A BROWNIAN MOTION MODEL OF POLLUTANT CONCENTRATION DISTRIBUTIONS

WAYNE R. OTT

TECHNICAL REPORT NO. 46
APRIL 1981

PREPARED UNDER THE AUSPICES OF
SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
A BROWNIAN MOTION MODEL OF POLLUTANT CONCENTRATION DISTRIBUTIONS

by

WAYNE R. OTT
U.S. ENVIRONMENTAL PROTECTION AGENCY
AND
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 46
APRIL 1981

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER SUPPORT TO SIMS FROM

ENVIRONMENTAL PROTECTION AGENCY (EPA)
DEPARTMENT OF ENERGY (DOE)
SLOAN FOUNDATION
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
A BROWNIAN MOTION MODEL OF POLLUTANT CONCENTRATION DISTRIBUTIONS

Wayne R. Ott
U.S. Environmental Protection Agency
and
Stanford University

I. INTRODUCTION

A number of technical papers have appeared in the air pollution literature claiming that pollutant concentrations measured in the atmosphere generally always tend to be lognormally distributed. Applications of the lognormal distribution to ambient air quality data appeared in the literature as early as 1956, and Larsen [1] has developed a mathematical model of pollutant concentrations and averaging times, based on analyses of a large body of empirical air quality data, which assumes that "... concentrations are approximately lognormally distributed for all pollutants in all cities for all averaging times." The reader wishing to examine the literature on this topic in greater detail may wish to consult a technical report by Ott, Mage, and Randecker [2], which contains over 70 references on applications of various probability distributions to air quality data.

Air quality monitoring activities in urban areas generate concentration readings on a continuous basis, but it is customary to average the concentrations over one-hour periods in order to store the data conveniently. If, for a given season or year, the cumulative frequencies of hourly values are plotted on logarithmic probability paper - that is, the logarithm of concentration versus cumulative normal probits - the resulting line usually looks quite straight. Hourly carbon monoxide (CO)
observations obtained from the San Jose air monitoring station for the one year period from June 1, 1970 to May 31, 1971 illustrate this finding. The histogram of the observations has the characteristic lognormal shape - a single positive mode close to the origin (right offset mode) and a long tail extending to the right (right skewness; see Figure 1). When the cumulative frequencies of these observations are plotted on logarithmic probability paper (Figure 2), the resulting plot looks reasonably straight, although some minor curvature is evident.

The study of CO concentrations by Ott, Mage, and Randecker [2], which covers data sets from a representative cross section of air monitoring stations in U.S. cities, notes that a line can appear quite straight on logarithmic probability paper, while the same data set seriously fails standard statistical tests for lognormality. Furthermore, successive hourly air quality observations are serially correlated and are not independent. As a result, the observed concentration of a pollutant cannot be treated as a stationary lognormally distributed random variable in the traditional sense, and Ott, Mage, and Randecker [2] describe it as having a "pseudo-lognormal" distribution. Despite the lack of perfect lognormality exhibited by actual air quality data, the characteristic shape of the histogram - a single mode to the right of the origin and a long right-skewed tail - is so ubiquitous and striking that this shape demands explanation. It would give great insight if some unique property of the diffusion process could be identified which explains why this distributional form is so common. Can very simple processes generate a concentration distribution with this same characteristic shape? This paper examines this question in detail, showing that a relatively simple Brownian motion process yields similar right-skewed - although not necessarily lognormal - distributions.
Figure 1. Histogram of hourly CO concentrations measured at the San Jose air monitoring station over the 1-year period from June 1, 1970, to May 31, 1971, showing characteristic shape of the distribution of air quality data: a single mode near the origin and right-skewed tail.
Figure 2. Logarithmic-probability plot of hourly CO concentrations measured at the San Jose air monitoring station over the 1-year period from June 1, 1970, to May 31, 1971, showing "straightness" of resulting line.

\[ \mu_p = 3.5373 \]
\[ \sigma_p = 2.2710 \]
\[ n = 8573 \]
II. BACKGROUND LITERATURE

The air pollution literature contains several theoretical arguments attempting to explain how a lognormal distribution of air quality data might arise naturally. Unfortunately, these explanations seem rather general, for the authors do not develop their models in great detail or specificity. In addition, different investigators have proposed different rationales for lognormality, and the alternative explanations in the literature may sometimes seem contradictory.

Kahn [3] was one of the first investigators to point out a possible explanation of the apparent lognormality of observed air quality data. He was appreciative, however, of the complexities involved in the air pollution transport process:

The complexity of the emission-transport-receptor process makes it difficult, if not impossible, to prove that air pollutants are lognormal . . .

... The concentration of air pollution at any point in time is the result of many diverse meteorological, physiographic and emission factors acting simultaneously. These factors are governed by random processes which, if they were all to be considered, would be very difficult, if not impossible, to describe analytically.

Kahn [3] treats the concentrations measured at regular time intervals, such as every five minutes, as a positive discrete random process. The observed concentration at these regular time intervals is considered to have values $C_0, C_1, \ldots, C_n$, where $C_0$ is the value at some arbitrary starting point, and $C_n$ is the value of the $n^{th}$ observation.

In this model, the change in pollutant concentration from one observation to the next is assumed to be expressed as follows:

\[
C_i - C_{i-1} = \rho_i C_{i-1} \quad \text{for} \quad i=1, 2, \ldots, n,
\]
where \( C_i \) and \( C_{i-1} \) are values of the process at times \( i \) and \( i-1 \), respectively, and \( \rho_i \) is a random variable of proportionality which "... represents the effect of the many diverse, simultaneously occurring random processes on the pollution concentration observed at time \( i-1 \)" [3].

Equation 1 can be written as follows:

\[
(2) \quad \frac{C_i - C_{i-1}}{C_{i-1}} = \rho_i .
\]

If both sides of Equation 2 are summed over \( n \) states, the result is written as follows:

\[
(3) \quad \sum_{i=1}^{n} \frac{C_i - C_{i-1}}{C_{i-1}} = \sum_{i=1}^{n} \rho_i .
\]

Assuming that the incremental change in concentration is small at each successive state in time, the left-hand side of Equation 3 can be approximated by an integral:

\[
(4) \quad \sum_{i=1}^{n} \frac{C_i - C_{i-1}}{C_{i-1}} \approx \int_{C_0}^{C_n} \frac{\text{d}c}{c} = \log C_n - \log C_0 .
\]

Combining Equations 3 and 4, we can solve for \( \log C_n \) as follows:

\[
(5) \quad \log C_n = \log C_0 + \rho_1 + \rho_2 + \ldots + \rho_n .
\]

Assuming that the \( \rho \)'s are independent, Kahn [3] invokes the additive form of the central limit theorem to argue that the right-hand side of Equation 5 is asymptotically normally distributed:

Central limit theory says that \( \log C_n \) is asymptotically normally distributed regardless of the distribution of \( \rho_1 \) and that \( C_n \) is asymptotically lognormally distributed.
The above model requires two basic assumptions: (1) a proportionality factor, as indicated in Equation 1, and (2) behavior of this proportionality factor as an independent random variable, permitting the central limit theorem to be invoked. As described by Aitchison and Brown [4], this model originally was proposed in a book by Kapteyn [5] in 1903 and is called the law of proportional effect, which is described as follows:

A variate subject to a process of change is said to obey the law of proportionate effect if the change in the variate at any step of the process is a random proportion of the previous value of the variate. [4].

In the past, this model has been used as a rationale for generating the skewed frequency distributions often observed with biological variables, such as the size of organisms comprising a population. These applications assume that the growth of each organism is a random proportion of the size it already has attained, and the logical consequence of this model is that the resulting size distribution of the population should be asymptotically lognormal.

A problem with Kahn's application of the law of proportionate effect to air pollution concentrations [3] is that he does not show any physical mechanism that justifies the appropriateness of the model to atmospheric transport and diffusion processes. Rather, he suggests this process simply as a heuristic model. Why should the law of proportionate effect be applicable to the processes that generate air pollution concentration distributions?

Gifford [6] has suggested that the law of proportionate effect may be attributed to the strong dependencies that exist in sequential air pollutant concentrations:
... The irregular change in these numbers from one sampling interval to the next reflects all the obviously complex variability of source, meteorological, and other factors. But at any time it will most strongly depend on the existing air pollution concentration levels. This is true for at least two reasons. First, relevant meteorological factors, principally the wind speed and direction, tend to be strongly self-correlated. Second, urban air pollution is more uniform than that from isolated rural sources because the urban source is distributed more uniformly, over a very large area. [6].

This reasoning does not explain why the randomness should be reflected by a proportionate term, as it is in the law of proportionate effect, rather than an additive term, as it is in most time series models.

Furthermore, this reasoning does not explicitly describe why the serial correlation arises in the first place.

