RMetS

Royal Meteorological Society

Representation of 3D heterogeneous cloud fields using copulas: Theory for water clouds

Peter M. Norris,[a,b]* Lazaros Oreopoulos,[c,d] Arthur Y. Hou,[b†]
Wei-Kuo Tao[d†] and Xiping Zeng[a,d]

[a]*Goddard Earth Science and Technology Center, University of Maryland, Baltimore County, Baltimore, USA*
[b]*Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA*
[c]*Joint Center for Earth Systems Technology, University of Maryland, Baltimore County, Baltimore, USA*
[d]*Laboratory for Atmospheres, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA*

ABSTRACT: It is shown that a general representation of GCM column cloud fraction within probability density function (PDF)-based statistical cloud parametrizations can be obtained using statistical functions called copulas that encapsulate the dependence structure of rank statistics in a multivariate system. Using this theory, a new Gaussian copula formulation of GCM cloud overlap is obtained. The copula approach provides complete flexibility in the choice of the marginal PDF of each layer's moisture and temperature, and, compared with earlier approaches, including the 'generalized overlap' approach, allows a far more general specification of the correlation between any pair of layers. It also allows easy addition of new layer variables, such as temperature, into the modelled grid-column statistics. As a preliminary test of this formulation, its ability to statistically describe a cloud-resolving model simulation of a complex multi-layer case-study, including both large-scale and convective clouds, is examined. The Gaussian copula cloud fraction is found to be significantly less biased than other common cloud overlap methods for this case-study. Estimates of several nonlinear quantities are also improved with the Gaussian copula model: the variance of condensed water path and the fluxes of solar and thermal radiation at atmospheric column boundaries. This first paper, though limited to the simpler case of water clouds, addresses subgrid-scale variability in both moisture and temperature. This work is envisaged as a first step towards developing a generalized statistical framework for GCM cloud parametrization and for assimilating statistical information from high-resolution satellite observations into GCMs and global analyses. Copyright © 2008 Royal Meteorological Society

KEY WORDS   statistical cloud parametrizations; cloud overlap; cloud heterogeneity; radiative transfer; rank statistics

Received 13 September 2007; Revised 20 June 2008; Accepted 15 August 2008

1. Introduction

A GCM cloud parametrization typically predicts the cloud fraction and mean cloud properties in each model layer containing cloud. To calculate the radiative impact of these clouds in a GCM column requires the specification of the overlap of the different cloud layers. Early schemes assumed random overlap between model cloud layers, but this suffers from an unphysical dependence on model vertical resolution, and even randomly overlaps adjacent cloudy model layers within the same physical cloud. An improvement is the maximum-random overlap assumption (Tian and Curry, 1989), which assumes maximal overlap between adjacent cloudy model layers. (The more commonly used variant of the maximum-random overlap scheme due to Geleyn and Hollingsworth (1979) has a somewhat more complicated interpretation.

It is discussed in Appendix D.) However, the scheme still retains some resolution dependence because it relaxes to random overlap between non-contiguous cloud layers, regardless of the thickness of the intervening clear layer. Furthermore, it is reasonable to expect some loss of correlation with height within deeper clouds due to vertical wind shear, for example.

Recently, based on radar observations (Hogan and Illingworth, 2000; Mace and Benson-Troth, 2002) and cloud-resolving model studies (Oreopoulos and Khairout-dinov, 2003; Räisänen *et al.*, 2004; Pincus *et al.*, 2005), a more physical approach has been proposed, called 'generalized overlap', in which the combined cloud fraction of any two layers can assume any value between those defined by the extreme assumptions of perfect maximum and random overlap. This is achieved through the introduction of a correlation parameter that can take values between zero for random overlap and one for maximum overlap and whose value drops exponentially as a function of the separation distance between the cloudy layers.

This concept can be also extended to the rank correlation of cloud condensate amount between layers. This governs the likelihood that a large water content in one

---

Correspondence to: Dr. Peter Norris, Global Modeling and Assimilation Office, NASA/GSFC, Code 610.1, Greenbelt, MD 20771, USA. E-mail: peter.m.norris@nasa.gov

layer of a GCM column (relative to that layer's water content range) will be paired with a relatively large water content in another layer, and likewise for relatively small water contents. Räisänen *et al.* (2004) and Pincus *et al.* (2005) find that this rank correlation also decreases exponentially with layer separation. The existence of such condensate amount correlations between cloudy layers affects both radiative transfer and precipitation/re-evaporation processes (Jakob and Klein, 1999).

This paper is concerned with formalizing and generalizing these concepts of 'generalized overlap' and 'condensate rank correlation' within the context of probability density function (PDF)-based statistical cloud parametrizations (e.g. Smith, 1990; Xu and Randall, 1996; Larson *et al.*, 2001; Tompkins, 2002), which represent the unresolved variability within a model grid box statistically using PDFs. In doing so, we hope to provide a framework for representing realistic three-dimensional subgrid-scale variability in GCM grid columns, leading to improved cloud and radiation parametrizations, and as a first step towards transferring the statistical information content of high-resolution satellite cloud observations into GCMs and global analyses.

Very recently, Larson (2007) has presented an even more general approach for treating PDF overlap, by modelling the joint PDF of p variables (e.g. temperature and moisture) and n model layers in a high-dimensional $p \times n$ phase space. Larson presents several possible models for this PDF and settles on a mixture of a small number of pn-dimensional Gaussians, with the added condition that each marginal distribution (i.e. the univariate PDF of a given variable and layer) is the weighted sum of at most two normal distributions. Our approach, which Larson also briefly discusses, is the use of statistical functions called *copulas*. These are essentially joint cumulative distribution functions (CDFs) of the *ranks* of the pn layer-variables within their respective marginal distributions. Compared with Larson's approach, the copula approach provides complete flexibility in the choice of the marginal PDF of each layer's moisture and temperature, but is potentially less easy to use in cases where estimates of the linear correlations between variables are provided rather than their rank correlations. Compared with earlier approaches, including the 'generalized overlap' approach of Räisänen *et al.* (2004) and Pincus *et al.* (2005), the copula approach allows a far more general specification of the correlation between any pair of layers, and also allows for easy addition of new layer variables, such as temperature, into the modelled grid-column statistics.

The paper is organized as follows. Section 2 introduces some key notation and discusses our assumptions. Section 3 introduces the connection between grid-column cloud fraction and copulas of the moisture saturation ratio of each model layer. Section 4 generalizes our use of copulas to include subgrid-scale temperature variability as well, which allows us to move beyond cloud fraction to more complicated grid-column quantities, such as radiative fluxes. This section introduces the Gaussian copula in particular, and presents a simple method for inferring an unbiased estimate of the Gaussian copula from

grid-column temperature and moisture samples. Section 5 discusses the generation of random subcolumns from our copula-based framework, which can be used for Monte-Carlo evaluation of grid-column average properties. Subcolumn generators are provided for the Gaussian copula, as well as for existing overlap schemes: random and maximum overlap, Geleyn and Hollingsworth (1979) maximum–random overlap, and for a version of generalized overlap due to Räisänen *et al.* (2004). Section 6 provides a detailed analysis of the application of our copula-based framework to a single CRM-generated frontally disturbed test case over the ARM (Atmospheric Radiation Measurement) Southern Great Plains site and highlights many practical issues in implementing the new method. Gaussian copula estimates of grid-column cloud fraction, liquid water path variance and radiative fluxes are compared with other overlap methods. Section 7 summarizes our conclusions and provides a suggested roadmap for practical application of the copula-based framework to GCM cloud-radiation parameterization and high-resolution satellite cloud data assimilation. It concludes with a discussion of the copula-based method in comparison to the recent Larson (2007) 'peg-hat' approach and a recommendation on how elements of both methods can be combined.

2. Preliminaries

2.1. Notation

Consider an air parcel of volume V containing masses m_d, m_v and m_c of dry air, water vapour, and water condensate. The corresponding 'densities' are $\rho_{d,v,c} \equiv m_{d,v,c}/V$. The 'total water density' is $\rho_t \equiv \rho_v + \rho_c$ and the 'parcel density' is $\rho \equiv \rho_d + \rho_t$. The partial pressures of dry air and water vapour are $p_d = \rho_d R_d T$ and $e_v = \rho_v R_v T$, where R_d and R_v are the gas constants of dry air and water vapour and T is the parcel temperature. The parcel pressure is $p = p_d + e_v = \rho_* R_d T$, where $\rho_* \equiv \rho_d + \rho_v/\varepsilon$ will be called the 'virtual density' and $\varepsilon \equiv R_d/R_v \approx 0.622$.

In this paper we will use 'moisture contents'

$$q_{v,c,t} = \rho_{v,c,t}/\rho_*, \tag{1}$$

since the saturation vapour content $q_s \equiv \rho_s/\rho_* = \varepsilon e_s(T)/p$, where $e_s(T)$ is the saturation vapour pressure over a plane pure liquid water surface, has a particularly simple but exact form in this normalization.

Next, define the total saturation ratio (or simply 'saturation ratio') as

$$S \equiv q_t/q_s, \tag{2}$$

i.e. the total moisture content scaled by the saturation vapour content. This is similar to the 'relative humidity', expressed as a fraction, but uses the *total* water content, not the vapour content in the numerator. We will use S instead of q_t in our analysis. The reasons for this choice will become evident in later sections.

Finally, note that we will use boldface to represent vectors, e.g. **X**, and the sans serif font to represent matrices, e.g. X.

2.2. Assumptions

This work has an intentionally narrow focus defined by the following assumptions, most of which are quite common in GCM cloud parametrizations:

(A1) Consider a single GCM grid column, or a contiguous vertical section thereof, comprised of K layers, numbered $k = 1 \dots K$, bottom to top. Assume each layer to be sufficiently shallow that *all* its properties can be taken as *vertically* uniform. Clearly this is unphysical for pressure, but assume that no significant error is made by using the layer mid-point pressure in all calculations that follow.

(A2) Assume the layer pressure p is also horizontally uniform, whereas the layer temperature T and saturation ratio S are horizontally non-uniform and distributed according to cumulative distribution functions F_{T_k} and F_{S_k} for layer k. Specifically, $F_{T_k}(t)$ and $F_{S_k}(s)$ are the fractions of the layer volume for which $T_k \leq t$ and $S_k \leq s$. By assumption (A1), these volume fractions are identical to the respective areal fractions of the layer, as seen from directly above (i.e. from zenith). We therefore employ the shorthand $F_{T_k}(t) = \mathrm{Fr}(T_k \leq t)$ and $F_{S_k}(s) = \mathrm{Fr}(S_k \leq s)$, where $\mathrm{Fr}(A)$ is the areal fraction of the grid column, as viewed from zenith, for which condition A holds. Note that F_{T_k} is non-decreasing and has $F_{T_k}(0) = 0$ and $F_{T_k}(\infty) = 1$, and similarly for F_{S_k}. We further assume that both F_{T_k} and F_{S_k} are continuous.

(A3) We restrict our attention to contiguous vertical subsections of a grid column that contain only liquid phase condensate and for which each contained layer has $T \geq 0\,^{\circ}\mathrm{C}$ everywhere. We make this assumption in order to simplify development of our basic theory, as discussed in (A4) below. We will include mixed phase and ice clouds in a future paper.

(A4) Assume that q_{v} nowhere exceeds its saturation value $q_{\mathrm{s}} = \varepsilon e_{\mathrm{s}}(T)/p$, and that any excess water is present as condensed liquid water

$$q_{\mathrm{c}} = (q_{\mathrm{t}} - q_{\mathrm{s}})H(q_{\mathrm{t}} - q_{\mathrm{s}}) = (S - 1)\,q_{\mathrm{s}}H(S - 1),$$

where H is the Heaviside step function. ($H(x)$ is 0 for $x < 0$, 1 for $x > 0$ and $1/2$ for $x = 0$.) This so-called 'bulk condensate' assumption is often used in GCMs, since in most warm clouds vapour supersaturations are very small. (In reality, cloud condensate is composed of hydrometeors which may or may not be instantaneously in equilibrium with the surrounding vapour field. Nevertheless, at least for a typical warm cloud condition, our calculations with a microphysical droplet growth

model indicate such equilibrium typically occurs within a minute, much shorter than the typical time step for GCM phase change routines. The situation for ice-phase clouds is definitely more complicated, and will be addressed in a future paper. For the liquid water clouds currently under consideration, assume that the liquid phase is in equilibrium with the vapour field, and that curvature and solute effects, which lead to departures of the saturation vapour pressure from that over a plane pure liquid surface, are not significant in the bulk, hydrometeor-size-integrated sense.) This partially explains why we have limited this paper to liquid water clouds – in ice clouds, supersaturations in excess of 10% or 20% are often seen.

One consequence of (A4) is that S is a direct indicator of cloudiness, without regard to T, with $S > 1$ for cloudy air and $S \leq 1$ for clear air. This fact will allow us to represent the grid-column cloud fraction in terms of the vertical correlations of S alone.

3. Cloud fraction and copulas

3.1. Definition of clear/cloud fraction

Define the *clear fraction* f_k' of layer k as the areal fraction of the layer, when viewed from zenith, which has zero condensate at any level within the layer. This is technically an *areal* clear fraction, but under the shallow-layer assumption (A1) it is identical to the so-called *volumetric* clear fraction, which is the fractional volume of the layer with zero condensate. We will therefore drop the prefix 'areal' or 'volumetric' and speak only of the clear fraction f_k'. Using assumptions (A1)–(A4), we may therefore write:

$$f_k' = \mathrm{Fr}(q_{\mathrm{t}k} \leq q_{\mathrm{s}k}) = \mathrm{Fr}(S_k \leq 1) = F_{S_k}(1). \quad (3)$$

The corresponding *cloud fraction* is just $f_k \equiv 1 - f_k'$.

The *column clear fraction* f' is likewise defined as the areal fraction of the entire grid column, as viewed from zenith, which has zero condensate at any level. The determination of f' is a major subject of this paper. The *column cloud fraction* f is likewise defined by $f \equiv 1 - f'$.

Apart from being mathematically interesting quantity, f' has been used extensively in the past in simple plane-parallel radiative transfer schemes as a proxy for the portion of a GCM grid column unaffected by cloud. Also, from a satellite data assimilation perspective, f' is related (albeit imperfectly) to the fraction of pixels in a region that are not 'masked' as cloudy, a common product of many satellite cloud retrievals. In the real world there are many complicating issues: the existence of sideways photon transport in broken cloud fields, the occurrence of non-zenith solar and satellite viewing angles, etc. Nevertheless, for a start, we wish to explore the determination of f' under the assumptions discussed above and for the simple reasons just outlined. It will turn

out, as we proceed, that our analysis can also be used to model complex inter-variable and inter-layer correlations within a grid column, and will therefore have far wider application to GCM parametrization and data assimilation than our initial focus on column cloud fraction suggests.

3.2. Column clear fraction and copulas

Let $F_{\mathbf{S}}$, where $\mathbf{S} \equiv (S_1, \ldots, S_K)^{\mathrm{T}}$, be the joint cumulative distribution function of saturation ratio for all layers in the grid column. Specifically,

$$F_{\mathbf{S}}(s) = \mathrm{Fr}(S_1 \leq s_1, \ldots, S_K \leq s_K),$$

where $\mathrm{Fr}(A, B)$ means $\mathrm{Fr}(A \cap B)$. It therefore follows from the definition of f' that

$$f' = F_{\mathbf{S}}(1, \ldots, 1). \tag{4}$$

Define the 'kth margin' of $F_{\mathbf{S}}(s)$ as the one-dimensional function obtained from it by setting $s_i = \infty \ \forall \ i \neq k$. From the definition of $F_{\mathbf{S}}$ we see that its kth margin is none other than the cumulative distribution function of layer k alone, i.e. F_{S_k} (A2 above).

