplines [*Kohonen*, 2013], they are particularly effective at developing climatologies [*Hewitson and Crane*, 2002] and investigating weather patterns and extremes [*Cassano et al.*, 2015; *Jolly et al.*, 2016]. They have also been widely used in identifying the impact of circulation patterns on extremes and trends [*Cavazos*, 1999; *Horton et al.*, 2015]. The recent review by *Sheridan and Lee* [2011] provides a good overview of SOMs in climate research. Clustering methods in general, and SOMs specifically, also have great potential for validating atmospheric model output against observations [*Coggins et al.*, 2014]. A more detailed mathematical treatment of the SOM scheme is detailed in *Kohonen* [1990] and *Kohonen* [2013].

This study is the first to our knowledge that has applied the SOM technique to satellite cloud data to identify multiple cloud types or WS globally. Within the context of this study, the purpose of the SOM is to cluster regions and periods with similar cloud conditions together into classes. The size and shape of a SOM are particularly important factors to consider before training: too many possible classes will result in low frequencies of occurrence for each class, while too few will result in classes that are "averages" of what may have been two similar, but distinct classes in a larger SOM [*Cassano et al.*, 2015].

After testing various configurations and examining ancillary information, such as the geographical and seasonal distribution of the various states and ERA-Interim reanalyses data, a 5×3 SOM was selected to be optimum. The 5×3 grid was selected as this shape and size showed sufficient intercluster variability while maintaining relatively high frequencies for each cluster. A Sammon map [*Sammon*, 1969] was also produced which showed approximately even separation between the SOM nodes (see later discussion) with some minor distortion which is a good indication that the SOM was well constructed and trained for the purpose of this study [*Cassano et al.*, 2015]. As an additional check, multiple SOMs were trained using different years of data and generally showed excellent correspondence with the results presented (not shown).

3. Results

Figure 1 displays the joint histogram for the 15 nodes identified via application of the SOM technique to all the daytime histograms observed globally during the year 2000. The color-coded bins identify the cloud fraction within each cloud top pressure/cloud optical depth bin. The relative frequency of occurrence and the total cloud cover associated with each node is identified in the titles. An examination of Figure 1 shows that the top left node in the SOM (node 1) is predominantly related to clouds with high tops (low pressure) and moderate to high optical depths, the total cloud cover also being high. The bottom right node in the grid (node 15) is related to low-level (high cloud top pressure) clouds with low to medium optical depths, though it should be noted that the total cloud cover is again high. Thus, this diagonal relates to polar opposites in terms of cloud top pressure. Similarly, the top right and bottom left nodes relate to clouds with rather different properties, but interestingly again, both relate to high cloud covers. This structure is not surprising given that *Reusch et al.* [2007] has previously noted that nodes at the ends of one diagonal are often similar to the positive and negative phases of the first principal component of the input data, while the other diagonal has corresponding similarities to the second principal component. This may also be a function of the initialization process used in this study. However, in general, even SOMs initialized randomly often display this property and the SOM pattern formed in this work is insensitive to different initializations.

An alternative way of viewing the ordering of the SOM is to examine the ordering of the nodes based on their cloud top pressure and optical depth independently. It is then possible to observe that the cloud top pressure for the highest occurrence states within the joint histogram increases from left to right in the grid, while the optical depth decreases from top to bottom in the grid. Interestingly, within this SOM the nodes linked to low total cloud covers (less than 50%) are in the center of the SOM (nodes 7–9, and 13). This may reflect the fact that cloud states with higher cloud covers will likely have higher variance than those for smaller cloud covers and thus the most different states will most naturally act as endpoints in the classification. We also note that the occurrence of thin cirrus cloud identified by the cloud fraction in the top left grid point in each joint histogram also reduces from left to right, showing that the ordering within the SOM can be interpreted in a

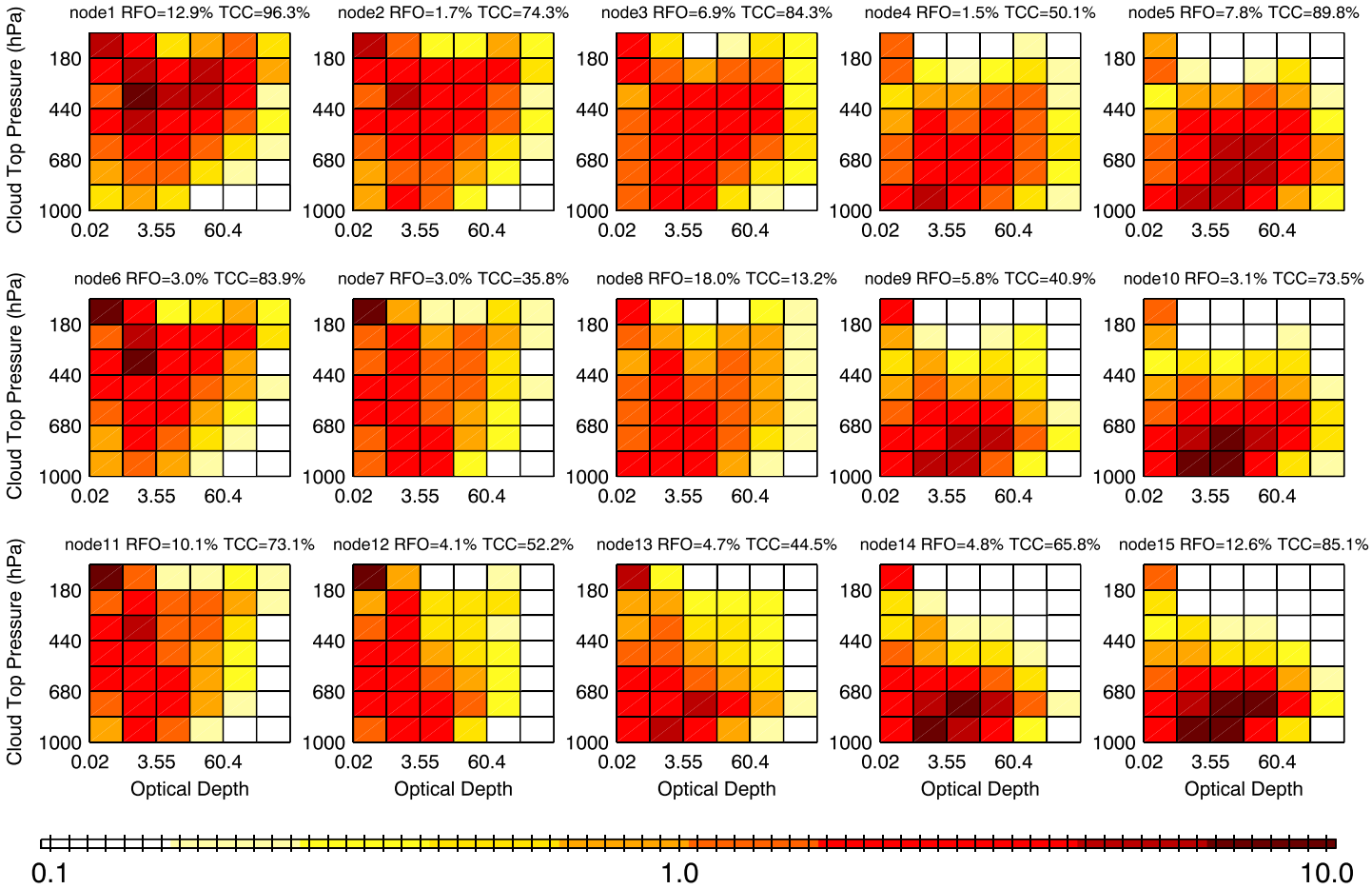

Figure 1. The SOM reference vectors displayed as joint histograms of cloud optical depth versus cloud top pressure, representing the 15 nodes identified by the self-organizing map. The relative frequency of occurrence (RFO) and the total cloud cover (TCC) associated with each node in the SOM are indicated in the legend above each histogram. Note the logarithmic color scale.

physically meaningful way. The objective ordering of states provided by the SOM technique is potentially an advantage relative to previous studies in certain applications. A comparison of the various nodes in the SOM with the WS in *Tselioudis et al.* [2013], derived using the *k*-means clustering algorithm applied to the global ISCCP data set, shows a number of corresponding states again highlighting the physically meaningful nature of this clustering. This correspondence also acts to independently verify the robustness of classes derived using the *k*-means clustering technique. For example, WS 1 in *Tselioudis et al.* [2013] compares well with node 1 in the current analysis. For the sake of clarity, a more detailed comparison of the differences and similarities between this analysis and previous work is left until the geographical distributions associated with each node have been introduced.

To understand the relationship between the various nodes in the SOM, and also to identify an interesting property of the SOM, a Sammon map is displayed in Figure 2. A Sammon map projects multidimensional vectors into a two-dimensional space allowing a visualization of the relationships between the different states. The reference vectors for each generalized SOM node, in this case identifying the cloud fraction in each grid point within the joint histogram, are projected onto a two-dimensional surface allowing the SOM nodes to be plotted on the basis of relative neighbor-to-neighbor similarity. Figure 2 represents the Sammon map for the SOM used in this study, where the separation between the points represents the Euclidean distance between individual nodes in the SOM and therefore the relationship between the neighboring nodes. The geometric shape of the Sammon map is important in defining how the nodes making up the SOM are interrelated. *Cassano et al.* [2015] suggests that a uniform Sammon map indicates little distortion which allows the physical behavior depicted in the SOM to be interpreted more easily. The pattern displayed by the Sammon map in Figure 2 is relatively regular, with each row and column having the same ordering as on the grid. However,

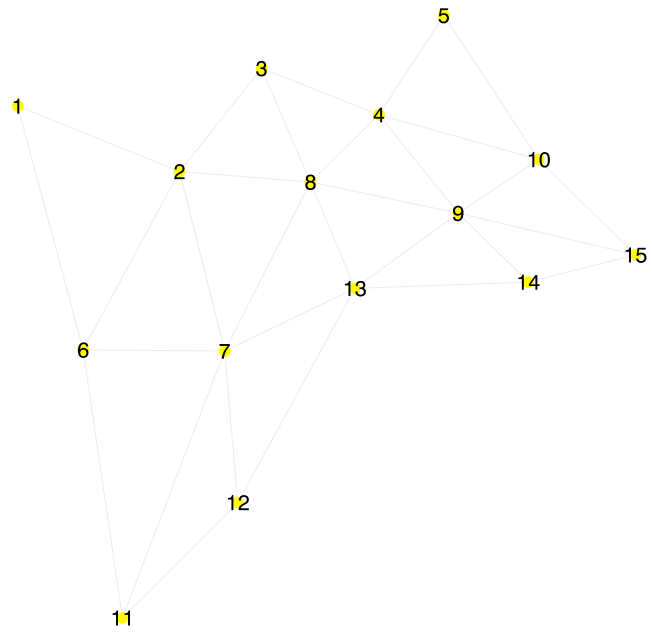

Figure 2. Sammon map displaying the relative position of each SOM node in the 2-D projection of the SOMs reference vectors (i.e., an estimate of the data's probability density function). The distances between nodes represent the Euclidean distances between SOM node reference vectors.

the proximity of some nodes to their neighbors varies, with nodes 1, 6, and 11 being very well separated in this two-dimensional representation. The fact that the Sammon map is not a simple grid demonstrates that the SOM technique distributes the nodes according to the density of the input data which is potentially an extremely beneficial trait in this application.