Knox and Pollack [7] manipulate the equations of diffusion and suggest that the final result can be written in a form that is consistent with the law of proportionate effect. They point out that the distribution is not necessarily lognormal if the measured concentration is comparable in magnitude with the source concentration. A report by Pollack [8] represents atmospheric diffusion processes as a "box model" which can be written in a form similar to the law of proportionate effect. In this view, lognormality will be typical for well-mixed, urban atmospheres:

This reasoning is most easily justified for a well mixed urban region. It is not clear that the lognormal distribution will fit as well for the non-urban, poorly mixed areas. We do feel, however, that the characteristics of an area's topography and typical meteorology would have to be highly unusual ... that the lognormal distribution would fit poorly. [8].

Pollack [8] also notes that the law of proportionate effect can be transformed into a first-order autoregressive process by first simplifying Equation 1 as follows, where $\varepsilon_1 = 1 + \rho_1$,

\begin{equation}
C_i = C_{i-1} \varepsilon_i .
\end{equation}
Then, taking logarithms, he obtains the following:

\[ \ln C_i = \ln C_{i-1} + \ln \varepsilon_i. \]

He argues that the autocorrelation obtained by taking natural logarithms of CO concentrations observed in San Francisco in 1970 exhibits exponential decay better than the untransformed data and that it has some of the properties that would be expected for the autocorrelation function of Equation 7.

Mage and Ott [9, 11] show that a simple, first order linear differential equation, if a random coefficient is permitted with no forcing function, can be written as Kapteyn's law of proportionate effect. Using discrete mathematics to simplify their presentation, they also show that, if the random coefficient is highly correlated with a forcing function input, a censored, three-parameter lognormal distribution results. It is not clear how realistic these assumptions are for actual atmospheric diffusion processes. The censored, three-parameter lognormal model has been shown by Mage [10] to explain the downward curvature often observed in plots of air pollutant concentrations on logarithmic-probability paper.

The above explanations of lognormality of air pollutant concentrations all rely on the central limit theorem and on unproved assertions about the independence of random phenomena affecting concentrations. The overall models are very general and do not allow specific predictions that can be tested. For example, none of these papers attempts to calculate the parameters of the resulting lognormal distribution that is assumed to arise, and there is no attempt to relate the parameters to specific atmospheric phenomena. Thus, the models must be taken on "faith",
and they seem to offer little insight into the underlying processes involved. Furthermore, these models simply may be applied incorrectly as a representation of physical reality.

The vagueness of models based on Kapteyn's law can be illustrated by noting that no requirements are specified for the distribution of \( \varepsilon = \rho + 1 \). The general solution of Equation 7 for any \( n \geq 1 \) is as follows:

\[
C_n = C_0 \prod_{i=1}^{n} \varepsilon_i .
\]

If \( \varepsilon_i > 1 \) for all \( i \), then \( C_n \) is unbounded and will approach \( \infty \) as \( n \to \infty \). Conversely, if \( \varepsilon_i < 1 \) for all \( i \), then \( C_n \) will approach 0 as \( n \to \infty \). The only way to avoid these limits is for \( \varepsilon \) to have the following property:

\[
\lim_{n \to \infty} \left\{ \frac{n}{\prod_{i=1}^{n} \varepsilon_i} \right\} = 1 .
\]

This condition might be met by requiring that the geometric mean of \( \varepsilon \) be 1. Even if this requirement is met, the variance of the concentrations will increase with time. Actual observed time series of air pollutant concentrations do not behave in this manner.

Unfortunately, it appears that these investigators have applied the law of proportionate effect to air pollutant concentrations in a sequential manner that is inconsistent with Kapteyn's original formulation and may not be aware of this disparity. If the stochastic processes represented by Equation 1 begins at \( i=1 \) and is allowed to generate observations until \( i=n \), then \( C_{n1} \) denotes the first member of a distribution. If the process then is repeated and allowed to generate values from \( i=1 \) to \( i=n \) once more, then \( C_{n2} \) denotes the second member of a distribution.
Kapteyn's law is applied as it originally was published, then the values from \( m \) repeated trials, \( C_{n1}, C_{n2}, \ldots, C_{nm} \), form a population which is asymptotically lognormal. Because \( n \) is the same for each member, these members are identically distributed random variables whose mean and variance are constant and are a function of \( n \) and the mean and variance of \( \rho \). This result is analogous to examining the distribution of the lengths of the leaves of a tree at some instant of time, assuming that the growth of each leaf was governed by the law of proportionate effect. In the published air pollution papers, however, the authors claim that the concentrations at successive, equally spaced times form a distribution that is asymptotically lognormal. These air pollution applications are analogous to examining the distribution of the lengths of a single leaf at successive, equally spaced points in time. Although the resulting distribution may or may not be lognormal, it is clear that each successive sample results from a larger number of random increments, and therefore the variance of the concentration increases with time, a property not shared by actual air pollution concentrations.

Another rationale for lognormality appearing in the literature is based on the assumed distributional form of one of the independent variables affecting concentration: wind speed. If \( U \) denotes the wind speed, atmospheric dispersion models such as the Gaussian plume model assume that pollutant concentrations are proportional to the source strength \( Q \) and inversely proportional to \( U \). That is, \( C = kQ/U \), where \( k \) is a constant of proportionality. If \( U \) is a lognormally distributed random variable, then \( 1/U \) also is a lognormally distributed random variable.
Benarie [12] has argued that wind speeds are basically lognormally distributed, thus causing the distribution of observed concentrations to appear lognormal:

It was shown by Benarie [13, 14] ... that the lognormal distribution is strictly valid for concentration frequency distributions in any given direction around a point source. This is a consequence of the facts that wind velocity distributions in any given direction may be approximated by a lognormal and of some very general mathematical properties of this function.

It may be objected that winds of zero knots often occur, and the logarithmic model is unable to handle these cases because the logarithm of zero is undefined. However, Benarie [12] argues that observed winds never really are zero, because, even during very calm periods, some air motion is present, and the zero values that do appear really are errors of the measuring process.

This review of the literature shows that several very different, and often con conflicting, reasons for the apparent lognormality of air pollutant concentrations have been suggested. The arguments resting on the law of proportionate effect and the central limit theorem are very general, may be applied incorrectly, and give a time series with properties inconsistent with actual observed concentrations. The arguments that rest on the assumption that winds are lognormally distributed also have defects, because this assumption has not been tested rigorously and because, even if the assumption is true, it is not clear why it is true. If winds are lognormally distributed, then a mechanism for the lognormality must be identified.

A final problem is that most investigators claim that concentrations are "approximately lognormal", a description that is ambiguous. Ott, Mage, and Randecker [2] have shown that a large cross section of U.S. CO air quality data sets fail standard goodness-of-fit tests for lognormality,
even though the cumulative frequencies plot as nearly straight lines on logarithmic-probability paper. They refer to these data sets as "pseudo-lognormal" to emphasize the ambiguous nature of the distributions. It should be noted that pollutant concentrations usually are serially correlated and are not independent, adding further to the ambiguity. All that can be said with certainty about air quality data is that the histograms of observed concentrations generally always are right-skewed and generally always have a single mode near the origin.

One of the problems with the published theories about lognormality of air quality data is that the processes involved - diffusion and transport of pollutants in the atmosphere from a number of sources - are very complex. If a much simpler diffusion process were found that also produces right-skewed, unimodal distributions, the resulting model could give insight into the processes that occur on a more complex level in the atmosphere, particularly if the process could be described in detail. In this paper, I examine the concentration distributions that result from a very simple diffusion process - individual particles released from a single point source subjected only to Brownian motion without drift.
III. APPROACHES FOR MONITORING AIR POLLUTANT CONCENTRATIONS

When pollutants are measured in the ambient atmosphere, the monitoring instrument usually is located some distance from the source, and the intake probe is at a fixed point in space. By the time the pollutant reaches the receptor probe, many random factors have acted on it as it is transported and diluted (for example, vertical air movements, horizontal air movements, turbulence, and mixing). Resulting concentrations are many orders of magnitude lower than when they leave the source and usually vary considerably with time.

At the monitoring location, pollutant concentration is measured either as mass of the pollutant per unit volume of air (for example, micrograms per cubic meter, μg/m³) or as volume of the gaseous pollutant per unit volume of the total mixture (for example, parts-per-million, or ppm). Because the observed concentration C(t) varies with time, and because storage of the instantaneous time series of concentration in these circumstances would be difficult, it is customary to average the concentration over a specified averaging time τ (for example, a one hour period) and to store only the resulting average value \( \bar{C} \):

\[
(10) \quad \bar{C} = \frac{1}{\tau} \int_{0}^{\tau} C(t) \, dt .
\]

If "consecutive" averaging periods are used, as is usually the case, then this process generates a series of n average values, each averaged over time interval \( \tau \):

\[
(11) \quad \bar{C}_i = \frac{1}{\tau} \int_{(i-1)\tau}^{i\tau} C(t) \, dt \quad \text{for} \quad i=1, 2, \ldots, n .
\]
As indicated earlier, it is the resulting sequence of \( n \) average values \( \bar{C}_1, \bar{C}_2, \ldots, \bar{C}_n \) that authors of various air pollution articles have regarded as lognormally distributed. Whether or not the distribution actually is lognormal, most monitoring data show that it is highly skewed to the right. In general, air pollution concentration data yield unimodal, right-skewed distributions for nearly all air pollutants, all cities, and all averaging times.