Then, by Sklar's Theorem (Nelsen, 2006, Theorem 2.10.9), $\exists$ a unique 'K-copula' $C_{\mathbf{S}}$ (to be defined shortly) such that $\forall \ s \in [0, \infty]^K$,

$$F_{\mathbf{S}}(s) = C_{\mathbf{S}}(F_{S_1}(s_1), \ldots, F_{S_K}(s_K)). \tag{5}$$

Then, by (4) and (3), we may write

$$f' = C_{\mathbf{S}}(F_{S_1}(1), \ldots, F_{S_K}(1)) = C_{\mathbf{S}}(f'_1, \ldots, f'_K). \tag{6}$$

Hence, we see that this K-copula $C_{\mathbf{S}}$ is the special function which relates the set of layer clear fractions $\{f'_k\}$ to the grid-column clear fraction f'. At this point, $C_{\mathbf{S}}$ provides a completely general expression of this relationship, given the assumptions of section 2.2. In section 5 we will derive the particular copula functions for many well known overlap assumptions: random overlap, maximum overlap, the maximum-random overlap method due to Geleyn and Hollingsworth (1979), and a total water version of the Räisänen *et al.* (2004) 'generalized overlap' method. As general as the latter method is, it is still restricted to non-negative correlations between layers and has a very specific form for the correlation between layers that is built from correlations between adjacent layers only (section 5.6). One of the advantages of the copula approach is that it allows for a very general specification of the correlation between arbitrary pairs of layers. For example, in section 4.3 we will derive the so-called 'Gaussian copula', which is able to remove the above restrictions of the total water Räisänen *et al.* (2004) method.

What exactly is a copula? From (5), it is a function which joins a joint CDF to its one-dimensional margins. It describes the scale-free dependence among a set of variables in the sense that it is a joint CDF of the *ranks* of the variables within their respective distributions.

The theory of copulas is discussed extensively in a monograph by Nelsen (2006). Formal details aside, a K-copula can be thought of as a cumulative distribution function on $\mathbf{I}^K \equiv [0, 1]^K$ with uniform margins, meaning that $\forall \ r = (r_1, \ldots, r_K)^{\mathrm{T}} \in \mathbf{I}^K$, $C(r) = 0$ if at least one coordinate of r is 0, and $C(r) = r_k$ if all coordinates of r are 1 except r_k.

We will use the following results from copula theory:

(R1) Two simple copulas with special significance in copula theory are the 'product copula' $\prod^K(r) = \prod_{k=1}^K r_k$ and the 'M copula' $M^K(r) = \min(r_1, \ldots, r_K)$.

(R2) The $\binom{K}{k}$ 'k-margins' of a K-copula C, formed by setting $K - k$ of the arguments of C to one, are themselves k-copulas.

(R3) A copula is non-decreasing in each of its arguments.

3.3. Simple cases

Two common special cases of (6) are:

(a) If all layers are completely independent we will have $f' = \prod_{k=1}^K f'_k$, which corresponds to $C_{\mathbf{S}} = \prod^K$, the 'product copula'. This leads to the so-called 'random overlap' cloud fraction

$$f_{\mathrm{RAN}} = 1 - \prod_{k=1}^K f'_k. \tag{7}$$

Random overlap is unrealistic when applied to closely separated model layers within a contiguous cloud layer.

(b) If $C_{\mathbf{S}} = M^K$, the 'M-copula', then we obtain $f' = \min(f'_1, \ldots, f'_K)$ and the so-called 'maximum overlap' cloud fraction

$$\begin{aligned} f_{\mathrm{MAX}} &= 1 - \min(f'_1, \ldots, f'_K) \\ &= \max(f_1, \ldots, f_K). \end{aligned} \tag{8}$$

Though more realistic than random overlap when applied within a cloud, we do expect some transition towards random overlap between distant model layers within a deep cloud.

Appendix C gives an example of the combination of these different cloud fraction copulas for the case of multiple cloud layers.

4. Beyond cloud fraction

4.1. The joint distribution of $\mathbf{T}$ and $\mathbf{S}$

So far we have a way of evaluating the layer and column cloud fractions using (3) and (6) via the statistical properties of $\mathbf{S}$. But we also need a way to characterize the *joint* distributions of $\mathbf{T}$ and $\mathbf{S}$ so that we may estimate more complex quantities than cloud fraction, such as the radiative or precipitation forming properties of a collection of cloud layers, which depend on a

knowledge of $\mathbf{T}$, $\mathbf{q}_v$ and $\mathbf{q}_c$, among others. Let $F_{\mathbf{T},\mathbf{S}}$ be the joint distribution function of $\mathbf{T}$ and $\mathbf{S}$. Then by Sklar's Theorem $\exists$ a unique $2K$-copula $C_{\mathbf{T},\mathbf{S}}$ such that $\forall \, \boldsymbol{t}, \boldsymbol{s} \in [0, \infty]^K$,

$$
\begin{aligned}
F_{\mathbf{T},\mathbf{S}}(\boldsymbol{t}, \boldsymbol{s}) = C_{\mathbf{T},\mathbf{S}}(F_{T_1}(t_1), \ldots, F_{T_K}(t_K), \\
F_{S_1}(s_1), \ldots, F_{S_K}(s_K)).
\end{aligned} \quad (9)
$$

Note that the saturation ratio K-copula $C_{\mathbf{S}}$ of (5) is just the K-margin of $C_{\mathbf{T},\mathbf{S}}$ that marginalizes the temperatures. Namely,

$$
C_{\mathbf{S}}(r_1, \ldots, r_K) = C_{\mathbf{T},\mathbf{S}}(1, \ldots, 1, r_1, \ldots, r_K). \quad (10)
$$

4.2. Multivariate copulas

There are numerous bivariate copulas discussed in Nelsen (2006). However, in order to model the cloud overlap properties of multiple layers, we need to find suitable copulas of higher order. Unfortunately, it is usually not possible to extend a bivariate copula to higher dimensions while retaining the flexibility it has in two dimensions.

There are several different approaches to building copulas of order three and above. One strategy (Chakak and Koehler, 1995) starts with bivariate marginal copula and uses conditional probability arguments to iteratively construct higher order margins that include those of lower order but also capture the joint rank dependence of progressively larger numbers of layers. We have not found this method easy to use and will not discuss it further.

The approach we will take is to investigate transformations of $(\mathbf{T}, \mathbf{S})$ space that yield known multivariate distribution functions. These analytic $F_{\mathbf{T},\mathbf{S}}$ then translate to analytic copulas via Sklar's Theorem (9).

4.3. Transformed multinormal distributions and the Gaussian copula

Suppose there exist *monotonic increasing* transformations

$$
\left.\begin{aligned}
Z_{Tk} = G_{Tk}(T_k) \\
Z_{Sk} = G_{Sk}(S_k)
\end{aligned}\right\}, \quad k \in \{1, 2, \ldots, K\}, \quad (11)
$$

such that $\mathbf{Z} \equiv (Z_{T1}, \ldots, Z_{TK}, Z_{S1}, \ldots, Z_{SK})^{\mathrm{T}} \in \mathbb{R}^{2K}$ has a multivariate normal (i.e. multinormal) distribution with zero mean ($\mathrm{E}(\mathbf{Z}) = \mathbf{0}$) and covariance $\mathbf{C} = \mathrm{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}})$, or using common statistical shorthand, $\mathbf{Z} \sim N_{2K}(\mathbf{0}, \mathbf{C})$, and that $\mathbf{C}$ has unit variances (ones on its diagonal) and is therefore a correlation matrix.

(For the existence of such transformations we actually require only that there exist monotonic increasing transformations such that $\mathbf{Z}$ has an arbitrary multinormal distribution, i.e. $\mathbf{Z} \sim N_{2K}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, except only that $\boldsymbol{\Sigma}$ has all non-zero variances $\sigma_k^2 \equiv \Sigma_{kk} > 0$, since then the further monotonic increasing transformations $(Z_k - \mu_k)/\sigma_k \mapsto Z_k$, yield $\mathbf{Z} \sim N_{2K}(\mathbf{0}, \mathbf{C})$, where $\mathbf{C}$ is the correlation matrix $\mathbf{C}_{ij} \equiv \Sigma_{ij}/(\sigma_i \sigma_j)$.)

The implication of these transformations is that the density of areal fraction in $\mathbf{Z}$ space is

$$
p_{\mathbf{Z}}(\mathbf{z}; \mathbf{C}) = (2\pi)^{-K}|\mathbf{C}|^{-1/2} \exp(-\mathbf{z}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{z}/2). \quad (12)
$$

Using the monotonic increasing property of the G_k,

$$
\begin{aligned}
F_{\mathbf{T},\mathbf{S}}(\boldsymbol{t}, \boldsymbol{s}) & \\
= \mathrm{Fr}\,(T_1 \leq t_1, & \ldots, T_K \leq t_K, \\
& S_1 \leq s_1, \ldots, S_K \leq s_K) \\
= \mathrm{Fr}\,(Z_{T1} \leq G_{T1}(t_1), & \ldots, Z_{TK} \leq G_{TK}(t_K), \\
& Z_{S1} \leq G_{S1}(s_1), \ldots, Z_{SK} \leq G_{SK}(s_K)) \\
= F_{\mathbf{Z}}\big(G_{T1}(t_1), & \ldots, G_{TK}(t_K), \\
& G_{S1}(s_1), \ldots, G_{SK}(s_K); \mathbf{C}\big),
\end{aligned} \quad (13)
$$

where

$$
F_{\mathbf{Z}}(z_1, \ldots, z_{2K}; \mathbf{C}) = \int_{-\infty}^{z_1} \ldots \int_{-\infty}^{z_{2K}} p_{\mathbf{Z}}(\mathbf{z}'; \mathbf{C}) \, \mathrm{d}\mathbf{z}' \quad (14)
$$

is the multinormal cumulative distribution function (discussed further in section 5).

It is a property of the multinormal $\mathbf{Z} \sim N_{2K}(\mathbf{0}, \mathbf{C})$ that each of the $2K$ margins is normally distributed as $Z_k \sim N(0, \mathbf{C}_{kk}) = N(0, 1)$, i.e. each Z_k is a standard normal variate. Then

$$
\begin{aligned}
F_{Z_{Tk}}(z) = F_{Z_{Sk}}(z) &= \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right] \\
&\equiv \Phi(z),
\end{aligned} \quad (15)
$$

which is the standard normal cumulative distribution function, expressed in terms of the error function

$$
\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2} \mathrm{d}t.
$$

Its inverse is

$$
\Phi^{-1}(p) = \sqrt{2}\,\mathrm{erf}^{-1}(2p - 1). \quad (16)
$$

Since G_{Tk} is monotonic increasing,

$$
\begin{aligned}
F_{T_k}(t) = \mathrm{Fr}(T_k \leq t) &= \mathrm{Fr}(Z_{Tk} \leq G_{Tk}(t)) \\
&= F_{Z_{Tk}}(G_{Tk}(t)) = \Phi(G_{Tk}(t)),
\end{aligned}
$$

and similarly for G_{Sk}, so that

$$
\begin{aligned}
G_{Tk}(t) &= \Phi^{-1}(F_{T_k}(t)) \\
\text{and} \quad G_{Sk}(s) &= \Phi^{-1}(F_{S_k}(s)).
\end{aligned} \quad (17)
$$

Combining these with (13), we have

$$
\begin{aligned}
F_{\mathbf{T},\mathbf{S}}(\boldsymbol{t}, \boldsymbol{s}) = C_{\mathbf{Z}}\big(F_{T_1}(t_1), & \ldots, F_{T_K}(t_K), \\
& F_{S_1}(s_1), \ldots, F_{S_K}(s_K); \mathbf{C}\big),
\end{aligned} \quad (18)
$$

as in (9), with

$$
\begin{aligned}
C_{\mathbf{Z}}(r_1, \ldots, r_{2K}; \mathbf{C}) = & \\
& F_{\mathbf{Z}}\big(\Phi^{-1}(r_1), \ldots, \Phi^{-1}(r_{2K}); \mathbf{C}\big),
\end{aligned} \quad (19)
$$

$\forall r_k \in [0, 1]$. This $C_{\mathbf{Z}}$ is the so-called 'Gaussian copula' (e.g. Cherubini *et al.*, 2004) and can be used in place of $C_{\mathbf{T},\mathbf{S}}$. This Gaussian copula is simply the description in rank space of what we have described in $\mathbf{Z}$ space using the multinormal $N_{2K}(\mathbf{0}, \mathbf{C})$. Once the $K(2K - 1)$ unique elements of the correlation matrix $\mathbf{C}$ are specified, we can evaluate the column clear fraction f' using (10) and (6). Or if we know the form of the $2K$ single layer margins F_{T_k} and F_{S_k}, we can evaluate the full joint distribution $F_{\mathbf{T},\mathbf{S}}$ using (18).

The above theory assumes that the $(\mathbf{T}, \mathbf{S})$ field can be represented as a transformed multinormal $\mathbf{Z} \sim N_{2K}(\mathbf{0}, \mathbf{C})$. In practice, this will never be exactly true. It *is* exactly true that each $Z_k \sim N(0, 1)$, since F_{T_k} and F_{S_k} are by definition uniformly distributed and so the Z_{Tk} and Z_{Sk} follow standard normal distributions per (11) and (17). However, a distribution that has all normal margins is not necessarily multinormal, even though the converse is true. In practice, we can adjust $\mathbf{C}$ (and the parameters of the F_{T_k} and F_{S_k} distributions if they are modelled) to give the best fit to observed cloud fraction or $(\mathbf{T}, \mathbf{S})$ data. We can then judge the quality of the best fit model, either by looking at the quality of the model predictions, or, more directly, by examining the fit in $\mathbf{Z}$ or $(\mathbf{T}, \mathbf{S})$ space.

4.4. Inference of Gaussian copula parameters

Say we are supplied with N random samples $\{\mathbf{T}^{(1)}, \dots, \mathbf{T}^{(N)}\}$ and $\{\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(N)}\}$ from a grid column and wish to model the scale-free dependence between layers using the Gaussian copula $C_{\mathbf{Z}}(r_1, \dots, r_{2K}; \mathbf{C})$. We will assume that the marginal distributions F_{T_k} and F_{S_k} are known exactly or have been estimated in a prior step. Then the sample ranks are given by

$$R_k^{(n)} = F_{T_k}(T_k^{(n)}), \quad R_{k+K}^{(n)} = F_{S_k}(S_k^{(n)}), \quad (20)$$

for $k = 1, \dots, K$. From these ranks, we seek an unbiased estimate of the copula parameter, call it $\widehat{\mathbf{C}}$, such that subsequent generation of samples from the copula (section 5) will be as consistent with the underlying population as possible.

Since $\mathbf{C}$ is the correlation matrix of $\mathbf{Z} \equiv (\Phi^{-1}(R_1), \dots, \Phi^{-1}(R_{2K}))^{\mathrm{T}} \sim N_{2K}(\mathbf{0}, \mathbf{C})$, the 'sample correlation matrix' (Snedecor and Cochran, 1980) of the $\mathbf{Z}^{(n)}$ provides the required unbiased estimate:

$$\widehat{\mathbf{C}}_{ij} \equiv \Sigma_{ij}^* / \sqrt{\Sigma_{ii}^* \Sigma_{jj}^*}, \quad (21)$$

where $\quad \Sigma_{ij}^* \equiv \dfrac{1}{N-1} \displaystyle\sum_{n=1}^{N} (Z_i^{(n)} - \overline{Z}_i)(Z_j^{(n)} - \overline{Z}_j)$

and $\quad \overline{Z}_k \equiv \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} Z_k^{(n)}$.

