# Scale Localization of Cloud Particle Clustering Statistics

MICHAEL L. LARSEN

*Department of Physics and Astronomy, College of Charleston, Charleston, South Carolina*

(Manuscript received 4 January 2012, in final form 26 June 2012)

ABSTRACT

Recent work has examined the spatial distribution of droplets within a cloud. Experimentally, most studies analyze interevent times from static probes flown linearly through a cloud, allowing the spatial information to be conveyed through a time series of particle detections. Analysis of these data has shown unequivocally that most clouds have nontrivial spatial structure. There is still debate as to which of many different possible statistical descriptions is most appropriate for characterizing this spatial structure and for identifying spatiotemporal scales of physical significance from the data record. This paper seeks to clarify some pervasive misunderstandings and to outline more carefully the range of validity for several of these statistical tools. Simulations are used to explore the relative scale-localizing capabilities of various commonly used statistical tools.

## 1. Introduction

The spatial distribution of cloud droplets on scales from microns to kilometers has implications for many different processes in cloud physics including enhanced droplet growth via collision–coalescence (e.g., Kostinski and Shaw 2005) and nonexponential radiative transmission (e.g., Kostinski 2001). Deviations from perfect spatial randomness (often called clustering) may be generated by a number of different physical mechanisms including turbulent mixing, particle shattering, and inertial clustering; for a detailed discussion of these mechanisms, see Shaw (2003).

Although the basic understanding of the physical mechanisms causing clustering is constantly improving due to direct numerical simulations and laboratory investigations, the characterization of deviations from perfect randomness for real clouds in the atmosphere remains primarily statistical in nature. On some of the smallest scales of physical significance, the discreteness of the particles becomes important—both in terms of the processes involved and the description of the spatial distribution. Therefore, any ''mean field'' understanding that we gain through investigating underlying physical mechanisms becomes of secondary importance to quantifying existing variability (due to a combination of multiple physical clustering mechanisms in conjunction with discrete shot noise).

Characterizing existing deviations from perfect randomness for cloud particle time series goes back at least as far as Baker (1992) and has been an active field for discussion ever since. Other papers on the general topic of characterizing deviations from perfect spatial randomness among cloud droplets include—but are certainly not limited to—Shaw et al. (1998), Kostinski and Jameson (2000), Chaumat and Brenguier (2001), Kostinski and Shaw (2001), Pinsky and Khain (2001), Shaw et al. (2002), Shaw (2003), Knyazikhin et al. (2005), Marshak et al. (2005), Larsen (2006), Lehmann et al. (2007), and Baker and Lawson (2010). Additionally, characterizing deviations from perfect randomness has become part of the existing literature in the aerosol community (e.g., Preining 1983; Larsen et al. 2003; Larsen 2006, 2007) and rain communities (e.g., Kostinski and Jameson 1997; Lavergnat and Gole 1998; Uijlenhoet et al. 1999; Jameson and Kostinski 2000; Lovejoy et al. 2003; Larsen et al. 2005; Kostinski et al. 2006; Larsen 2006; Larsen et al. 2010).

Some of the above-cited papers are devoted to discussing the theoretical tools that can be used to characterize particulate clustering, whereas other papers focus on characterizing deviations from randomness in experimental data. Methods used to interpret a time

*Corresponding author address:* Michael L. Larsen, Department of Physics and Astronomy, College of Charleston, Charleston, SC 29424.
E-mail: larsenml@cofc.edu

DOI: 10.1175/JAS-D-12-02.1

series of droplet detection times include the Fishing statistic, the clustering index, the autocorrelation function, the pair-correlation function, the volume-averaged pair-correlation function, power spectral analysis, and various forms of fractal data analysis. There have also been some studies (e.g., Shaw et al. 2002; Shaw 2003; Larsen 2006; Baker and Lawson 2010) that have attempted to examine multiple statistical techniques in the same paper so as to show connections between the methods or to argue for the primacy of one tool over others for a particular set of measured or simulated data. Given the large amount of attention this problem has received from a varied list of investigators proposing and comparing assorted tools to describe what are often subtle statistical deviations, several misunderstandings have crept into the literature.

This paper is presented to (i) review some of the basic statistical tools used to characterize deviations from perfect randomness in clouds while maintaining an internally consistent and simple notational structure, (ii) correct some of the pervasive mistakes in definitions used in the prior literature, and (iii) analyze and interpret the ability of different statistical tools to identify scale-localized deviations from perfect spatial randomness. To accomplish these goals, we first review the commonly used discrete measures of statistical structure in the atmosphere. Then, mathematical connections between these measures are summarized. Finally, these measures are used to analyze simulated systems with inherent scale-dependent behavior to compare the relative utility of these tools in identifying relevant embedded scales in a physical system. We find that a careful understanding of statistical stationarity/homogeneity as well as particular care in interpreting statistical results is critical when examining cloud droplet structures.

## 2. Review of frequently used tools

A complete review of all the common theoretical statistical tools used for particle characterization is beyond the scope of this paper; an excellent summary can be found in Shaw (2003). Here, we choose just a few discrete measures to review that have been recently used by Baker and Lawson (2010). A more complete description of these tools can be found in Shaw et al. 2002, Shaw 2003, Larsen 2006, and Baker and Lawson 2010. Note that there are some measures discussed in these cited papers that are not presented here; we have explicitly limited ourselves to *discrete* measures of particle clustering for this treatment. The other measures (including the autocorrelation function, power spectrum, and scaling analysis approaches) may have substantial merit—depending on context—but are beyond the scope of systems that are easily examined using the simulation tools utilized later in this manuscript.

For notational consistency, we will assume that cloud drop detections form a time series $\{x_i\}$, where $x_{i+1} > x_i$: there are a total of $N$ drops detected; and the time between the first drop detection and the final drop detection (e.g., $x_N - x_1$) is equal to $T$. The number of detections per time $N/T$ will be written as $\lambda$, and $t$ represents some variable amount of time $t < T$.

### a. Clustering index

The clustering index, written $\mathrm{CI}(t)$, is the first of several statistical tools that rely on binning the data at a variable scale. The total measurement time $T$ is divided into $m$ equal intervals of duration $T/m = t$. After binning the data into these intervals, we define the quantities $V_t$ and $M_t$ as the variance and mean associated with each interval (note that $M_t$ can also be written as $\lambda t$). The clustering index, then, can be written as

$$\mathrm{CI}(t) = \left(\frac{V_t}{M_t} - 1\right). \tag{1}$$

For a perfectly random (Poisson) dataset, $V_t = M_t \; \forall \; t$ and, therefore, $\mathrm{CI}(t) \equiv 0$ for a statistically stationary Poisson time series.

### b. Fishing statistic

The Fishing statistic was introduced to the cloud droplet detection community by Baker (1992). It is structured similarly to the clustering index, but scaled differently. It is designed to be a hypothesis-testing statistic to determine whether a set of data points is Poisson distributed (Baker and Lawson 2010). Using the same notation developed above, we can write the Fishing statistic at scale $t$ via

$$F(t) = \left[\frac{V_t/M_t - 1}{\sqrt{2/(m-1)}}\right] = \left[\frac{V_t/M_t - 1}{\sqrt{2/(T/t-1)}}\right] = \left(\frac{V_t}{M_t} - 1\right)\sqrt{\frac{T-t}{2t}}. \tag{2}$$

The Fishing statistic also gives a value of 0 at all values of $t$ for a stationary Poisson distribution, but has a $t$-dependent weighting factor to help $F$ from increasing with $t$ for non-Poisson (clustered) systems.