Figure 3 displays the geographical distribution of the various nodes, which in turn are related to particular cloud types, as a relative frequency of occurrence at each location. The RFO identified in Figure 1 is the global average; thus, it is possible to have small regions with high RFO compared to the node RFO. For example, node 1 has a global mean of 12.9% but can have regional values greater than 40%. To aid in understanding the physical significance of the patterns in Figure 3, it is useful to also examine the joint histograms displayed in Figure 1. For the sake of brevity we describe only a subset of the individual nodes and their relationships to the work of *Tselioudis et al.* [2013]. Examination of node 1 in Figures 1 and 3 suggests that this node represents clouds related to tropical deep convection, as it includes mostly high top optically thick clouds and is found primarily along the equator, over the ocean in the Pacific warm pool, the South Pacific and South Atlantic convergence zones and the Indian Ocean. This node also has high relative occurrence rates above land over the Amazon, equatorial Africa, and the Himalayas. The geographical and joint histogram for node 1 has a strong qualitative correspondence to WS 1 identified in *Tselioudis et al.* [2013], with the exception of the high occurrence rate over the Himalayas. We speculate that the pattern in the Himalayas may be related to the seasonal monsoon. It should be noted that node 1 also has some similar features to WS 3 in *Tselioudis et al.* [2013], but given that those states have a combined relative frequency of occurrence of 12.5%, it may be that this node actually represents a combination of these two states given that its frequency of occurrence is 12.9%. Previous work by *Tselioudis et al.* [2013], *Tan et al.* [2013], and *Tan et al.* [2015] have shown that the difference between WS 1 and WS 3 are physically important; thus, this point is considered later in section 3.3.

The joint histogram for node 2 in Figure 1 shows significant similarities to node 1 and the low global relative frequency of occurrence of this node might suggest that it should not be classified distinctly from node 1. However, examination of the Sammon map in Figure 2 suggests that node 2 is actually more similar to node 3 than any other node. In addition, node 2 displays a very distinct geographic structure with maxima linked to the Amazon, equatorial Africa, and the Himalayas (see Figure 3). This analysis therefore suggests that high top and moderately optically thick clouds observed over land are highlighted in this class. Whether this node is related to clouds with relatively lower cloud covers (74.3%) for which the retrieval of optical thickness has been impacted by the underlying surface type does come to mind. It should be noted that this type does not display any strong correspondence with any of the WS identified in *Tselioudis et al.* [2013].

Node 3 is related to midlevel clouds with a wide range of optical depths based on Figure 1 with a geographical distribution weighted toward latitudes poleward of 45° in both hemispheres, with a particularly high occurrence in the Southern Ocean storm track near the edge of Antarctica (see Figure 3). This type therefore bears some resemblance to WS 4 and WS 5 in *Tselioudis et al.* [2013]. Node 5 also has a similar joint histogram (see Figure 1) and similar geographical features (see Figure 3) to WS 4 in *Tselioudis et al.* [2000]. In particular, the enhancement over the West Antarctic in both WS 4 and node 5 is noticeable.

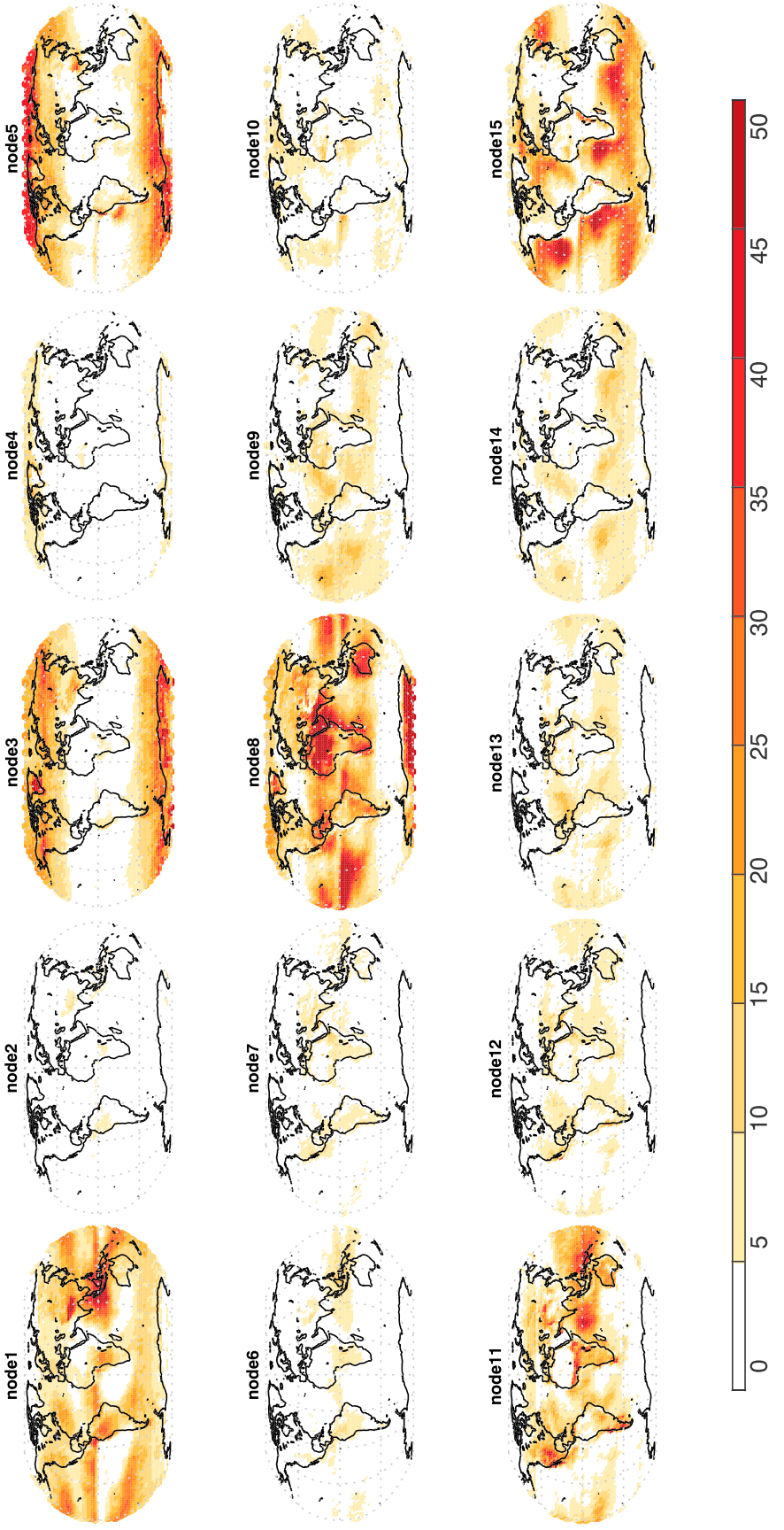

Figure 3. Maps of the annual mean relative frequency of occurrence (%) of the cloud types linked to the 15 nodes identified by the SOM displayed in Figure 1.

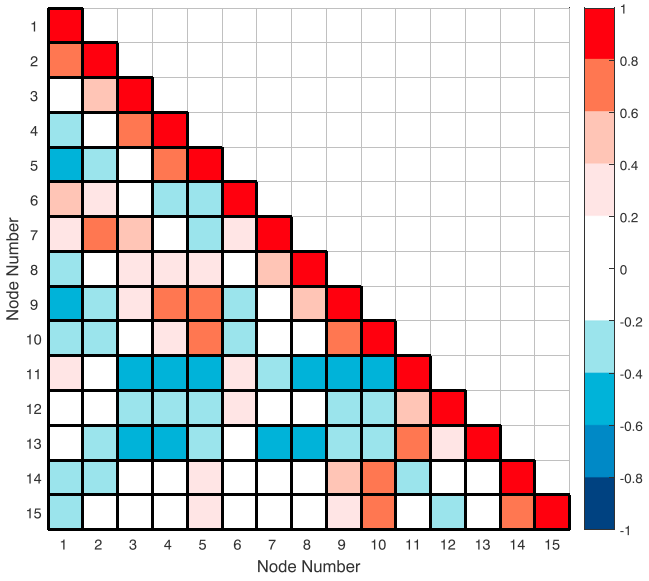

Figure 4. Correlation between the geographic relative frequency of occurrence maps for the different nodes.

As previously identified, the nodes with the lowest total cloud covers (below 50%) are related to nodes 7–9 and to a lesser extent 13. The geographic patterns and the joint histograms of these nodes display distinct similarities with WS 7. In particular, the node 8 geographical pattern bears a striking relationship to WS 7 with a widespread distribution over the tropical and subtropical oceans away from the convective regions and over the ice-capped landmasses of Antarctica and to a lesser extent Greenland. Interestingly, the geographical pattern of high occurrence for node 7 matches rather well with the pattern for node 8 over land regions apart from the Sahara and the ice-covered regions; while the node 9 pattern looks rather similar to the node 8 pattern over the oceans and the Sahara. Node 13 is also dominated by cloud cover over the tropical ocean regions. Given the very high relative frequency of occurrence of Weather State 7 in *Tselioudis et al.* [2013] (32.5%) and the lower occurrence rates of nodes 7–9 and 13, it seems that the current analysis has more finely separated the fair weather state in the previous work, highlighting differences in the geographic distribution of low cloud cover types, likely related to anticyclonic regions, over land and ocean [*Tselioudis et al.*, 2000].

This finer separation corresponds with previous work, with *Cassano et al.* [2006] indicating that the SOM classification approach is characterized by a tendency to categorize the distributions more uniformly than traditional cluster analysis algorithms. For example, *Michaelides et al.* [2001] compare an agglomerative hierarchical clustering with a SOM classification applied to precipitation data and find that the latter has a greater ability to identify minor characteristics within the distributions. *Michaelides et al.* [2001] also find that precipitation classification using the SOM categorizes the distributions more uniformly, whereas the hierarchical agglomerative clustering scheme classifies the distribution into a few distinctive classes. The latter behavior can result in grouping rarer data points into larger cluster classes that are not necessarily representative of those data points. However, it should be noted that the SOM method creates a gridded representation of a continuous data space. Thus, one issue with the SOM classification is that occasionally the scheme identifies transition nodes which may not be physically meaningful because the SOM creates a continuous gridded representation of the data space. Thus, some patterns will span relatively empty portions of the physical data space and these nodes show up as being relatively low frequency. Node 4, 6, 10, and also possibly node 12 in the current analysis may be transition states. This might be seen as a flaw, but it is also possible to consider these patterns as representing relatively rare features in the physical data space which might be an advantage when considering transitions. We also identify later that these transition states have different meteorological and radiative properties linked to them.

The cloud fraction joint histogram linked to node 11 (bottom left in grid) is dominated by cirrus (top left grid point) in Figure 1, and in Figure 3 is distributed around the Pacific warm pool, Indian Ocean, and equatorial Africa regions, and thus has distinct geographic similarities with the tropical convective cloud identified in node 1 of the SOM. However, unlike the node 1 geographic pattern there is a low frequency of this type above the Amazon. This node matches extremely well with WS 6 in *Tselioudis et al.* [2013], even including the low occurrence rate over the Amazon which *Tselioudis et al.* [2013] attributes to the strong seasonality of cirrus in this region.