4.5. Estimation of grid-column averages

Let ψ be some scalar grid-column quantity that can be expressed in terms of grid-column temperature $\mathbf{T}$ and saturation ratio $\mathbf{S}$ and which therefore has a horizontal distribution within the grid column. We define the grid-column average (expectation value) of ψ as

$$\langle \psi \rangle \equiv \int_0^\infty \cdots \int_0^\infty \psi(\mathbf{t}, \mathbf{s})\, p_{\mathbf{T},\mathbf{S}}(\mathbf{t}, \mathbf{s})\, \mathrm{d}\mathbf{t}\, \mathrm{d}\mathbf{s}, \quad (22)$$

where $p_{\mathbf{T},\mathbf{S}}$ is the density of areal fraction within $(\mathbf{T}, \mathbf{S})$ phase space, defined such that

$$F_{\mathbf{T},\mathbf{S}}(\mathbf{t}, \mathbf{s}) = \int_0^{s_1} \cdots \int_0^{s_K} \int_0^{t_1} \cdots \int_0^{t_K} p_{\mathbf{T},\mathbf{S}}(\mathbf{t}, \mathbf{s})\, \mathrm{d}\mathbf{t}\, \mathrm{d}\mathbf{s},$$

for all $\mathbf{t}, \mathbf{s} \in [0, \infty]^K$. Note that this average is a linear operator, namely $\langle a\psi \rangle = a\langle \psi \rangle$, and $\langle \psi + \phi \rangle = \langle \psi \rangle + \langle \phi \rangle$. Also, if ψ is only a function of a single layer T_k and S_k, then the other layers are marginalized and we find

$$\langle \psi(T_k, S_k) \rangle = \int_0^\infty \int_0^\infty \psi(t, s)\, p_{T_k, S_k}(t, s)\, \mathrm{d}t\, \mathrm{d}s, \quad (23)$$

where $p_{T_k, S_k}(t, s) = \partial^2 F_{T_k, S_k}(t, s) / \partial t\, \partial s$.

For some ψ, such as the total or condensed water paths, the layer contributions appear linearly and so the $\langle \cdot \rangle$ passes through to individual layer evaluations in $\langle \psi \rangle$. This means that $\langle \psi \rangle$ can be evaluated as a weighted sum of single-layer marginal averages as in (23). But for many other quantities, such as radiative transfer integrals, or surface precipitation flux, the ψ is a nonlinear function of the layer contributions, and so $\langle \psi \rangle$ cannot be evaluated in terms of layer averages, $\langle \cdot \rangle_k$. We will consider the numerical evaluation of such $\langle \psi \rangle$ in section 5.

Finally, note that the expectation value $\langle \psi \rangle$ is invariant under any arbitrary transformation of $(\mathbf{T}, \mathbf{S})$. In particular, for any one-to-one transformation to space $\mathbf{X} \subset \mathbb{R}^{2K}$ we may rewrite (22) as

$$\langle \psi \rangle = \int_{\mathbf{X}} \psi(\mathbf{x})\, p_{\mathbf{X}}(\mathbf{x})\, \mathrm{d}\mathbf{x}. \quad (24)$$

We may choose any transformation that is convenient for the evaluation of $\langle \psi \rangle$.

5. Monte Carlo subcolumn generation

There are many quantities, such as the short-wave cloud transmittance, for which the grid-column average integral in (22) or (24) does not have a simple analytic form and therefore requires evaluation by numerical methods. One such method involves the Monte Carlo generation of a finite number of 'subcolumns' from the grid column and then approximation of $\langle \cdot \rangle$ as the average over this population of subcolumns. By a 'subcolumn' we mean a description of all K layers at one particular horizontal location in the grid column, as specified by a $2K$-vector

$(\mathbf{T}, \mathbf{S})$. The Monte Carlo subcolumn generator must produce a set of such subcolumns in a manner consistent with the underlying areal distribution function $F_{\mathbf{T},\mathbf{S}}$ of the grid column. While there exist other numerical methods that are more accurate, we will focus on Monte Carlo methods here because of the recent interest in Monte Carlo radiative transfer parametrizations for GCMs (e.g. Räisänen and Barker, 2004; Räisänen *et al.*, 2005).

In the following subsections we will present subcolumn generators for the Gaussian copula and for a number of existing overlap assumptions. In section 6 we will compare the performance of these generators using cloud-resolving model (CRM) data.

5.1. A Gaussian copula subcolumn generator

For the Gaussian copula (GCOP), we may use the monotonic increasing transformations to $\mathbf{Z}$, per (11), so that

$$\langle \psi \rangle = \int_{\mathbb{R}^{2K}} \psi(\mathbf{z}) \, p_{\mathbf{Z}}(\mathbf{z}; \mathsf{C}) \, \mathrm{d}\mathbf{z}, \qquad (25)$$

where $p_{\mathbf{Z}}(\mathbf{z}; \mathsf{C})$ is defined by (12).

This suggests the following Monte Carlo method to estimate $\langle \psi \rangle$:

(1) Select N_s random sample vectors $\mathbf{Z} \in \mathbb{R}^{2K}$ from the distribution $N_{2K}(\mathbf{0}, \mathsf{C})$, i.e. from a population with probability density $p_{\mathbf{Z}}(\mathbf{z}; \mathsf{C})$, using Appendix A;

(2) Form a rank vector $\mathbf{R}$ for each $\mathbf{Z}$ using $R_k = \Phi(Z_k)$;

(3) Form a $\mathbf{T}$ and $\mathbf{S}$ for each $\mathbf{R}$ using the inverse of the marginal distributions
$$T_k = F_{T_k}^{-1}(R_k) \qquad \text{and} \qquad S_k = F_{S_k}^{-1}(R_{k+K}),$$
$k = 1, \dots, K$;

(4) Form a $\mathbf{q}_c$ for each $\mathbf{T}$ and $\mathbf{S}$ using
$$q_{ck} = (S_k - 1)q_s(T_k)H(S_k - 1) \quad \text{and any other}$$
intermediate quantities needed to evaluate a ψ for each sample;

(5) Finally, $\langle \psi \rangle_{N_s}^{\mathrm{GCOP}}$ is the average of the N_s such ψ.

5.2. 'S-only' generation

The Gaussian copula method is able to model the interdependence of T and S, both within and between layers. This flexibility is one of the advantages of the Gaussian copula method. However, it is not immediately obvious how to include joint T–S generation in many of the other generators we will use for comparison purposes. For example, consider so-called 'maximum overlap'. We will demonstrate in section 5.4 that a maximum overlap generator is very simple for S alone. However, if the S fields among layers are maximally overlapped, and if the intra-layer T–S relationship varies among layers, then it is clear that the T fields cannot also be maximally overlapped. Since a careful treatment of the incorporation of joint T–S variability into existing generators is outside the scope of this paper, we will use S-only generators (i.e. we will generate S fields only,

and use the average T for each layer; the S-only version of the Gaussian copula generator is simple – $\mathbf{Z}$ and C are restricted to the K-space of S values alone) for almost all of the intercomparisons in section 6. We would also add that S-only generators are also a more natural next step for current GCMs, some of which have cloud parametrizations with subgrid-scale moisture PDFs (e.g. Rienecker *et al.*, 2007), but few, if any, have joint PDFs of moisture and temperature.

Having justified this use of S-only generators on practical grounds, at least two questions remain: how significant are the errors made by modelling only S variability and using layer average temperatures, and why do we choose to examine S-only variability, and not, say, q_t-only variability? We will answer both these questions in section 6, where we will show that $S = q_t/q_s(T)$ – which under section 2.2(A4) is a direct indicator of cloudiness, without regard to T – is able to capture the most significant effects of T variation implicitly, via its dependence on $q_s(T)$. Furthermore, we will show that the grid-column-averaged condensed water path and column radiative properties have smaller errors if the full S variability is retained and layer averages used for explicit T, than if full q_t variability is retained and layer-average T is used. In other words, it is more important in the evaluation of these quantities to get the cloudiness in each layer correct but make small errors in the water contents than it is to get the q_t correct but to make errors in the extent of the cloudiness. Finally, we note that the physical extent of cloudiness is usually more accurately retrievable from remote sensing measurements than is q_t.

5.3. The 'random overlap' copula and generator

So-called 'random cloud overlap' (RAN) can be modelled using K independent layers and the product copula, $C_{\mathbf{S}} = \prod^K$, as per section 3.3(a). The generator of this copula is trivial: for each sample $n \in \{1, \dots, N_s\}$, start with a sample rank vector $\mathbf{R}^{(n)} = (U_1^{(n)}, \dots, U_K^{(n)})^{\mathrm{T}}$, each element of which is independently and uniformly distributed on $[0, 1]$. Then, the generated sample is

$$\mathbf{S}^{(n)} = (F_{S_1}^{-1}(U_1^{(n)}), \dots, F_{S_K}^{-1}(U_K^{(n)}))^{\mathrm{T}}.$$

5.4. The 'maximum overlap' copula and generator

In section 3.3(b) we noted that 'maximum cloud overlap' (MAX) in K layers can be represented by the 'M-copula', i.e. using $C_{\mathbf{S}}(r_1, \dots, r_K) = \min(r_1, \dots, r_K)$. The areal density in rank space, $\partial^K C_{\mathbf{S}}(r_1, \dots, r_K)/(\partial r_1 \dots \partial r_K)$, is therefore zero off the main diagonal of $\mathbf{I}^K$, suggesting that the generation scheme for the ranks should use the same uniform rank for each dimension. Namely if we select random rank vectors $\mathbf{R}^{(n)} \equiv (U^{(n)}, \dots, U^{(n)})^{\mathrm{T}} \in \mathbf{I}^K$, where $U^{(n)}$ is uniformly distributed on $[0, 1]$, then for infinite samples,

$$\mathrm{Fr}\,(R_1^{(n)} \le r_1, \dots, R_K^{(n)} \le r_K)$$
$$= \mathrm{Fr}\,(U^{(n)} \le \min(r_1, \dots, r_K)) = \min(r_1, \dots, r_K),$$

as required. It is clear, then, that the M-copula gives the maximum possible correlation between all the K ranks.

5.5. The Geleyn and Hollingsworth (1979) generator

We also considered several combination maximum-random overlap (MRO) variants. The 'standard' version, due to Tian and Curry (1989), is discussed in Appendix C. The method is identical to maximum overlap for a single contiguous cloud layer at model layer resolution, even if the cloud is really comprised of multiple sources, as in our case-study in section 6. Instead, we will use the more commonly implemented and more realistic variant MROGH, due to Geleyn and Hollingsworth (1979), which is described in detail in Appendix D.

5.6. The Räisänen *et al.* (2004) generator

Finally, we consider a version of the types of 'generalized overlap' method described in the Introduction. Specifically, we use the total water generator described in Appendix B of Räisänen *et al.* (2004), but adapted to saturation ratio S rather than q_t. The method (RAIS) is discussed in detail in our Appendix B. In summary, it yields a linear combination of maximum and random overlap between any two layers k and l in the column, with the weighting depending on the *product* of the rank correlation coefficients of S between each pair of *adjacent* layers between k and l. Note that this is a far more restrictive overlap specification than the Gaussian copula, which permits very general rank correlations between pairs of layers, including anticorrelations (i.e. elements of so-called 'minimal overlap').

5.7. Note on cloud fraction evaluation

For each of the generators presented in this section (and supporting Appendices B and D) we have derived an associated copula. This permits direct evaluation of column cloud fraction using (6). That being said, not all of the copulas are necessarily efficient to evaluate directly, and a simple alternative is to generate a large number of subcolumns, and to count the fraction of which have cloud in any layer. Since generating an ensemble of subcolumns will usually be necessary to evaluate radiative fluxes anyway, this may be an efficient option for moderate accuracy of f in many cases. (Further discussion is in section 6.4.)

For example, for the Gaussian copula, there is no closed form for the multinormal distribution $F_{\mathbf{Z}}$ of (14), so it must be evaluated numerically (Drezner and Wesolowsky, 1989; Drezner, 1994; Genz and Bretz, 1999, 2002; Genz, 2004). A MATLAB implementation mvncdf.m is based on these papers. For dimensions greater than three, the MATLAB implementation uses a quasi-Monte Carlo integration algorithm based on methods developed by Genz and Bretz, as above. These methods can be time consuming for high accuracy and so the more simplistic subcolumn approach just discussed is also attractive.

6. A CRM case-study

This paper is mainly intended as a theoretical outline of our new approach to dealing with cloud overlap. We will show here only some preliminary numerical results using synthetic data from a cloud-resolving model (CRM) simulation and present a more extensive testing in a follow-up paper.

6.1. Goddard Cumulus Ensemble model

We use output from a Goddard Cumulus Ensemble (GCE) model simulation of the Atmospheric Radiation Measurement (ARM) Spring 2000 Intensive Operation Period (IOP) over the Southern Great Plains (SGP) site. The simulation has 128×128 grid points per layer with a 1 km grid spacing, and 41 levels in the vertical, spread non-uniformly from the surface to 22 km. The GCE model is non-hydrostatic with a bulk microphysics scheme including three species of ice: crystals, snow and graupel. Details of the model and an analysis of this particular simulation can be found in Zeng *et al.* (2007).

6.2. A preliminary 12-layer test

We consider the overlap of the GCE model's lowest twelve layers from a single snapshot of the GCE simulation output at 1831 UTC on 16 March 2000, about 15 days into a 20-day simulation, and about one day after the passage of a cold front through the domain. The lowest 12 layers are precisely the ones that contain no grid points with $T < 0\,^{\circ}\mathrm{C}$ and no cloud ice in this snapshot. Each of these layers contains $N_{\mathrm{H}} \equiv 128 \times 128 = 16\,384$ grid points, and the twelve layers together form a simplified 'grid column' for the cloud fraction tests to follow. The average profile of temperature for these layers, which extend from the surface to about 2 km, is shown in Figure 1(a).

For each grid box in our simplified grid column, we form a total water content q_t by summing the model vapour and condensed (but non-precipitating) cloud water contents. The average profile of q_t is shown in Figure 1(b) and together with the temperature profile shows a relatively warmer, moister air mass overlying a poorly mixed surface layer up to 800 m. The upper air mass is the air that has been lifted by the passing cold front, and the lower layer is the colder, drier air that swept under it. The bulk condensed water content, q_c, and saturation ratio, S, are then calculated according to section 2.2 (A4). Figure 1(c) shows the presence of some condensed water in each of the twelve layers.

Figure 2 shows horizontal slices of T and q_c for the GCE snapshot for three selected layers: 3, 5, and 12, with mid-layer heights at 214, 470, and 2122 m above the surface.

Layer 3 has a complex structure containing narrow, string-like perturbations with high T and q_c (shown) and also high q_t and strong upward motion (not shown). These are cumulus cloud bands associated with surface-driven warm, moist convective updraughts. These form
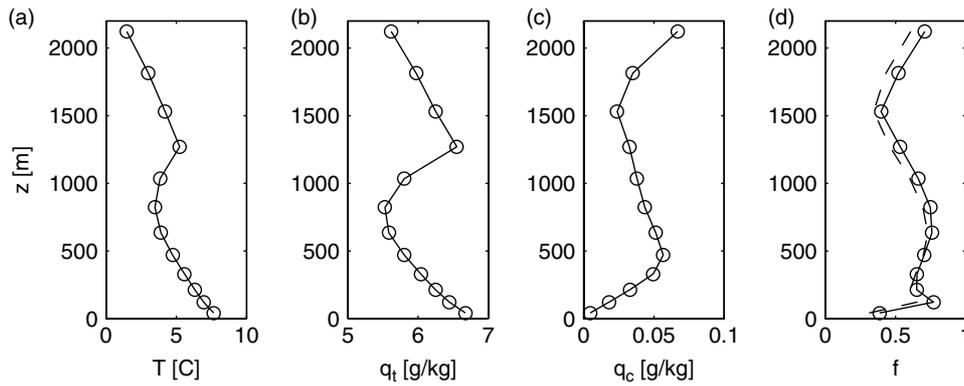
Figure 1. Layer average profiles for the lowest twelve layers of the GCE snapshot (see text) of (a) temperature, T, (b) total water content, q_t, and (c) condensed water content, q_c. (d) shows the f_k^{EDF} (solid) and f_k^{GEV} (dashed) layer cloud fractions (see text).

readily in the lower region because the surface has been recently wet by frontal rain and the latent heat flux is high. Layer 12, conversely, is representative of the warmer, moister, frontally raised air mass and is characterized by a much more slowly varying, large-scale cloud field surrounding a distinct clear region. Finally, layer 5 is in the transition between these two air masses and shows elements of both.