To construct a Brownian motion model that is analogous to the transport processes affecting pollutants in the atmosphere, the atmospheric transport system must be simplified greatly. If, however, a relatively simple Brownian motion model can be shown to generate right-skewed distributions similar to those observed in real ambient measurements, then the way in which real atmospheric processes generate these distributions may become more obvious. The overall purpose of such an analogue is to gain insight into the phenomena that affect actual atmospheric diffusion, transport, and dilution processes.
III. A BROWNIAN MOTION MODEL

The following sections present three cases in which the movement of molecules (or particles) is assumed to be governed by Brownian motion: (1) a single particle released from the origin at time \( t=0 \); (2) a number of particles (that is, a "puff" of particles or molecules) released from the origin at \( t=0 \); (3) a single particle released from the origin at a sequence of successive, equally spaced times (that is, a stream of particles emitted at a constant flow rate).\(^\dagger\)

In all three cases, the origin denotes the location of a single point source of air pollution, such as a single smoke stack, and zero drift of the carrier medium (that is, zero wind speed) is assumed. As an alternative model, one can visualize a closed room or experimental chamber, in which the pollutant (for example, a molecule of tracer gas) is released from a single point (the orifice of a nozzle) at the center of the room, and the pollutant is measured at a receptor at some arbitrary point in the room. Both the source and the receptor are stationary relative to each other, and \( X_j(t) \) denotes the horizontal distance of the \( j^{th} \) molecule from the source nozzle. Zero exit velocity from the nozzle is assumed, and the molecule or particle, once it leaves the source nozzle, is free to move throughout the room in response to collisions from other molecules present within the air mixture.

Relationship of Concentration to Counting Processes

Ordinarily, observed pollutant concentrations are treated as continuous variables, while counts of molecules or particles are discrete

\(\dagger\) Throughout this paper, the terms "molecule" and "particle" will be used interchangeably.
integer variables. For purposes of this paper, observed pollutant concentration will be related to discrete counting process variables in a relatively direct fashion. Because concentration is really the quantity of mass of the pollutant (or quantity of volume of a pollutant) per unit volume of the total mixture, the concentration will be proportional to the number of molecules counted in a particular carrier medium divided by the quantity of total material included in the count.

Suppose, for example, that one examines a volume of air $V$ to determine the concentration of a particular chemical species. If we were able to count $N_V$ molecules of the species in the volume $V$, we then could express the concentration as proportional to $N_V/V$. Thus, if $\gamma > 0$ is introduced as a proportionality factor, then the concentration $C_V$ in the sample of volume $V$ is given by:

$$C_V = \gamma \frac{N_V}{V}.$$  

The proportionality factor becomes necessary because concentration usually is expressed not in units of the number of molecules per unit volume, which gives very large values, but in units designed to give values in the range of $0.01 \leq C_V \leq 1000$, such as ppm or $\mu g/m^3$.

Although Equation 12 applies to a three-dimensional system, it is difficult to represent Brownian motion in three dimensions, so a simpler system commonly is used which can be viewed as an analogue to a three-dimensional model. If a two-dimensional system were used, pollutant concentration would be represented by an analogous quantity based on the number of particles per unit area:

$$C_A = \gamma' \frac{N_A}{A}.$$  

17
In Equation 13, if the concentration is not uniform throughout the carrier medium, $C_V$ will vary with the volume selected for measurement. Similarly, if the pollutant is assumed to be not uniformly mixed in the two-dimensional analogue, the concentration will vary with the particular area that is selected.

Finally, in one dimension, the concentration will be proportional to the number of particles counted in some segment of the line. In a linear reference system, if $(x_1, x_2)$ represent two points on the line such that $x_2 - x_1 = L$, and if $N$ molecules are counted in that segment, then the concentration $C_L$ measured in that segment will be given by:

\[
C_L = \gamma'' \frac{N_L}{L}.
\]

In equations 12, 13 and 14, nothing has been said about time, and it is assumed that $C_V$, $C_A$ and $C_L$ are instantaneous concentrations. To obtain the average concentration, it is necessary to integrate $C_L(t)$ over time $\tau$ and divide by $\tau$.

Because one is not interested in the actual numerical values of $C_L$, concentration in one dimension can be redefined using the units "number of molecules per unit length". With these units, $\gamma'' = 1$ and one-dimensional concentration can be written as follows, where $\delta = x_2 - x_1$, the distance between two points $(x_1, x_2)$ measured at some linear distance from the origin, where $x_2 > x_1$, and $N_\delta$ is the number of particles counted in this segment.

\[
C_\delta = \frac{N_\delta}{\delta} = \frac{N_\delta}{x_2 - x_1}.
\]

18
In the following discussion, \( x_1 = a \), a point some arbitrary distance from the origin, and \( x_2 = a + \delta \), where \( \delta \) is finite but can be arbitrarily small.

Here, \( x_1 = a \) denotes the observation or receptor point. Because a point has zero length, the concentration at \( x_1 = a \) is undefined, and one must speak of the concentration in the segment \( (a < X < a + \delta) \). Thus, the concentration is the number of particles present in the line segment \( x_1 = a \) and \( x_2 = a + \delta \), divided by the length of the segment. If \( N(x) \) denotes the number of particles present per unit length, then

\[
C_\delta = \frac{1}{\delta} \int_{a}^{a+\delta} N(x) \, dx.
\]

Here \( \delta \) can be arbitrarily small, for it is usual to measure pollutant concentration in a very small quantity of the carrier medium. In a three-dimensional system, the volume of the carrier medium would be relatively small, and, in a one-dimensional system, \( \delta \) becomes a correspondingly tiny segment of distance. In this paper, the sampling interval \( (a \leq X \leq a + \delta) \) will be called the "receptor segment". Because \( X \) is a continuous variable, the receptor segment can be expressed either by strict inequalities or regular inequalities; that is, \( (a \leq X \leq a + \delta) = (a < X < a + \delta) = (a < X < a + \delta) \).

In Equation 16, nothing is said about time, and it is assumed that \( C_\delta \) is the instantaneous concentration measured at some time \( t \) in the receptor segment. If \( N(x,t) \) denotes the number of particles present at location \( x \) and time \( t \), then the average concentration \( \overline{C}_\delta \) measured in the receptor segment would be as follows:

\[
\overline{C}_\delta = \frac{1}{T\delta} \int_{0}^{T} \int_{a}^{a+\delta} N(x,t) \, dx \, dt.
\]
Path of a Single Particle in One Dimension

Consider a one-dimensional reference system in which \( X(t) \) denotes the distance of a particle from the origin as a function of time, and \( (a \leq X \leq a+\delta) \) denotes the receptor segment (Figure 3). The particle is released at the origin at time \( t=0 \) such that \( X(0) = 0 \), and the trajectory of the particle after its release results from numerous collisions with various molecules in the carrier medium. As discussed in various texts on the topic of Brownian motion [15, 16, 17], three basic assumptions apply to the movement of the particle; (a) the increments \( X(t+s) - X(s) \) for all \( t, s \in \mathbb{R}_+ \) are normally distributed with zero mean and variance \( t \); that is, \( X(t+s) - X(s) \) is \( \mathcal{N}(0,t) \); (b) \( X(t) \) is a process with independent increments; that is, \( X(t+s) - X(s) \) is an independent, identically distributed random variable for all \( t, s \in \mathbb{R}_+ \); (c) \( X(t) \) is a continuous function of \( t \) for all \( t \in \mathbb{R}_+ \). Although the sample path \( X(t) \) is continuous, its derivative exists nowhere.