Node 15 is related to the most commonly occurring low-level cloud type with predominantly lower optical depths (see Figure 1). An inspection of Figure 3 shows that this node is widespread geographically, but has the largest occurrence frequencies off the western coasts of North and South America, Africa, and Australia

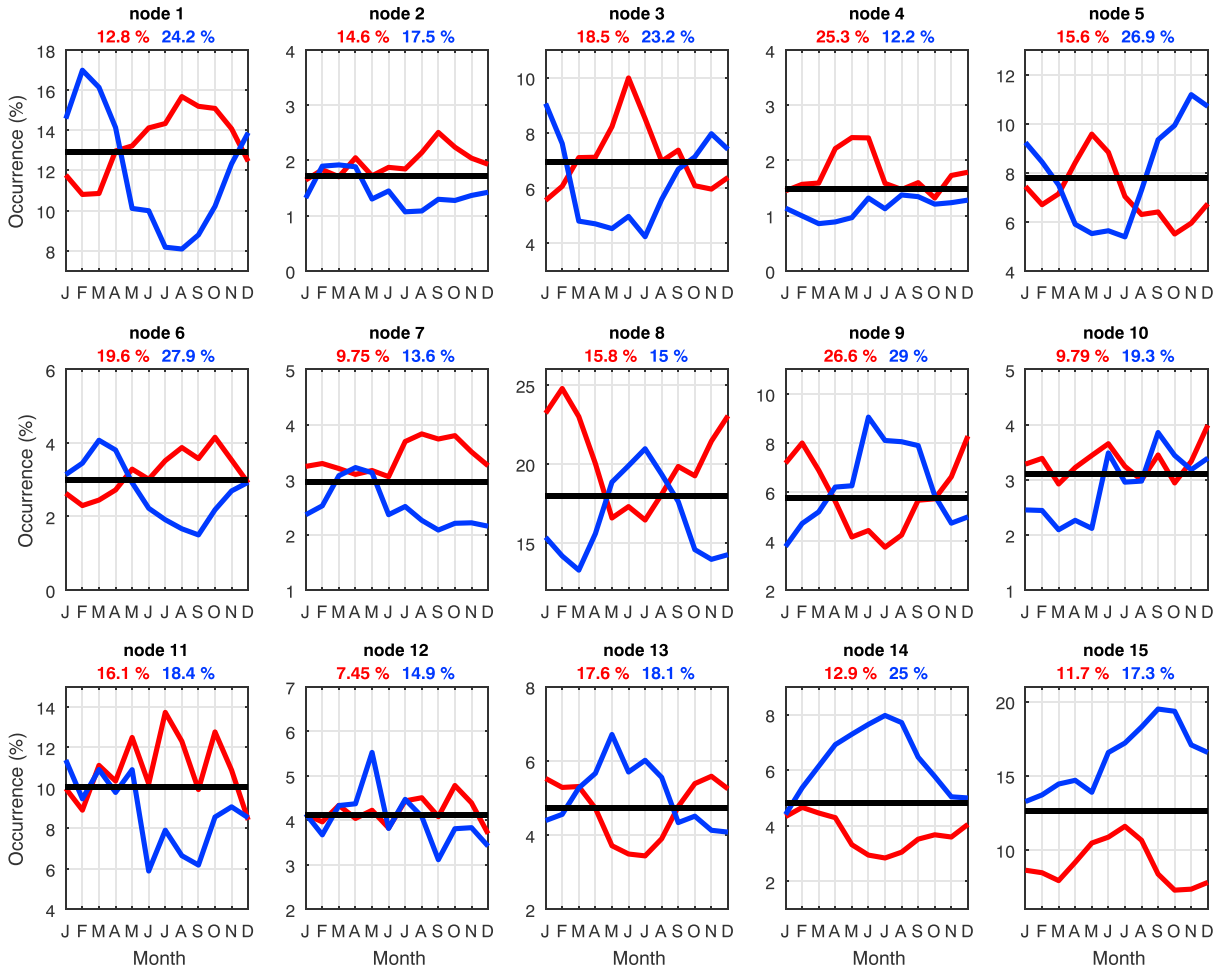

Figure 5. The monthly variation in the frequency of occurrence (%) of the various nodes in the SOM displayed in Figure 1 for the Northern and Southern Hemispheres (red and blue lines, respectively). The global mean for each node is identified as a black horizontal line in each subfigure. The standard deviation as a percentage over the year relative to the annual mean relative frequency of occurrence for each node is identified in each legend for the Northern and Southern Hemispheres (red and blue text).

and is widespread at approximately 40°S. The spatial structure and the joint histogram of this node therefore match extremely well with WS 10 in *Tselioudis et al.* [2013], though the global relative frequency of occurrence of this node in this analysis is almost double that in the previous work. This means that the five nodes with the highest relative frequency of occurrence have similarities with weather states in *Tselioudis et al.* [2013].

Note that the joint histograms (Figure 1) for Node 14 and 15 display distinct similarities which in previous studies may have meant a reduction in the number of states in the classification. However, examination of the Sammon map (Figure 2) shows that these states are close but distinct, and an inspection of the corresponding geographic maps (Figure 3) shows that node 14 is less tightly focused on the western coasts of North and South America, Africa, and Australia than the node 15 pattern. We will explore these nodes, their seasonal progression, and their meteorological context later in this analysis. There is also no corresponding enhancement at 40°S in the node 14 pattern. Therefore, to further test the robustness of the number of nodes in the SOM, the correlation coefficients between the different geographical patterns related to each node were derived. Figure 4 displays a matrix of the correlation coefficients between the geographic distributions for the various nodes in the SOM identified from the ISCCP joint histograms and displays a predominance of positively correlated nodes, with negatively correlated nodes being distinctly rarer. However, using the 0.8 correlation coefficient threshold from previous studies [*Tselioudis et al.*, 2013], the nodes are all identified to be distinct. It is interesting to note that the largest nonself correlations in Figure 4 consistently occur between

neighboring nodes. This further validates the objective ordering of the nodes geographically as a result of the SOM algorithm, in addition to the objective ordering previously displayed in the Sammon map (Figure 2). We also note that nodes 3 and 5 do not occur in the regions where the fair weather states are common.

3.1. Temporal Variability

Figure 5 displays seasonal variations of the relative frequency of occurrence for the different nodes in the Northern and Southern Hemispheres (red and blue lines, respectively). We display these distribution for both hemispheres and normalize the relative frequency data such that the impact of continual daylight or night in the polar regions has a small impact on the analysis. Examination of the Northern Hemisphere patterns in Figure 5 (red lines) shows a range of complicated seasonal behavior associated with the various nodes. In particular, nodes 1, 2, and 6 predominately occur in August, September, and October, respectively. Nodes 3 and 15 have a maximum in June and July. A number of the fair weather patterns have a minimum in Northern Hemisphere summer (nodes 8, 9, and 13) and nodes 4 and 5 have a maximum in May. Nodes 7, 10, and 12 also have relatively muted seasonal variation. The titles in Figure 5 for each of those nodes shows the range of variation for the Northern and Southern Hemispheres (red and blue text, respectively).

Inspection shows that the relative frequency of occurrence varies by less than 20% of the annual mean in the Northern Hemisphere as measured by the standard deviation of the monthly relative frequency of occurrence, apart from nodes 4 and 9. It is also noteworthy that nodes 14 and 15 have occurrence frequencies lower than the global mean in every month in the Northern Hemisphere, suggesting that these clouds are predominantly observed in the Southern Hemisphere which is also clear from inspection of Figure 3.

Figure 5 also displays the seasonal pattern of the frequency of occurrence for each node in the Southern Hemisphere (blue lines). A number of the nodes are seen to produce almost the opposite pattern to the Northern Hemisphere cycles, that is, the seasonal cycle is shifted by approximately 6 months in the Southern Hemisphere relative to the Northern Hemisphere. For example, nodes 1 and 6 now have minima in August and September. We also see that nodes 7 and 12 have muted seasonal cycles in the Southern Hemisphere. Interestingly, the seasonal variation is larger in the Southern Hemisphere than in the Northern Hemisphere with only nodes 4 and 8 having larger relative variations in the Northern Hemisphere (cf. red and blue standard deviation values in the titles). This is particularly clear in nodes 1, 5, 10, 12, and 14 where the seasonal variation is close to a factor of 2 larger than in the Northern Hemisphere. Node 15 is particularly interesting as it does not display the expected 6 month shift in the timing of maximum occurrence between the hemispheres. Thus, the 5×3 SOM created displays coherent structures associated with the joint histograms and both the geographical and seasonal patterns, suggesting that the patterns selected are physically meaningful.

As previously hinted at, one potential advantage of the SOM technique relative to previous studies, which nearly all used *k*-means clustering, is that the SOM scheme creates a reproducible organization of the various nodes. Studies using the *k*-mean scheme require a subjective ordering of the various states [*Tselioudis et al.*, 2013; *Mason et al.*, 2015], which means that this type of analysis is more difficult to use for studies focusing on the transitions between states, an area that has therefore received little attention previously. The exceptions are a minor application in *Jakob et al.* [2005] and recent work by *Tan and Jakob* [2013] and *Tan et al.* [2015]. *Tan and Jakob* [2013] developed a convective regime data set based on ISCCP infrared-only retrievals. They demonstrated that these regimes capture the essential properties of the original weather states, but can be used to track states and transitions between states. This allows them to examine the diurnal cycle of convection in the framework of their regimes, for example. A scheme which objectively identifies states into an ordering could be beneficial for similar studies.

Figure 6 displays a set of grids, one for each node, representing information about transition frequencies between nodes. These transition frequencies have been derived from data one day apart over the entire globe. We decided to compare the daily transitions because of the significant potential for creating a biased sample if we considered 3 h transitions given that only daytime histograms can be utilized in our analysis. The application of SOMs to ISCCP output for both day and nighttime data as used in *Tan and Jakob* [2013] is beyond the scope of the current study, but clearly is an area worthy of future work. Examination of the grid linked to node 1 in Figure 6 shows a number of interesting features. First, the highest transition probability for node 1 is itself suggesting that node 1 is a persistent type. We also note that when a particular geographic position in the node 1 state transitions to another node, this is most commonly node 11. We also see that all the other corner nodes (nodes 5, 11, and 15) are persistent and are most likely to remain in that type for more than 1 day. In addition, for nodes 5, 11, and 15 when moving to another node, the node transitioned

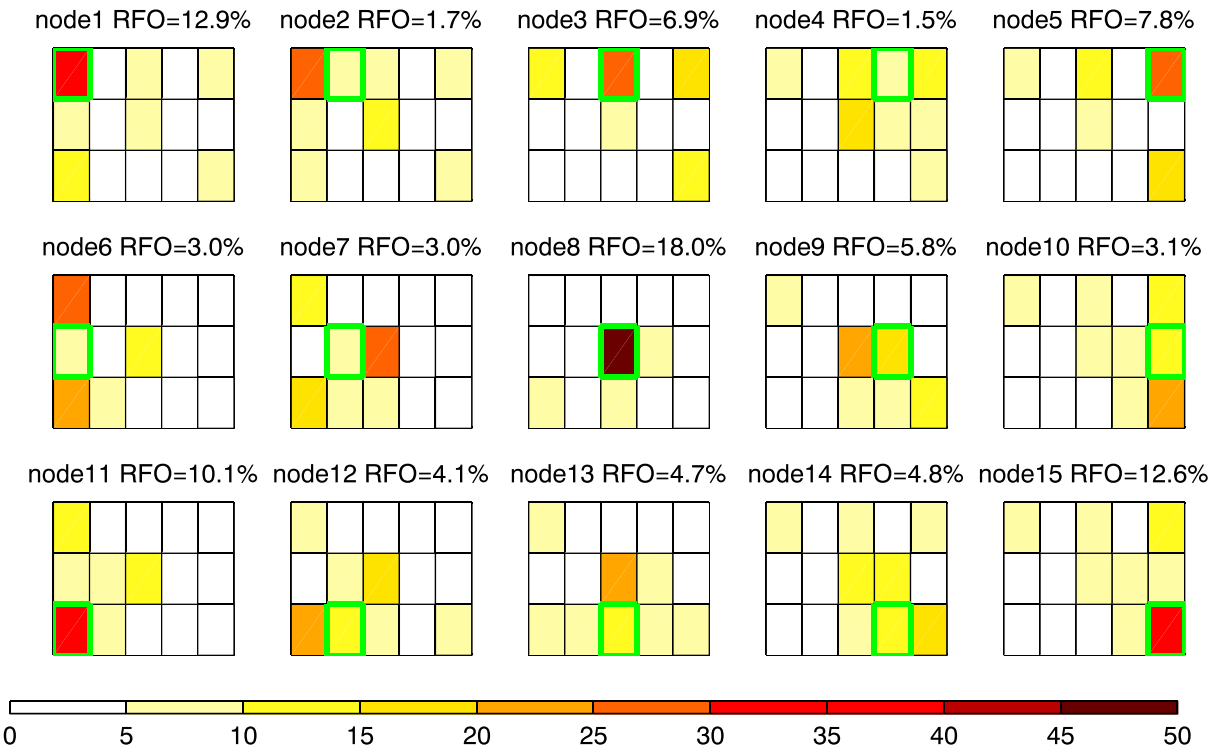

Figure 6. The set of grids in this figure identify the transition probability (%) for each node to another node. The green box identifies the node considered for the transition probabilities in each grid. The relative frequency of occurrence associated with each node in the SOM is indicated in the legend above each grid.

to is the nearest adjacent corner. The most common transitions away from a node are also associated with nodes close to the original node. It is interesting that the nodes with higher relative frequencies of occurrence also generally have higher persistence. For example, the nodes with the five largest relative frequencies of occurrence (1, 5, 8, 11, and 15) all have high persistence from one day to another.

