Stochastic Analysis for Context Modeling

Paul Pukite, Steve Bankes, Dan Challou
BAE Systems

Abstract: Models of the physical environment play an important role in supporting Model Based Engineering (MBE). This paper describes how fundamental principles of thermodynamics and statistical physics can be applied to create compact parameterized models capable of statistically capturing the patterns exhibited in wide range of environmental contexts. Such models will allow more efficient and systematic assessment of the strengths and weaknesses of candidate designs than is possible using benchmark datasets or test tracks, and this can play an important role in using computer simulation to produce better designs of complex cyber-physical systems more quickly, affordably, and reliably. This is the first in a series of papers on the statistical description of natural processes. Subsequent papers in this series will treat specific examples in greater detail than provided in this overview.

Preface

Models of natural and man-made environmental contexts are important for a variety of purposes, in particular to support computational assessment of the performance of candidate designs for engineered systems. Vehicles perform on roads or other terrain, and both mobile and static engineered structures must contend with wind, waves, precipitation, and corrosion. Much of the value of engineered systems comes from their performance in the context of the environments in which they operate. Consequently, reliable assessment of candidate engineering designs requires models of potential operational contexts that accurately capture both the variability and overarching patterns presented by these environments.

Without computer simulation, candidate designs must be physically tested on sample deployed environments, for example test tracks. One aspiration for Model Based Engineering (MBE) is to achieve cost and other savings by replacing physical testing with testing through simulation. This requires validated models of the context environment that perform at least as well as a physical test in revealing design problems. However, simulation based testing can provide value beyond the elimination of physical testing. A significant cause of design failure often relates to providing an insufficient safety margin to deleterious environmental factors and contextual variables, such as an exceedingly rough terrain or corrosive atmospheric elements. Environmental context models that parametrically span the range of possible environmental challenges can provide a means for establishing the level of challenge at which a design will fail, generating considerably more information than produced by benchmark tests.

Environmental models that can replicate data derived from a wide range of actual environments are highly valuable, both through potential reuse across many design activities, and by supporting the creation of more robust and adaptable systems. Many of the patterns observable in contextual environments stem from fundamental processes derived from statistical mechanics and thermodynamics. Consequently, significant insight can be gained by approaching the patterns from first-principles physics. A combination of data and physical theory can result in models that extrapolate and accurately infer behavior outside of the data sets used for modeling. This approach results in models that are much more reliable for prediction and verification tasks than can be achieved through generic statistical modeling uninformed by physical considerations

This paper introduces a set of research results relating to stochastic modeling of environmental contexts. We first provide some background to probability theory and for formulating the building blocks for stochastic analysis — in particular, that of *probability density functions* and Markov and semi-Markov processes. These building blocks comprise the core fundamentals of stochastic modeling, whereby we can reason about probabilities, sampling, and uncertainty. The more sophisticated techniques of autocorrelation and power spectrum densities are elaborated in an associated report on terrain

characterization, and models of dispersive growth and uncertainty propagation are described within an associated report on diffusion process context models.

Background: Environmental models for cyber-physical systems

Two criteria are central in crafting probabilistic models for observed behavior: (1) the importance or impact of observed events and (2) the associated likelihood of the event. For MBE, context models are needed for phenomena that play a significant role either through their frequency of occurrence or the severity of their impact. Context models portray the distribution of a metric of the phenomenon of interest. Metrics must be measurable or countable and could involve extensive variables such as volumes or consist of ratios such as rate (i.e. volume over time).

Table 1 : The use of probabilities can describe high likelihood and high impact events.

	High Likelihood	High Impact
Terrain	RMS roughness	Steep slopes
Rainfall	Humid climates	Heavy downpours
Wind	Prevailing winds	Gusts
Waves	Chop and swell	Rogue waves
Particulates	Aerosols	High density volcanic dust

The fundamental building block in the creation of context models is the use of probability density functions (PDF) to model sample spaces. These facilitate the characterization and modeling of natural phenomena that are prevalent in human environmental contexts, including distribution of terrain slopes, wind velocities, rainfall amounts, etc. These can both model the high likelihood events through sampling of the meat of the distribution curves, but also provide for the rare cases through the concept of exceedance probabilities [1] . Models that can be expressed as fairly simple analytical forms will be more broadly useful. For disordered systems and data containing uncertainty, techniques such as the *maximum entropy principle* (MaxEnt)[2] and *superstatistics*[3] will be applied; these often have a more formal basis than the heuristic fractal models[4] often employed.