If the path of a particle obeys the above assumptions of independence, stationarity, and continuity, then the probability distribution of the location of the particle at any time \( t > 0 \) is Gaussian with mean 0 and variance \( t \):

\[
P(X \leq b) = \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} \, dx.
\]

The probability that the particle is found within the receptor segment \( (a \leq X \leq a+\delta) \) is obtained by integrating over the length of the segment:

\[
P(a \leq X \leq a+\delta) = \int_{a}^{a+\delta} \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} \, dx.
\]
Figure 3. One-dimensional path of a particle released at the origin at $t=0$ and subjected to Brownian motion, showing the receptor segment ($a \leq X \leq a+\delta$) somewhat enlarged.
The probability density function \( f_x(x) \) for the position of the particle is given by:

\[
(20) \quad f_x(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}.
\]

This function satisfies the diffusion equation, which can be shown by differentiating Equation 20 once with respect to \( t \) and twice with respect to \( x \):

\[
(21) \quad \frac{\partial f}{\partial t} = \frac{1}{2} \left( \frac{x^2}{t^2} - \frac{1}{t} \right) \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}
\]

\[
(22) \quad \frac{\partial^2 f}{\partial x^2} = \left( \frac{x^2}{t^2} - \frac{1}{t} \right) \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}.
\]

Thus,

\[
(23) \quad \frac{\partial f}{\partial t} = \frac{1}{2} \frac{\partial^2 f}{\partial x^2}.
\]

We shall assume that each time the particle enters the receptor segment \((a \leq X \leq a+\delta)\) it is counted once and thereby contributes to the observed concentration given in Equation 15. If the particle does not enter the receptor segment at all, then \( N_\delta = 0 \), giving \( C_\delta = 0 \). For the particle to enter the receptor segment, it is necessary for the particle first to reach the location \( X=a \). Let \( T_a \) denote the "first hitting time", which is the time at which \( X(t) \) first attains the value \( X=a \). As shown elsewhere in the literature, \([16, 17]\), the probability distribution of \( T_a \) can be derived by means of the reflection principle, giving the following relationship:
(24) \[ P\{T_a \leq t_a\} = \int_0^{t_a} \frac{a}{\sqrt{2\pi t^3}} e^{\frac{-a^2}{2t}} \, dt. \]

If a given particle enters the receptor segment \((a \leq X \leq a+\delta)\) and then exits, then \(N_0 = 1\). Once the particle leaves the receptor segment, however, the same particle can return. At the boundaries of the receptor segment, counting the number of departures and returns is not possible due to the fundamental properties of the process. At times immediately after the first hitting time \(t > t_a\), the particle can cross the point \(X=a\) an infinite number of times, causing an infinite number of entries and departures. To obtain a sense of the distribution of the number of return visits into the receptor segment, it is necessary to define a "return visit" as one in which the particle first travels beyond some distant point \(X=b\), where \(b > a\), prior to its return into the receptor segment.

At the first hitting time \(t = t_a\), it is possible for the particle to overshoot beyond the point \(X=a\), since \(X(t_a) > a\). For simplicity, we shall assume that the particle is located approximately at \(X=a\) at the first hitting time \(t = t_a\); that is, \(X(t_a) = a\). By the first assumption of the Brownian motion process, each increment is independent of the previous increment. Thus, the process exhibits the Markov property that the position of the particle at \(X(t_a + s)\) depends only on its position at \(X(t_a)\) and not on previous positions. Because of this "memoryless" property, it is possible, once the position of the particle is known, to treat the process as beginning over again, or as "updated". If this process is assumed to be updated once the particle hits the boundary of the receptor segment \(X(t_a) = a\), then one can treat the
origin as translated from its original position to the new position at \( X(t') = a \), where \( t' = t - t_a \), and \( X' = X - a \). Substituting \( t' \) for \( t \) and \( X' \) for \( X \) into the original probability distribution given by Equation 18, the probability that the particle is present in the region \( X > b \), given that it originates at \( X(t_a) = a \), is expressed as follows:

\[
(25) \quad P\{b < X(t) | X(t_a) = a\} = \int_b^\infty \frac{1}{\sqrt{2\pi(t-t_a)}} \exp\left(-\frac{(x-a)^2}{2(t-t_a)}\right) \, dx
\]

for \( t > t_a \).

Now suppose that, once attaining the position \( X(t) > b \), the particle again returns to the receptor segment \((a < X < a+\delta)\). If the particle reaches \( X=a \) at the first return time \( t = t_{r1} \), then the process once again is updated. A "return time" is the time at which the particle returns to \( X=a \), given that it has visited the region \( X > b \). Because the process is updated at each return time, the times between successive returns \((t_{r2} - t_{r1}), (t_{r3} - t_{r2}), \ldots, (t_{rm} - t_{r(m-1)})\) will be independent. Let \( p_r \) denote the probability that a particle returns to \( X=a \) over some time \( T > t_a \), given that it has visited \( X > b \). Then the probability of two returns in time \( T \) will be \( p_r^2 \), and the probability of three returns in time \( T \) will be \( p_r^3 \).

Thus, the number of returns in time \( T \) will have a geometric probability distribution with parameter \( p_r \):

\[
(26) \quad P\{k \text{ returns in time } T\} = (1 - p_r)(p_r)^k
\]

for \( k = 0, 1, \ldots, \infty \).
Each time the particle returns to $X=a$ and crosses the receptor segment, $N_0$ is incremented by one. Thus, the average concentration $\bar{C}_0$ measured at the receptor segment ($a \leq X \leq a+\delta$) over some time $\tau$ will be proportional to the total number of return visits and will have a geometric probability distribution. The geometric probability distribution is right-skewed, and thus the average concentration will be right-skewed.

Although the above argument applies only to a particle that moves to the region $X \geq b$ and returns, a similar argument can be given for a particle that moves in the opposite direction to some point $X \leq c$, where $c < a$. By symmetry, the two cases are analogous. Thus, the average concentration will have a right-skewed probability distribution. That is, if the process is allowed to operate for some averaging time $\tau$ and if the process involving a single particle is repeated over and over, the resulting average concentrations should be right-skewed. Because the geometric probability distribution has its mode at $k=0$, however, this reasoning is not sufficient to explain why concentrations should have an offset mode.

**Single Puff of Many Particles**

In the previous section, it was assumed that a single particle was released from the origin at $t=0$, and its sample path, or one of many possible realizations of the trajectory of the particle, was described by the Brownian motion process $\{X(t): t \geq 0\}$. Now consider a one-dimensional Brownian motion process in which a very large number of particles, which can be described as a "single puff", is released from the origin at $t=0$ (see Figure 4). Let $N_0$ denote the number of particles released at $t=0$. In addition to the previous assumptions of independent increments,
Figure 4. Paths of many particles of a pollutant released from the origin at $t = 0$ in a "single puff" and subjected to Brownian motion, showing the receptor segment ($a \leq X \leq a + \delta$).
stationarity, and continuity, assume also that the sample paths are independent of each other and that, even if they cross, they do not interact in any fashion. That is, each particle \( j \) is a separate realization of the Brownian motion process \( \{X_j(t) : t \geq 0\} \), and the path \( X_j(t) \) is independent for all \( j \in N_0 \). Thus, the probability density function for the \( j^{th} \) particle is given by

\[
f_{X_j}(x_j) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x_j^2}{2t}}, \quad \forall j \in N_0.
\]

From the earlier discussion of Equations 16 and 17, the observed instantaneous concentration at time \( t \) and receptor point \( x=a \) is proportional to the number of particles present in the tiny receptor segment between \( x=a \) and \( x = a+\delta \). If \( N(x,t) \) denotes the number of particles present at any distance and at any time \( t \geq 0 \), then the number of particles \( N_\delta(t) \) contained in the receptor segment at time \( t \) is given as follows:

\[
N_\delta(t) = \int_a^{a+\delta} N(x,t) \, dx.
\]

Because of the stochastic nature of the process, the exact number of particles present in the interval \( x \in (a, a+\delta) \) at any instant of time cannot be computed. However, if one knows the probability that a particle is within the receptor segment at time \( t \); that is, \( P(a < X \leq a+\delta|t) \), then the expected number of particles present \( E[N_\delta(t)] \) is computed as follows:

\[
E[N_\delta(t)] = N_0 P(a < X \leq a+\delta|t) = N_0 \int_a^{a+\delta} \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} \, dt.
\]
Here $E[N_\delta(t)]$ is an "intensity"; it is the expected number of particles present within the small line segment $X=a$ and $X = a+\delta$ at time $t$. For $N_0$ very large, the pollutant concentration $C_\delta(t)$ will be roughly proportional to the expected number of particles in the receptor segment:

$$
C_\delta(t) \approx \frac{E[N_\delta(t)]}{\delta} = \frac{N_0}{\delta} \int_a^{a+\delta} \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} \, dx
$$

for $N_0 >> 0$.