To gain a better understanding of the transitions and to reduce the impact of the varying relative frequency of occurrence of the different types, Figure 7 displays the difference between the transition probability and the node's global relative frequency of occurrence (identified in the titles in Figure 7). The underlying assumption in this analysis is that, if transitions are random, then the transition probability should have the same value as that node's relative frequency of occurrence. Figure 7 demonstrates that the vast majority of nodes either stay in the same node or shift to near neighbors after a day. In particular, nodes 1, 3, 5, 8, 9, 11, 13, and 15 are most likely to stay in the current node than to transition to any other node. We also see that node 8 is most likely to transition to nodes 7, 9, or 13 which are all nodes associated with low total cloud covers or fair weather states. All of the corner nodes (1, 5, 11, and 15) also show strong ordering with the probability of transitions between diametrically opposite corners being unlikely in every case. The combination of Figures 6 and 7 strongly demonstrates that the SOM has identified a clustering which is ordered in a physically meaningful way in terms of transition probabilities between cloud types between days and that this has occurred without the need for subjective ordering using ancillary data as in *Tan and Jakob* [2013]. Also, highlighting that the nodes linked to transitions, such as nodes 4 and 12, would be useful in future investigations of temporal variability.

To obtain further insight on the seasonal variability and to further highlight the quality of the SOM classification, Figure 8 displays the geographical relative frequency of occurrence as a function of season for node 15. Node 15 was selected for examination because of its high annual mean occurrence rate and the large seasonal variability displayed. As previously identified, the low-level clouds that this node represents are widely spread over the ocean regions. However, an examination of the relative frequency of occurrence of this node in Figure 8 shows that the distinct seasonal cycle in this type is associated with variations in the occurrence rates off the western coasts of North and South America, Africa, and Australia and to lesser extent Europe. In particular, the high occurrence rates off the western coast of North America have a strong seasonal cycle with a maximum in boreal summer, while in the Southern Hemisphere the regions of high

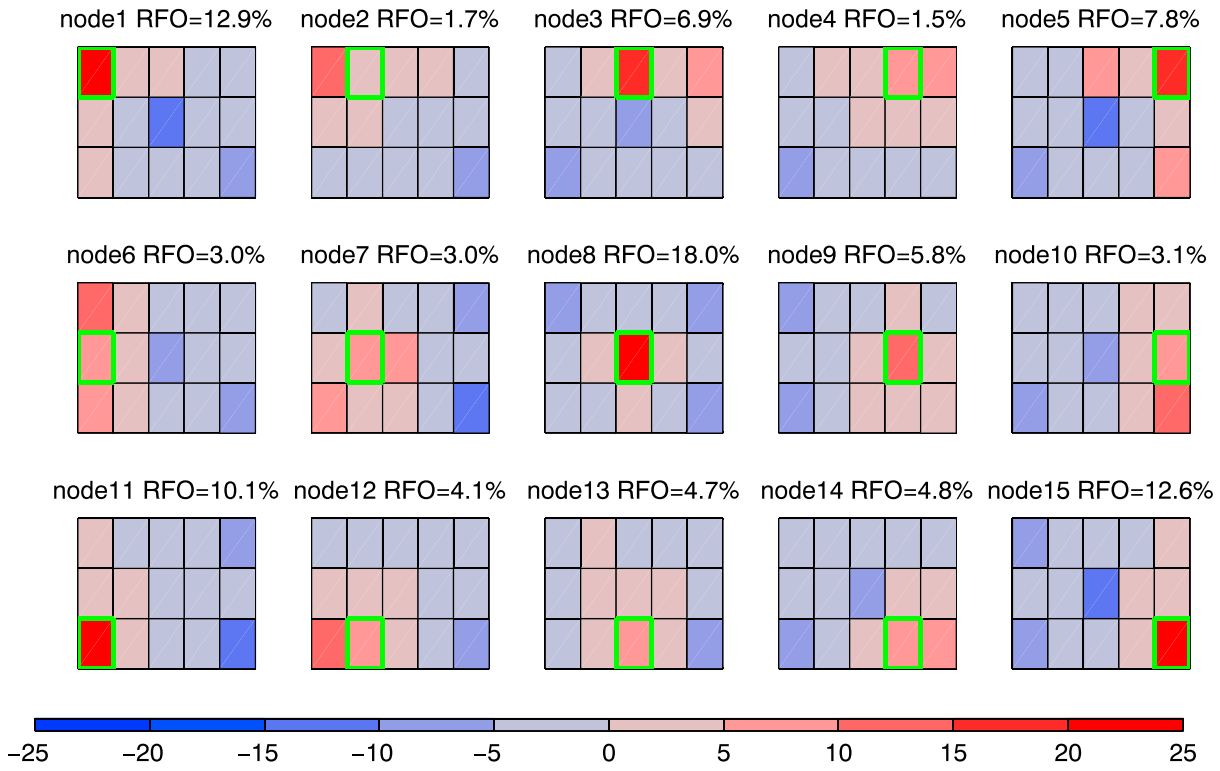

Figure 7. As in Figure 6 but for the difference between the transition probability and the relative frequency of occurrence (%).

occurrence off the west coasts of South America, Africa, and Australia minimize in March to May. The similarity of node 15 with WS 10 in *Tselioudis et al.* [2013] suggests that this node is related to marine stratocumulus and stratus cloud decks. This interpretation is supported by the almost exclusive occurrence of this type over oceans, particularly off the western coastlines in the subtropics, locations that are linked with extensive stratocumulus decks [*Klein and Hartmann*, 1993]. Work in *Muhlbauer et al.* [2014] has previously examined marine stratocumulus globally and found a very similar seasonal variation in the North Pacific and Atlantic to those identified for node 15 in Figure 8. Similarly, the seasonal variation and geographic position of maxima in occurrence match rather well with the results of *Muhlbauer et al.* [2014]. That study identified that the maxima near

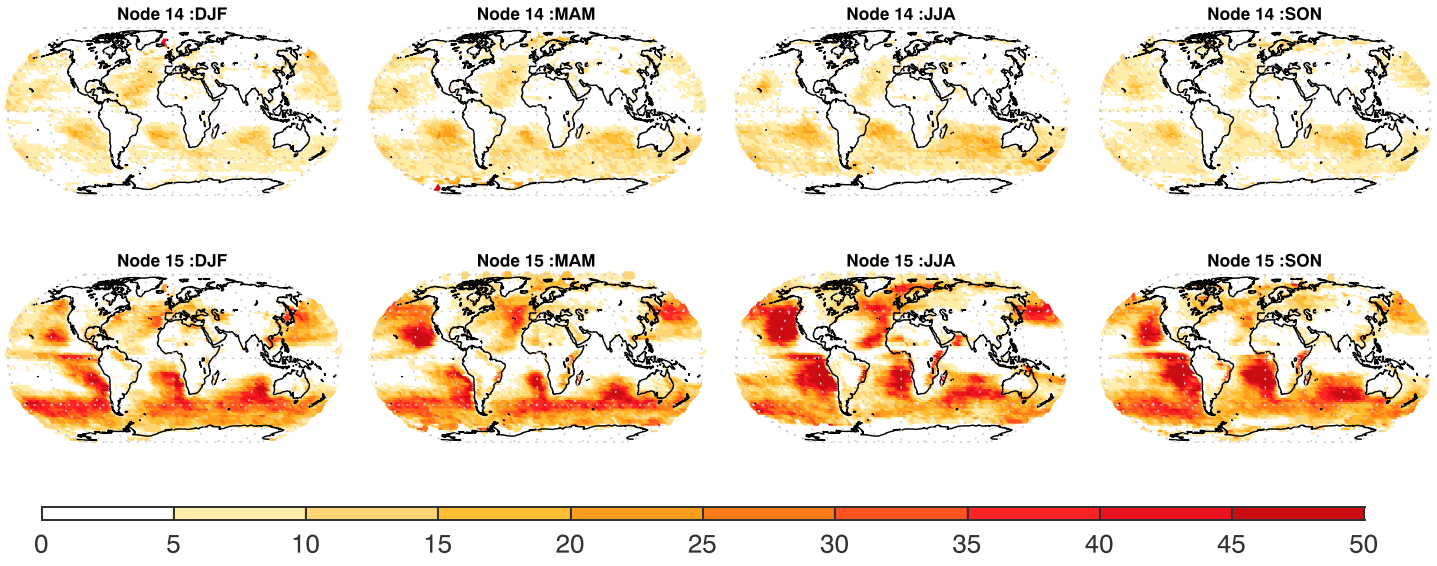

Figure 8. Maps of the seasonal mean relative frequency of occurrence (%) for nodes 14 and 15 identified by the SOM displayed in Figure 1.

45°S is more consistent than the variability at lower latitudes which is also a feature of this nodes seasonal geographic pattern. Further examination of the work of *Muhlbauer et al.* [2014] also suggests that node 14 has distinct similarities with the open mesoscale cellular convection (MCC) class in that work (see Figure 8). Examination of Figure 8 shows many similar features in the two nodes; however, there is a very clear enhancement in June, July, and August which corresponds with the open MCC classification in Figure 5 of *Muhlbauer et al.* [2014]. This is surprising as this is a rather subtle categorization of stratocumulus, particularly given the complexity of the analysis utilized in that study to identify the different spatial structures of the MCC. This demonstrates that the ISCCP data set is classified very effectively with the SOM scheme in this region of the phase space.

Examination of the results in *Muhlbauer et al.* [2014] also suggests that the total cloud cover is generally higher in the closed MCC than the open MCC class and that while the cloud top heights are rather similar, the closed MCC tend to have higher cloud optical depths. These differences are also observed in nodes 14 and 15, with higher cloud covers in node 15 than 14, little difference in cloud top pressure and lower optical depths in node 14. This good correspondence therefore again supports the subtlety of the SOM classification. This rather subtle classification could be considered irrelevant; however, *Muhlbauer et al.* [2014] show that the instantaneous shortwave cloud radiative forcing is about twice as high in the closed MCC case than the open MCC case. We show later in this study that a similar change occurs for our classification using the ISCCP FD data. This supports our interpretation that the SOM classification scheme is identifying physically meaningful classes.

3.2. Comparison With Ancillary Data

Applications of the cloud regime information derived in previous classification studies include understanding cloud feedbacks and the evaluation of regime-dependent errors in GCMs. As identified in *Jakob et al.* [2005], these applications rely on the assumption that the various regimes or weather states relate to a distinct atmospheric context. For example, *Jakob et al.* [2005] consider thermodynamic and radiative characteristics. A number of studies [*Klein and Hartmann*, 1993; *Medeiros and Stevens*, 2011; *Tan et al.*, 2013, 2015] have also shown that the midtropospheric vertical velocity and lower tropospheric stability (LTS) are important parameters for identifying the impact of the dynamic and thermodynamic state of the atmosphere on cloud. In addition, work in *Jakob et al.* [2005] has used the total column water vapor (TCWV) to characterize the overall water vapor content in each regime. We therefore examine the vertical velocity profile, the LTS, and the TCWV values related to the various nodes. Given the significant latitudinal variability in the TCWV, we display the perturbations of that variable relative to the zonal mean to allow greater specification. In this analysis, ERA-Interim reanalysis data are used and subsampled to a square 2.5° by 2.5° grid which aligns with a similar grid derived from the ISCCP data's equal area coordinate system. The equal area ISCCP data were interpolated to a 2.5° by 2.5° latitude-longitude grid using a nearest neighbor scheme. Given the 6-hourly output available from the ERA-Interim reanalysis, only the corresponding subset from the ISCCP 3-hourly data set is used in this analysis.