In Baker (1992) and Baker and Lawson (2010), the algorithm for computing and presenting the Fishing statistic was modified so as to filter out some of the artificial signal introduced when both the particle placement simulation and analysis code share the same time origin. For example, if a nonstationary signal with inherent periodicity

$\tau$ is simulated, then strange—but understandable—signals for $F(t)$ can be observed; a local maximum or minimum in $F(t)$ at $t = \tau$ (and multiples of $\tau$) is often found. This observed minimum or maximum is present at $t = \tau$ due to both properties of the time series itself as well as arising spuriously due to the binning intervals coinciding exactly with the changes in local concentration of particles that create the statistical change in the local mean (nonstationarity). To remove some of the spurious signal arising from both the simulated time series and the binning-based analysis matching, the modified Fishing statistic averages $F(t)$ calculated by defining the time origin at several different locations (only one of which corresponds to the time origin for the simulation). Further explanation can be found in Baker (1992) and Baker and Lawson (2010). Following these previous studies, the modified Fishing statistic will be used in this work.

### c. *Pair-correlation function*

The pair-correlation function has been discussed at great length elsewhere [see, in particular, Shaw et al. (2002); Larsen (2006) for thorough introductions to the statistical tool and its use]. Most physically, it can be introduced as the scale-localized deviation from a stationary Poisson distribution.

Given a particle detected at some time $t_\circ$, the probability of a particle detection in some later time window $(t_\circ + t, t_\circ + t + dt)$ would, for a stationary Poisson distribution, be given by

$$[p(t_\circ + t \,|\, t_\circ)dt]_{\text{Poisson}} = \lambda dt, \tag{3}$$

since the presence or absence of a particle at $t_\circ$ has no bearing on the presence of a particle in $(t_\circ + t, t_\circ + t + dt)$ for a stationary Poisson distribution. For a statistically stationary (but not Poisson) distribution, however, we may still have an enhanced or diminished probability of finding a second particle in the specified window if the system is ''clumpy'' or ''correlated.'' We therefore modify the relationship to read, in general,

$$p(t_\circ + t \,|\, t_\circ)dt = \lambda dt[1 + \eta(t)], \tag{4}$$

with $\eta(t)$ being the pair-correlation function evaluated at $t$. This definition can be inverted to formulate an expression for $\eta(t)$:

$$\eta(t) = \frac{p(t_\circ + t|t_\circ)}{\lambda} - 1. \tag{5}$$

Finding $p(t_\circ + t|t_\circ)$ takes some computational effort, but there are straightforward links to obtaining an expression for this via use of the empirical $k$th nearest-neighbor functions (see, e.g., Picinbono and Bendjaballah 2005; Larsen and Kostinski 2009). For an empirical or analytical $k$th nearest-neighbor function, we can write the pair-correlation function as (Larsen and Kostinski 2009)

$$\eta(t) = -1 + \frac{1}{\lambda dt} \sum_{k=1}^{\infty} f_k(t), \tag{6}$$

where $f_k(t)$ is the probability that the $k$th particle detection posterior to a given particle detection can be found in the interval $(t, t + dt)$ after the given particle detection. Though formally this requires adding together an infinite number of distribution functions, empirical results obtained are very accurate if $k \approx 20t/\tau$ is used. (Note that this requires a different number of terms to the summation for longer $t$. In the results that follow, this only requires $k \sim 40$ for excellent convergence. Throughout all simulation results, $k = 1000$ was used so that this limit was not a concern.) With current computer hardware, this is by no means a difficult task for even the longest experimental data records (MATLAB code for calculating the pair-correlation function for a time series with this algorithm can be obtained by contacting the author).

Note that this method of calculating the pair-correlation function is slightly different than that described in Eq. (4) of Baker and Lawson (2010), but it does not rely on the use of the volume-averaged pair-correlation function developed in Shaw et al. (2002). This is important because, as described in the following section, there are existing misconceptions about the use of the volume-averaged pair-correlation function and the links between the pair-correlation function and the basic statistics of $V_t$ and $M_t$ as defined above.

As can be seen from the physical definition or by using the closed-form expressions for the $k$th nearest-neighbor distribution functions for a stationary Poisson system, the pair-correlation function $\eta(t)$ is identically 0 for all $t$ in the case of a stationary Poisson system. It bears noting, however, that the pair correlation is *scale localized*: $\eta(t)$ depends only on the presence or absence of particles separated by $t \pm dt$ in time. Though both the clustering index and Fishing index give values of zero for all scales from a stationary Poisson dataset, they both can have inflated values at scale $t$ due to structure on scales smaller than $t$ [this will be shown explicitly through example in the appendix and has been demonstrated in previous works, including Shaw et al. (2002) and Larsen (2006)].

Physically, if $\eta(t) > 0$ for some $t$, it indicates that there is a larger-than-Poisson expectation to find particles separated by times $t \pm dt$; similarly, for $\eta(t) < 0$, there is a smaller-than-Poisson expectation to find particles separated by times $t \pm dt$. When mutual detection is

impossible (due to, e.g., instrumental dead time), $\eta(t) = -1$ (Larsen and Kostinski 2009).

### d. Scaled clustering index

The volume-averaged pair-correlation function (VAPC) statistic was introduced in Shaw et al. (2002) and used again in several other studies including Shaw (2003) and Baker and Lawson (2010). The measure, as defined by Baker and Lawson, can be written as

$$\mathrm{VAPC}(t) = \frac{1}{M_t}\left(\frac{V_t}{M_t} - 1\right) = \frac{1}{M_t}[\mathrm{CI}(t)]. \tag{7}$$

This measure has been previously used because it is computationally easier to measure than the pair-correlation function (at the expense of not being a localized measure of deviations from pure randomness). The motivation for introducing the volume-averaged pair-correlation function was through the correlation-fluctuation theorem.

As was shown in Larsen (2006) and Larsen and Kostinski (2009), however, use of the volume-averaged pair correlation function [sometimes written as $\overline{\eta}(t)$] was, at the very least, a poorly chosen term. In truth, Eq. (7) is misleading; VAPC must be defined differently than stated in Shaw et al. (2002) for Eq. (7) to be true. Several other papers have propagated this error, and it has caused some confusion in this discussion.

To help clarify the issue, the term ''scaled clustering index'' [SCI($t$)] for the quantity CI($t$)/$M_t$ is proposed here. Following the development in Larsen (2006), it can be shown that this can be written (in one dimension) as

$$\mathrm{SCI}(t) = \frac{\mathrm{CI}(t)}{M_t} = \frac{2}{t^2}\int_0^t (t - t')\eta(t')\,dt'. \tag{8}$$

SCI($t$)—like CI($t$), $F(t)$, and $\eta(t)$—takes a value of zero for a stationary Poisson (perfectly random) system.

Practically speaking, most previous studies that referred to the volume-averaged pair-correlation function [or, equivalently, $\overline{\eta}(t)$] ought to have used SCI($t$) instead. Use of the inaccurate formula given in Shaw et al. (2002) and elsewhere leads to some qualitative mistakes in interpretation: SCI($t$) actually has even more scale memory than $\overline{\eta}(t)$ since the integrand of SCI($t$) is weighted with a more substantial contribution from $t'$ near 0 than from $t'$ near $t$. [In $\overline{\eta}(t)$ all scales smaller than $t$ contribute equally to the statistic]. Many other studies that refer to VAPC or $\overline{\eta}(t)$ have some formula errors when connecting $\eta(t)$ to variance/mean ratio through the use of $\overline{\eta}$. Nevertheless, the qualitative results from these studies are still often valid. In fact, owing to the different weighting factor, SCI($t$) has even *more* pronounced memory effects than previous discussions of VAPC/$\overline{\eta}$ would lead the reader to believe. This author urges the reader not to hastily dismiss results that used the erroneous link between $\eta$ and variance/mean ratio through $\overline{\eta}(t)$ as being wholly wrong, but to reinterpret these results carefully to determine what results still hold true by using the correct relationship stated in terms of SCI($t$) above.