If $\delta$ is very close to zero, $E[N_\delta(t)]$ will be approximately proportional to the density function contained within the integral in Equation 19. Thus, if $\delta$ is allowed to be very small, then the concentration $C_\delta(t)$ observed at the receptor segment $X \in (a, a+\delta)$ will be approximately proportional to the probability density function $f_X(x)$, where $\gamma$ is a proportionality constant that includes $N_0$:

$$
C_\delta(t) = \gamma f_X(x) = \frac{\gamma}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}.
$$

It is of interest, therefore, to examine the behavior of this density as a function of $t$ for selected values of $X=a$. To emphasize that this density is a function of $t$, and that $X=a$ is fixed, it will be written as $f_X(a|t)$:

$$
f_X(x|t) \bigg|_{x=a} = f_X(a|t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{a^2}{2t}}.
$$

For small values of $a > 0$, the function rises very abruptly from the origin at $t=0$, reaching a single maximum $f_{\text{max}}$ at $t = t_m$, thereafter
declining very slowly with time (see Figure 5). The mode \( t_m \) is found by setting the first derivative with respect to time, Equation 21 equal to zero. Thus,

\[
0 = \frac{1}{2} \left( \frac{x^2}{t_m^2} - \frac{1}{t_m} \right) \frac{1}{\sqrt{2\pi t_m}} e^{-\frac{x^2}{2t_m}}.
\]

or

\[
t_m = x^2 = a^2.
\]

At the mode \( t_m = a^2 \), the value of the maximum is obtained by substituting \( t_m \) into Equation 23:

\[
f_{\text{max}} = \frac{1}{a\sqrt{2\pi}} e^{-1/2} = 0.24197 \ a^{-1}.
\]

Thus, the maximum of this function is inversely proportional to \( x=a \) and occurs at \( t_m = a^2 \).

The curves in Figure 5 were plotted to give insight into the concentration \( C_t(t) \) as a function of time at various distances \( a=0, a=1, a=2, a=3, \) and \( a=4 \) from the origin. Except for the case of \( a=0 \), for which \( f_x(0|t) \) approaches \( \infty \) as \( t \) approaches 0, all of the plots of \( f_x(a|t) \) have a single mode at \( t_m = a^2 \). However, as the receptor point is placed greater and greater distances from the origin, the shape of the curve near its maximum becomes very smooth and loses its sharpness (i.e., "peakedness") while the maximum, \( f_{\text{max}} \), decreases as \( 1/a \). Even for the relatively small change in distance from \( a=1 \) to \( a=4 \) in Figure 5, the curve becomes very smooth and gradual, and it is difficult to determine by inspection where the maximum lies. For \( t >> a \), \( e^{-a^2/2t} \) asymptotically
Figure 5. Plot of $f(a|t)$ as a function of $t$ for selected values of $a$.

$$f(a|t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{a^2}{2t}}$$
approaches $(2\pi t)^{-1/2}$. This function has a very long right tail, decreasing toward the origin very slowly. If $e^{-a^2/2t}$ is expanded as a Taylor series, Equation 32 can be written as follows:

$$f_X(a|t) = \frac{1}{\sqrt{2\pi t}} \left[ 1 - \frac{a^2}{2t} + \frac{a^4}{212t} - \ldots \right].$$

Thus, for $t >> a$, all terms in the brackets that involve $a$ and $t$ can be disregarded:

$$f_X(a|t) \approx \frac{1}{\sqrt{2\pi t}} \quad \text{for } t >> a.$$  

Because $e^{-a^2/2t} = 1$ for $a=0$, but $e^{-a^2/2t} < 1$ for all $a > 0$, the following function can be considered an "outer envelope" for all the functions for which $a \geq 0$:

$$f_X(0|t) = \frac{1}{\sqrt{2\pi t}} \quad 0 < t < \infty.$$  

Consider the case in which $a=0$ and the receptor point is located at the origin. At $t=0$, all particles are contained within the infinitesimal segment located between $X=0$ and $X=\delta$. Thus, integration over the receptor segment must give all $N_o$ molecules, and the concentration should be $C_\delta(0) = N_o/\delta$. However, as $t$ approaches 0, $f_X(0|t)$ approaches infinity and is not defined, so this must be treated as a special case. For cases in which $a \neq 0$, the concentration, as $t$ increases, rises to a maximum value at $t_m = a^2$, decreasing thereafter very slowly toward $X=0$ (but never reaching it), and ultimately declining as $1/\sqrt{t}$ for $t >> a$.

If one considers some overall time period $\tau$ of interest, then the average concentration over time $\tau$ is computed using Equation 1 as follows:
(38) \[ \bar{C}_\delta = \frac{1}{\tau} \int_0^\tau C_\delta(t) \, dt. \]

For the simplified case of the outer envelope in Figure 5, substituting Equation 31 for \( X=0 \) into Equation 38, the average concentration is computed as follows:

(39) \[ \bar{C}_\delta = \frac{\gamma}{\tau} \int_0^\tau \frac{1}{\sqrt{2\pi t}} \, dt = \frac{\gamma}{\delta} \left[ -\frac{t^{1/2}}{2\sqrt{2\pi}} \right]_0^\tau = \frac{\gamma}{2\sqrt{2\pi}}. \]

Now consider the time series of concentration resulting from a single puff of many particles released from the origin at \( t=0 \) and the histogram that is generated by this time series. What will be the shape of this distribution of concentrations over some finite observation time, and will it be right-skewed? For illustrative purposes, let \( \gamma=1 \) in Equation 31, and assume that time and concentration are rescaled for simplicity. For example, if \( N_0 \) equals 1 million particles, then we assume that \( C_\delta \) and \( t \) are expressed in appropriate concentration and time units such that \( \gamma=1 \). Assume that one receptor point is located at \( a=0 \) and another is located at \( a=1 \).

Suppose that the total observation time of interest is \( T = 100 \) time units and that one chooses each class interval of the histogram equal to \( \Delta C = 0.05 \) concentration units. To construct the histogram of concentrations during the observation time, one is interested in the projection of successive concentration intervals on the time axis. In Figure 6, the axes of Figure 5 have been interchanged, and the histogram is constructed from this figure by determining the amount of time that the concentration lies in specified ranges (Figure 7). That is, the magnitude of the \( k^{th} \) bar comprising the histogram will be determined by finding
Figure 6. Plot of f(a|t) for a = 0 and a = 1 with the axes of Figure 5 interchanged, showing how a histogram of concentrations can be constructed by reading Δt from the figure for each successive ΔC, where ΔC = 0.05 concentration units.
Figure 7. Histogram generated at location $a = 0$ by a single puff of many particles observed from time $t = 0$ to $t = 100$, with intervals of size $\Delta C = 0.05$. 
the time during which \( C_\delta(t) \) lies in the interval \( (k-1)\Delta C < C_\delta(t) < k\Delta C \) for \( k = 1, 2, 3, \ldots, m \), where \( m \) is the number of intervals comprising the histogram. For the first interval and for \( a = 0 \), for example, the time during which \( 0 < C_\delta(t) < 0.05 \) is found by solving Equation 37 for \( t \), giving \( t(C_\delta) = \left[ \frac{2\pi C_\delta^2}{2} \right]^{-1} = \left[ \frac{2\pi(0.05)^2}{2} \right]^{-1} = 63.7 \) time units. Thus, if the particles are released at \( t = 0 \), then \( C_\delta(t) \) first declines below \( C_\delta = 0.05 \) concentration units at time \( t = 63.7 \) time units. If the observation period continues until \( t = 100 \), then the total time during which \( C_\delta(t) \) lies in the interval \( k = 1 \) is given by \( \Delta t_1 = 100 - 63.7 = 36.3 \) time units. Similarly, the time spent in the interval \( k = 2 \) is determined by first solving Equation 37 for \( t(0.1) = \left[ \frac{2\pi(0.1)^2}{2} \right]^{-1} = 15.9 \) time units and then subtracting this result from the previous concentration interval crossing time, giving \( \Delta t_2 = 63.7 - 15.9 = 47.7 \) time units. The result of these computations for 20 intervals (Table 1) reveals that \( \Delta t_k \) becomes successively smaller as \( k \) approaches 20 but never quite reaches zero. The result is to produce a distinctly right-skewed histogram with a long tail to the right (Figure 7).