In a similar manner to *Gordon et al.* [2005], *Jakob et al.* [2005], and *Gordon and Norris* [2010], we have derived the mean vertical velocity profile for each cloud regime for pressure levels between 1000 and 100 hPa. Figure 9 displays vertical velocity profiles for each node in black, and gray lines identify the vertical velocity profiles connected to all the other nodes for reference. Inspection of Figure 9 shows that the top left node in the SOM (node 1) is predominantly related to large negative pressure vertical velocities (connected to upward motion) with a maximum value at around 400 hPa. Thus, a node predominantly related to clouds with high tops in Figure 1 is linked to a region of strong ascent. The bottom right node in the grid (node 15) is related to positive vertical velocities throughout the profile and is therefore linked to a region of strong descent. In Figure 1, node 15 (bottom right in grid) is related to low-level clouds as might be expected. This diagonal relates to polar opposites in terms of vertical velocity sign and cloud top pressure which reiterates the fact that there is a physically meaningful ordering. The gradual shift from positive (descent) to negative (ascent) vertical velocities as we move from right to left in the SOM structure is also clear. Interestingly, the vertical velocity profile for nodes 9, 14, and 15 are so similar that they are indistinguishable in Figure 9, which might suggest an issue. However, separation of different cloud regimes via the mean vertical velocity in the midtroposphere (500 hPa) as completed in previous work by *Tan and Jakob* [2013] and *Tselioudis et al.* [2013] also shows that certain states have little difference in their vertical velocity distributions. For example, the vertical velocity distributions for WS 9 to WS 11 are near identical in *Tselioudis et al.* [2013]. The use of profiles does provide slightly more information. For example, they highlight that nodes 2, 3, and 6 have different vertical structures, but near-identical mean values at 500 hPa. Though, Figure 9 also clearly identifies the fact that different clouds

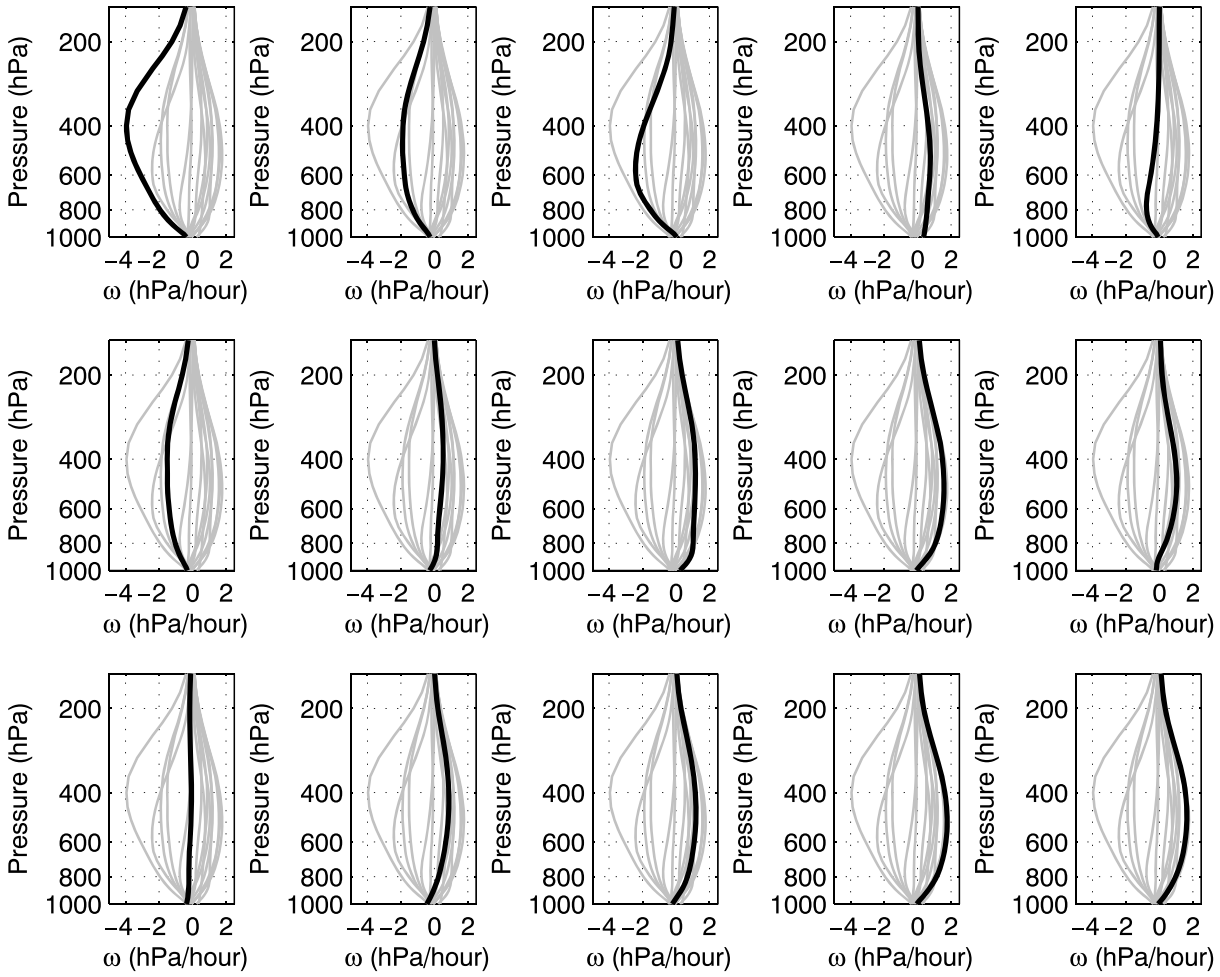

Figure 9. Vertical profiles of the pressure vertical velocity (hPa/h). Negative values correspond to upward motion and positive values relate to downward motion.

can form in very similar vertical velocity regimes. We will therefore examine other factors relevant to clouds below. Notably, all the transition states (nodes 4, 6, 10, and 12) have different vertical velocity profiles than their nearest neighbors based on Figure 7, suggesting that the transition states may be physically meaningful, but rare.

Before moving to other variables, it is useful to compare our patterns with previous work, comparison of the patterns in *Gordon and Norris* [2010] associated with joint histograms and vertical velocity profiles shows a good deal of correspondence with the distributions in our analysis. For example, their Cluster 7 (associated with strong frontal activity in their midlatitude study) is linked to large negative vertical velocities and optically thick cloud with low cloud top pressures which are very similar structures to those observed for our node 1. The vertical velocity profile for node 1 is also similar to the CD regime in *Jakob et al.* [2005], which they identified as their regime that is closest to the classic tropical convection profile. Thus, node 1 seems to link to both enhanced vertical velocity regimes in the tropics and midlatitudes as might be expected from the geographic distribution displayed in Figure 3. Cluster 1 in *Gordon and Norris* [2010] is related to near-zero vertical velocities and a joint histogram related to clear sky conditions which corresponds well with our node 6. Clusters 2 and 3 in *Gordon and Norris* [2010] are linked to positive vertical velocity profiles and optically thin low-level cloud, which they identify as cumulus and stratocumulus, which corresponds nicely with nodes in the bottom right of the SOM grid. Similarly, our node 1, previously identified to be connected to WS 1 and WS 3, has larger negative vertical velocities at 500 hPa than all other states in line with the pattern in *Tselioudis et al.* [2013].

To further assess the importance of the dynamic and thermodynamic environment. Figure 10 displays the mean values of the TCWV perturbation associated with specific ranges of the midtroposphere vertical velocity

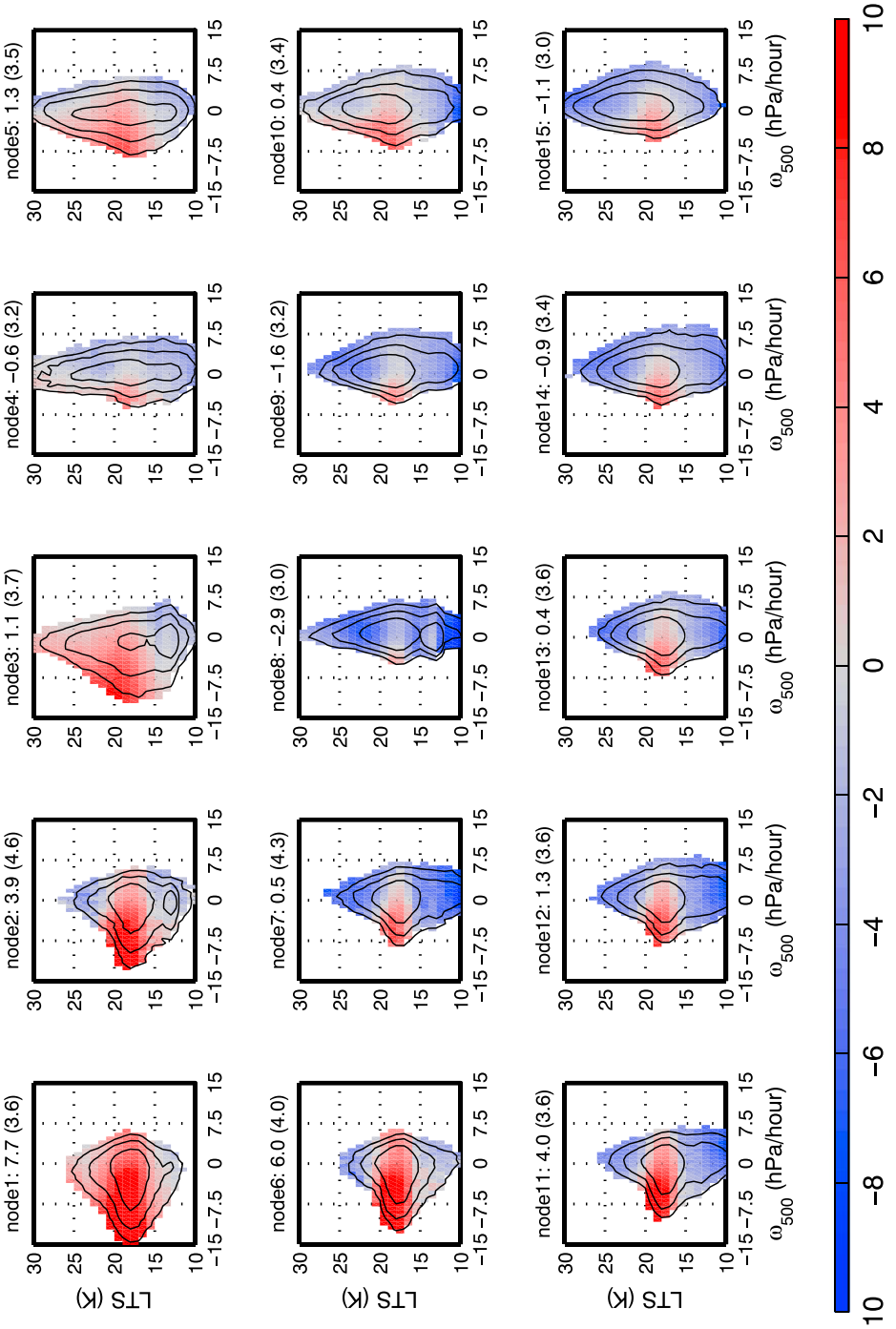

Figure 10. Color contours identify the mean TCWV perturbation as a function of middle tropospheric vertical velocity (ω_{500}) and lower tropospheric stability for individual nodes. Note that negative values of ω_{500} correspond to upward motion and positive values relate to downward motion. Only regions where the density of observations is above a threshold of 0.075% are included in each diagram to demonstrate the most relevant range of TCWV, ω_{500}, and LTS. The three contour lines in each figure relate to a density level of 0.1%, 0.3%, and 0.5% for the particular grid square. The first number in the legend of each subfigure identifies the weighted mean total column water vapor perturbation, while the number in brackets represents the standard deviation of the TCWV perturbation distribution.

and the lower tropospheric stability. Only TCWV perturbation values in regions where the density of observations is greater than 0.075% of the total number of observations within that node are displayed. Contour lines in each node identify the 0.1%, 0.3%, and 0.5% density values, with the most central line always representing the highest densities. This allows us to effectively identify the joint histogram related to the density of observations as a function of vertical velocity and LTS, similar to analysis in *Tan et al.* [2013]. Examination of Figure 10 shows a clear ordering in the density patterns with the smallest values of the LTS and most negative vertical velocities (highest rates of ascent) observed in the left-hand column of the grid. Notably, the highest density states (linked to specific ranges of vertical velocity and LTS) for node 1 are linked to negative and near-zero vertical velocities and are linked to a well-defined region with LTS values between 15 and 20 K. The opposite diagonal (node 15) has a prevalence of larger LTS values and almost exclusively positive vertical velocities. Thus, the joint histograms have a clear ordering linked to these opposite diagonals. We also note that as we move from left to right in the grid the range of LTS values increases and the vertical velocity values become more positive. Interestingly, box-whisker plots of LTS and vertical velocity for the regimes identified in *Tan and Jakob* [2013] show a similar ordering, namely, more negative vertical velocities connected to lower values of LTS. This fact is supportive of both the ordering of the SOM and the fact that it is physically meaningful.