## 3. Connections between $F$, CI, SCI, and $\eta$

The principle goal of Shaw et al. (2002) was to demonstrate and establish the connections between the statistics commonly used in characterizing the structure of cloud droplet detection time series. Because of the confusion/errors in the correlation-fluctuation theorem that have propagated through the literature, repeating the final (corrected) relationships connecting $F$, CI, SCI, and $\eta$ would be useful in future research. Following the development established in Larsen (2006), the corrected relationships (for a time series) are

$$F(t) = [\mathrm{CI}(t)]\sqrt{\frac{T-t}{2t}} = \lambda[\mathrm{SCI}(t)]\sqrt{\frac{t(T-t)}{2}} = \frac{\lambda}{t}\sqrt{\frac{2(T-t)}{t}}\int_0^t (t-t')\eta(t')\,dt', \tag{9}$$

$$\mathrm{CI}(t) = F(t)\sqrt{\frac{2t}{T-t}} = \lambda t[\mathrm{SCI}(t)] = \frac{2\lambda}{t}\int_0^t (t-t')\eta(t')\,dt', \tag{10}$$

$$\mathrm{SCI}(t) = \frac{F(t)}{\lambda}\sqrt{\frac{2}{t(T-t)}} = \frac{\mathrm{CI}(t)}{\lambda t} = \frac{2}{t^2}\int_0^t (t-t')\eta(t')\,dt'. \tag{11}$$

Inverting these relationships to develop expressions for $\eta$ in terms of $F$, CI, and SCI proves daunting (see discussion in Larsen 2006). Even if closed-form expressions can be written, we would find that $\eta(t)$ would be related to the time derivative of $F$, CI, and SCI. Since numerical derivatives are generally not well structured, obtaining $\eta(t)$ from measurements of the other variables is not usually practical. We find below, however, that we can often write $\eta(t)$ itself based on physical considerations and then perform a fairly straightforward integration to get all of the other variables of interest [including those not discussed in detail here, see, e.g., Shaw et al. (2002)].

## 4. Simulated clouds

Baker and Lawson (2010) explored some simulated datasets in an effort to more fully develop and understand the practical usage of $F(t)$, CI($t$), SCI($t$), and $\eta(t)$. The

conclusion was drawn that for one particular simulated system (one with periodic concentration fluctuations), the Fishing statistic most effectively identified the embedded scale of fluctuations. This system of periodic concentration change was likely used because (i) it was easy to simulate and understand and (ii) there was an unambiguous temporal scale of relevance that the different statistical tools could be used to try and identify. Below, we reexplore the same simulation; we also examine the conclusions drawn previously about this system. Finally, we extend the analysis technique to a different system that still has an unambiguous temporal scale of relevance to identify but is statistically stationary (the issue of statistical stationarity, and how it influences the validity and utility of the different tools in question here is addressed in appendix B).

### *a. Fluctuating concentration*

The starting point for analysis in Baker and Lawson (2010) is a model that attempts to simulate airplane drop concentration data with a clustering structure on scales smaller than clearly visible from examination of the raw time series. In this model, there is a fluctuating concentration that takes on a value $\lambda_1$ for the first centimeter of simulated flight, $\lambda_1 + \lambda_2$ for the second centimeter, $\lambda_1$ for the third centimeter, and $\lambda_1 + \lambda_2$ for the fourth centimeter, etc. The simulated data was 2 m long with this embedded statistical structure. Though Baker and Lawson left their analysis in terms of distances, here we convert the dataset to a measured time series for consistency with the notation in the previous sections. (In practice, this is often what is actually done with cloud particle data. A time series of particle ''arrivals'' at the instrument is recorded and then inverted using the plane's constant velocity to infer the spatial positions along the flight path where each particle was found. Here, we take the spatial data and interpret it as a time series in the same way. Instead of a fluctuation concentration on scales equal to 1 cm, we have the concentrations occurring at time on scales of $\tau = 0.01$.)

Using this basic model, Baker and Lawson (2010) examined the (modified) Fishing statistic, clustering index, scaled clustering index, pair-correlation function, and power spectrum of this simulated data [see Figs. 2 and 12 in Baker and Lawson (2010)]. Based on the computational results shown there, Baker and Lawson (2010, p. 3357) claimed that ''...we show that for periodic structure the PC [pair correlation] is less local than *F* or the power spectrum.'' Below, we will show that this conclusion is misleading and more of a consequence of the parameters chosen for the model than true in general for a periodic concentration fluctuation. In fact, a closed-form expression for the pair-correlation function

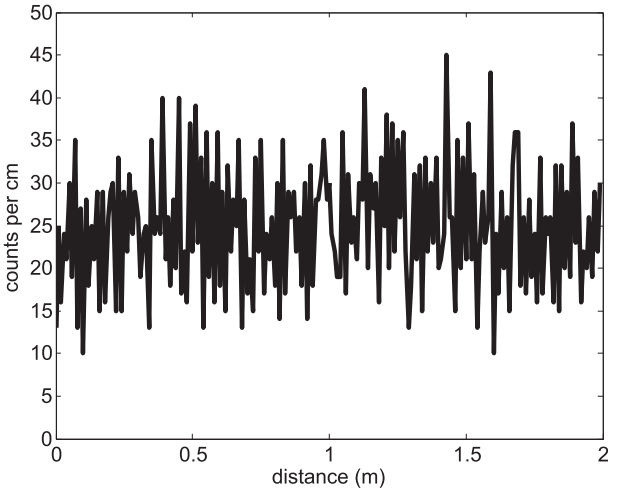


FIG. 1. One realization of a simulated one-dimensional collection of cloud droplets. The *y* axis shows how many particles are found in the current centimeter of the cloud. Following Baker and Lawson (2010), the underlying fluctuating concentration varies from 2000 to 3000 $m^{-1}$ every centimeter. At this resolution, the structure is not necessarily evident [cf. to Fig. 1b of Baker and Lawson (2010)].

for this model can be developed that clearly shows the scales of physical relevance (see appendix A). Through the relations above linking the pair-correlation function with the Fishing statistic, clustering index, and scaled clustering index–closed-form expressions for these variables, too, can be developed. These expressions, which result from *physical considerations/interpretations of the statistical structure*, yield far deeper insight than any statistical treatment involving tests of significance could allow. The closed-form expression even reveals structures clearly visible on the figures displayed in Baker and Lawson that were left without comment in their manuscript, such as the periodicity in $\eta(t)$ clearly visible in Fig. 12 in Baker and Lawson (2010).

It bears repetition, however, that this computational model is explicitly statistically nonstationary. Formally speaking, this means that some of the common statistical tools (e.g., autocorrelation function, pair-correlation function) ought not be used on this system. This is a distinction that is widely ignored, but important (the author here will not be immune to this error, either—in appendix A and below, analysis is done using the pair-correlation function for this system where we *know* the time series is statistically nonstationary). Interpretation of the consequences of using statistically stationary tools on a statistically nonstationary dataset is described in appendix B.

Figure 1 shows a realization of simulated drop counts as a function of distance traveled through a simulated cloud using the same parameters designed in Baker and Lawson (2010) ($\lambda_1 = 2000\ m^{-1}, \lambda_2 = 1000\ m^{-1}$, $T = 2$ m,

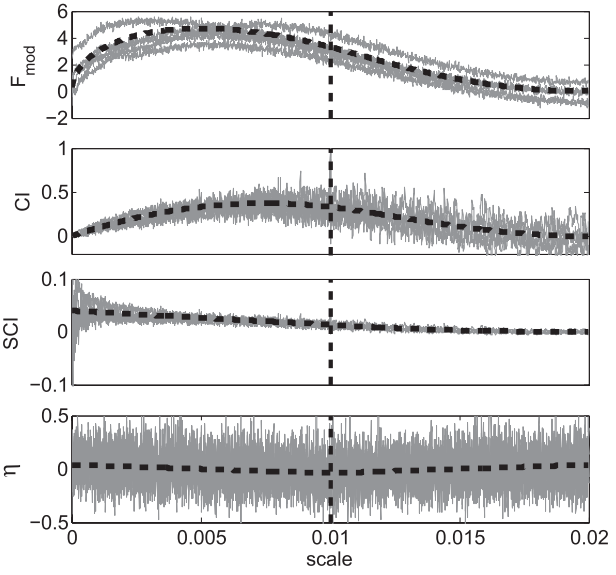


FIG. 2. Plots of the modified (from top to bottom) Fishing statistic, clustering index, scaled clustering index, and pair-correlation function as a function of scale for the fluctuating concentration model. The scale of interest $\tau$ is marked with a dashed vertical line. The light gray curves are five different realizations with the same model parameters, and the dotted lines are the theoretical expressions derived in appendix A. These results are essentially identical to the results shown in Figs. 2 and 12 of Baker and Lawson (2010), though on a linear (instead of logarithmic) $x$ axis.