Thus, this Brownian motion model will generate a distinctly right-skewed histogram when the receptor point is located at the origin, \( a = 0 \). For the case of \( a = 1 \), the distribution will be right-skewed, but the highest concentration interval will be \( C = 0.25 \) concentration units, and the histogram no longer will have a long tail to the right. This happens because the extremely high concentrations that occur at \( t = 0 \) when the receptor point is very close to the origin are greatly attenuated when the receptor point is moved to a finite distance from the origin, thus reducing the magnitude of the maximum concentration encountered. Notice that the lowest concentration encountered in the time range \( 0 < t < 100 \)
Table 1

Histogram Generated by a Single Puff of Many Particles
Observed from $t = 0$ to $t = 100$ at Location $a = 0$, $\Delta C = 0.05$

<table>
<thead>
<tr>
<th>$k$</th>
<th>$k\Delta C$</th>
<th>$[2\pi(k\Delta C)^2]^{-1}$</th>
<th>$\Delta t_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>63.662</td>
<td>36.338</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
<td>15.915</td>
<td>47.747</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
<td>7.074</td>
<td>8.842</td>
</tr>
<tr>
<td>4</td>
<td>0.20</td>
<td>3.979</td>
<td>3.095</td>
</tr>
<tr>
<td>5</td>
<td>0.25</td>
<td>2.546</td>
<td>1.432</td>
</tr>
<tr>
<td>6</td>
<td>0.30</td>
<td>1.768</td>
<td>0.778</td>
</tr>
<tr>
<td>7</td>
<td>0.35</td>
<td>1.299</td>
<td>0.469</td>
</tr>
<tr>
<td>8</td>
<td>0.40</td>
<td>0.995</td>
<td>0.305</td>
</tr>
<tr>
<td>9</td>
<td>0.45</td>
<td>0.786</td>
<td>0.209</td>
</tr>
<tr>
<td>10</td>
<td>0.50</td>
<td>0.637</td>
<td>0.149</td>
</tr>
<tr>
<td>11</td>
<td>0.55</td>
<td>0.526</td>
<td>0.110</td>
</tr>
<tr>
<td>12</td>
<td>0.60</td>
<td>0.442</td>
<td>0.084</td>
</tr>
<tr>
<td>13</td>
<td>0.65</td>
<td>0.377</td>
<td>0.065</td>
</tr>
<tr>
<td>14</td>
<td>0.70</td>
<td>0.325</td>
<td>0.052</td>
</tr>
<tr>
<td>15</td>
<td>0.75</td>
<td>0.283</td>
<td>0.042</td>
</tr>
<tr>
<td>16</td>
<td>0.80</td>
<td>0.249</td>
<td>0.034</td>
</tr>
<tr>
<td>17</td>
<td>0.85</td>
<td>0.220</td>
<td>0.028</td>
</tr>
<tr>
<td>18</td>
<td>0.90</td>
<td>0.196</td>
<td>0.024</td>
</tr>
<tr>
<td>19</td>
<td>0.95</td>
<td>0.176</td>
<td>0.020</td>
</tr>
<tr>
<td>20</td>
<td>1.00</td>
<td>0.159</td>
<td>0.017</td>
</tr>
</tbody>
</table>
occurs at time $t = 100$, giving $C_0(100) = 1/\sqrt{2\pi(100)} = 0.04$ concentration units. Thus, if the histogram intervals are smaller than this quantity, at least the first interval will be empty. Considerable time must elapse before the concentration becomes very close to zero, and thus the concentration distribution over any arbitrary time period generally will exhibit a lower boundary that is greater than zero. Because the slope $t'(C)$ rapidly increases as C approaches the origin, or as C approaches its lower boundary, the lowest nonzero bars of the histogram will tend to be higher than bars further to the right, and the histogram will tend toward right-skewness. In the case for $a >> 0$, a variety of histograms are possible, depending on the values of a and T. When the receptor point is very far from the origin, the maximum concentration will be attenuated greatly, and the range between the highest and lowest concentration will be narrowed.

The above analysis has demonstrated that a purely symmetrical process, such as Brownian motion with Gaussian increments, can produce distinctly right-skewed distributions. A right-skewed distribution of concentration can be observed by measuring concentration at a fixed location. Here, the total population of average concentrations computed over time for a single puff released at time $t=0$ will be right-skewed. That is, the above argument applies not to repeated runs of this process but to a single run of the process. In this analysis, the receptor segment is intended to be analogous to a conventional fixed air monitoring station operating at a fixed distance from a single source, with no winds. It should be noted that the resulting distribution, although skewed to the right, is not lognormal in any sense. Thus, this analysis indicates that a relatively simple Brownian motion process can generate a right-skewed, nonlognormal distribution of average concentrations with an offset mode.
Successive Emissions of Single Particles

Sources of air pollution usually release their emissions on a "continuous" basis; that is, the gases or particulate matter leave the exhaust stack at a more-or-less constant flow rate, often measured in grams per second. The two cases discussed above - a single particle released at the origin and a puff of many particles released at $t=0$ - do not reflect the continuous nature of pollutant emissions. To construct a Brownian motion model that is analogous to continuous source emissions, we shall assume that a constant rate of emissions can be decomposed into single events by dividing time sufficiently fine. Furthermore, to simplify the emission process, we assume that each event, consisting of a single particle released at the origin, is equally spaced in time. Finally, we assume that the paths of the particles are independent of each other and are subject to the assumptions of Brownian motion discussed above.

Consider the one-dimensional Brownian motion process $\{X_j(t): t \geq 0, \ j \in \mathbb{N}\}$ in which $n$ denotes the total number of particles released at the origin. The $j$th particle is released at the origin at times $(t_0, t_1, t_2, \ldots, t_{j-1}, \ldots, t_{n-1})$, and the difference between any two successive times $\Delta t = t_j - t_{j-1}$ for $j=1, 2, \ldots, n$ is constant (see Figure 8). The three Brownian motion assumptions of independent Gaussian increments, stationarity, and continuity apply to this process. Furthermore, it is assumed that particles cannot collide or interact with each other in any fashion and the paths $X_j(t)$ are independent for all $j$. As before, a receptor segment exists such that a particle entering the interval $X_j(t) \in (a, a+\delta)$ contributes to observed concentration.

To simplify the model further, we assume that, as $\delta$ becomes very small and approaches zero, each particle that reaches the receptor
Figure 8. Brownian motion paths of particles emitted from the origin at successive, equally spaced times.
segment is counted only once. That is, particle \( j \) is counted as reaching
the receptor segment when it first attains the position \( X_j(t) = a \), and no
return visits to the receptor segment by particle \( j \) are permitted. In
effect, the boundary \( X=a \) of the receptor segment behaves as an absorbing
barrier. The result of these assumptions is to focus attention on the
first hitting time of each particle. If \( N_\tau \) denotes the number of particles
that reach the receptor segment for the first time during the averaging
period \( 0 \leq t \leq \tau \), then the average concentration \( \bar{c}_0(\tau) \) over time \( \tau \) will
be proportional to \( N_\tau \):

\[
(40) \quad \bar{c}_0(\tau) = \gamma N_\tau .
\]

Here, \( \gamma \) is a constant of proportionality differing in value from \( \gamma \) in the
previous section.

One also can visualize successive averaging times \( \tau_1, \tau_2, \ldots, \tau_m \),
each separated by a fixed averaging period \( \Delta \tau = \tau_k - \tau_{k-1} \) for \( k-1, 2, \ldots, m \), where \( \Delta \tau > \Delta t \). We are interested in the distribution of average
concentrations \( c_1, c_2, \ldots, c_m \) generated at these successive times.

Let \( T \) denote the first hitting time of a particle that is emitted
from the origin at \( t=0 \). From Equation 24, which is based on the reflection
principle, the probability density function \( f_T(t|a) \) for the random
variable \( T \), given the position of the receptor point \( X=a \), is as follows:

\[
(41) \quad f_T(t|a) = \frac{a^2}{\sqrt{2\pi t^3}} e^{-\frac{a^2}{2t}} .
\]

Thus function has a very gradual decay for times \( t >> 0 \) (Figure 9). For
t > 0 in this model, each successive particle is emitted from the origin
Figure 9. Plot of the density $f_T(t|a)$ of the first hitting time $T$ of a particle first attaining the position $X = a$, for selected values of $a$. 

$$f_T(t|a) = \frac{a}{\sqrt{2\pi t^3}} e^{-\frac{a^2}{2t}}$$
at times $\Delta t$, $2\Delta t$, ..., $(j-1)\Delta t$, ..., $(n-1)\Delta t$, and the probability that particle $j$ reaches the receptor point before time $t$ is obtained by translating the time axis in Equation 41 by $j\Delta t$ time units:

$$P\{T_j < \tau\} = \int_0^{\tau} \frac{a}{\sqrt{2\pi[t-(j-1)\Delta t]^3}} e^{-\frac{a^2}{2t}} dt$$

for $j=1, 2, ..., n$.

Suppose averaging times $\tau_1$, $\tau_2$, ..., $\tau_k$, ..., $\tau_m$ are considered. Then the probability that the $j^{th}$ particle arrives at $X=a$ in the $k^{th}$ time period, $p_{jk}$, is given as follows:

$$p_{jk} = P\{\tau_{k-1} < T_j < \tau_k\} = \int_{(k-1)\Delta t}^{k\Delta t} \frac{a}{\sqrt{2\pi[t-(j-1)\Delta t]^3}} e^{-\frac{a^2}{2t}} dt$$

for $k=1, 2, ..., m$ and $j=1, 2, ..., n$.

By Equation 40, the average concentration over any averaging period will be proportional to the number of particles arriving over the period. Let $N_{\tau_k}$ denote the number of particles that arrive during the $k^{th}$ averaging period, $(k-1)\Delta t < t < k\Delta t$, for $k=1, 2, ..., m$. Then the average concentration observed in the $k^{th}$ averaging period, $\bar{c}_{\delta k}$, will be proportional to $N_{\tau_k}$. Because analytical treatment becomes difficult for large values of $j$ and $k$, it is useful to consider a simple example to gain insight into this model.