We will estimate a Gaussian copula using the T and S data from this twelve-layer snapshot, as described below, and then see how our copula-based cloud fraction, cloud water path, and column radiative properties compare with the exact snapshot properties and with the properties predicted by other overlap methods.

6.3. Estimating marginal distributions

There are two different approaches we take to estimating the underlying marginal distributions F_{T_k} and F_{S_k} of the GCE grid-column data. The first is to estimate a non-parametric empirical distribution function (EDF). To do this, we assign an equal probability $1/N_{\mathrm{H}}$ to each subcolumn value, as a delta function in the areal fraction density, namely,

$$p_{T_k}^{\mathrm{EDF}}(t) = \frac{1}{N_{\mathrm{H}}} \sum_{n=1}^{N_{\mathrm{H}}} \delta(t - T_k^{(n)}),$$

where the $T_k^{(n)}$ are the GCE subcolumn temperatures for layer k, and similarly for S_k. Then, the areal EDF is just

$$F_{T_k}^{\mathrm{EDF}}(t) = \frac{1}{N_{\mathrm{H}}} \sum_{n=1}^{N_{\mathrm{H}}} H(t - T_k^{(n)}), \qquad (26)$$

where H is the Heaviside step function (section 2.2 (A4)). (The rationale for a delta function assignment of probability about a data point $T_k^{(n)}$ is that such data are only available at some finite precision and that T values from the underlying population that are near $T_k^{(n)}$ will all present in the data as $T_k^{(n)}$.)

For almost all $t \in \mathbb{R}$, specifically, for $t \notin \mathbb{T}_k \equiv \{T_k^{(1)}, \ldots, T_k^{(N_{\mathrm{H}})}\}$, this $F_{T_k}^{\mathrm{EDF}}(t)$ is just the fraction of subcolumns with $T_k^{(n)} < t$, since there is zero probability of $T_k^{(n)}$ being equal to any such t. Similarly, none of the N_{H}
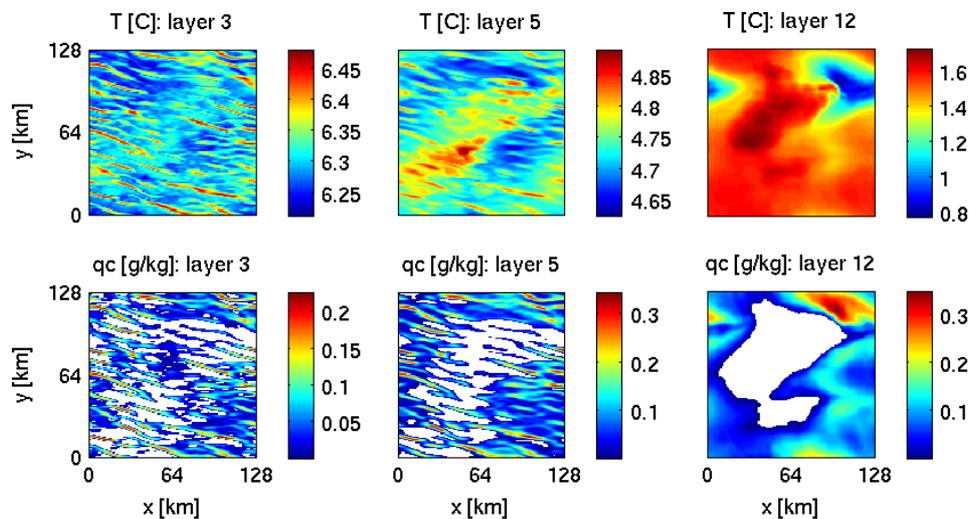

Figure 2. Horizontal slices of the GCE snapshot at layers 3, 5 and 12, for (top row) temperature and (bottom row) condensed water content. This figure is available in colour online at www.interscience.wiley.com/journal/qj

grid boxes in any layer of our grid column has exactly $S = 1$ and so the EDF clear and cloud fractions, $f_k^{\prime\,\text{EDF}} \equiv F_{S_k}^{\text{EDF}}(1)$ and $f_k^{\text{EDF}} \equiv 1 - f_k^{\prime\,\text{EDF}}$, are the fractions of sub-columns with $S_k < 1$ and $S_k > 1$, respectively. By the same token, the clear and cloudy *column* fractions, $f^{\prime\,\text{EDF}}$ and $f^{\text{EDF}} = 1 - f^{\prime\,\text{EDF}}$, are the fractions of the sub-columns with $S < 1$ for every layer and $S > 1$ for at least one layer, respectively. Conversely, consider the empirical ranks assigned to the data samples themselves. If there are no ties in $\mathbb{T}_k$, then $R_k^{(n)EDF} \equiv F_{T_k}^{\text{EDF}}(T_k^{(n)}) = (n^* - 1/2)/N_\text{H}$, where $n^* \in \{1, 2, \dots, N_\text{H}\}$ is the ordinal position of $T_k^{(n)}$ within an ascending reordering of $\mathbb{T}_k$. Thus the empirical ranks are easily generated from an ascending sort of $\mathbb{T}_k$. Similar comments apply to the ranks $R_{k+K}^{(n)EDF} \equiv F_{S_k}^{\text{EDF}}(S_k^{(n)})$ associated with S.

(By 'ties', we mean repeated values. For finite precision experimental data, or even for simulation output such as the GCE data in which temperature is output in Kelvin in 'single precision', ties can exist in the $\mathbb{T}_k$. An algorithm for assigning a common rank to each repeated value while satisfying (26) is not difficult and will not be discussed further.)

The second approach to margin estimation is parametric and involves fitting a known distribution function to the T_k and S_k data. Various distributions were tried, but the Generalized Extreme Value (GEV) distribution proved to give the best fits. This distribution is defined by

$$F_X^{\text{GEV}}(x; \mu, \sigma, \xi) = \text{e}^{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}} \qquad (27)$$

for $1 + \xi(x - \mu)/\sigma > 0$, where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter and $\xi \in \mathbb{R}$ is the shape parameter. It can handle both positively and negatively skewed distributions. For $\xi > 0$, the distribution is bounded below, and has an infinite tail in the positive x direction. Conversely for $\xi < 0$, the distribution is bounded above, and has an infinite tail in the negative x direction. However, for the distributions of positive T and S we are studying, there will, in practice, be negligible probability for $T < 0$ or $S < 0$. The maximum likelihood fits are shown in the top panels of Figures 3 and 4 for the same three selected layers shown in Figure 2. The layer 3 and 5 fits are generally very good, although the narrow localized peak at $S = 1$ is not captured for layer 3. The layer 12 fits are poorer – it seems that although the GEV distribution is generally reasonable, it cannot model a broad distribution with an additional sharp peak. Future work should be done to select an improved marginal distribution, possibly one that is the sum of two distributions, one to handle the main spread, and another to model an additional peak. Note also from the figures how both positively and negatively skewed marginal distributions are needed for T_k and S_k. The marginals are clearly non-Gaussian (see section 7 for further discussion).

In the following analysis we shall use both the EDF margins $F_{T_k}^{\text{EDF}}$ and $F_{S_k}^{\text{EDF}}$ and the GEV margins $F_{T_k}^{\text{GEV}}$ and $F_{S_k}^{\text{GEV}}$ and compare the results. We prefer the EDF margins, since they are more directly tied to the data, and since the GEV margins are a less than perfect fit and therefore introduce an extra source of error into our analysis that obscures the interpretation of the errors associated with the Gaussian copula model of the rank dependence structure. Nevertheless, we retain the parametric GEV results, since our ultimate plan is to use the new copula method in tandem with PDF-based cloud parametrizations within GCMs, in which case a parametrized marginal distribution function is needed. We will therefore conduct parallel calculations with the two
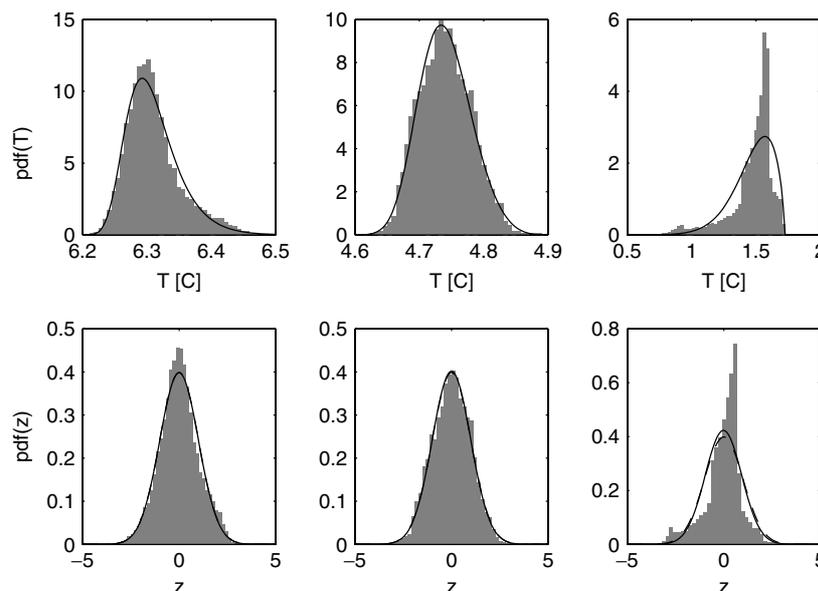

Figure 3. Top row: binned frequency distribution of T_k data for (left to right) layers 3, 5 and 12 expressed in probability density units (grey shading) and the probability density function (solid line) corresponding to the maximum likelihood fit of the Generalized Extreme Value (GEV) distribution to the T_k data. Bottom row: the corresponding frequency distributions (grey shading) of Z_{Tk} transformed from T_k using (11) and (17) and the GEV-fitted margins $F_{T_k}^{\text{GEV}}$, and a maximum likelihood normal fit (solid line). The PDF for the standard normal $N(0, 1)$ is shown (dashed line) for comparison, but it is only just distinguishable from the maximum likelihood fit in layer 12.
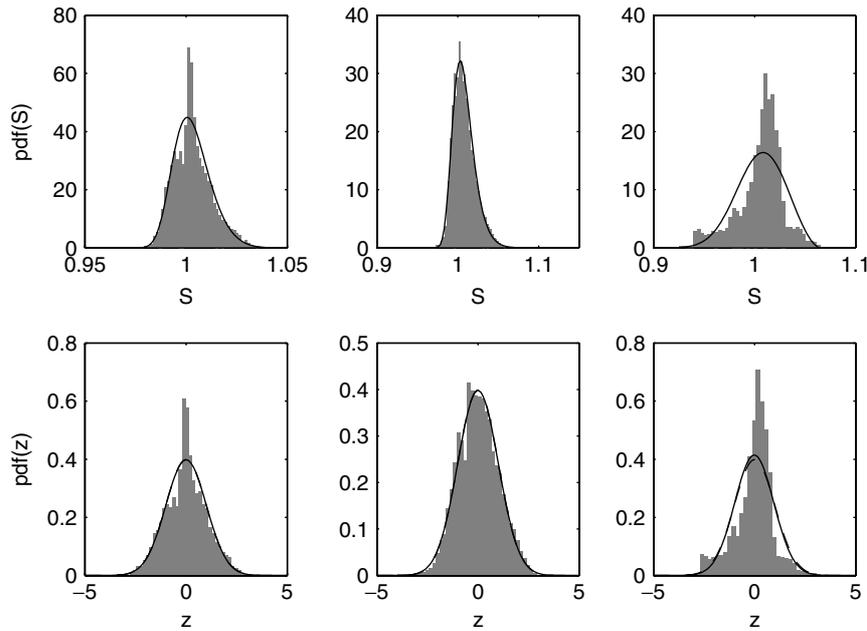
Figure 4. As Figure 3, but for the distributions of saturation ratio, S_k for layers 3, 5 and 12.

approaches, treating the EDF results as our reference, and using the GEV results to indicate how much potential improvement can be expected from a better and more flexible parametric model of the margins.

One final comment is in order with regard to tail behaviour in the margins. It could be argued that the EDF margin does a poor job of modelling the tail behaviour in the underlying distribution, due to the limited number of tail samples available. However, the same can also be true of a parametric model, unless it has a form that is based on extensive analysis of tail behaviour in large-sample studies. Tail behaviour should form an important part of future margin modelling studies, especially if the resulting copula model will be used to parametrize processes such as precipitation that are expected to have strong tail dependence.

6.4. Cloud fraction

The EDF layer cloud fractions f_k^{EDF} are shown in the solid line of Figure 1(d). Notice again that there is cloud in every layer, even though we have inferred from the temperature and moisture profiles above that we are likely looking at two different air masses. This observation gives further reason why a 'maximum overlap' assumption should not automatically be applied to a vertically contiguous set of cloudy layers.

The EDF *column* cloud fraction f^{EDF} is 0.9650. This is the fraction of the N_H horizontal points for which at least one layer has some condensate. By comparison, the column cloud fractions formed by the random overlap and maximum overlap assumptions, using the f_k^{EDF} as input, are 1.0000 and 0.7699, respectively. We will call these $f_{\mathrm{RAN}}^{\mathrm{EDF}}$ and $f_{\mathrm{MAX}}^{\mathrm{EDF}}$. They are in error by 3.6% and -20.2% respectively (Table I). Again, while neither random or maximum overlap is realistic, in this case maximum

overlap is very unrealistic, even though the whole layer is 'contiguously cloudy' in the vertical.

We also consider several maximum–random overlap variants. The 'standard' version, due to Tian and Curry (1989) is discussed in Appendix C, but is exactly the same as maximum (MAX) overlap for this case-study, since there is no completely clear layer in our 12-layer grid column. The Geleyn and Hollingsworth (1979) variant, described in section 5.5 and Appendix D, yields $f_{\mathrm{MROGH}}^{\mathrm{EDF}} = 0.9230$, in error by -4.3%.

Finally, the Räisänen *et al.* (2004) generator discussed in section 5.6 and Appendix B yields a cloud fraction $f_{\mathrm{RAIS}}^{\mathrm{EDF}} = 0.8768$, in error by -9.1%. In fairness to Räisänen *et al.*, the total water version we use is not their preferred implementation. Unlike their main version, which allows separate control of the cloud fraction overlap and the cloud condensate amount correlations, the total water version only allows control of the vertical correlations of q_t (or, in our case, S) within the grid column. Nevertheless, the total water (or S) version is most consistent with our 'bulk condensate' approach (section 2.2 (A4)), and so we will not use the full Räisänen *et al.* (2004) generator, which would likely yield improved results.

The GEV layer cloud fractions are evaluated from the GEV best fit margins, $f_k^{\mathrm{GEV}} = 1 - F_{S_k}^{\mathrm{GEV}}(1)$, using (3). These f_k^{GEV} values are shown as the dashed line in Figure 1(d). In the upper air mass, especially, these f_k^{GEV} tend to underestimate f_k^{EDF} by up to 20%. For layer 12 shown in Figure 4, we can see that this underestimate stems from failing to capture the sharp peak in S just above $S = 1$. The random and maximum overlap column fractions calculated using the f_k^{GEV} are as follows: $f_{\mathrm{RAN}}^{\mathrm{GEV}} = 1.0000$ and $f_{\mathrm{MAX}}^{\mathrm{GEV}} = 0.7284$, which are in error by 3.6% and -24.5% of f^{EDF}. Likewise, $f_{\mathrm{MROGH}}^{\mathrm{GEV}} = 0.8577$ and $f_{\mathrm{RAIS}}^{\mathrm{GEV}} = 0.8357$, in error by

f	GCOP	MAX	RAN	MROGH	RAIS
EDF	–1.9	–20.2	3.6	–4.3	–9.1
GEV	–5.8	–24.5	3.6	–11.1	–13.4

The EDF row uses the EDF ranks, while the GEV row uses the GEV-derived ranks. Both RAN results have f maximized at one and so have equal biases.

–11.1% and –13.4%, respectively. All these results are compared in Table I.