The weighted mean values of the TCWV perturbations identified in the legend of each subfigure are not so clear, with the largest values on the left-hand side of the grid. But the smallest values distributed throughout the right-hand side and center of the grid. In particular, node 1 (top left in grid) has a weighted arithmetic mean TCWV perturbation of 7.7 mm, node 8 has a mean TCWV perturbation of −2.9 mm, and the final node on the diagonal (node 15) has a value of −1.1 mm. The most positive TCWV anomalies relate to regions of low LTS and large negative vertical velocity and therefore likely relate to regions of horizontal moisture convergence and upward vertical motion, while the largest negative TCWV anomaly relates to node 8, a clear sky state.

Comparison of the nodes linked to clear skies (node 7–9 and 13) show some differences in their vertical velocity versus LTS joint histograms. For example, node 8 has a wide range of LTS but a more constrained set of vertical velocities than the other clear sky states. We also see that node 8 has a large negative TCWV anomaly related to it. Thus, this cloud regime seems to correspond to low vertical velocities, a wide range of LTS values, and relatively dry regions of the atmosphere. For node 7, associated with slightly positive TCWV perturbations, the lack of cloud appears to be related to very weak vertical motion, while larger LTS and dry air could be factors in the absence of clouds in node 9. Figure 10 shows that there is a good deal of overlap in the joint histograms for vertical velocity and LTS, but given the results in *Tan and Jakob* [2013] this is to be expected. Thus, similar cloud regimes can be found for relatively wide ranges of LTS and vertical velocity, suggesting that other controlling factors play a role.

To further support the usage of self-organizing maps on the ISCCP joint histogram data, the radiative effects of the various nodes are displayed in Figure 11. The analysis here is similar to that detailed in *Oreopoulos and Rossow* [2011]. Their study quantified the cloud radiative effect associated with the cloud regimes identified in *Rossow et al.* [2005] using the ISCCP FD data set for shortwave and longwave wavelengths. Figure 11 displays the median SW and LW CRE for each node in the SOM. The error bars in Figure 11 indicate one quarter of the interquartile range of the distributions used to calculate the composite means, similar to the method of display used in *Oreopoulos and Rossow* [2011]. Examination of the ordering of the nodes in the CRE LW versus SW space shows that nodes 1, 6, and 11 are the closest nodes to the top right of the diagram, all of these nodes are in the leftmost column of the SOM grid displayed in Figure 1. Nodes 5, 10, and 15 bound the bottom left portion of the range observed and are linked to the rightmost column in Figure 1. This organization is likely associated with the fact that optically thicker cloud has a greater SW TOA CRE. While higher cloud tops are associated with larger LW TOA CRE for a particular optical thickness. For reference, it was noted previously that the cloud top pressure for the highest occurrence states within the joint histogram increase from left to right in the grid in Figure 1, while the optical depth decreases from top to bottom in the grid. We also note that the most positive TCWV perturbations occur for large vertical velocities and intermediate values of the LTS in every case demonstrating the importance of these parameters on TCWV anomalies.

It is also notable that there are a large number of nodes with similar CRE properties linked to the top left region of Figure 11, effectively clouds which have low CREs. This clustering of nodes is related to nodes 4, 7–9, and 12–14. This is likely a feature of the organization of the SOM in Figure 1 which attempts to distribute evenly over the phase space, which means that low cloud fraction states (nodes 7–9 and 13) will be clustered, for example. The color coding in Figure 11 shows the logarithm of the density of ISCCP FD states over the

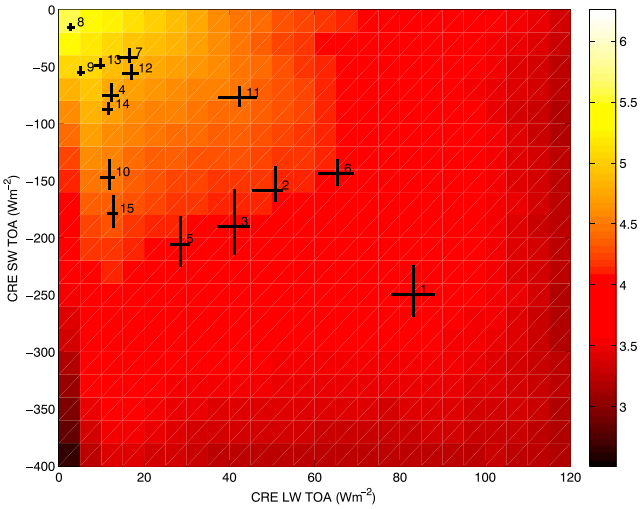

Figure 11. ISCCP FD daily LW and SW TOA CREs for the year 2000 composited using the SOM nodes displayed in Figure 1. The horizontal and vertical error bars indicate one quarter of the interquartile range of the distributions used to calculate the composite means; distance from median to 25th percentile is represented by the error bars below and to the left of the symbol, while that to the 75th percentile is represented by the error bars above and to the right. Color contours show the $\log_{10}$ of the number of observations within a CRE LW TOA versus CRE SW TOA grid square, effectively identifying observational density.

CRE SW/LW phase space. This clearly shows that most observations occur in this region; therefore, the SOM has placed more nodes in a region connected to the most populous region in this case. We therefore argue that this clustering is justified. For example, node 8 is clearly separated from all the other nodes. Given the difference in the CRE of nearly 100 W m^{-2} in the shortwave, separation of the clear sky states into a set of nodes makes some sense physically. In particular, the small individual CRE differences are counteracted by the relatively large relative frequency of occurrence and the wide variation in cloud fraction.

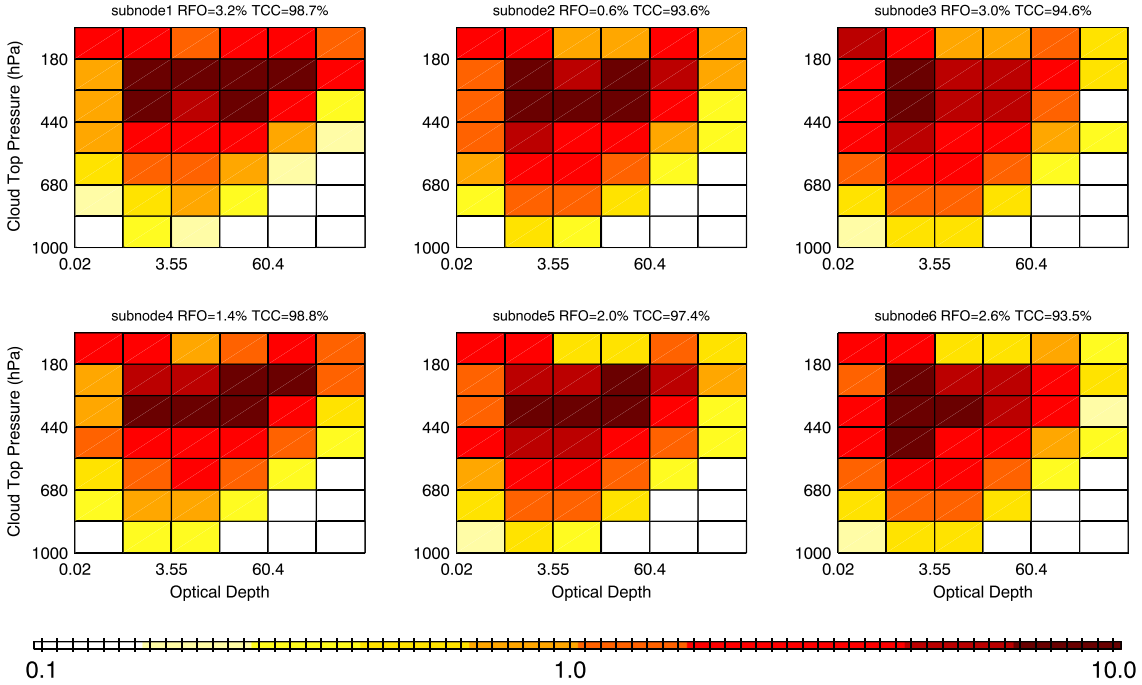

Figure 12. Joint histograms of cloud optical depth versus cloud top pressure, representing six subnodes identified by applying the SOM method to only those data classified in node 1 in the original SOM displayed in Figure 1. The relative frequency of occurrence (RFO) and the total cloud cover (TCC) associated with each subnode in the sub-SOM is indicated in the legend above each histogram. Note the logarithmic color scale.

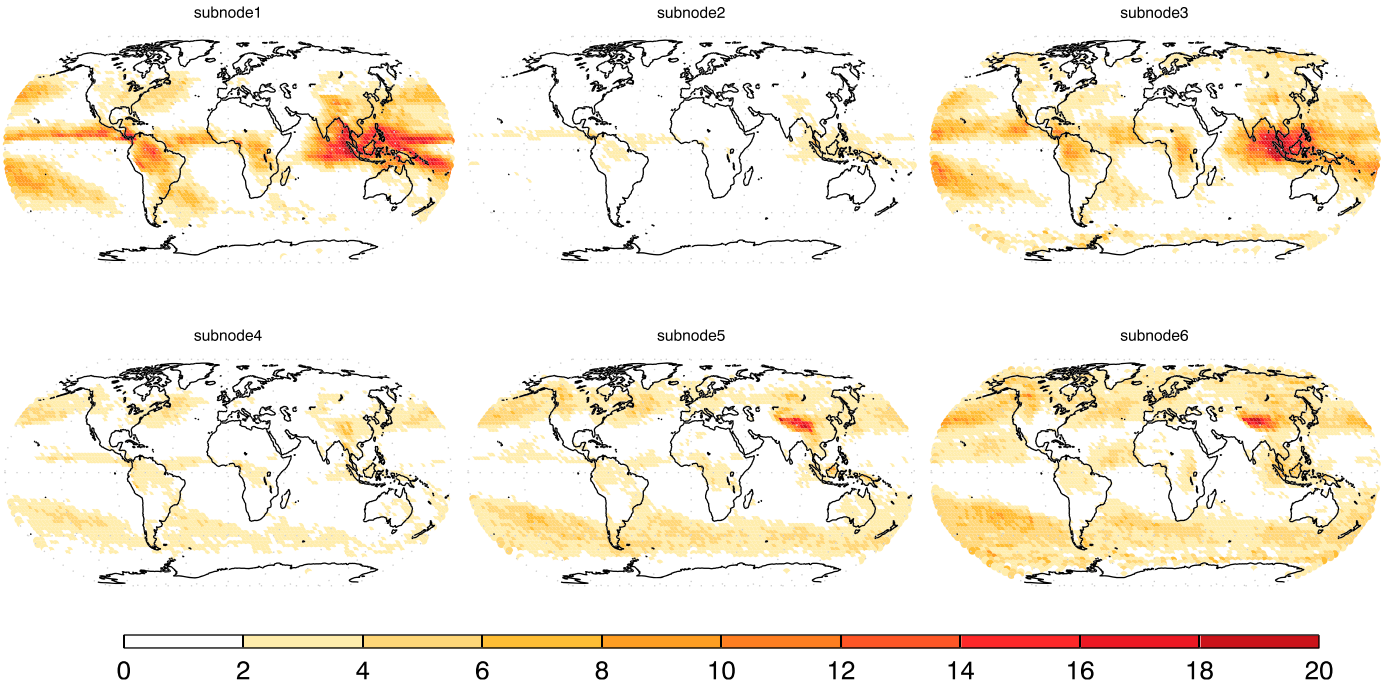

Figure 13. Maps of the annual mean relative frequency of occurrence (%) of the cloud types linked to the six nodes identified by the sub-SOM displayed in Figure 12.