$\tau = 0.01$ m). Figure 2 shows (from top to bottom) $F$, CI, SCI, and $\eta$ for five different realizations with the same parameters in gray. The dotted lines in each panel correspond to the theoretical results derived in appendix A (the $x$ axis for all of the panels is the same).

Figure 2 accurately recreates information from Figs. 2 and 12 in Baker and Lawson (2010), though there are some minor differences; here, the $x$ axis was chosen to be linear and only data from $0 \leq t \leq 2\tau$ is shown to more clearly identify the scale of interest ($\tau$). Note that the qualitative and quantitative results here are very comparable to those shown in the earlier work.

It is of particular note that, though it was argued in Baker and Lawson (2010) that $F$ accurately is able to identify the scale of physical relevance, the curve for $F$ actually reaches a maximum near $\tau/2$. This is harder to see on a semilog plot, but it is clear here that none of the curves has a maximum at $t = \tau$ (this could also be demonstrated by looking for the zero of the derivative of the closed-form expression derived in appendix A for each of the variables in question, which results in a value very close to $\tau/2$ when using these parameters).

To further explore this system, five more realizations of the simulation were carried out and analyzed. Figure 3 shows the spatial fluctuation on centimeter scales (similar to Fig. 1), and Fig. 4 shows the same statistical measures

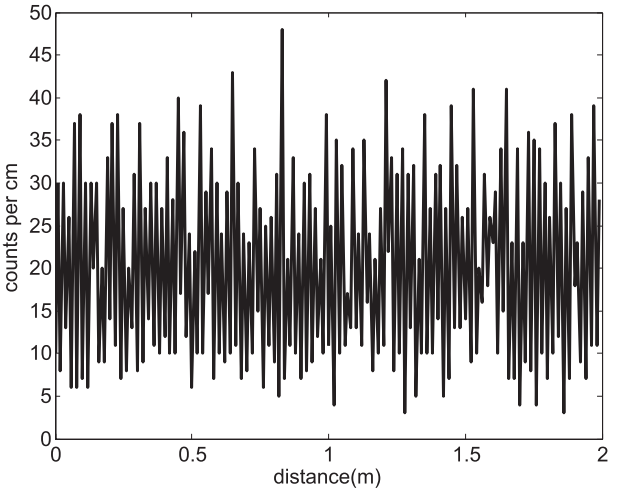


FIG. 3. As in Fig. 1, but the concentration varies from 1000 to 3000 $m^{-1}$.

used above. These realizations are structured similarly to those discussed in Baker and Lawson, except that here $\lambda_1 = 1000\ m^{-1}$ and $\lambda_2 = 2000\ m^{-1}$. This has the result of lowering the total number of particles and making the clustering more pronounced.

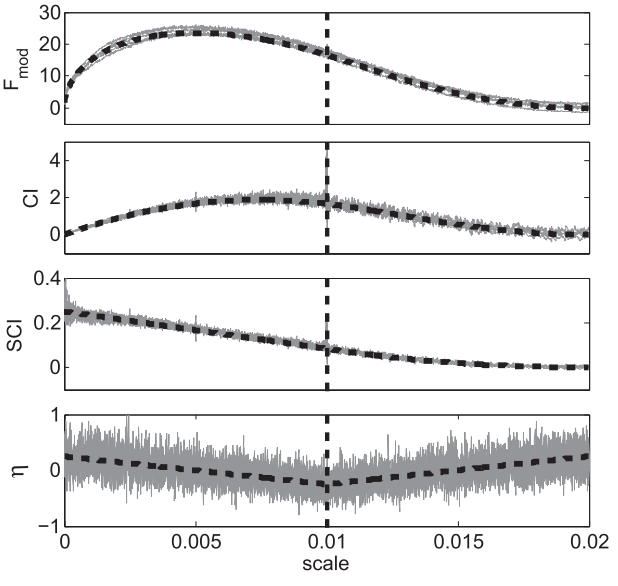


FIG. 4. A plot of $F$, CI, SCI, and $\eta$ as a function of scale for the fluctuating concentration simulations with the modified concentration fluctuations described in Fig. 3. Note that the more stringent clustering makes some statistical properties more apparent and significantly narrows the statistical noise in all of the panels. As before, five independent realizations are given in gray and the theoretical expression derived in appendix A is shown as a dotted black line. The scale of interest is marked with a dashed vertical line.

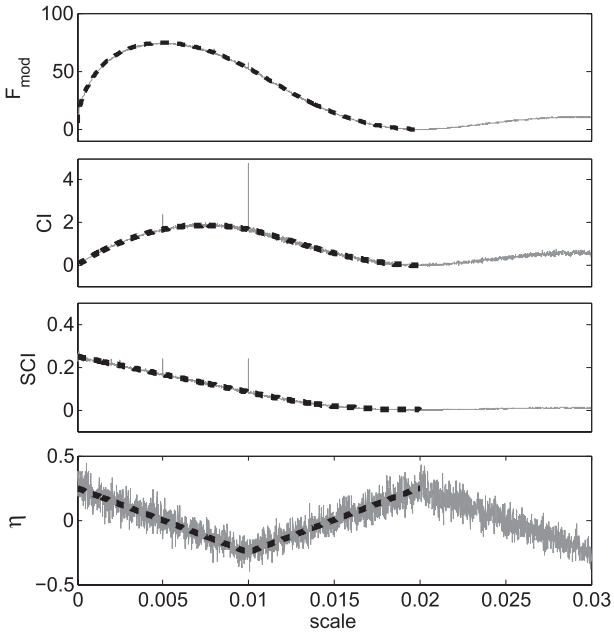


FIG. 5. As in Fig. 4, but for a single realization with a length of 20 units (instead of the 2 units used for Figs. 2 and 4). Note that the deviations from the expected closed-form solutions are less pronounced, yet this system still has very realistic $4 \times 10^4$ particles (many experimental datasets are far longer than this). These plots show clearly that the pair-correlation function is the only statistic with meaningful behavior change at the relevant scale $\tau = 0.01$. The spikes in CI and SCI for the simulated data occur for the same reason that $F_{\mathrm{mod}}$ is used instead of $F$: the artificial coalignment of the analysis and simulation origins cause a spurious spike at the simulation scale.

Merely by changing $\lambda_1$ and $\lambda_2$ in the simulation, we can clearly see that some of the statistical significance tests used in Baker and Lawson (2010) are of questionable validity. Figures 5, 12, 13, and 15 of that work imply that the Fishing statistic was a useful measure of clustering on a particular scale because whenever $F$ exceeded a relatively constant value (nominally 3), there was solid evidence for clustering on the scale in question. In Fig. 4 here, however, we note that $F$ exceeds this value of 3 over almost the entire domain examined. It is also with these simulations that we see more clearly the structure in the pair-correlation function that was not easily observed in Fig. 2. Note that, of the four statistics of interest, *only the pair-correlation function* has an unambiguous behavior change at the scale of relevance (again, at scale = 0.01).

Figure 5 shows a single realization of a simulation with the same parameters used in Figs. 3 and 4, with the only change that the distribution was allowed to have a length of 20 units (meters) instead of 2 m. Comparison of Figs. 5 and 4 shows some interesting behavior; note that, despite that the underlying distribution is functionally identical, the magnitude of the Fishing statistic grows substantially, while the other three statistics maintain their previous shape/values (this is due to the $\sqrt{T-t}$ term in the Fishing statistic, which causes the entire function to grow by about $\sqrt{20\,\mathrm{m}/2\,\mathrm{m}} \sim 3.2$). Note also that the difficulty in effectively identifying trends in the pair-correlation function is diminished with a longer sample [further noisiness in the pair-correlation function could be reduced by a smoothing technique similar to that suggested by Baker and Lawson (2010) if necessary]. The fits to the theoretical expressions are so ideal that the $x$ axis was extended beyond the range calculated to help the reader see the simulated curve.