Next, sample ranks are assigned from (20), both for the EDF and GEV marginal distributions, and these are then transformed to **Z** space using (11) and (17). The binned PDFs of the Z_{Tk} and Z_{Sk} for the GEV case are shown in the lower panels of Figures 3 and 4. As expected, these distributions are close to $N(0, 1)$, at least for layers 3 and 5 where the GEV fit is good. The binned Z_k for the EDF case need not be shown – they are distributed almost exactly as $N(0, 1)$, since the EDF ranks have essentially exact uniform distributions.

The column cloud fraction f depends on the copula of **S** alone via (6). In particular, the Gaussian copula (GCOP) estimate is

$$f_{\text{GCOP}} = 1 - C_{\mathbf{Z}}(f_1', \ldots, f_K'; \widehat{\mathbf{C}}_{\mathbf{S}}),$$

where $C_{\mathbf{Z}}$ in this case is the K-dimensional only version of (19) and where $\widehat{\mathbf{C}}_{\mathbf{S}}$ is the $K \times K$ sample correlation matrix (section 4.4) formed from the Z_{Sk} samples only. The values of f_{GCOP} for the two different marginals are:

$$f_{\text{GCOP}}^{\text{EDF}} = 1 - C_{\mathbf{Z}}(f_1^{'\,\text{EDF}}, \ldots, f_K^{'\,\text{EDF}}; \widehat{\mathbf{C}}_{\mathbf{S}}^{\text{EDF}}) = 0.9467,$$

and

$$f_{\text{GCOP}}^{\text{GEV}} = 1 - C_{\mathbf{Z}}(f_1^{'\,\text{GEV}}, \ldots, f_K^{'\,\text{GEV}}; \widehat{\mathbf{C}}_{\mathbf{S}}^{\text{GEV}}) = 0.9089.$$

These are in error by –1.9% and –5.8% of f^{EDF}. For the EDF case, this GCOP error is less than the respective errors associated with all the other methods (Table I) and demonstrates, at least for this simple test case, the utility of the Gaussian copula cloud fraction method.

For the GEV case, however, the RAN error is actually smaller than the GCOP error. We suspect this is only because our case-study is so close to overcast, thereby capping the RAN column fraction at one. In any case, the GEV results are clearly poorer than the EDF results, due to the less than ideal fit to the margins provided by the GEV distribution. It seems likely that an improved marginal model (such as the sum of two distributions, as suggested earlier) would further improve the parametric results and make them comparable to the EDF results.

Note that if we use the GEV-based Gaussian copula, but with EDF-based clear fractions as arguments, we get a cloud fraction

$$1 - C_{\mathbf{Z}}(f_1^{'\,\text{EDF}}, \ldots, f_K^{'\,\text{EDF}}; \widehat{\mathbf{C}}_{\mathbf{S}}^{\text{GEV}}) = 0.9464$$

which is almost as accurate as the pure EDF result $f_{\text{GCOP}}^{\text{EDF}} = 0.9467$. This implies that the chief sensitivity of $f_{\text{GCOP}}^{\text{GEV}}$ to the poor GEV margins comes through the layer clear fractions, $f_k^{'\,\text{GEV}}$, not from the copula itself (i.e. not from $\widehat{\mathbf{C}}_{\mathbf{S}}^{\text{GEV}}$). This makes sense because the estimation of $\widehat{\mathbf{C}}_{\mathbf{S}}^{\text{GEV}}$ depends on all the GEV rank data, and so errors are suppressed by averaging. The $f_k^{'\,\text{GEV}} = F_{S_k}^{\text{GEV}}(1)$, however, is sensitive only to the rank error at saturation ($S_k = 1$), which can be large, as we have seen from Figure 4.

The cloud fraction results in this section are evaluated exactly from their analytic copulas and quoted at four significant digits, except for the GCOP results, which are evaluated using the quasi-Monte Carlo integration MATLAB algorithm discussed in section 5.7, but are also accurate to the four significant figures quoted. Alternatively, we can estimate the column cloud fractions by generating subcolumn ensembles from the S-only generators of section 5 and counting the fraction of those subcolumns which contain cloud. For each generator method we have performed this procedure for a range of ensemble sizes N_s. For each N_s we perform 100 different realizations to characterize the random error for various ensemble sizes. We find that the root-mean-square (r.m.s.) error in the ensemble column cloud fraction goes as $N_s^{-1/2}$, as expected. At $N_s = 16$, the absolute (relative) errors in f are about 0.1 (10%), while at $N_s = 16384$ they are about 0.003 (0.3%).

At the end of section 4.3, we noted that the Gaussian copula model would only be as good as the degree to which the (**T**, **S**) or **S** data transformed to **Z** space could be modelled as a multinormal. Any statistical model will entail some loss of information and will involve an approximation to the dependencies between variables found in the data. In our case, it is certainly possible when we evaluate the correlation matrix $\widehat{\mathbf{C}}$ (section 4.4), thereby fitting a multinormal to the **Z**, that we will destroy some important dependence information. One way to examine this possibility, is to regenerate new (**T**, **S**) or **S** data from the model (section 5.1), and compare the dependence structures found in the regenerated data to those in the original (**T**, **S**) or **S** data.

Figure 5 shows such a comparison for the inter-layer S dependencies between (a) layers 3 and 5 and (b) layers 5 and 12. For each pair of layers, a joint PDF $p(S_k, S_l)$ is estimated:

$$p(S_k, S_l) = \frac{1}{N_{\text{H}}} \sum_{n=1}^{N_{\text{H}}} \omega\left(S_k - S_k^{(n)}, S_l - S_l^{(n)}\right), \quad (28)$$

where $n = \{1, \ldots, N_{\text{H}}\}$ indexes the GCE subcolumns and $\omega(x, y)$ is a kernel PDF, obeying $\iint \omega(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$. In this way, each of the GCE subcolumns is assigned a probability $1/N_{\text{H}}$ which is distributed in (S_k, S_l) space according to the kernel. (The particular kernel we used is a radially symmetric spline kernel $\omega(x, y) = \omega^*(r/a)$ from Monaghan and Lattanzio (1985), where $r = \sqrt{x^2 + y^2}$ and a is a smoothing scale length.
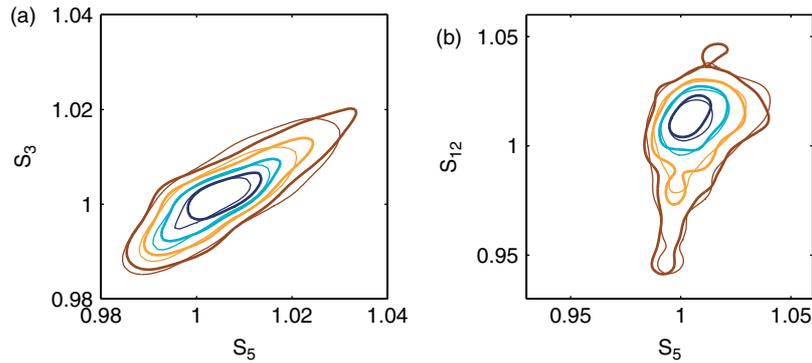
Figure 5. Contours (brown, orange, cyan, and blue) contain 80, 60, 40, and 20% of the joint inter-layer S probability, such that the probability densities within each contour are larger than the probability densities outside of it. Bold contours are for the 16 384 raw GCE subcolumns, and thin contours are for 16 384 subcolumns generated from an S-only Gaussian copula and empirical (EDF) margins. (a) is for layers 3 and 5 (separated by 256 m), and (b) for layers 5 and 12 (separated by 1652 m). The panels are consistent with a general reduction in inter-layer correlation with increasing layer separation (i.e. contours are aligned more along the diagonal of equal values in (a).) The underlying probability density maps have been kernel smoothed (see text) for ease of interpretation, but exactly the same smoothing parameters have been applied to both the GCE- and copula-generated data. The Gaussian copula appears to model the S dependencies well. This figure is available in colour online at www.interscience.wiley.com/journal/qj

The kernel has compact support, with $\omega = 0$ for $r \geq 2a$. We used $a = \min(\sigma_k, \sigma_l)$, where σ_k is the standard deviation of S_k from the raw GCE subcolumns. This a was found to provide a suitably smooth PDF and contours.)

This technique allows us to generate a smoothed empirical PDF at a regular grid in (S_k, S_l) space. A simpler alternative would have been to bin the $(S_k^{(n)}, S_l^{(n)})$ into a two-dimensional frequency histogram, but this produces a rather more noisy PDF which detracts from easy interpretation of the figure. An important point is that the same kernel density method, with the exact same smoothing scale, is applied to both the raw GCE subcolumns and to N_H Gaussian copula-generated subcolumns, so the figure does provide a fair comparison between these two ensembles.

The contours in the figure require some further explanation. They are not direct contours of $p(S_k, S_l)$. Rather, the contours (brown, orange, cyan, and blue) contain 80, 60, 40, and 20% of the total joint inter-layer S probability, and the probability densities $p(S_k, S_l)$ within each contour are larger than all the $p(S_k, S_l)$ outside of it. This contour method is the same for the GCE and regenerated data, so the comparison is again fair. From the figure, the comparison is very good, with the thick contours (GCE data) and thin contours (copula-regenerated data) tracking one another very well. This gives us confidence that the Gaussian copula method models the dependence structures of S very well for this cloud field.

6.5. Condensed water path

The condensed water path is

$$W_c \equiv \int_0^\infty \rho_c \, dz = \sum_{k=1}^K b_k \lfloor S_k - 1 \rfloor_0 , \qquad (29)$$

where $\lfloor X \rfloor_0 \equiv \max(X, 0)$, $b_k \equiv \varepsilon e_s(T_k) \Delta z_k / (R_d T_k)$ and Δz_k is the layer thickness, and where we have used (1), (A1), (A4), and the definition of ρ_* from section 2.1. The grid-column average is therefore

$$\langle W_c \rangle = \sum_{k=1}^K \langle b_k \lfloor S_k - 1 \rfloor_0 \rangle . \qquad (30)$$

Note how the layer contributions to W_c appear linearly, so $\langle \cdot \rangle$ passes through to individual layer evaluations. This means that no inter-layer correlations are needed to evaluate $\langle W_c \rangle$, so all the previous methods of cloud overlap are irrelevant for its calculation. This is not the case, however, for the evaluation of the mean square water path, since

$$\langle W_c^2 \rangle = \sum_{k=1}^K \sum_{l=1}^K \langle b_k b_l \lfloor S_k - 1 \rfloor_0 \lfloor S_l - 1 \rfloor_0 \rangle . \qquad (31)$$

In this case, the correlation between every pair of layers plays a role.

Using all N_H subcolumns from our GCE case-study, we obtain the following average and standard deviation: $\langle W_c \rangle = 99.04 \, \mathrm{g \, m^{-2}}$ and $s_{W_c} = 102.6 \, \mathrm{g \, m^{-2}}$, where $s_x^2 \equiv \langle x^2 \rangle - \langle x \rangle^2$. If we re-evaluate these using the layer average temperatures in the calculation of the b_k, we obtain $\langle W_c \rangle = 99.79 \, \mathrm{g \, m^{-2}}$ and $s_{W_c} = 104.1 \, \mathrm{g \, m^{-2}}$, which deviate from the exact values by 0.8% and 1.5%, respectively. Given the smallness of these deviations, we will proceed using layer average temperatures, meaning that we will only consider inter-layer correlations in S, not in T. In the previous section on cloud fraction, we also only considered inter-layer S correlations, which were formally sufficient to define the grid-column cloud fraction. We will now evaluate $\langle W_c \rangle$ and s_{W_c} using various S-only generators, since it appears that adding explicit T variability has a small effect for condensed water path, at least for the grid column under study.

The above 'S-only' approximation uses the exact GCE temperature in the evaluation of $S = q_t/q_s(T)$, thus retaining the full implicit effect of T variability

within S. The layer average temperature, $\overline{T}$, is used everywhere else, i.e. in the evaluation of b_k in the current context. This approximation preserves the exact *locations* of clouds in the grid column, which depend only on S, but effectively rescales the water contents q_t and q_c in subsequent recalculations of these quantities from S and $\overline{T}$, i.e. $q_t = S \cdot q_s(\overline{T})$ and $q_c = \lfloor S - 1 \rfloor_0 \cdot q_s(\overline{T})$. Alternatively, we can use a 'q_t-only' approximation, in which q_t variability is exactly preserved and $\overline{T}$ is used uniformly in *all* calculations, including the evaluation of $S = q_t/q_s(\overline{T})$. This means that the boundaries of the clouds will change somewhat, but the exact q_t field is preserved. Under this 'q_t-only' approximation, $\langle W_c \rangle = 82.88 \, \text{g m}^{-2}$ and $s_{W_c} = 64.9 \, \text{g m}^{-2}$, which are much more in error with respect to the exact GCE values than the S-only approximation values presented earlier. Thus, we uniformly use the S-only approximation rather than the q_t-only approximation.

For each of the previously discussed MAX, RAN, MROGH, RAIS and GCOP generators, we produce $N_H = 16\,384$ subcolumns of $\mathbf{S}$ and combine these with the layer average temperatures to produce fair comparisons with the S-only GCE values $\langle W_c \rangle = 99.79 \, \text{g m}^{-2}$ and $s_{W_c} = 104.1 \, \text{g m}^{-2}$ discussed above. Each generator produces a series of ranks, $V_k^{(n)}, k \in \{1, \dots, K\}, n \in \{1, \dots, N_H\}$ from which the $S_k^{(n)} = F_{S_k}^{-1}(V_k^{(n)})$ are evaluated using the inverse marginal distribution for each layer. For the GEV margins, this process is simple, since $F_{S_k}^{\text{GEV}}$ is continuous and easily inverted. For the EDF margins, $F_{S_k}^{\text{EDF}}$, which are not continuous (cf. (26)), we use the following procedure:

Project $V_k^{(n)}$ onto $\{1, \dots, N_H\}$ using $n_k = \text{ceil}(N_H V_k^{(n)})$, where ceil rounds up a non-integer to the next integer; $S_k^{(n)}$ is then the n_kth element in a sorted (ascending) list of the N_H GCE values of S_k.

The generation procedures described above are Monte Carlo in nature, and produce slightly different $\langle W_c \rangle$ and s_{W_c} values for each realization. We therefore conduct 100 separate realizations to characterize the spread in these quantities. Table II presents, for each generator, the relative biases (with respect to the GCE reference calculations) of the mean $\langle W_c \rangle$ and s_{W_c} values over the 100 realizations.

As expected, the EDF biases for $\langle W_c \rangle$ are all zero to within sampling error, since $\langle W_c \rangle$ does not depend on the overlap assumption. The GEV $\langle W_c \rangle$ are all about 4.5% too high. This may seem strange, since the GEV margins underestimate the cloud fraction f, but the GEV fit tends to overestimate larger S probabilities (Figure 4, layer 12), which are more highly weighted in $\langle W_c \rangle$, which depends on the first moment of $\lfloor S - 1 \rfloor_0$. (As noted in the caption of Table II, the numbers inside the parentheses are actually standard deviations over the 100 realizations. To treat them as errors in the quantity to the left of the parentheses, we must assume normality in the distribution of the 100 results and then divide the standard deviations by $\sqrt{100} = 10$. Even after doing so, the EDF biases for $\langle W_c \rangle$ are all zero to within about one standard error.)

The s_{W_c} biases vary considerably for the different generators. The GCOP bias is about 0.3% for EDF and about 7% for GEV. It clearly outperforms the other generators for this case study. The RAIS generator, which also directly uses the rank correlations in the GCE snapshot, but only between adjacent layers, comes next at about 11 and 18% for EDF and GEV respectively. The MAX and MROGH generators perform about the same, with biases of about 19 and 26% for EDF and GEV. The RAN generator is worst, underestimating s_{W_c} by about 60%. This is because random overlap tends to homogenize the W_c field, destroying strong vertical correlations in cloud water present in the data (e.g. Figure 2, layers 3 and 5).