In addition, we identified that nodes 14 and 15 seem to have strong similarities with the open MCC and closed MCC classes in *Muhlbauer et al.* [2014], respectively. That work identified that the instantaneous shortwave cloud radiative forcing is about twice as high in the closed MCC case than in the open MCC case. Inspection of Figure 11 shows an almost identical doubling relationship, with node 15 having a CRE SW TOA of approximately −175 W m^{-2} and node 14 approximately −80 W m^{-2}. We are therefore very confident that nodes 14 and 15 are physically distinct.

3.3. Issues With the SOM Representation

We have already identified strong points in favor of the SOM analysis, but the fact that they form a gridded representation of a continuous data space can cause issues. For example, node 1 in Figure 1 is likely related to WS 1 and WS 3 from *Tselioudis et al.* [2013] which relates to different physical regimes according to *Tselioudis et al.* [2013] and *Tan et al.* [2013]. One way to mitigate this issue is to form a sub-SOM, effectively applying a SOM to the data related to that particular node. Figure 12 displays the joint histogram for six nodes identified via application of the SOM technique to ISCCP data linked to node 1 in Figure 1. This sub-SOM, as might be expected, displays much subtler variations than those identified in Figure 1. But some similarities to that previous SOM in terms of a distinct ordering can be identified. For example, the cloud top pressure increases from top to bottom in the 3 × 2 grid, while the optical depth decreases from left to right in the grid. Comparison of specific subnodes within the sub-SOM shows relationships with the WS in *Tselioudis et al.* [2013]. For example, subnodes 1 and 2 in the sub-SOM are now very similar to WS 1, while subnode 3 in the sub-SOM is very similar to WS 3 in *Tselioudis et al.* [2013]. The corresponding geographical patterns linked to the sub-SOM are displayed in Figure 13 and for subnodes 1–3 show good correspondence with WS 1 and WS 3 patterns in *Tselioudis et al.* [2013], respectively. The lower row in the grid show similarities to WS 1–WS 3. Subnode 5 in the sub-SOM in particular bears a resemblance to WS 2 in *Tselioudis et al.* [2013]. This suggests that the different classification methodologies can find a number of similar cloud regimes but that the classes formed are not identical. Given the differences between previous individual weather state classifications, for example, compare *Rossow et al.* [2005] and *Tselioudis et al.* [2013], we might expect this result. Regardless, we feel that this study demonstrates that SOMs can identify physically meaningful cloud regimes and that further work to examine the classification of cloud data using SOMs is warranted.

4. Discussion and Conclusion

This work demonstrates that clustering joint histogram data from the ISCCP D1 data set using self-organizing maps is a promising methodology for use in cloud classification. Data exploration identified that a 5×3 SOM applied to the ISCCP joint histograms for the year 2000 provided a good representation of the range of joint histograms observed globally and also a set of clusters which passed separation criteria used in previous studies. We also demonstrated that the 5×3 SOM created has a number of nodes which display a close resemblance to cloud regimes (weather states) identified using the *k*-means clustering algorithm in other studies [*Jakob and Tselioudis*, 2003; *Rossow et al.*, 2005; *Williams and Webb*, 2009; *Tselioudis et al.*, 2013]. We find that both the joint histograms and the geographical patterns of the relative frequency of occurrence display significant similarities to those previous studies, further enhancing our confidence in the output of the SOM scheme and previous work. Effectively this analysis demonstrates that very different clustering methodologies can produce some very similar states. The identified nodes also display coherent seasonal patterns in the Northern and Southern Hemispheres (see Figure 5).

We also show that the SOM scheme automatically creates an organization of the various nodes. For example, Figure 8 appears to show that nodes 14 and 15 relate to two distinct types of MCC based on comparison with results in *Muhlbauer et al.* [2014]. This organization also groups nodes in such a way that neighboring nodes represent cloud states that the current node is likely to transition to over the period of 1 day. This type of ordering is completed in a subjective manner in other schemes. It is also interesting to note that nearly half the nodes within the SOM show long-term persistence, with the highest transition probability anomaly being related to that node staying in the same state for the 1 day period examined. This may be related to the fineness of the clustering used in this study, our expectation being that larger SOMs would have the same property over shorter time scales. These properties are extremely interesting and could be used as a further test on model representations. The transition frequency of nodes at shorter time scales (between scenes measured at 3-hourly intervals) will be the subject of future work, potentially using a variation of the methodology in *Tan et al.* [2015].

Previous studies have identified that other clustering schemes tend to produce a smaller numbers of states which sometimes means that portions of the data distribution which are rare can be poorly represented [*Michaelides et al.*, 2001]. Other schemes also often lead to a high occurrence state which can be a catch all for a range of distinct states. For example, the WS linked to clear sky conditions identified in *Tselioudis et al.* [2013] could be an example of this preference. In the SOM created in this study, a number of nodes are identified with low total cloud covers (below 50%) and are shown to have distinct distributions in the joint histogram and geographical patterns. They also display subtler differences in their seasonality (node 7 has a small seasonal cycle, while nodes 9 and 13 have clear maxima in Northern Hemisphere winter). In addition, the SW CRE displayed in Figure 11 associated with the clear sky nodes (7–9 and 13) are separated with node 8 having distinctly different properties than nodes 7, 9, and 13. It is also clear that there is a distinct ordering of the magnitude of the subsidence identified in the vertical velocity profiles displayed in Figure 11 connected with each of these nodes.

The fine-scale structure that can be identified with the SOM scheme is best exemplified by focusing on two closely spaced nodes within the 5×3 classification. Nodes 14 and 15 in the derived SOM were shown to have significant similarities to the classifications associated with stratocumulus regions linked to closed and open mesoscale cellular convection in *Muhlbauer et al.* [2014]. These nodes also show distinct similarities with the seasonality of the classifications and differences in the shortwave CRE which are consistent with the previous work. This result supports the view that the SOM technique provides a finer scale classification than is readily possible with *k*-means clustering for this application.

Composite mean vertical velocity profiles linked to the various SOM nodes also show a distinct structure with a shift from profiles dominated by positive (descent) to negative (ascent) vertical velocities as we move from right to left in the SOM structure (see Figure 9). Joint histograms of the density of states linked to midtroposphere vertical velocity and the lower tropospheric stability also display a distinct organization (see Figure 10). In particular, the smallest values of the LTS and most negative vertical velocities (highest rates of ascent) are observed in the left-hand column of the grid with the opposite properties toward the right. The weighted mean values of the TCWV perturbations identified in the titles of Figure 10 show less ordering with the largest values on the left-hand side of the grid, but the smallest values are distributed in the right-hand column and the center of the grid. Though, as might be expected, the most positive TCWV anomalies seem to relate to

regions of horizontal moisture convergence and upward vertical motion, while the largest negative TCWV anomaly suggests that the aridity of the air is important for node 8, the most common clear sky state.

For the sake of balance, we should identify that the SOM classification is not identical to previous *k*-means clustering-based classifications and that regimes with physically meaningful separations based on the work in *Tselioudis et al.* [2013], *Tan and Jakob* [2013], and *Tan et al.* [2015] are missing in the 5 × 3 SOM. However, Figures 12 and 13 show that we can use the sub-SOM methodology to mitigate this issue. There is also the possibility that transition states (4, 6, 10, and 12), states with low relative frequency of occurrence, exist in the SOM analysis. However, we would argue that these may be beneficial in identifying rare states in studies focused on transitions and also identify that they are linked to distinct meteorological and radiative states. In addition, the clustering methodology used in *Mason et al.* [2015] which clusters based on both the satellite data and model output simultaneously could be utilized if these states were considered disadvantageous in particular applications. This study therefore demonstrates that SOMs can identify physically meaningful cloud regimes and that further work using SOMs in this field is warranted.

It should be noted that the interpretation of cloud classifications based on the ISCCP joint histograms alone is potentially fraught with issues. For example, *Marchand et al.* [2010] has discussed the cloud top height versus optical depth histograms derived from the ISCCP, MODIS, and MISR data sets. That work demonstrates that while there are broad similarities among the data sets, there are also large differences. This clearly therefore increases the uncertainty on the analysis from any one data set. However, because the different data sets have different strengths and weaknesses, *Marchand et al.* [2010] find that a combination of data sets can provide more information than any individual data set. An area of future work could therefore be to perform a SOM analysis on the combined data sets simultaneously.

Another clear expansion of this work is to compare joint histograms created from a SOM analysis of both satellite data and output from the COSP simulator [*Bodas-Salcedo et al.*, 2011] in a similar way to that detailed recently in *Mason et al.* [2015] for GCM validation. One possibility which would also expand on this technique is to examine the intercluster and intracluster variations as a function of time in climate data, a technique which has previously been applied to SOM and cluster analysis data in general [*Cassano et al.*, 2007; *Coggins et al.*, 2014; *Coggins and McDonald*, 2015].

Acknowledgments
We would like to thank the numerous people that helped to create the ISCCP data set. The ISCCP D1 data used were obtained from the NASA Langley Research Center Atmospheric Science Data Center (http://eosweb.larc.nasa.gov). ERA-Interim data provided courtesy of ECMWF (http://apps.ecmwf.int). ISCCP FD data were accessed via the ISCCP NASA Goddard Institute for Space Studies page (http://isccp.giss.nasa.gov). The SOM states needed to create all the information are available from the corresponding author. This work was funded as part of two subprograms in the Deep South National Science Challenge. We would like to thank Greg Bodeker, Marwan Katurji, and Fraser Dennison for valuable discussions on early drafts. We would also like to express our gratitude to the three anonymous reviewers who have helped to improve this work.

References

Bankert, R. L., and J. E. Solbrig (2015), Cluster analysis of A-train data: Approximating the vertical cloud structure of oceanic cloud regimes, *J. Appl. Meteorol. Climatol.*, *54*(5), 996–1008, doi:10.1175/jamc-d-14-0227.1.

Bodas-Salcedo, A., et al. (2011), COSP satellite simulation software for model assessment, *Bull. Am. Meteorol. Soc.*, *92*(8), 1023–1043, doi:10.1175/2011bams2856.1.

Bodas-Salcedo, A., K. D. Williams, P. R. Field, and A. P. Lock (2012), The surface downwelling solar radiation surplus over the Southern Ocean in the MET office model: The role of midlatitude cyclone clouds, *J. Clim.*, *25*(21), 7467–7486, doi:10.1175/jcli-d-11-00702.1.

Bony, S., J. L. Dufresne, H. Le Treut, J. J. Morcrette, and C. Senior (2004), On dynamic and thermodynamic components of cloud changes, *Clim. Dyn.*, *22*(2–3), 71–86, doi:10.1007/s00382-003-0369-6.

Cassano, E. N., J. M. Glisan, J. J. Cassano, W. J. Gutowski, and M. W. Seefeldt (2015), Self-organizing map analysis of widespread temperature extremes in Alaska and Canada, *Clim. Res.*, *62*(3), 199–218, doi:10.3354/cr01274.

Cassano, J. J., P. Uotila, and A. Lynch (2006), Changes in synoptic weather patterns in the polar regions in the twentieth and twenty-first centuries, Part 1: Arctic, *Int. J. Climatol.*, *26*(8), 1027–1049, doi:10.1002/joc.1306.