The scale of relevance for Figs. 2, 4, and 5 are all 0.01—only the pair-correlation function shows any change in behavior at that scale, which makes sense since $F$, CI, and SCI are weighted integrals of the pair-correlation function and therefore will have maxima at different times. Although the pair-correlation function is noisier than the other statistics (because, in fact, it is not smoothed through integration and is scale localized), there are ways to mitigate this in data analysis—for example, manipulation of the value of $dt$ used ($1 \times 10^{-6}$ throughout this entire report) and smoothing the curve. Use of the other statistics may result in erroneous attribution of scales of physical significance, though all seem to be sufficient to identify scale localization to within an order of magnitude (at least for this particular system).

### b. *Matérn cluster process*

As repeatedly mentioned above, one of the weaknesses in using a periodically fluctuating concentration model is that—by construction—it is inhomogeneous/nonstationary. Baker and Lawson attempted to rectify this error by using a ''random vortices'' model, but careful analysis reveals that this model also falls prey to the same ultimate criticism—the fluctuations occur on a discrete time grid with a coincident origin as the simulation system. Ultimately, the random vortices model is also statistically nonstationary.

Fortunately, there is a class of doubly stochastic spatial point-process models that are explicitly homogeneous (or, when simulated in one dimension and assigned to a time domain, stationary). These doubly stochastic Poisson point processes [often called Cox processes, e.g., Grandell (1976); Martinez and Saar (2002)] have a subcategory of Neyman–Scott fields (following Neyman and Scott 1958). Neyman–Scott fields are convenient for simulating spatial and temporal systems because they have an analytically tractable conditional probability structure that allows for the explicit calculation of the closed-form pair-correlation

function. These systems are *statistically stationary/homogeneous*, yet still have a *unique spatial scale of interest* and therefore are ideal for the analysis carried out here.

In an effort to explore which of the four statistics in question does the best job of localizing this scale, the Matérn cluster process was simulated. The basic model is rather simple. First, some number $N_{\text{par}}$ of "parent" particles are distributed perfectly randomly (following Poisson statistics) in the interval $(0, T)$ and ultimately placed at positions $x_i$. In the interval $(x_i \pm S, S$ being a range around each parent particle relevant to the system) a random number $(N_D)_i$ of daughter particles are placed at random. The number of daughter particles for each parent is drawn from a Poisson distribution with a mean value $D$ specified in advance. The daughter particles (alone) make up the final collection of particles [more information about the simulation and properties of a Matérn cluster process—and other Neyman-Scott fields—can be found in Cox and Isham (1980), Stoyan and Stoyan (1994), Martinez and Saar (2002), and Larsen (2006)].

The parameters of this model are $T$, $D$, $S$, and $N_{\text{par}}$. Periodic boundary conditions are utilized so that the system remains statistically stationary (otherwise, the times near the origin and $T$ would be disfavored). The total number of particles in this system is close to $N_{\text{par}}\, D$, and there will be clusters sprinkled throughout the volume of size/duration $2S$ (the unequivocal characteristic time scale of this system). Clusters may overlap, and the cluster centers—since they arrived out of a random distribution in the domain—can occur anywhere, not on a preset grid. The global mean $\lambda$ for this system is $N_{\text{par}}\, D/T$. By specifying the number of parents, we specify the number of clusters in the system—but the clusters may be anywhere.

Physically, a Matérn cluster process may have some physical significance in describing galaxy formation (Martinez and Saar 2002) or, of more atmospheric interest, droplet breakup (Larsen 2006). For our purposes here, this approach is chosen as a convenient statistically stationary system that is easily simulated and with a closed-form conditional probability development. The pair-correlation function and other statistics for this system can be found in appendix A.

A realization of this system is shown in Fig. 6. To match the fluctuating concentration system described above and used in Baker and Lawson (2010), we chose $T = 2$ m and $N_{\text{par}} \cdot D/T$ was chosen to be 2500 $\text{m}^{-1}$; $N_{\text{par}}$ was chosen to be 200 (corresponding to the 200 "high" regions in the earlier simulation), which left $D = 50$. A quick examination of Fig. 6 demonstrates that this system is a bit more "clumpy" or "clustered" than the periodic concentration model, and since there was no embedded "background" concentration, the count record is a bit more intermittent. Comparison of this figure to actual cloud drop time series data (e.g., Shaw et al. 2002) suggests that, though this simulation model is not particularly physically motivated here, the time series look similar when viewed on a reasonably coarse spatial scale.

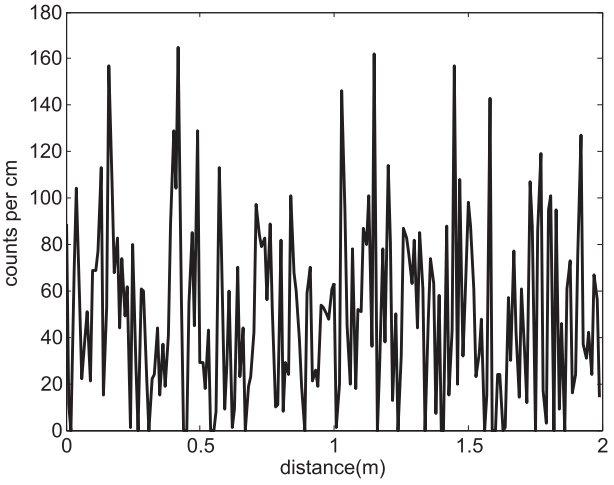


FIG. 6. A single realization of a Matérn cluster process with $N_{\text{par}} = 200$, $D = 50$, and $S = 0.005$. This is an example of a statistically stationary but correlated/clumpy system with an inherent scale of relevance equal to $2S = 0.01$.

Figure 7 shows the four statistics for this system. The gray curves show five realizations of the simulation, and the dotted black curves show the closed-form expressions expected based on the development in appendix A: $2S$ (the scale of interest) is again equal to 0.01.

Note that $\eta$ once again shows a clear change in behavior at the scale of interest, whereas SCI and CI seem to have no change in behavior anywhere near this scale. The modified Fishing statistic does show a peak in the general vicinity of $2S$. However, this peak is not sharply defined, and—once again—$F$ exceeds the "threshold value" of 3 for basically all scales (note that this is particularly egregious here because *by construction,* there is *no statistical structure at all* on scales larger than $2S$; this can be seen by noting that $\eta(t > 2S) = 0$: the probability of finding a particle at $t_\circ + t$ given a particle at $t_\circ$ if $t > 2S$ is not enhanced nor diminished in this system). The pair-correlation function properly identifies that there is coherence/structure on scales $t < 2S$ $[\eta(t) > 0]$ and no structure on scales $t > 2S$ $[\eta(t) = 0]$.

## 5. Conclusions

Through use of a few simple computational models, it can be seen that the pair-correlation function—when

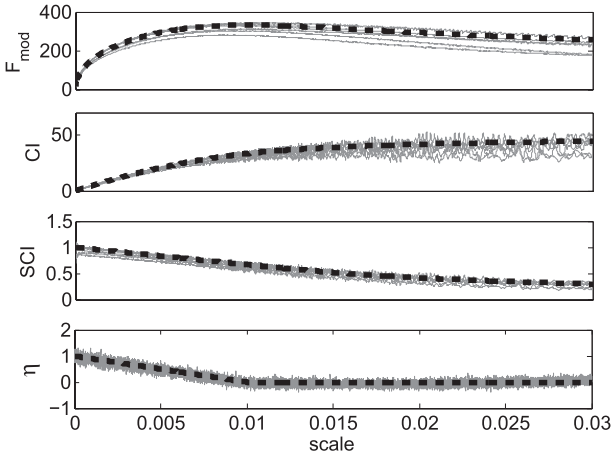


FIG. 7. A plot of $F$, CI, SCI, and $\eta$ as a function of scale for the Matérn cluster simulation. Simulation results for five realizations are in light gray, while the theoretical expressions given in appendix A are shown with a dotted black line. Note the clear change in behavior of the pair-correlation function at scale 0.01.

carefully and properly applied—is able to unequivocally identify the localized scales of relevance in simulated data. The statistic is not a panacea: for short data records, the statistic is often noisy. However, the direct link of the statistic to physical statistical understanding is a strength that should not be underestimated.