All the results in Table II derive from subcolumn ensembles of $N_H = 16\,384$ members. We have also performed the generator calculations for ensemble sizes $N_s < N_H$ and find that the standard deviations (numbers inside the parentheses) scale as $N_s^{-1/2}$, much as expected.

The XOR and XOH methods in Table II have not yet been introduced. These are the 'eXact Overlap Randomized and Homogenized' methods. Both work with an S-only (i.e. temperature-averaged) version of the GCE subcolumn ensemble in which the S values at all clear points are exactly preserved. The cloudy points have their S values altered, as described below, but remain cloudy (i.e. $S > 1$). Both methods therefore have the exact geometrical cloud overlap present in the S-only GCE data. In this sense, these methods are not traditional cloud generators, since they use exact overlap information from the case-study that we cannot hope to parametrize in a GCM. Nevertheless, they are useful methods for isolating the relative importance of pure geometrical cloud overlap, against the effects of full horizontal and vertical variability of cloud condensate amount.

Specifically, XOR randomly interchanges the cloudy points in each layer, and does so independently for each layer. This preserves the exact EDF of S (and q_c) in each layer, but destroys subgrid-scale vertical condensate correlations within and between clouds. This introduces a large -38% bias in s_{W_c}, demonstrating that such vertical correlations are very important to model, even if the geometrical overlap of cloud boundaries is perfect.

The XOH method replaces the cloudy S (and q_c) values in each layer by their average value. This simulates a common GCM approach in which clouds are horizontally uniform in each model layer, but overlapped geometrically according to some overlap assumption. In the XOH method, these uniform cloud layers are magically overlapped in precisely the correct way. This produces an even worse s_{W_c} bias of -47%.

Taken together, the XOR and XOH results point to the importance of modelling both horizontal and vertical variability of clouds, even on small scales. It is not surprising that modelling such variability is necessary to capture the variance in W_c, but this variance is directly related to radiative transfer and therefore very important for the estimation of average radiative fluxes in a GCM grid column, as we shall see in the next section.

Table II. Percent biases in the grid-column average and standard deviation of condensed water path, $\langle W_c \rangle$ and s_{W_c}, with respect to the GCE reference values, 99.79 and 104.1 g m^{-2}, for methods GCOP, MAX, RAN, MROGH and RAIS, and $N_s = 16384$.

	GCOP	MAX	RAN	MROGH	RAIS	XOR	XOH
$\Delta \langle W_c \rangle$ (%)							
EDF	−0.04 (0.78)	0.08 (0.95)	0.03 (0.31)	0.10 (0.92)	−0.03 (0.90)	0	0
GEV	4.47 (0.90)	4.42 (1.12)	4.55 (0.33)	4.69 (1.04)	4.57 (0.97)		
Δs_{W_c} (%)							
EDF	0.29 (0.80)	19.47 (0.95)	−60.55 (0.22)	18.77 (0.95)	10.97 (1.01)	−38.05 (0.16)	−46.68
GEV	6.59 (0.71)	26.81 (0.89)	−57.38 (0.22)	26.00 (0.90)	18.09 (0.87)		

The EDF and GEV denote the type of marginal S distributions used. The exact overlap methods, XOR and XOH, preserve exactly the clear point S values from the GCE snapshot, but respectively randomize and homogenize the cloudy point S values in each layer. See the text for further details. All results use layer-average temperatures (i.e. the S-only method). Cells containing parentheses summarize the results of 100 realizations to quantify the statistical variability: on the left of the parentheses is the bias in the mean value over the realizations, and inside the parentheses is the standard deviation of the realizations, both normalized as a percentage of the GCE reference value for the quantity.

Finally, as noted earlier, the Gaussian copula is able to model the full joint (T, S) variability, not just the S-only variability shown so far in this section. We find that the biases of the full Gaussian copula with respect to the exact GCE $\langle W_c \rangle$ and s_{W_c} are similar to but slightly larger than the biases of the S-only copula results to the S-only GCE data (the GCOP column of Table II). For example, the full Gaussian copula s_{W_c} with EDF margins is biased by 0.47 % with respect to the exact GCE value, as compared to the bias of 0.29 % presented in Table II. This indicates that the Gaussian copula has a slightly harder time capturing the (T, S) joint variability than it does the variability of S values alone. That being said, the S-only GCE s_{W_c} is 1.5% higher than the exact GCE value, so the bias of the S-only copula is about $1.5 + 0.29 \approx 1.8\%$ with respect to the exact GCE results. In other words, if one wants to model the full (T, S) joint variability in the GCE data, it is still better to use the full (T, S) Gaussian copula. Either way, both the S-only and full (T, S) Gaussian copula biases are very small compared to all the other generators.

6.6. Radiative transfer

We will study the short-wave (SW) transmittance, TRN, and reflectance, RFL, (i.e. the downwelling SW flux at the surface and the upwelling SW flux at the top of the atmosphere (TOA), both normalized by the downwelling SW flux at the TOA) and the downwelling long-wave (LW) flux at the surface, DLR, and outgoing LW flux at TOA, OLR, as evaluated by the SW (Chou et al., 1998; Chou and Suarez, 1999) and LW (Chou et al., 2001) column radiation models used in various NASA–GSFC large-scale models. (Clearly the overlap assumptions of these column radiation codes are turned off, since they are being applied iteratively to single subcolumns for which each layer is either overcast or clear.)

The surface was assumed black for both SW and LW calculations (i.e. zero surface albedo and unit emissivity across the spectrum) with a temperature of 281.73 K from the GCE. The temperature and water vapour profile used above the lowest 12 layers comes from the average GCE fields up to a pressure of 37 mb (the top of the GCE domain) and from a standard midlatitude summer (MLS) profile above that level. The profile of ozone molecular concentrations also comes from the same MLS profile, while the CO_2 concentration was set to 370 ppm throughout the domain. The solar code used an overhead sun. An effective radius for cloud droplets of 10 μm was assumed.

As in previous studies of this type (Räisänen et al., 2004), our reference calculation $\langle \psi \rangle^{GCE}$ is the Independent Column Approximation (ICA) calculation in which ψ is evaluated for each of the N_H subcolumns comprising the GCE snapshot and the results averaged. Besides a full 3D calculation, which does not seem like a realistic prospect for GCMs in the foreseeable future, ICA provides the greatest accuracy in domain average fluxes. This $\langle \psi \rangle^{GCE}$ is evaluated for the S-only GCE version (sections 5.2 and 6.5), since this yields only very small biases with respect to the full GCE $\langle \psi \rangle$. (Compared with the exact GCE $\langle \psi \rangle$, the S-only GCE $\langle \psi \rangle$ are biased by about 0.1% in absolute terms for TRN and RFL, and about 0.01% for DLR and OLR. The q_t-only GCE $\langle \psi \rangle$ are all between one and two orders of magnitude more biased.) The $\langle \psi \rangle^{GCE}$ values are presented in the caption of Table III. The exact overlap averages $\langle \psi \rangle^{XOR}$ and $\langle \psi \rangle^{XOH}$ are also ICA averages over the N_H GCE subcolumns, but after a layer-by-layer randomization and homogenization, respectively, of the cloudy point S values, as discussed earlier. For the XOR, each random re-arrangement generates a slightly different $\langle \psi \rangle^{XOR}$, so we produce 100 realizations to quantify the expected spread in $\langle \psi \rangle^{XOR}$ from single randomizations. Next, for the generators χ = GCOP, MAX, RAN, MROGH, and RAIS, we generate N_H subcolumns to evaluate the grid-column averages $\langle \psi \rangle^\chi$. The generation procedure is likewise repeated 100 times, to quantify the expected spread in $\langle \psi \rangle^\chi$ that will be obtained from single realizations of N_H subcolumns. The spread for smaller subcolumn ensembles $N_s < N_H$ is expected to scale as $N_s^{-1/2}$ as for $\langle W_c \rangle$ in section 6.5. Details of the quality of such subsampling will be presented shortly.

Table III presents, for each grid-column radiative average $\langle \psi \rangle$ and for each generator, the relative bias (with

Table III. Percent biases in grid-column average radiative quantities, $\langle TRN \rangle$, $\langle RFL \rangle$, $\langle DLR \rangle$, and $\langle OLR \rangle$, with respect to the GCE reference values, 0.4529, 0.3364, 336.94 and 236.43 W m^{-2}, for methods GCOP, MAX, RAN, MROGH and RAIS, and $N_s = 16384$.

	GCOP	MAX	RAN	MROGH	RAIS	XOR	XOH
$\Delta\langle TRN\rangle$ (%)							
EDF	0.01 (0.38)	8.12 (0.44)	−21.19 (0.17)	7.03 (0.40)	4.26 (0.44)	−11.91 (0.05)	−14.64
GEV	0.78 (0.44)	10.29 (0.51)	−22.77 (0.17)	8.54 (0.45)	5.84 (0.47)		
$\Delta\langle RFL\rangle$ (%)							
EDF	−0.02 (0.44)	−9.35 (0.51)	24.54 (0.20)	−8.20 (0.47)	−4.91 (0.52)	13.83 (0.06)	17.04
GEV	−0.86 (0.51)	−11.81 (0.60)	26.42 (0.21)	−9.88 (0.52)	−6.69 (0.55)		
$\Delta\langle DLR\rangle$ (%)							
EDF	0.05 (0.05)	−2.66 (0.06)	3.42 (0.00)	−0.83 (0.05)	−1.35 (0.07)	1.44 (0.01)	1.74
GEV	−0.40 (0.06)	−3.36 (0.08)	3.46 (0.01)	−1.58 (0.07)	−1.89 (0.07)		
$\Delta\langle OLR\rangle$ (%)							
EDF	−0.01 (0.01)	0.26 (0.01)	−0.51 (0.01)	0.05 (0.01)	0.14 (0.01)	−0.18 (0.00)	−0.33
GEV	0.19 (0.01)	0.47 (0.01)	−0.40 (0.01)	0.26 (0.01)	0.34 (0.01)		

The EDF and GEV denote the type of marginal *S* distributions used. The exact overlap methods, XOR and XOH, and the use of layer-average temperatures, are as in Table II. Cells containing parentheses summarize the results of 100 iterations to quantify the statistical variability: on the left of the parentheses is the bias in the mean value over the iterations, and inside the parentheses is the standard deviation of the iterations, both normalized as a percentage of the GCE reference value for the quantity.

respect to the reference $\langle\psi\rangle^{GCE}$ value) of the mean of $\langle\psi\rangle$ over 100 independent realizations.

For the SW quantities *TRN* and *RFL*, the GCOP biases are *much* smaller than for the other methods and strongly support the utility of the GCOP method, which nicely models the effects of both geometrical cloud overlap and vertical correlation of in-cloud properties.

Both XOH and XOR exhibit the so-called 'plane-parallel bias' in which homogenizing the condensed water path in a cloud makes it less transmissive and more reflective in the SW. Note that in XOR, randomizing each layer's cloudy grid points does *not* homogenize the cloud on a layer-by-layer basis, since we ignore all non-plane-parallel effects in our radiative transfer calculations, but it *does* tend to homogenize vertically integrated cloud properties, such as the condensed water path, in regions where multiple cloudy GCE layers contribute. Maximum cloud overlap (MAX) produces the maximum column clear fraction and therefore promotes transmittance over reflectance. Furthermore, the maximum generator we have employed produces highly inhomogeneous cloud water paths, in the sense that saturation ratio is maximally rank correlated in the vertical. This produces the opposite of the plane-parallel bias just described, yielding larger cloudy transmittance and smaller reflectance. Random overlap (RAN) underestimates column clear fraction and therefore transmittance. It also has a homogenizing effect on the condensed water path, as described above, and therefore also underestimates transmittance. MROGH and RAIS mix maximum and random overlap, and from the sign of the biases in Table III appear to be more heavily weighted in this particular case-study towards maximum overlap.

The following conclusions (for this test case) can be drawn from the magnitudes of the SW biases:

(1) It is very important to consider the vertical correlation of in-cloud quantities, not just the correct geometrical cloud overlap, since even with exact overlap, the XOR SW biases exceed 10%. Compare this with the very small biases for GCOP; evidently the correct modelling of inter-layer correlations of in-cloud properties by GCOP, which XOR destroys, more than compensates for its non-exact column cloud fraction.

(2) The accuracy of prediction of the column cloud fraction is a poor indicator of the quality of the solar transmittance and reflectance. First, as above, a perfect column cloud fraction in XOR still yields in excess of 10% biases in *TRN* and *RFL*. Second, as per Table I, the RAN cloud fraction only has a 4% error but SW biases in excess of 20% while the MAX cloud fraction is in error by more than 20% but has significantly smaller SW biases, about one third those of RAN. Judging from the cloud fraction alone, it appears that RAN is better than MAX, but the exact reverse is true for *TRN* and *RFL*. This value judgment about *f* comes as no surprise really: there is broad awareness in the GCM parametrization community that the cloud fraction is a rather imprecise quantity, especially as it relates to radiative transfer – a thin cloud may have negligible radiative effect even though being technically classified as cloud.

Next consider the LW biases for *DLR* and *OLR*. Again, GCOP gives the smallest biases, but the results are not nearly as definitive as for SW, and especially for *OLR*, which is never in error by more that 0.5%. The analysis of the LW biases is complicated (cf. SW) by the contributions of cloud top and base temperature variability, but we make the following observations:

(1) The exact cloud fraction (XOR) biases are smaller than for MAX and RAN, indicating that the accuracy of the cloud fraction f is more important in the LW than in the SW. This is consistent with the fact that clouds become black in the LW faster than they become saturated in SW reflectance, and so thinner clouds, which contribute as much to f as thicker clouds, become more important.

(2) The larger biases for *DLR* compared to *OLR* can be explained by the greater contrast between the brightness temperatures of the clouds and the clear sky above than between the clouds and the surface.

(3) The signs of the XOR, RAN, and MAX biases in Table III are consistent with the LW version of the 'plane-parallel bias', namely that the homogenization of cloud water path causes an increase in emissivity. Thus the XOR and RAN methods, which tend to homogenize cloud water path, act to blacken the clouds and therefore cause them to radiate at an effective emitting height closer to the boundary of the cloud. For *DLR* this causes the LW radiation to come from closer to the warmer cloud base (Figure 1(a)). Conversely, homogenization produces an effective emitting height for OLR closer to the colder cloud top. The signs of the MAX biases are just the opposite, since the MAX method tends to increase cloud inhomogeneity, as described above for SW.

(4) The signs of the LW biases for MAX and RAN are also consistent with a simple column cloud fraction effect: RAN overestimates column cloud fraction and thereby shields the cold sky for *DLR* and warm surface for *OLR*. The opposite is true for MAX which underestimates column cloud fraction.

(5) As with the SW results, the sign of the LW biases for MROGH and RAIS are more weighted towards maximum overlap than random overlap.

Finally, we have looked into how smaller sample sizes degrade the quality of the GCOP grid-column average when compared to large sample number averages from the other generators. Specifically, Tables IV and V present the number of iterations out of a total of 100 in which $\langle RFL \rangle$ and $\langle DLR \rangle$ from a GCOP generator with a reduced ensemble size, N_s, are less biased, in absolute terms, that the biases from the other methods, which are based on $N_H = 16384$ subcolumns.

It appears that for the GEV margins, which are more representative of a short-term GCM implementation, it is sufficient to use only 32 subcolumn ensembles to equal or beat all the other generator methods and the exact overlap methods. This is encouraging, since 32 subcolumns is well within the bounds of practical implementation with a Monte-Carlo ICA (McICA) strategy (Räisänen and Barker, 2004) to calculate grid-column-average radiative fluxes. Similar performance was found for $\langle TRN \rangle$ and $\langle OLR \rangle$.

Table IV. For various sized sample ensembles, N_s, the table presents the number of realizations out of a total of 100 for which $\langle RFL \rangle^{\mathrm{GCOP}}$ is less biased, in absolute terms, than $\langle RFL \rangle$ for the methods in the other columns, each of which has its bias evaluated using $N_H = 16384$ subcolumns. All methods use GEV margins.