Cassano, J. J., P. Uotila, A. H. Lynch, and E. N. Cassano (2007), Predicted changes in synoptic forcing of net precipitation in large Arctic river basins during the 21st century, *J. Geophys. Res.*, *112*, G04S49, doi:10.1029/2006JG000332.

Cavazos, T. (1999), Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas, *J. Clim.*, *12*(5), 1506–1523, doi:10.1175/1520-0442(1999)012<1506:lscact>2.0.co;2.

Coggins, J. H. J., and A. J. McDonald (2015), The influence of the Amundsen Sea Low on the winds in the Ross Sea and surroundings: Insights from a synoptic climatology, *J. Geophys. Res. Atmos.*, *120*, 2167–2189, doi:10.1002/2014JD022830.

Coggins, J. H. J., A. J. McDonald, and B. Jolly (2014), Synoptic climatology of the Ross Ice Shelf and Ross Sea region of Antarctica: *K*-means clustering and validation, *Int. J. Climatol.*, *34*(7), 2330–2348, doi:10.1002/joc.3842.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*(656), 553–597.

Field, P. R., and R. Wood (2007), Precipitation and cloud structure in midlatitude cyclones, *J. Clim.*, *20*(2), 233–254, doi:10.1175/jcli3998.1.

Gordon, N. D., and J. R. Norris (2010), Cluster analysis of midlatitude oceanic cloud regimes: Mean properties and temperature sensitivity, *Atmos. Chem. Phys.*, *10*(13), 6435–6459, doi:10.5194/acp-10-6435-2010.

Gordon, N. D., J. R. Norris, C. P. Weaver, and S. A. Klein (2005), Cluster analysis of cloud regimes and characteristic dynamics of midlatitude synoptic systems in observations and a model, *J. Geophys. Res.*, *110*, D15S17, doi:10.1029/2004JD005027.

Hewitson, B. C., and R. G. Crane (2002), Self-organizing maps: Applications to synoptic climatology, *Clim. Res.*, *22*(1), 13–26, doi:10.3354/cr022013.

Horton, D. E., N. C. Johnson, D. Singh, D. L. Swain, B. Rajaratnam, and N. S. Diffenbaugh (2015), Contribution of changes in atmospheric circulation patterns to extreme temperature trends, *Nature*, *522*(7557), 465–469, doi:10.1038/nature14550.

Huth, R., C. Beck, A. Philipp, M. Demuzere, Z. Ustrnul, M. Cahynova, J. Kysely, and O. E. Tveito (2008), Classifications of atmospheric circulation patterns recent advances and applications, *Ann. N. Y. Acad. Sci.*, *1146*, 105–152, doi:10.1196/annals.1446.019.

Jakob, C. (2003), An improved strategy for the evaluation of cloud parameterizations in GCMs, *Bull. Am. Meteorol. Soc.*, *84*(10), 1387–1401, doi:10.1175/bams-84-10-1387.

Jakob, C., and G. Tselioudis (2003), Objective identification of cloud regimes in the Tropical Western Pacific, *Geophys. Res. Lett.*, *30*(21), 2082, doi:10.1029/2003GL018367.

Jakob, C., G. Tselioudis, and T. Hume (2005), The radiative, cloud, and thermodynamic properties of the major Tropical Western Pacific cloud regimes, *J. Clim.*, *18*(8), 1203–1215, doi:10.1175/jcli3326.1.

Jolly, B., A. J. McDonald, J. H. J. Coggins, P. Zawar-Reza, J. Cassano, M. Lazzara, G. Graham, G. Plank, O. Petterson, and E. Dale (2016), A validation of the Antarctic mesoscale prediction system using self-organizing maps and high-density observations from SNOWWEB, *Mon. Weather Rev.*, *144*(9), 3181–3200, doi:10.1175/MWR-D-15-0447.1.

Klein, S. A., and D. L. Hartmann (1993), The seasonal cycle of low stratiform clouds, *J. Clim.*, *6*(8), 1587–1606, doi:10.1175/1520-0442(1993)006<1587:tscols>2.0.co;2.

Klein, S. A., Y. Y. Zhang, M. D. Zelinka, R. Pincus, J. Boyle, and P. J. Gleckler (2013), Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator, *J. Geophys. Res. Atmos.*, *118*, 1329–1342, doi:10.1002/jgrd.50141.

Kohonen, T. (1990), The self-organizing map, *Proc. IEEE*, *78*(9), 1464–1480, doi:10.1109/5.58325.

Kohonen, T. (2013), Essentials of the self-organizing map, *Neural Networks*, *37*, 52–65, doi:10.1016/j.neunet.2012.09.018.

Lau, N. C., and M. W. Crane (1995), A satellite view of the synoptic-scale organization of cloud properties in midlatitude and tropical circulation systems, *Mon. Weather Rev.*, *123*(7), 1984–2006, doi:10.1175/1520-0493(1995)123<1984:asvots>2.0.co;2.

Marchand, R., T. Ackerman, M. Smyth, and W. B. Rossow (2010), A review of cloud top height and optical depth histograms from MISR, ISCCP, and MODIS, *J. Geophys. Res.*, *115*, D16206, doi:10.1029/2009JD013422.

Mason, S., J. K. Fletcher, J. M. Haynes, C. Franklin, A. Protat, and C. Jakob (2015), A hybrid cloud regime methodology used to evaluate Southern Ocean cloud and shortwave radiation errors in ACCESS, *J. Clim.*, *28*(15), 6001–6018, doi:10.1175/jcli-d-14-00846.1.

Medeiros, B., and B. Stevens (2011), Revealing differences in GCM representations of low clouds, *Clim. Dyn.*, *36*(1–2), 385–399, doi:10.1007/s00382-009-0694-5.

Michaelides, S., C. S. Pattichis, and G. Kleovoulou (2001), Classification of rainfall variability by using artificial neural networks, *Int. J. Climatol.*, *21*(11), 1401–1414, doi:10.1002/joc.702.

Muhlbauer, A., I. L. McCoy, and R. Wood (2014), Climatology of stratocumulus cloud morphologies: Microphysical properties and radiative effects, *Atmos. Chem. Phys.*, *14*(13), 6695–6716, doi:10.5194/acp-14-6695-2014.

Oreopoulos, L., and W. B. Rossow (2011), The cloud radiative effects of international satellite cloud climatology project weather states, *J. Geophys. Res.*, *116*, D12202, doi:10.1029/2010JD015472.

Oreopoulos, L., N. Cho, D. Lee, S. Kato, and G. J. Huffman (2014), An examination of the nature of global MODIS cloud regimes, *J. Geophys. Res. Atmos.*, *119*, 8362–8383, doi:10.1002/2013JD021409.

Oreopoulos, L., N. Cho, D. Lee, and S. Kato (2016), Radiative effects of global MODIS cloud regimes, *J. Geophys. Res. Atmos.*, *121*, 2299–2317, doi:10.1002/2015JD024502.

Reusch, D. B., R. B. Alley, and B. C. Hewitson (2007), North Atlantic climate variability from a self-organizing map perspective, *J. Geophys. Res.*, *112*, D02104, doi:10.1029/2006JD007460.

Rossow, W. B., and R. A. Schiffer (1991), ISCCP cloud data products, *Bull. Am. Meteorol. Soc.*, *72*(1), 2–20, doi:10.1175/1520-0477.

Rossow, W. B., and R. A. Schiffer (1999), Advances in understanding clouds from ISCCP, *Bull. Am. Meteorol. Soc.*, *80*(11), 2261–2287, doi:10.1175/1520-0477(1999)080<2261:aiucfi>2.0.co;2.

Rossow, W. B., G. Tselioudis, A. Polak, and C. Jakob (2005), Tropical climate described as a distribution of weather states indicated by distinct mesoscale cloud property mixtures, *Geophys. Res. Lett.*, *32*, L21812, doi:10.1029/2005GL024584.

Rossow, W. B., Y. C. Zhang, and G. Tselioudis (2016), Atmospheric diabatic heating in different weather states and the general circulation, *J. Clim.*, *29*(3), 1059–1065, doi:10.1175/jcli-d-15-0760.1.

Sammon, J. W. (1969), A nonlinear mapping for data structure analysis, *IEEE Trans. Comput.*, *18*(5), 401–409.

Sassen, K., and Z. Wang (2008), Classifying clouds around the globe with the CloudSat radar: 1-year of results, *Geophys. Res. Lett.*, *35*, L04805, doi:10.1029/2007GL032591.

Sheridan, S. C., and C. C. Lee (2011), The self-organizing map in synoptic climatological research, *Prog. Phys. Geog.*, *35*(1), 109–119, doi:10.1177/0309133310397582.

Stubenrauch, C. J., et al. (2013), Assessment of global cloud datasets from satellites: Project and database initiated by the GEWEX radiation panel, *Bull. Am. Meteorol. Soc.*, *94*(7), 1031–1049, doi:10.1175/bams-d-12-00117.1.

Su, H., J. H. Jiang, D. G. Vane, and G. L. Stephens (2008), Observed vertical structure of tropical oceanic clouds sorted in large-scale regimes, *Geophys. Res. Lett.*, *35*, L24704, doi:10.1029/2008GL035888.

Tan, J., and C. Jakob (2013), A three-hourly data set of the state of tropical convection based on cloud regimes, *Geophys. Res. Lett.*, *40*, 1415–1419, doi:10.1002/grl.50294.

Tan, J., C. Jakob, and T. P. Lane (2013), On the identification of the large-scale properties of tropical convection using cloud regimes, *J. Clim.*, *26*(17), 6618–6632, doi:10.1175/jcli-d-12-00624.1.

Tan, J., C. Jakob, W. B. Rossow, and G. Tselioudis (2015), Increases in tropical rainfall driven by changes in frequency of organized deep convection, *Nature*, *519*(7544), 451–454, doi:10.1038/nature14339.

Tselioudis, G., Y. C. Zhang, and W. B. Rossow (2000), Cloud and radiation variations associated with northern midlatitude low and high sea level pressure regimes, *J. Clim.*, *13*(2), 312–327, doi:10.1175/1520-0442(2000)013<0312:carvaw>2.0.co;2.

Tselioudis, G., W. Rossow, Y. C. Zhang, and D. Konsta (2013), Global weather states and their properties from passive and active satellite cloud retrievals, *J. Clim.*, *26*(19), 7734–7746, doi:10.1175/jcli-d-13-00024.1.

Webb, M., C. Senior, S. Bony, and J. J. Morcrette (2001), Combining ERBE and ISCCP data to assess clouds in the Hadley Centre, ECMWF and LMD atmospheric climate models, *Clim. Dyn.*, *17*(12), 905–922, doi:10.1007/s003820100157.

Williams, K. D., and M. J. Webb (2009), A quantitative performance assessment of cloud regimes in climate models, *Clim. Dyn.*, *33*(1), 141–157, doi:10.1007/s00382-008-0443-1.

Wood, R., and D. L. Hartmann (2006), Spatial variability of liquid water path in marine low cloud: The importance of mesoscale cellular convection, *J. Clim.*, *19*(9), 1748–1764, doi:10.1175/jcli3702.1.

Zelinka, M. D., S. A. Klein, and D. L. Hartmann (2012), Computing and partitioning cloud feedbacks using cloud property histograms. Part I: Cloud radiative kernels, *J. Clim.*, *25*(11), 3715–3735, doi:10.1175/jcli-d-11-00248.1.

Zhang, Y. C., W. B. Rossow, A. A. Lacis, V. Oinas, and M. I. Mishchenko (2004), Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: Refinements of the radiative transfer model and the input data, *J. Geophys. Res.*, *109*, D19105, doi:10.1029/2003JD004457.

Zhang, Y. Y., S. Klein, G. G. Mace, and J. Boyle (2007), Cluster analysis of tropical clouds using CloudSat data, *Geophys. Res. Lett.*, *34*, L12813, doi:10.1029/2007GL029336.