The term *volume-averaged pair-correlation function* used in much of the cloud particle clustering literature should be avoided, as several misconceptions about how the averaged pair-correlation function relates to first- and second-order statistics (variance and mean) seem to be pervasive in the literature; here we suggest the alternative ''scaled clustering index'' for the quantity $(V_t/M_t - 1)M_t^{-1}$. In an effort to remove any confusion, one-dimensional versions linking the pair-correlation function, Fishing statistic, clustering index, and the scaled clustering index were presented directly for the first time. Use of the relationships was verified by developing closed-form expressions for systems and confirming them through simulation.

Care should be taken when arguing for the primacy of one statistical tool over another for cloud physics use, as quite often the tools are applied outside their stated range of validity. A tool should not be criticized or discarded when it fails in a regime where it has no expectation of success—rather justifications for assuming a particular statistical structure should be given and used to justify what tools are used.

*Acknowledgments.* This work was supported with start-up funds from the School of Science and Mathematics at the College of Charleston and with a Cottrell College Science Award from the Research Corporation for Scientific Advancement. The author would also like to thank the reviewers for insightful comments that improved the manuscript.

## APPENDIX A

### Closed-Form Expressions for Simulated Systems

In the main text it was insinuated that it is often possible to write down the pair-correlation function (and, therefore, the other statistics used in this study) from stochastic considerations alone. Here we demonstrate by example. Although the algebraic expressions are somewhat bulky, the formulas are found by using direct appeal to the physical definition of the pair-correlation function. The robustness of the results obtained is clearly shown when comparing these theoretical expressions (shown with dotted lines on the figures in the main text) to the simulation outputs. Note that the final expressions are closed form: there are no fitting coefficients beyond the simulation parameters.

#### *a. Periodic simulation*

To obtain $\eta(t)$ for the periodic simulation described in section 4, we return to the basic definition of the pair-correlation function: if an event is known to be in a time series at time $t_\circ$, then the probability of a second event measured in the time interval $(t_\circ + t, t_\circ + t + dt)$ is given by

$$p(t_\circ + t \,|\, t_\circ)dt = \lambda\, dt(1 + \eta(t)). \tag{A1}$$

Following the description given earlier, let there be a time series of total duration $T$ where the number of counts per unit time $\lambda(t)$ is defined via

$$\lambda(t) = \begin{cases} \lambda_1 & \text{for} \quad (0 \le t \le \tau, 2\tau \le t \le 3\tau, 3\tau \le t \le 5\tau, \ldots) \\ \lambda_1 + \lambda_2 & \text{for} \quad (\tau \le t \le 2\tau, 3\tau \le t \le 4\tau, 5\tau \le t \le 6\tau, \ldots) \end{cases} \tag{A2}$$

with $\tau$ some (constant) scale of interest. For this system we have a total number of particles equal to

$$N_{\text{total}} = T\overline{\lambda(t)} = \frac{T}{2}\lambda_1 + \frac{T}{2}(\lambda_1 + \lambda_2), \tag{A3}$$

$$\overline{\lambda(t)} = \frac{2\lambda_1 + \lambda_2}{2}. \tag{A4}$$

The probability that a random particle is in a region of "high" intensity is given by $(\lambda_1 + \lambda_2)(2\lambda_1 + \lambda_2)^{-1}$, and the probability that a random particle is in a region of "low" intensity is given by $\lambda_1(2\lambda_1 + \lambda_2)^{-1}$.

Using the expression for $\overline{\lambda(t)}$ as $\lambda$ in the equation defining the pair-correlation function above, we can find an expression for $\eta(t)$ if we use some physical reasoning to establish $p(t_\circ + t|t_\circ)$.

First, we note

$$p(t_\circ + t \,|\, t_\circ) = p(t_\circ \in H)p(t_\circ + t \,|\, t_\circ \in H) + p(t_\circ \in L)p(t_\circ + t \,|\, t_\circ \in L), \tag{A5}$$

where $p(t_\circ \in H)$ is the probability that the base particle (known to be present) is in a region of high concentration and $p(t_\circ \in L)$ is the probability that the base particle (known to be present) is in a region of low concentration.

Further, we can break down and write expressions for $p(t_\circ + t|t_\circ \in H)$ and $p(t_\circ + t|t_\circ \in L)$ as follows:

$$p(t_\circ + t \,|\, t_\circ \in H)dt = [(\lambda_1 + \lambda_2)dt]p(t_\circ + t \in H \,|\, t_\circ \in H) + \lambda_1 dtp(t_\circ + t \in L \,|\, t_\circ \in H), \tag{A6}$$

$$p(t_\circ + t \,|\, t_\circ \in L)dt = [(\lambda_1 + \lambda_2)dt]p(t_\circ + t \in H \,|\, t_\circ \in L) + \lambda_1 dtp(t_\circ + t \in L \,|\, t_\circ \in L). \tag{A7}$$

To evaluate these expressions, we need to write a relation for $p(t_\circ + t \in H|t_\circ \in H)$ and similar terms in the above expression. For a particle known to be located in a region of high intensity, the particle is equally probable anywhere in an interval of duration $\tau$. For a given delay time $t < \tau$, we note that the probability that $t_\circ + t \in H$ is equal to the probability that $t_\circ$ is between the beginning of the high-intensity region and $\tau - t$. For delay times $\tau < t < 2\tau$, we note that, if $t_\circ + t$ is in the (next) region of high intensity, we need $t_\circ$ to be within $t - \tau$ of the end of the high-intensity region. These simple observations lead us to

$$p(t_\circ + t \in H \,|\, t_\circ \in H) = \begin{cases} 1 - \dfrac{t}{\tau} & \text{for} \quad t \le \tau \\ \dfrac{t}{\tau} - 1 & \text{for} \quad \tau \le t \le 2\tau. \end{cases} \tag{A8}$$

Based on an identical geometrical argument, we can state that

$$p(t_\circ + t \in L \,|\, t_\circ \in L) = \begin{cases} 1 - \dfrac{t}{\tau} & \text{for} \quad t \le \tau \\ \dfrac{t}{\tau} - 1 & \text{for} \quad \tau \le t \le 2\tau. \end{cases} \tag{A9}$$

If we note that $t_\circ + t$ must be in either a region of high intensity or a region of low intensity, then we can write

$$p(t_\circ + t \in L \,|\, t_\circ \in H) = \begin{cases} \dfrac{t}{\tau} & \text{for} \quad t \le \tau \\ 2 - \dfrac{t}{\tau} & \text{for} \quad \tau \le t \le 2\tau, \end{cases} \tag{A10}$$

$$p(t_\circ + t \in H \,|\, t_\circ \in L) = \begin{cases} \dfrac{t}{\tau} & \text{for} \quad t \le \tau \\ 2 - \dfrac{t}{\tau} & \text{for} \quad \tau \le t \le 2\tau. \end{cases} \tag{A11}$$