N_s	MAX	RAN	MROGH	RAIS	XOR	XOH
1	8	18	6	6	9	12
2	26	46	25	19	26	33
4	31	60	25	17	35	47
8	43	79	37	25	53	62
16	51	87	42	27	57	72
32	76	100	66	50	79	89
64	87	100	79	61	93	98
128	97	100	91	79	97	100
256	100	100	100	91	100	100
512	100	100	100	96	100	100
1024	100	100	100	100	100	100

Table V. As for Table IV, but for $\langle DLR \rangle$.

N_s	MAX	RAN	MROGH	RAIS	XOR	XOH
1	40	41	8	9	8	8
2	41	44	11	16	8	14
4	71	72	28	36	20	33
8	84	85	56	64	50	60
16	90	90	53	64	48	54
32	100	100	83	91	78	85
64	100	100	91	94	84	92
128	100	100	99	100	98	99
256	100	100	100	100	100	100
512	100	100	100	100	100	100
1024	100	100	100	100	100	100

7. Conclusions, discussion and applications

We have shown that a general representation of GCM column cloud fraction within the PDF-based statistical cloud parametrization context can be obtained by the use of statistical functions called 'copulas' that encapsulate the dependence structure of rank statistics in a multivariate system. Using this theory and cloud-resolving model (CRM) simulations for guidance, a new formulation of GCM cloud overlap has been obtained. Compared with earlier overlap methods, including the 'generalized overlap' approach of Räisänen *et al.* (2004) and Pincus *et al.* (2005), the copula approach allows a far more general specification of the correlation between pairs of layers. It also allows for easy addition of new layer variables, such as temperature, into the modelled grid-column statistics.

We find that Gaussian copula estimates of column cloud fraction for a 12-layer test case using synthetic data from a Goddard Cumulus Ensemble simulation are an improvement over both random and maximum overlap estimates and over several maximum–random combination methods. Furthermore, Gaussian copula Monte

Carlo estimates of cloud water path variance and radiative fluxes showed very significant improvement over all the other generators for this test case. The Gaussian copula generator method also outperforms two exact overlap methods described in the text. It is further found that as few as 32 randomly selected subcolumns from the Gaussian copula generator are sufficient to radiatively outperform all the other generators with an essentially infinite number of subcolumns. These results suggest significant potential for the copula-based parametrization of cloud overlap in future GCM cloud radiation parametrizations. In a follow-up paper, we will present a more extensive testing of the method using GCE simulations for a range of synoptic conditions.

This paper has concentrated on the general theory of the application of copulas to the description of the horizontal and vertical distribution of temperature and water content within a GCM-like grid column. We provide this theoretical basis in the hope that it will be used to develop a new set of GCM cloud parametrizations. Our initial thoughts on how this might be accomplished are as follows:

(1) Our starting point is that a good GCM parametrization should be governed by a reasonably small number of parameters which are prognosticated or diagnosed. The method described in this paper requires a small set of marginal distribution parameters for F_{T_k} and F_{S_k} for each of the K model layers, and a $2K \times 2K$ correlation matrix describing the Gaussian copula of **T** and **S**. The $K(2K-1)$ unique elements of this correlation matrix **C**, as calculated directly from the GCE data for this study, are excessive for the purposes of parametrization, so we will seek an appropriate model of **C** with one or several length-scale parameters, following the work of Gaspari and Cohn (1999) and Gaspari *et al.* (2006).

(2) The GEV distributions used for F_{T_k} and F_{S_k} each have three parameters (a location, scale, and shape parameter) which are roughly equivalent to the specification of the mean, variance, and skewness for each of temperature and saturation ratio. These will be prognosticated following the lead of Tompkins (2002). A suitable diagnostic or prognostic parameterization for the correlation length scales used to model **C** will also be needed.

(3) The marginal-estimated layer cloud fractions and the copula-estimated column cloud fraction can be output as diagnostics. The radiative transfer calculations will use the Gaussian copula generator (section 5.1) as an alternative to the more empirical generator of Räisänen *et al.* (2004) in the implementation of the Monte Carlo ICA method of Räisänen and Barker (2004).

(4) This paper has been limited to the simpler case of water clouds, in which the saturation with respect to water is typically very small and the excess total water above saturation provides a good estimate of the liquid water concentration (section 2.2).

The parametrization of ice clouds is more difficult, but as a first-order extension of the method presented here, we will replace the saturation vapour content with respect to water, $q_s(T)$ in this paper, with a generalized temperature-dependent condensation formation point, $q_s^*(T)$, defined such that the total (liquid + ice) condensate content is approximated by $q_l + q_i = (q_t - q_s^*)H(q_t - q_s^*)$. The split between q_l and q_i will be given by a temperature-dependent liquid/ice fraction, as in many existing GCMs. A parametrized form for this fraction, and for $q_s^*(T)$, will be sought based on analysis of GCE simulations and field data.

We also hope to extend the usefulness of such a copula-based GCM cloud parametrization to the context of data assimilation for NWP applications. The high horizontal and vertical resolution cloud data available from current satellites (e.g. from MODIS on EOS Terra and Aqua, and from CloudSat and other A-Train satellites) and anticipated from future satellite missions (e.g. NPOESS) holds a wealth of statistical information on the distribution of cloud water inside GCM-sized grid columns. We envisage the use of this cloud data to update the marginal and copula parameters of the new GCM parametrization using a parameter estimation approach (e.g. Norris and da Silva, 2007; Dee and da Silva, 1999; Dee *et al.*, 1999). In this way the information from high-resolution satellite observations can be incorporated into global analyses.

The main reason for introducing copulas in cloud applications is their direct relevance to the evaluation of grid-column cloud fraction. Specifically, under the assumptions of section 2.2, the copula of the layer total saturation ratios is a special function that converts the layer clear fractions to the grid-column clear fraction, as in (6). Thus, in section 5 and related appendices, we showed that it is possible to derive the copula associated with a number of existing cloud fraction methods: random overlap, maximum overlap, Geleyn and Hollingsworth (1979) maximum–random overlap, and a total water version of the Räisänen *et al.* (2004) generator.

More generally, copulas can be used to conveniently construct multivariate distributions with *a priori* specified margins via Sklar's theorem (5). This is because a copula is a representation of the dependence structure between variables in rank space and, when combined with marginal distributions for each variable, a complete multivariate distribution can be formed. This approach allowed us to construct a joint distribution function for the saturation ratio (and temperature) among the model layers, using GEV or EDF margins and a Gaussian copula. One important advantage of the copula approach is that it allows complete flexibility in the choice of the analytic form of the marginal distributions. Nothing in principle should prevent replacement of the GEV distribution, which provided the best fit to our data, with better performing single distributions, or even a mixture of distributions. Whatever the chosen margin, it may be combined with a copula choice, in our case the Gaussian

copula, and tested for its skill in representing quantities of interest via stochastic generation, as in section 6.

Such a flexibility may not be inherent to the peg-hat approach of Larson (2007), which uses mixtures of multinormals in the phase space consisting of the total water mixing ratio and liquid water potential temperature for every layer. The choice of multinormal mixtures is necessary in order to get margins which are mixtures of Gaussians, and that are further constrained to be double Gaussians to make the method tractable. While double Gaussian margins are a reasonable fit to many observed and CRM-simulated variables, they may not be completely adequate for the highly skewed distributions sometimes found in simulated or observed data (e.g. for the PDFs of T_{12} and S_5 from our GCE case-study, in Figures 3 and 4, or Figure 2d of Larson *et al.* (2001), from an aircraft leg through cumuli that were rising into broken stratocumuli.) Specifically, double Gaussians tend to have trouble capturing a sharp increase on one side of a PDF that smoothly transitions to a gradual decrease at the other tail. It is thus our understanding that, in order to be tractable, the peg-hat approach limits itself to very particular kinds of marginal distributions, when the data itself may be suggesting more general distributions. No such restriction is imposed on the copula approach which allows complete freedom to choose the margins of each variable in each layer, from any analytic form deemed appropriate, to even an empirical margin directly derived from the data (as in our EDF). A copula-based algorithm essentially splits the problem into two tractable parts: the margins are modelled directly from the layer data, while the joint distribution of the ranks is modeled with a copula.

Of course, the particular choice of copula function is very important, since the best description of the correlations among variables and layers in rank space remains to be determined. Our use of the Gaussian copula has been based on its simplicity and ease of use. Nevertheless, the Gaussian copula has been able to yield estimates of cloud fraction, condensate path variance, and grid-column-averaged radiative fluxes that are very close to the GCE 'truth' and are considerably more accurate than other recent overlap methods (section 6). Other copulas, such as the Student-t copula (Demarta and McNeil, 2005), should be investigated in the future. Such an investigation is beyond the scope of the current article, and will require an examination of the many other fields, such as hydrology and finance, where copulas have proven useful (Cherubini *et al.*, 2004; Dupuis, 2007; Genest and Favre, 2007; Salvadori and De Michele, 2007).

A criticism of copula methods (Larson, 2007) revolves around the greater difficulty of formulating prognostic equations for the rank correlations between variables, say moisture and temperature, than for the linear correlations in physical space, since the filtered Navier–Stokes equations naturally produce correlation terms in the physical variables, not their ranks (e.g. Golaz *et al.*, 2002). This could indeed be a potential drawback of the copula approach for those GCMs that choose to provide the correlation terms prognostically via some sort of turbulent closure. An analysis of this problem will require further work, but we note the following:

(1) There are still many contexts, such as analysis of high-resolution modelling data, from CRM or large-eddy simulation (LES) models, where the copula approach can provide a useful method of analysis of the 'subgrid-scale' variability among variables and layers.
(2) In the context of GCM parametrizations, it may be possible to construct simpler diagnostic specifications of the intra-layer rank correlations, coupled with diagnostic length-scale parametrizations of the inter-layer correlations as discussed earlier.

Ultimately, further studies will be required to compare the peg-hat and copula methods, particularly once both approaches have matured into full GCM parametrizations. In the end, it may be that some combination of elements from both approaches will produce the most benefit. For example, if we consider that grid-column variability is largely composed of large-scale and convective components, we could use a weighted sum of two multivariate distribution functions, each with its own margins and copula. The weighting factor would be height dependent and would need to be parametrized in some way. This approach would allow greater simplicity in the margins and the copula of each component (large-scale and convective). For example, it would allow bimodality in the margins, but the component contributions to those margins could have *unimodal but non-Gaussian* forms. For example, in the case of cumuli rising into a largely decoupled environment, we might expect positively skewed moisture PDFs from the convective component, and perhaps Gaussian PDFs for the large-scale component. Another potential component could be added for convective outflow (anvils), as an extension of schemes which separately model anvil clouds (e.g. Rienecker *et al.*, 2007). Such weighted multi-component systems contain similarities to the mixture of Gaussians employed by Larson (2007), but each component could project non-Gaussian margins, as in this paper, by using copulas.

Acknowledgements

Appendix A. Multinormal random number generation

A multinormal random vector $\mathbf{Z} = (Z_1, \dots, Z_{2K})^{\mathrm{T}} \sim N_{2K}(\mathbf{0}, \mathsf{C})$, where C is a strictly positive definite correlation matrix, can be generated as follows:
$\mathbf{Z} = \mathsf{H}\mathbf{G}$ where H is the lower-triangular Cholesky decomposition of C, such that $\mathsf{C} = \mathsf{H}\mathsf{H}^{\mathrm{T}}$ and H has strictly positive diagonal entries, and $\mathbf{G} = (G_1, \dots, G_{2K})^{\mathrm{T}}$ is a vector of independent standard normal variates, i.e. each $G_k \sim N(0, 1)$ independently.

For the two-dimensional case ($K = 1$),

$$\mathsf{C} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad \text{and} \quad \mathsf{H} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1 - \rho^2} \end{pmatrix}, \quad (32)$$

where $\rho \in (-1, 1)$ is the correlation coefficient, and $\mathbf{Z} = (G_1, \rho G_1 + \sqrt{1 - \rho^2} G_2)^{\mathrm{T}}$ has a particularly simple form.

Appendix B. The Räisänen *et al.* (2004) generator

Räisänen *et al.* (2004), in their Appendix B, describe a version of their stochastic grid-column generator for a total-water scheme. A corresponding version of this generator for the saturation ratio S is as follows:
For sample $n \in \{1, \dots, N_s\}$, start with a sample rank vector $\mathbf{R}^{(n)} = (V_1^{(n)}, \dots, V_K^{(n)})^{\mathrm{T}}$, each element of which is uniformly distributed on $[0, 1]$. Then proceeding from $k = 2, \dots, K$, if $V_k^{(n)} \leq r_{k-1,k}^S$, replace $V_k^{(n)}$ with $V_{k-1}^{(n)}$, otherwise, replace $V_k^{(n)}$ with a new uniform variate on $[0, 1]$. Finally, the generated $\mathbf{S}$ sample is

$$\mathbf{S}^{(n)} = (F_{S_1}^{-1}(V_1^{(n)}), \dots, F_{S_K}^{-1}(V_K^{(n)}))^{\mathrm{T}}.$$

The $r_{k-1,k}^S$ are the linear correlation coefficients between the marginal ranks of S in adjacent layers, otherwise known as rank correlations. Heuristically, the first replacement above maximally overlaps layers $k - 1$ and k, while the second replacement randomly overlaps them, and the two are chosen between based on the specified rank correlation between the layers. As Räisänen *et al.* note, their schemes assume non-negative inter-layer correlations, i.e. in this context, $r_{k-1,k}^S \in [0, 1]$. We will discuss this assumption below.

Let us investigate the joint S distribution and copula implied by this algorithm. For the first k layers, 1 to k, define

$$\begin{aligned} F_S^{1:k}(s_1, \dots, s_k) &\equiv \mathrm{Fr}(S_1 \leq s_1, \dots, S_k \leq s_k) \\ &= \mathrm{Fr}(V_1 \leq F_{S_1}(s_1), \dots, V_k \leq F_{S_k}(s_k)) \\ &= \mathrm{Fr}(V_1 \leq r_1, \dots, V_k \leq r_k) \\ &= C_S^{1:k}(r_1, \dots, r_k), \end{aligned} \quad (33)$$

where $r_i \equiv F_{S_i}(s_i)$. Consider first the two-layer case $k = 2$. As $N_s \to \infty$, a fraction $r_{1,2}^S$ of subcolumns will have $V_2 = V_1$, which is uniformly random on $[0, 1]$, and a fraction $(1 - r_{1,2}^S)$ will have a new random V_2, uniform on $[0, 1]$ and independent of V_1. Therefore, by (33),

$$C_S^{1:2}(r_1, r_2) = r_{1,2}^S \min(r_1, r_2) + (1 - r_{1,2}^S) r_1 r_2, \quad (34)$$

which shows the familiar linear combination of maximum and random overlap components (section 3.3). For the general $k > 2$ case,

$$\begin{aligned} C_S^{1:k}&(r_1, \dots, r_k) \\ &= r_{k-1,k}^S \, C_S^{1:k-1}(r_1, \dots, r_{k-2}, \min(r_{k-1}, r_k)) \quad (35) \\ &\quad + (1 - r_{k-1,k}^S) \, C_S^{1:k-1}(r_1, \dots, r_{k-1}) \, r_k, \end{aligned}$$

which can be used iteratively to increase from $k = 2$ to higher k. For example,

$$\begin{aligned} C_S^{1:3}(r_1, r_2, r_3) &= r_{2,3}^S \, C_S^{1:2}(r_1, \min(r_2, r_3)) \\ &\quad + (1 - r_{2,3}^S) \, C_S^{1:2}(r_1, r_2) \, r_3 \\ &= r_{1,2}^S \, r_{2,3}^S \min(r_1, r_2, r_3) \\ &\quad + (1 - r_{1,2}^S) \, r_{2,3}^S \, r_1 \min(r_2, r_3) \\ &\quad + r_{1,2}^S \, (1 - r_{2,3}^S) \min(r_1, r_2) \, r_3 \\ &\quad + (1 - r_{1,2}^S)(1 - r_{2,3}^S) \, r_1 r_2 r_3. \end{aligned} \quad (36)$$