**Example.** Consider, for illustrative purposes, an example in which $\Delta t = 2\Delta t$, and only 2 particles are released at the origin during each
averaging period (Figure 10). During the first averaging period \( \tau_1 \), it is possible for only the two particles released at \( t_0 = 0 \) and \( t_1 = \Delta t \) to arrive. Using the notation of Equation 43 for \( k=1 \) and \( j=1, 2 \), the probabilities that each of these particles arrives during this period are given by the following equations, where \( \tau_1 = 2\Delta t \).

\[
(44) \quad P_{11} = \int_0^{\tau_1} \frac{a}{\sqrt{2\pi t^3}} e^{-\frac{a^2}{2t}} \, dt
\]

\[
(45) \quad P_{21} = \int_0^{\tau_1} \frac{a}{\sqrt{2\pi (t-\Delta t)^3}} e^{-\frac{a^2}{2(t-\Delta t)}} \, dt.
\]

Because particle arrivals are independent events, the probabilities that 0, 1, or 2 particles arrive are given as follows:

\[
(46) \quad P_0 = P\{N_{\tau_1} = 0\} = (1 - P_{11})(1 - P_{21})
\]

\[
(47) \quad P_1 = P\{N_{\tau_1} = 1\} = (1 - P_{11})P_{21} + (1 - P_{21})P_{11}
\]

\[
(48) \quad P_2 = P\{N_{\tau_1} = 2\} = P_{11}P_{21}.
\]

If the first arrival probabilities, \( P_{11} \) and \( P_{21} \), were equal for these two particles, which they are not, then the result would be a Bernoulli process, and \( N_{\tau_1} \) would have a binomial distribution. The result of unequal arrival probabilities is to generate a "weighted" Bernoulli process.

Under what conditions, if any, will \( N_{\tau_1} \) be right-skewed, giving a right-skewed distribution of average concentration in the first averaging period? This distribution will be right-skewed if \( P_0 > P_2 \) and if the
Figure 10. Example showing successive particles released at equally spaced times, with two particles released during each averaging period of duration $\Delta t$, giving $\Delta \tau = 2\Delta t$. 
mode is $p_1$; that is, the distribution is right-skewed if $p_1 > p_0 > p_2$.

If $p_0 > p_2$, then $(p_0 - p_2) > 0$, and, subtracting Equation 48 from Equation 46, $(1 - p_{11} - p_{21}) > 0$, giving the following condition:

(49) 
$$ (p_{11} + p_{21}) < 1. $$

If the above condition is met, and if $p_1 > p_0$, then $p_1$ will be the mode of this distribution. This second condition implies that $(p_1 - p_0) > 0$, and, subtracting Equation 46 from Equation 47, $(-1 + 2p_{11} + 2p_{21} - 3p_{11}p_{21}) > 0$, which implies that $(p_{11} + p_{21}) > \left(\frac{1}{2} + \frac{3}{2} p_{11}p_{21}\right)$. If the condition specified by Equation 49 is satisfied, then $1 > (p_{11} + p_{21}) > \left(\frac{1}{2} + \frac{3}{2} p_{11}p_{21}\right)$, giving $1 > \left(\frac{1}{2} + \frac{3}{2} p_{11}p_{21}\right)$, or $\frac{1}{2} > \frac{3}{2} p_{11}p_{21}$. This inequality gives the following condition which also must be satisfied:

(50) 
$$ p_{11}p_{21} < \frac{1}{3}. $$

Thus, if the conditions specified in Equations 40 and 41 both are satisfied, then $N_{\tau_1}$ will have a right-skewed probability distribution, and the average concentration $\overline{C_{\delta\tau_1}}$ in the first averaging period $\tau_1$ will have a right-skewed probability distribution.

Thus the probability distribution of the concentration averaged over time $t=0$ and $t = \tau_1$ will be a weighted binomial, and it will be right-skewed if $p_{11}$ and $p_{21}$ are small enough to satisfy Equations 49 and 50. If $p_{11} < 0.5$ and $p_{21} < 0.5$, these conditions will be more than satisfied. Because the density $f_{\tau}(t|a)$ is right-skewed (see Figure 9), the area under the curve to the left of the mode $\tau_m$ will be less than 0.5. Thus, if $\tau_1 < \tau_m$, $p_{11} < 0.5$ and $p_{21} < 0.5$. By differentiating Equation 41 and setting the result to zero, the mode can be found to be

45
$\tau_m = a^2/3$. Consequently, if $\tau_1 < a^2/3$, repeated runs of this process will yield a right-skewed, weighted binomial distribution for the average concentration $C_{\delta t_1}$.

If we consider the second averaging period in this example, for $\tau_1 < t < \tau_2$, where $\tau_1 = 2\Delta t$ and $\tau_2 = 4\Delta t$, we find that it is possible for four particles - $j=1, 2, 3,$ and $4$ - to arrive. The arrival probabilities for particle $j$ in time period $k=2$ are given as follows:

(51) \[ p_{12} = \int_{\tau_1}^{\tau_2} \frac{a}{\sqrt{2\pi t^3}} e^{-\frac{a^2}{2t}} \, dt , \]

(52) \[ p_{22} = \int_{\tau_1}^{\tau_2} \frac{a}{\sqrt{2\pi (t-\Delta t)^3}} e^{-\frac{a^2}{2t}} \, dt , \]

(53) \[ p_{32} = \int_{\tau_1}^{\tau_2} \frac{a}{\sqrt{2\pi (t-2\Delta t)^3}} e^{-\frac{a^2}{2t}} \, dt , \]

(54) \[ p_{42} = \int_{\tau_1}^{\tau_2} \frac{a}{\sqrt{2\pi (t-3\Delta t)^3}} e^{-\frac{a^2}{2t}} \, dt . \]

As for the first interval, the probability distribution of the number of particles that arrive in the second time interval is a weighted binomial. For example, the probability that all four particles arrive is given as follows:

(55) \[ P\{4 \text{ particles arrive in the second interval}\} = p_{12}p_{22}p_{32}p_{42} . \]
As time passes, more particles are available - simply because the number of particles released increases with time. For each time interval of length $\Delta t$, 2 additional particles are available. At the end of any time period $t$ which is an even multiple of $\Delta t$, $n = t/\Delta t$ particles will have been released. Thus, in every successive averaging period, it is possible for more particles to arrive. However, if any averaging period of length $\Delta t$ is considered by itself, the probability that all $n$ particles arrive during the period is quite small (assuming all $p$'s are less than 0.5), because this probability is the product of $n$ particle probabilities:

\[
(56) \quad P\{\text{all } n \text{ particles arrive in time } \tau_{k-1} < t < \tau_k \} = \prod_{i=1}^{n} p_{ik}.
\]

If $\Delta t$ is small relative to the time of the overall process, which usually will be true, then each particle probability will be quite small, and the probability given by Equation 56 will be very small. Since Equation 56 gives the probability for the maximum concentration that can occur within a given averaging period, the resulting distribution will be right-skewed, with a long tail to the right. In the limit, as $n$ approaches infinity, this probability of the maximum concentration will approach zero.

This analysis shows that the distribution of concentrations within any given averaging period will be right-skewed with a long right tail. Because particles are released "continuously" in this process, the concentration distributions will be greater in magnitude as successive averaging periods are considered, but the distribution of concentrations in any one averaging period on successive runs will be a weighted binomial that is right-skewed.
Multiple-Particle Case Revisited

In the case of a single puff discussed earlier, a large number of particles are released all at once at the origin at time \( t=0 \). The above analysis suggests additional insights into this single-puff, multiple particle case.

Let \( s \) denote any arbitrary time and \( \Delta \tau \) denote the duration of the averaging period over the interval \((s, s+\Delta \tau)\). Assume, as before, that the paths of the particles are independent of each other. Then the probability that any particle first arrives at the receptor position \( X=a \) during this averaging period is given by integrating Equation 41 in the following manner:

\[
p = P \left\{ \text{first arrival at } X=a \text{ in the interval } (s, s+\Delta \tau) \right\} = \int_{s}^{s+\Delta \tau} \frac{a}{\sqrt{2\pi t^3}} e^{-\frac{t^2}{2a}} \, dt.
\]

Because each arrival is an independent event, and because \( N_0 \) particles were initially released, the number of first arrivals \( K \) has a standard (unweighted) binomial distribution given as follows, where \( p \) is the probability given by Equation 57:

\[
P(K=k) = \binom{N_0}{k} p^k (1-p)^{N_0-k}.
\]

Assuming that the averaging time \( \Delta \tau \) is small relative to the overall time scale of the process, \( p \ll 1 \). The coefficient of skewness \( \sqrt{\beta_1} \) for this distribution is given as follows:

\[
\sqrt{\beta_1} = \frac{1 - 2p}{[np(1-p)]^{1/2}}.
\]
If the distribution is to be right-skewed, \( \sqrt{\lambda_1} \) must be positive, which implies that \( (1 - 2p) > 0 \), or \( p < 1/2 \). As can be seen from Figure 9 and the discussion in the previous section, this condition is readily met, particularly if sampling is done at a distant point from the origin such that \( |a| >> 0 \). For the binomial distribution, the expected number of particles arriving in this interval is given as follows:

\[
E[K] = N_0 p.
\]

Because the concentration measured in this averaging period is assumed to be proportional to \( K \), the concentration \( \bar{C}_{\delta_S} \) also will have a binomial distribution that is right-skewed.