Substituting these expressions back into the left-hand side of the definition of the pair-correlation function and simplifying, we have

$$p(t_\circ + t \,|\, t_\circ) = \begin{cases} \left(\dfrac{\lambda_1 + \lambda_2}{2\lambda_1 + \lambda_2}\right)\left[(\lambda_1 + \lambda_2)dt\left(1 - \dfrac{t}{\tau}\right) + \lambda_1 dt\dfrac{t}{\tau}\right] \\ \quad + \left(\dfrac{\lambda_1}{2\lambda_1 + \lambda_2}\right)\left[(\lambda_1 + \lambda_2)dt\dfrac{t}{\tau} + \lambda_1 dt\left(1 - \dfrac{t}{\tau}\right)\right], \quad 0 \le t \le \tau \\ \left(\dfrac{\lambda_1 + \lambda_2}{2\lambda_1 + \lambda_2}\right)\left[(\lambda_1 + \lambda_2)dt\left(\dfrac{t}{\tau} - 1\right) + \lambda_1 dt\left(2 - \dfrac{t}{\tau}\right)\right] \\ \quad + \left(\dfrac{\lambda_1}{2\lambda_1 + \lambda_2}\right)\left[(\lambda_1 + \lambda_2)dt\left(2 - \dfrac{t}{\tau}\right) + \lambda_1 dt\left(\dfrac{t}{\tau} - 1\right)\right], \quad \tau \le t \le 2\tau. \end{cases} \tag{A12}$$

Or, after some algebraic simplification,

$$p(t_\circ + t \,|\, t_\circ) = \begin{cases} \dfrac{dt}{2\lambda_1 + \lambda_2}\left[2\lambda_1(\lambda_1 + \lambda_2) + \lambda_2^2\left(1 - \dfrac{t}{\tau}\right)\right], & 0 \le t \le \tau \\ \dfrac{dt}{2\lambda_1 + \lambda_2}\left[2\lambda_1(\lambda_1 + \lambda_2) + \lambda_2^2\left(\dfrac{t}{\tau} - 1\right)\right], & \tau \le t \le 2\tau. \end{cases} \tag{A13}$$

Returning to the definition of the pair-correlation function to define $\eta(t)$, we have

$$\eta(t) = \frac{p(t_\circ + t|t_\circ)}{\overline{\lambda(t)}dt} - 1. \tag{A14}$$

Noting that $\overline{\lambda(t)}$ for this system is $(2\lambda_1 + \lambda_2)/2$, we then can simplify the above result for a closed-form expression for the pair-correlation function for the system described in Baker and Lawson (2010):

$$\eta(t) = \begin{cases} \dfrac{2[2\lambda_1(\lambda_1 + \lambda_2) + \lambda_2^2(1 - t/\tau)]}{(2\lambda_1 + \lambda_2)^2} - 1, & 0 \le t \le \tau \\ \dfrac{2[2\lambda_1(\lambda_1 + \lambda_2) + \lambda_2^2(t/\tau - 1)]}{(2\lambda_1 + \lambda_2)^2} - 1, & \tau \le t \le 2\tau. \end{cases} \tag{A15}$$

This result can then be integrated through the relations given starting with Eq. (9) to obtain closed-form expressions for $F(t)$, $\mathrm{CI}(t)$, and $\mathrm{SCI}(t)$. For the periodic simulation under discussion here, this integration is straightforward enough (though the integrand is piecewise defined for $t > \tau$). The results (for $0 \le t \le 2\tau$) yield

$$\mathrm{CI}(t) = \begin{cases} (2\lambda_1 + \lambda_2)t\left(\dfrac{Q}{2} + \dfrac{R}{2} - \dfrac{Rt}{6\tau}\right), & 0 \le t \le \tau \\ (2\lambda_1 + \lambda_2)\left[\tau R + \dfrac{t}{2}(Q - R) + \dfrac{Rt^2}{6\tau} - \dfrac{\tau^2 R}{3t}\right], & \tau \le t \le 2\tau, \end{cases} \tag{A16}$$

$$\mathrm{SCI}(t) = \frac{2\mathrm{CI}(t)}{(2\lambda_1 + \lambda_2)t}, \tag{A17}$$

$$F(t) = \mathrm{CI}(t)\sqrt{\frac{T - t}{2t}} \tag{A18}$$

with

$$Q = \frac{4\lambda_1(\lambda_1 + \lambda_2)}{(2\lambda_1 + \lambda_2)^2} - 1, \tag{A19}$$

$$R = \frac{2\lambda_2^2}{(2\lambda_1 + \lambda_2)^2}. \tag{A20}$$

Although these results were only calculated for $t \le 2\tau$, similar arguments can be made for $t > 2\tau$. [In fact, $\eta(t)$ for $t > 2\tau$ can be shown to be equal to $\eta(t - 2\tau)$, thus giving rise to the periodicity seen in the pair-correlation function shown in Fig. 12 of Baker and Lawson (2010). The other statistics, since they involve weighted integrals of $\eta(t)$, do not have quite so simple a periodic structure, but using the technique above allows for the calculations for expressions for $F$, CI, and SCI for $t > 2\tau$.]

*b. Matérn simulation*

The Matérn system used earlier in this paper also has a closed-form expression for the pair-correlation function and, by extension, $F(t)$, $\mathrm{CI}(t)$, and $\mathrm{SCI}(t)$. As shown in Santaló (1976) and Larsen (2006), $\eta(t)$ for this system can be written in terms only of the number of parents $N_{\mathrm{par}}$, the scale of interest $2S$, and $t$. Noting that the total number of particles in the system can be written as $N_{\mathrm{par}}D$, where $D$ is the number of daughters per parent, then we can write

$$\eta(t) = \begin{cases} \dfrac{T}{(2S)^2 N_{\mathrm{par}}}(2S - t), & 0 \le t \le 2S \\ 0, & t > 2S. \end{cases} \tag{A21}$$

Again, conducting a straightforward integration, we can obtain

$$\mathrm{CI}(t) = \begin{cases} \dfrac{D}{2S^2}\left(St - \dfrac{t^2}{6}\right), & 0 \le t \le 2S \\ \dfrac{2D}{t}\left(\dfrac{t}{2} - \dfrac{S}{3}\right), & t > 2S, \end{cases} \tag{A22}$$

$$\mathrm{SCI}(t) = \frac{T\mathrm{CI}(t)}{DN_{\mathrm{par}}t}, \tag{A23}$$

$$F(t) = \mathrm{CI}(t)\sqrt{\frac{T-t}{2t}}. \tag{A24}$$

## APPENDIX B

### On Statistical Homogeneity and Stationarity

A number of other studies have carefully noted that the pair-correlation function can only be interpreted as a scale-localized deviation from perfect randomness for a statistically stationary/homogeneous system [see, e.g., Shaw et al. (2002), p. 1055; Larsen et al. (2005), p. 473 and appendix A; Kostinski et al. (2006), p. 39; Larsen (2007), p. 808; or Larsen (2006), the extensive discussion in chapter 2]. The system analyzed in Baker and Lawson (2010) and reexplored here gives a particularly clear illustration of the importance of only interpreting the pair-correlation function under these conditions.

When used in a statistically stationary system, we can state that $\eta(t) > 0$ for some scale $t$ unequivocally identifies the presence of statistical correlations among infinitesimal volumes separated by scale $t$—it is an inherently scale-localized fact about the system. For the statistically inhomogeneous system described in Baker and Lawson (2010) and simulated in the main text, however, $\eta(t) > 0$ on some scale $t$ results—not from a statistical correlation/clustering between subvolumes of scale $t$—but rather due to concentration fluctuations that may occur at scales equal to, larger than, or less than $t$. With the insight supplied from knowledge of the nature of the statistical nonstationarity in this system, we are able to interpret this properly (see appendix A). A naive calculation of the pair-correlation function without the insight available to us here would lead us to the erroneous conclusion [like that drawn by Baker and Lawson (2010)] that the statistic fails to convey intelligible scale-localized information about the dataset. This is a case where the inappropriate use of the pair-correlation function on a nonstationary dataset in a sense "delocalizes" the information given by the pair-correlation function and ultimately led to the improper conclusion in Baker and Lawson that the pair-correlation function was generally an inadequate tool to identify scale-localized deviations from perfect spatial randomness.

The lesson to be learned from this is that the warning to use the pair-correlation function only for statistically stationary/homogeneous systems is important! Although physical scientists often do things that would make mathematicians cringe without too many practical consequences, here we see invalid conclusions being drawn due to the improper use of a statistical tool applied outside its intended range of validity.