Let $C_{S(i,j)}^{1:k}$, $i, j \in \{1, \dots, k\}$, $i < j$, denote the 2-copula formed from $C_S^{1:k}$ by marginalizing all layers but i and j, i.e. $C_{S(i,j)}^{1:k}(a, b) \equiv C_S^{1:k}(r_1, \dots, r_k)$, with $r_i = a$, $r_j = b$, and all the other r arguments equal to 1. Heuristically, $C_{S(i,j)}^{1:k}$ describes the dependence between the ranks of S in layers i and j, without regard for any of the other layers in $\{1, \dots, k\}$. Using this definition, and (36), it is easy to show that

$$C_{S(1,2)}^{1:3}(a, b) = r_{1,2}^S \min(a, b) + (1 - r_{1,2}^S) \, ab, \quad (37)$$

$$C_{S(2,3)}^{1:3}(a, b) = r_{2,3}^S \min(a, b) + (1 - r_{2,3}^S) \, ab, \quad (38)$$

and

$$C_{S(1,3)}^{1:3}(a, b) = r_{1,2}^S \, r_{2,3}^S \min(a, b) + (1 - r_{1,2}^S \, r_{2,3}^S) \, ab, \quad (39)$$

More generally, using (35), the definition of a copula in section 3.2, and a process of recursion, it can be shown that

$$C_{S(i,j)}^{1:k}(a, b) = r_{i,j}^S \min(a, b) + (1 - r_{i,j}^S) ab, \quad (40)$$

where for non-adjacent i and j, we use the generalization

$$r_{i,j}^S = r_{i,i+1}^S r_{i+1,i+2}^S \cdots r_{j-1,j}^S. \quad (41)$$

This means that any two layers have a linear combination of maximum and random overlap, with a weight equal to the product of all the interlayer rank correlations in between the layers. As we noted earlier, the Räisänen *et al.* generator assumes $r_{k-1,k}^S \in [0, 1]$, which is usually a very good assumption for adjacent layers. However, we can see that (41) therefore implies that $r_{i,j}^S \in [0, 1]$ for any pair of layers. This generator therefore prohibits anti-correlation between any layers, including well-separated

ones, even though it is at least possible for such anti-correlations to exist in nature due to complex dynamical and radiative interactions on different scales in the atmosphere. This is one advantage of the copula approach we present in this paper, since it is general enough to allow elements of minimal overlap as well.

Finally, let us consider the evaluation of the copula of the entire grid column, $C_S = C_S^{1:K}$, and the associated grid-column cloud fraction $f = 1 - C_S^{1:K}(f_1', \ldots, f_K')$. The exact method involves beginning with (35) with $k = K$ and doing recursive evaluations until $k = 2$ when (34) terminates the recursion. This will involve the calculation and summing of 2^{K-1} terms. The alternative for large K is instead to use the generator to evaluate a large number N_s of sample subcolumns and evaluate the column cloud fraction approximately from this ensemble. The error in such a cloud fraction will decrease as N_s increases, and will have similar statistical properties to errors in radiative properties calculated using the generator (e.g. Räisänen *et al.*, 2004; Räisänen and Barker, 2004).

Appendix C. Copulas for multiple cloud layers

One commonly applied assumption is that vertically separated cloud blocks (i.e. vertically contiguous blocks of at least partially cloudy model layers, separated by completely clear model layers) are randomly overlapped. We shall call this the 'Random Overlap Across Clear Layers' (ROACL) assumption. Say there are M such blocks of cloud, defined by

$$\left\{ f_1'^{(1)}, \ldots, f_{K^{(1)}}'^{(1)} \right\}, \ldots, \left\{ f_1'^{(M)}, \ldots, f_{K^{(M)}}'^{(M)} \right\},$$

such that all of the included model layers are at least partially cloudy (i.e. all $f_k'^{(j)} < 1$). The column clear fraction for block (j) is

$$f'^{(j)} = C_{\mathbf{S}}^{(j)} \left(f_1'^{(j)}, \ldots, f_{K^{(j)}}'^{(j)} \right),$$

where $C_{\mathbf{S}}^{(j)}$ is the marginal copula for block (j) layers only. Then, since the blocks are randomly overlapped, the clear fraction for the entire column is given by

$$f'_{\text{ROACL}} = \prod_{j=1}^{M} C_{\mathbf{S}}^{(j)} \left(f_1'^{(j)}, \ldots, f_{K^{(j)}}'^{(j)} \right).$$

One common subset of ROACL are the maximum–random overlap schemes. The Tian and Curry (1989) version, which we will denote MRO, specifies maximum cloud overlap within blocks, and so has

$$f'_{\text{MRO}} = \prod_{j=1}^{M} \min \left(f_1'^{(j)}, \ldots, f_{K^{(j)}}'^{(j)} \right).$$

Appendix D. Geleyn and Hollingsworth (1979) maximum-random overlap

We discussed the 'standard' version of maximum–random overlap, denoted MRO, in Appendix C. In fact, there is a much more commonly implemented variant, due to Geleyn and Hollingsworth (1979). This variant, denoted here by MROGH, has a grid-column clear fraction

$$f'_{\text{MROGH}} = f_1' \times \prod_{k=2}^{K} \frac{\min(f_k', f_{k-1}')}{f_{k-1}'}. \tag{42}$$

We propose a corresponding S generator using Räisänen *et al.* (2004) as a guide: for each sample $n \in \{1, \ldots, N_s\}$, a rank vector $\mathbf{R}^{(n)} = (V_1^{(n)}, \ldots, V_K^{(n)})^{\text{T}}$ is generated, from which $\mathbf{S}^{(n)} = (F_{S_1}^{-1}(V_1^{(n)}), \ldots, F_{S_K}^{-1}(V_K^{(n)}))^{\text{T}}$ is evaluated. For layer one, $V_1^{(n)}$ is uniformly random on [0, 1]. For each subsequent layer $k = 2 \ldots K$, if $V_{k-1}^{(n)} > f_{k-1}' = F_{S_{k-1}}(1)$, i.e. layer $k - 1$ is cloudy, then choose $V_k^{(n)} = V_{k-1}^{(n)}$, otherwise choose $V_k^{(n)}$ uniformly random on $[0, f_{k-1}']$.

Consider first the two-layer case $k = 2$. Using the same notation as in Appendix B, as $N_s \to \infty$, we seek the fraction $C_S^{1:2}(r_1, r_2)$ of subcolumns for which $V_1 \le r_1$ and $V_2 \le r_2$. $N_s f_1$ subcolumns will have $V_1 > f_1'$ and therefore $V_2 = V_1$. But only $N_s \max(0, \min(r_1, r_2) - f_1')$ subcolumns will obey the additional constraint that $V_1 \le \min(r_1, r_2)$. Conversely, $N_s f_1'$ subcolumns will have $V_1 \le f_1'$ and therefore V_2 uniformly random on $[0, f_1']$. Therefore, a fraction $\min(r_1, f_1')/f_1'$ of these subcolumns will have $V_1 \le r_1$, and an independent fraction $\min(r_2, f_1')/f_1'$ will have $V_2 \le r_2$. Consequently,

$$C_S^{1:2}(r_1, r_2) = \max(0, \min(r_1, r_2) - f_1') + \frac{\min(r_1, f_1') \min(r_2, f_1')}{f_1'}. \tag{43}$$

Hence, $f_{1:2}' \equiv C_S^{1:2}(f_1', f_2') = \min(f_2', f_1')$, as required by (42). For $k > 2$, using similar probability arguments, it can be shown that

$$\begin{aligned} C_S^{1:k}&(r_1, \ldots, r_k) \\ &= \max \Big[0, C_S^{1:k-1}(r_1, \ldots, r_{k-2}, \min(r_{k-1}, r_k)) \\ &\qquad - C_S^{1:k-1}(r_1, \ldots, r_{k-2}, f_{k-1}') \Big] \\ &\quad + C_S^{1:k-1}(r_1, \ldots, r_{k-2}, \min(r_{k-1}, f_{k-1}')) \\ &\qquad\quad \times \frac{\min(r_k, f_{k-1}')}{f_{k-1}'}, \end{aligned} \tag{44}$$

and hence

$$f_{1:k}' = C_S^{1:k}(f_1', \ldots, f_k') = \frac{f_{1:k-1}' \min(f_k', f_{k-1}')}{f_{k-1}'},$$

since any copula is non-decreasing in each of its arguments, per section 3.2 (R3). This form for $f_{1:k}'$ is consistent with (42).

References

Chakak A, Koehler KJ. 1995. A strategy for constructing multivariate distributions. *Commun. Stat. Simulat.* **24**: 537–550.

Cherubini U, Luciano E, Vecchiato W. 2004. *Copula methods in finance*. John Wiley & Sons, Ltd.

Chou M-D, Suarez MJ. 1999. 'A shortwave radiation parameterization for atmospheric studies'. NASA Tech Memo 15-104 606. NASA/Goddard: Greenbelt, Maryland, USA.

Chou M-D, Suarez MJ, Ho C-H, Yan MM-H, Lee K-T. 1998. Parameterizations for cloud overlapping and shortwave single-scattering properties for use in general circulation and cloud ensemble models. *J. Climate* **11**: 202–214.

Chou M-D, Suarez MJ, Liang X-Z, Yan MM-H. 2001. 'A thermal infrared radiation parameterization for atmospheric studies'. NASA Tech Memo 19-104 606. NASA/Goddard: Greenbelt, Maryland, USA.

Dee DP, da Silva AM. 1999. Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology. *Mon. Weather Rev.* **127**: 1822–1834.

Dee DP, Gaspari G, Redder C, Rukhovets L, da Silva AM. 1999. Maximum-likelihood estimation of forecast and observation error covariance parameters. Part II: Applications. *Mon. Weather Rev.* **127**: 1835–1849.

Demarta S, McNeil AJ. 2005. The *t* copula and related copulas. *Internat. Statist. Rev.* **73**: 111–129.

Drezner Z. 1994. Computation of the trivariate normal integral. *Math. Comput.* **63**: 289–294.

Drezner Z, Wesolowsky GO. 1989. On the computation of the bivariate normal integral. *J. Statist. Comput. Simulat.* **35**: 101–107.

Dupuis DJ. 2007. Using copulas in hydrology: Benefits, cautions, and issues. *J. Hydrol. Eng.* **12**: 381–393.

Gaspari G, Cohn SE. 1999. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **125**: 723–757.

Gaspari G, Cohn SE, Guo J, Pawson S. 2006. Construction and application of covariance functions with variable length-fields. *Q. J. R. Meteorol. Soc.* **132**: 1815–1838.

Geleyn J-F, Hollingsworth A. 1979. An economical analytic method for the computation of the interaction between scattering and line absorption of radiation. *Beitr. Phys. Atmos.* **52**: 1–16.

Genest C, Favre A-C. 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.* **12**: 347–368.

Genz A. 2004. Numerical computation of rectangular bivariate and trivariate normal and *t* probabilities. *Statist. Comput.* **14**: 251–260.

Genz A, Bretz F. 1999. Numerical computation of multivariate *t* probabilities with application to power calculation of multiple contrasts. *J. Statist. Comput. Simulat.* **63**: 361–378.

Genz A, Bretz F. 2002. Comparison of methods for the computation of multivariate *t* probabilities. *J. Comput. Graph. Statist.* **11**: 950–971.

Golaz J-C, Larson VE, Cotton WR. 2002. A PDF-based model of boundary layer clouds. Part I: Method and model description. *J. Atmos. Sci.* **59**: 3540–3551.

Hogan RJ, Illingworth AJ. 2000. Deriving cloud overlap statistics from radar. *Q. J. R. Meteorol. Soc.* **126**: 2903–2909.

Jakob C, Klein SA. 1999. The role of vertically varying cloud fraction in the parametrization of microphysical processes in the ECMWF model. *Q. J. R. Meteorol. Soc.* **125**: 941–965.

Larson VE. 2007. From cloud overlap to PDF overlap. *Q. J. R. Meteorol. Soc.* **133**: 1877–1891.

Larson VE, Wood R, Field PR, Golaz J-C, Vonder Haar TH, Cotton WR. 2001. Small-scale and mesoscale variability of scalars in cloudy boundary layers: One-dimensional probability density functions. *J. Atmos. Sci.* **58**: 1978–1994.

Mace GG, Benson-Troth S. 2002. Cloud-layer overlap characteristics derived from long-term cloud radar. *J. Climate* **15**: 2505–2515.

Monaghan JJ, Lattanzio JC. 1985. A refined particle method for astrophysical problems. *Astron. Astrophys.* **149**: 135–143.

Nelsen RB. 2006. *An Introduction to Copulas*. 2nd ed. Springer Science + Business Media, LLC: New York.

Norris PM, da Silva AM. 2007. Assimilation of satellite cloud data into the GMAO Finite Volume Data Assimilation System using a parameter estimation method. *J. Atmos. Sci.* **64**: 3880–3895.

Oreopoulos L, Khairoutdinov M. 2003. Overlap properties of clouds generated by a cloud-resolving model. *J. Geophys. Res.* **108**: 4479, DOI: 10.1029/2002JD003329.

Pincus R, Hannay C, Klein SA, Xu K-M, Hemler R. 2005. Overlap assumptions for assumed-PDF cloud schemes in large-scale models. *J. Geophys. Res.* **110**: D15S09.

Räisänen P, Barker HW. 2004. Evaluation and optimization of sampling errors for the Monte Carlo Independent Column Approximation. *Q. J. R. Meteorol. Soc.* **130**: 2069–2085.

Räisänen P, Barker HW, Cole JNS. 2005. The Monte Carlo Independent Column Approximation's conditional random noise: Impact on simulated climate. *J. Climate* **18**: 4715–4730.

Räisänen P, Barker HW, Khairoutdinov MF, Li J, Randall DA. 2004. Stochastic generation of subgrid-scale cloudy columns for large-scale models. *Q. J. R. Meteorol. Soc.* **130**: 2047–2068.

Rienecker MM, Suarez MJ, Todling R, Bacmeister J, Takacs L, Liu H-C, Gu W, Sienkiewicz M, Koster RD, Gelaro R, Stajner I, Nielsen E. 2007. 'The GEOS-5 Data Assimilation System – Documentation of versions 5.0.1, 5.1.0, and 5.2.0'. Technical Report Series on Global Modeling and Data Assimilation NASA/TM-2007-104 606, **27**: Suarez MJ (ed.) National Aeronautics and Space Administration: Available at http://gmao.gsfc.nasa.gov/pubs/docs/GEOS-5.0.1_-Documentation_r3.pdf.

Salvadori G, De Michele C. 2007. On the use of copulas in hydrology: Theory and practice. *J. Hydrol. Eng.* **12**: 369–380.

Smith RNB. 1990. A scheme for predicting layer clouds and their water content in a general circulation model. *Q. J. R. Meteorol. Soc.* **116**: 435–460.

Snedecor GW, Cochran WG. 1980. *Statistical Methods*, 7th ed. Iowa State Univ. Press: Ames, Iowa, USA.

Tian L, Curry JA. 1989. Cloud overlap statistics. *J. Geophys. Res.* **94**: 9925–9935.

Tompkins AM. 2002. A prognostic parameterization for the subgrid-scale variability of water vapor and clouds in large-scale models and its use to diagnose cloud cover. *J. Atmos. Sci.* **59**: 1917–1942.

Xu K-M, Randall DA. 1996. Evaluation of statistically based cloudiness parameterizations used in climate models. *J. Atmos. Sci.* **53**: 3103–3119.

Zeng X, Tao W-K, Zhang M-H, Peters-Lidard C, Lang S, Simpson J, Kumar S, Xie S-C, Eastman JL, Shie C-L, Geiger JV. 2007. Evaluating clouds in long-term cloud-resolving model simulations with observations. *J. Atmos. Sci.* **64**: 4153–4177.