This analysis indicates that a very simple Brownian motion process consisting of a puff of \( N_0 \) particles released at the origin at time \( t=0 \) and subject only to random collisions without drift will give an average particle concentration at any fixed receptor site that is binomial with parameters \( N_0 \) and \( p \). The binomial distribution is unimodal and, for \( p < 1/2 \), it is right-skewed. This result applies to repeated runs of the process in which the average concentration is measured over the same interval \( (s, s+\Delta T) \). The average value of these repeated concentrations will be proportional to \( N_0 p \). If \( N_0 >> 0 \), as will be the case if the puff contains a large number of particles, the distribution will appear almost continuous, and the resulting distribution may appear to some observers to be lognormal. However, the theoretical development presented in this report does not show it to be lognormal.
IV. CONCLUSIONS

This paper has sought to provide insights into the reasons why unimodal, right-skewed frequency distributions are so often observed in measurements of pollutants in environmental media (air, water, and soil) by exploring the distributions generated by a simple Brownian motion process. In this process, concentration is observed at a fixed location analogous to a pollutant monitoring location, and the pollutant, consisting of a finite number of particles, is released from a single point source. The trajectories of the particles are subject to the usual Brownian motion assumptions, consistent with the laws of diffusion. By examining the probability distributions of concentrations generated by several different cases, it is seen that right-skewed unimodal distributions seem to be the rule rather than the exception.

It is shown first that concentration is related to a counting process based on the number of particles present in a small volume in three dimensions, or a receptor line segment in one dimension. In the first case considered, a single particle subjected to Brownian motion (independent Gaussian increments resulting from collisions with other particles and molecules) is released from the origin, and the number of return visits to the receptor segment, once the particle arrives, appears to have a geometric probability distribution. This argument needs more analytical treatment than the present paper allows, and the concept of a "departure" and a "return" is somewhat vague, but it should serve to emphasize the difference between a return visit and an original visit to the receptor segment.

In the second case considered, a large number of particles (a single "puff") is assumed to be released from the origin at time $t=0$. By con-
sctructing a histogram derived from the expected number of particles present in the receptor segment during successive averaging periods from t=0 to some finite time, the resulting distribution of concentrations measured at or close to the origin is shown to be unimodal and right-skewed, but clearly not lognormal. Unfortunately, other receptor segment positions and times may generate distributions with other shapes, and this analysis needs further development. The analysis for this case considers the probability distribution of successive average concentrations, rather than the probability distribution of concentrations observed after the same elapsed time on repeated runs. It appears that the problem of successive concentrations is more difficult to handle than the problem of repeated runs, because the variance of the process changes with time.

In the third case considered in this paper, the particles are assumed to be released at successive, equally-spaced times, as if a "continuous stream" of particles is emitted from a single source. Although the average concentration tends to increase with time because more particles are present, it is shown that the concentration at any fixed elapsed time during repeated runs of the process is right-skewed. Because each particle is released at a different time, its probability of arriving in the receptor segment at a given time interval differs, and the resulting average concentration over the interval, on repeated runs, is shown to have a "weighted" binomial distribution.

Finally, returning to the case of a single puff of particles released at time t=0, it is shown that the number of particles making their first arrival at position X=a during some averaging period will be binomially distributed. If N₀ particles are released at the origin, then the binomial
distribution will have parameters \( N_0 \) and \( p \), where \( p \) is obtained by integrating the probability density of first hitting times from time \( s \) to \( s+\Delta \tau \), where \( s \) is the beginning of the averaging period of interest and \( \Delta \tau \) is its duration. The expected number of particles arriving during the averaging period will be \( N_0 p \), and if \( p < 1/2 \), a condition that is readily satisfied, then the resulting binomial distribution will be right-skewed. Assuming that the concentration is proportional to the number of first arrivals of particles at position \( x=a \), the concentration, on repeated runs of this process, also will have a binomial probability distribution that is right-skewed. In real situations, \( p \) ordinarily will be quite small, and \( N_0 \) ordinarily will be quite large because it represents the initial number of very small particles or molecules emitted from a smoke stack. As a result, most data analysts will treat concentration as a continuous variable, seeking to represent it by a continuous probability distribution. In some situations, it is likely that the lognormal model has been erroneously chosen over the right-skewed binomial distribution because the shapes of the two distributions look similar. It also is possible that the binomial distribution may asymptotically approach the lognormal model in some situations. Hopefully, these issues will be probed at length in the future.

The analyses presented in this paper reveal why the ubiquity of unimodal, right-skewed distributions of observed pollutant concentrations is not surprising. If a simple diffusion process consisting of nothing more than a puff of particles released at the origin and subjected to Brownian motion without drift can produce a strikingly right-skewed distribution such as the binomial with small \( p \) and large \( N_0 \), then it is not surprising why more complicated diffusion processes also would produce unimodal, right-
skewed distributions. In fact, such distributions would seem difficult to avoid. Although this result applies mainly to average concentrations observed after the same elapsed time on repeated runs of the same process, it is easy to imagine that successive puffs from this source would be essentially the same as repeated runs of this process, particularly if drift is introduced to keep the concentration from rising with time, thus making the process stationary. Thus, it is likely that successive average concentrations \( \bar{C}_1 \), \( \bar{C}_2 \), \ldots, \( \bar{C}_n \), each with averaging period \( \Delta t \), also could be shown to have a right-skewed binomial distribution. It is hoped that future investigators will examine these predictions in greater detail, and I believe that a simple experiment with a tracer gas in a windless environment would yield distributions similar to those presented in this paper.

Although the present investigation has shown how unimodal, right-skewed distributions might naturally arise from diffusion processes, it has not shown these distributions to be lognormal. In fact, some of the concentration distributions generated by these processes could be mistakenly described by some observers as lognormal (or perhaps 3-parameter lognormal) when they really are not and can be shown theoretically to be other distributions. This finding illustrates the difficulty of making inferences about the nature of underlying distributions when only an empirical frequency plot is available. If a process is known to be able to generate a variety of right-skewed, unimodal frequency distributions, all of which are similar in shape, there will be serious problems in differentiating one from the other only on the basis of frequency plots.

In real air pollution diffusion processes, it is likely that observed concentrations are a mixture of distributions, and representing them by a
single probability model, such as the lognormal or three-parameter lognormal, is an erroneous oversimplification. Obviously, such representations may be useful for some limited applications, but the user should be cautioned that there is no theoretical basis for making predictions or probability statements based on a particular probability model such as the lognormal, and such predictions are likely to be in error.

Suppose, for example, that the following prediction is made based on an assumption that concentrations are lognormally distributed: "With r% emission reduction, the probability that concentration C will exceed the air quality standard Cₜ is P." Notwithstanding the fact that there is little knowledge about how the parameters of this model will change when emission reduction is actually implemented, the computed value of P will be in error if concentrations are not actually lognormally distributed as assumed. Thus an incorrect source control level r will be implemented and will not bring about the anticipated degree of improvement in air quality. The seriousness of the error will depend on the extent to which the mixture of distributions behaves in a manner similar to the assumed lognormal distribution. Obviously, in view of the issues raised in the present investigation and the tenuousness of the assumptions on which such regulatory predictions rest, there is need to conduct more "before-and-after" empirical studies in which a given level of source reduction has been applied to an area and the effect of these reductions on observed concentrations over a number of years has been examined. Because more than a decade of experience with the control of various air pollutants now exists in the U.S., most U.S. cities provide ideal laboratories for examining the accuracy of predictions based on the assumption of a particular probability
model. Future investigations need to answer the following questions: "Has the assumption that pollutant concentrations arise from a particular distribution improved the accuracy of the predictions?" "Was the same predictive capability achieved using the raw observed frequency distribution, without any assumed model?" "If use of a model improved the predictive power, which model gave the best predicted result?" "What general conclusions can be drawn about the application of probability models to all locations, and what specific conclusions apply to air quality distributions observed in a specific locale under particular atmospheric and source conditions?"
REFERENCES