A difficulty arises, however, when an investigator then needs to choose which statistical tools to use/apply to field data. If the pair-correlation function (and related statistical tools) can only be meaningfully used on data that is statistically stationary/homogeneous, then the first step the investigator would be tasked to do is to determine whether the data gathered meets these criteria. Unfortunately, no unambiguous tool for determining statistical homogeneity/stationarity for real data exists. There have been some efforts to determine practical conditions for statistical homogeneity/stationarity (e.g., Larsen et al. 2005; Larsen 2006; Anderson and Kostinski 2010), but those fall short of the ultimate mark; although every statistically homogeneous/stationary system will pass these tests, some statistically inhomogeneous/nonstationary systems will pass these tests as well (formally speaking, these tests form necessary, but not sufficient, criteria for statistical stationarity). The difficulty in finding truly necessary and sufficient tests of stationarity for real data has proven to be tricky—at least in part because real data is finite and, therefore, formally statistically inhomogeneous/nonstationary even if the underlying system is inherently homogeneous/stationary.

This appendix does not have any panaceas, but serves mainly to remind other investigators of two main points: (i) the purpose of a statistical tool can sometimes matter and, when ignored, can lead to errors in interpretation of associated computed statistics and (ii), although we are sometimes able to definitively establish the inhomogeneity/nonstationarity of a dataset, no tool can unambiguously show us when a measured dataset comes from a statistically stationary/homogeneous system. Taken together, these two points suggest that interpretation of statistical results requires careful consideration and perhaps should include explicit discussion/consideration of the implications for both scenarios: if the system is actually statistically homogeneous/stationary or if it is statistically inhomogeneous/nonstationary.

## REFERENCES

Anderson, A., and A. Kostinski, 2010: Reversible record breaking and variability: Temperature distributions across the globe. *J. Appl. Meteor. Climatol.,* **49,** 1681–1691.

Baker, B., 1992: Turbulent entrainment and mixing in clouds: A new observational approach. *J. Atmos. Sci.,* **49,** 387–404.

——, and R. Lawson, 2010: Analysis of tools used to quantify droplet clustering in clouds. *J. Atmos. Sci.,* **67,** 3355–3367.

Chaumat, L., and J. Brenguier, 2001: Droplet spectra broadening in cumulus clouds. Part II: Microscale droplet concentration heterogeneities. *J. Atmos. Sci.,* **58,** 642–654.

Cox, D., and V. Isham, 1980: *Point Processes.* Chapman and Hall, 188 pp.

Grandell, J., 1976: *Doubly Stochastic Poisson Processes.* Springer-Verlag, 234 pp.

Jameson, A., and A. Kostinski, 2000: Fluctuation properties of precipitation. Part VI: Observations of hyperfine clustering and drop size distribution structures in three-dimensional rain. *J. Atmos. Sci.,* **57,** 373–388.

Knyazikhin, Y., A. Marshak, M. Larsen, W. Wiscombe, J. Martonchik, and R. Myneni, 2005: Small-scale drop size variability: Impact on estimation of cloud optical properties. *J. Atmos. Sci.,* **62,** 2555–2567.

Kostinski, A., 2001: On the extinction of radiation by a homogeneous but spatially correlated random medium. *J. Opt. Soc. Amer.,* **A18,** 1929–1933.

——, and A. Jameson, 1997: Fluctuation properties of precipitation. Part I: On deviations of single-size drop counts from the Poisson distribution. *J. Atmos. Sci.,* **54,** 2174–2186.

——, and ——, 2000: On the spatial distribution of cloud particles. *J. Atmos. Sci.,* **57,** 901–915.

——, and R. Shaw, 2001: Scale-dependent droplet clustering in turbulent clouds. *J. Fluid Mech.,* **434,** 389–398.

——, and ——, 2005: Fluctuations and luck in droplet growth by coalescence. *Bull. Amer. Meteor. Soc.,* **86,** 235–244.

——, M. Larsen, and A. Jameson, 2006: The texture of rain: Exploring stochastic microstructure at small scales. *J. Hydrol.,* **328,** 38–45.

Larsen, M., 2006: Studies of discrete fluctuations in atmospheric phenomena. Ph.D. dissertation, Michigan Technological University, 220 pp.

——, 2007: Spatial distributions of aerosol particles: Investigation of the Poisson assumption. *J. Aerosol Sci.,* **38,** 807–822.

——, and A. Kostinski, 2009: Simple dead-time corrections for discrete time series of non-Poisson data. *Meas. Sci. Technol.,* **20,** 095101, doi:10.1088/0957-0233/20/9/095101.

——, W. Cantrell, J. Kannosto, and A. Kostinski, 2003: Detection of spatial correlations among aerosol particles. *Aerosol Sci. Technol.,* **37,** 476–485.

——, A. Kostinski, and A. Tokay, 2005: Observations and analysis of uncorrelated rain. *J. Atmos. Sci.,* **62,** 4071–4083.

——, A. Clark, M. Noffke, G. Saltzgaber, and A. Steele, 2010: Identifying the scaling properties of rainfall accumulation as measured by a tipping-bucket rain gauge network. *Atmos. Res.,* **96,** 149–158.

Lavergnat, J., and P. Gole, 1998: A stochastic raindrop time distribution model. *J. Appl. Meteor.,* **37,** 805–818.

Lehmann, K., H. Siebert, M. Wendisch, and R. Shaw, 2007: Evidence for inertial droplet clustering in weakly turbulent clouds. *Tellus,* **59B,** 57–65.

Lovejoy, S., M. Lilly, N. Desaulniers-Soucy, and D. Schertzer, 2003: Large particle number limit in rain. *Phys. Rev.,* **E68,** 025301, doi:10.1103/PhysRevE.68.025301.

Marshak, A., Y. Knyazikhin, M. Larsen, and W. Wiscombe, 2005: Small-scale drop-size variability: Empirical models for drop-size-dependent clustering in clouds. *J. Atmos. Sci.,* **62,** 551–558.

Martinez, V., and E. Saar, 2002: *Statistics of the Galaxy Distribution.* Chapman and Hall/CRC, 432 pp.

Neyman, J., and E. Scott, 1958: Statistical approach to problems of cosmology. *J. Roy. Stat. Soc.,* **B20,** 1–43.

Picinbono, B., and C. Bendjaballah, 2005: Characterization of nonclassical optical fields by photodetection statistics. *Phys. Rev.,* **A71,** 013812, doi:10.1103/PhysRevA.71.013812.

Pinsky, M., and A. Khain, 2001: Fine structure of cloud droplet concentration as seen from the fast-FSSP measurements. Part I: Method of analysis and preliminary results. *J. Appl. Meteor.,* **40,** 1515–1537.

Preining, O., 1983: Optical single-particle counters to obtain the spatial inhomogeneity of particulate clouds. *Aerosol Sci. Technol.,* **2,** 79–90.

Santaló, L., 1976: *Integral Geometry and Geometric Probability.* Addison-Wesley, 404 pp.

Shaw, R., 2003: Particle-turbulence interactions in atmospheric clouds. *Annu. Rev. Fluid Mech.,* **35,** 183–227.

——, W. Reade, L. Collins, and J. Verlinde, 1998: Preferential concentration of cloud droplets by turbulence: Effects on the early evolution of cumulus cloud droplet spectra. *J. Atmos. Sci.,* **55,** 1965–1976.

——, A. Kostinski, and M. Larsen, 2002: Towards quantifying droplet clustering in clouds. *Quart. J. Roy. Meteor. Soc.,* **128,** 1043–1057.

Stoyan, D., and H. Stoyan, 1994: *Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics.* John Wiley and Sons, 389 pp.

Uijlenhoet, R., J. Stricker, P. Torfs, and J.-D. Creutin, 1999: Towards a stochastic model of rainfall for radar hydrology: Testing the Poisson homogeneity hypothesis. *Phys. Chem. Earth,* **24B,** 747–755.