To discover patterns in data, data analysis techniques such as rank histogram plots are useful. A set of data, binned according to frequency of occurrence for the parameter of interest reveals most of the structure of the probability density. These views can be manipulated or marginalized against conditional or joint probabilities.

One of the significant observations that one can make about typical environmental parameters is in the extent of their randomness. On occasion an environmental parameter, such as temperature, can exist within a narrow range of values and thus become well-suited for a normal Gaussian distribution model, but more commonly, skewed (i.e. asymmetric) and fat-tail distributions are much more applicable. In these cases, the data along with some physical reasoning will direct the modeler away from a normal distribution toward a higher variance distribution.

Environmental context modeling relies on knowledge of exogenous behavior – that behavior that exists outside the confines of the vehicle or other engineered system we are designing. Any behavior that we have little control over needs to be regarded as uncertain, and will in general require stochastic models. Recent advances in our understanding of stochastic phenomena have benefited greatly from the availability of data from a variety of sources. In the past, modeling of physical behavior has often been hampered by the lack of sufficient statistics to substantiate the original formulation. In combining stochastics and information elements for modeling, we can incorporate probability and information theory (Jaynes[2], Shannon[5]), pattern theory (Mumford[6], Grenander[7]), fat-tail statistics (Mandelbrot[8],

Taleb[9], Sornette[10]), and superstatistics and complexity theory (Beck[3], Gell-Mann[11]) and then apply these contemporary ideas to the characterization of environmental contexts.

Probability theory as advanced by E.T. Jaynes[2] suggests using probability as an extended logic, and we should consider inference and plausible reasoning under various levels of uncertainty. The key idea of Jaynes is to meld Shannon's information theory concept of entropy together with the statistical mechanics definition of entropy. Many important and non-trivial applications exist where Jaynes' maximum entropy principle is the only tool we may need, as it describes the minimal application of prior knowledge when appropriate — often a mean value is all that is required.

Pattern theory as advanced by Mumford[6] and Grenander[7] seeks to identify the hidden variables of a data set, characterizing real world observations as patterns. The approach uses the observed patterns to infer information about the unobservable factors, formulating prior distributions for those too complex or difficult to observe. If we can determine efficient inferential models for the observed variables by modeling observations as partly stochastic and partly deterministic, and apply the randomness and variability of these distributions along with considering their natural structures and constraints (i.e. symmetries, independences of parts, and marginals on key statistics) we can create classes of stochastic models by applying transformations to patterns. We can then synthesize (sample) from the models, and the stochastic factors affecting an observation exhibit strong conditional independence, making it easily decomposable. We will see this approach demonstrated when we consider terrain characterization.

The analysis of fat-tail statistics as advanced by Sornette[10] and Taleb[9] has shown promise for the prediction of crises and extreme events in complex systems and risk management, both for social[12] and natural systems. The general theory encompasses scale-free properties, fractals, and power-laws, and provides an alternative to normal or Gaussian statistics. This introduces black and gray swan terminology and the idea of rare dragon-kings which relates to extreme value analysis (EVA)[13] , and evidenced via the scarcity of 100-year events.

Considerations of complexity theory as advanced by Gell-Mann[11] leads to the idea that :"*when defining complexity, it is always necessary to specify a level of detail up to which the system is described, with finer details being ignored"*. Seemingly complex representations can often be represented by rather concise descriptions and we can apply concepts such as dispersion and coarse graining to simplify the complexity. One such idea is that of superstatistics[3], which ties in closely to the ideas of maximum entropy [14]. The essential approach here is to admit that randomness can exist on different scales and by combining these scales, the underlying real-world statistical distributions are revealed.

The awareness that we can indeed use probability to characterize larger scale phenomena has often been fought tooth-and-nail by opposing schools of thought. For example, Mumford[15] describes how classical statisticians opposed contextual Bayesian modeling when they claimed that *"this approach denies that statistical inference can have anything to do with real thought because real-life situations are always buried in contextual variables and cannot be repeated."* In this case, the contextual variables appear to get in the way of our understanding of the desired effect, whereas they should become part of the understanding of the *system*: i.e. the context of the vehicle within the environment. In reality, we can and have created very useful models by incorporating prior contextual knowledge to infer possible and potential behaviors. The key is that many of the contextual variables are governed by properties of nature that repeatedly occur under conditions of the thermodynamic arrow of time, which always leads to greater amounts of entropy. In this sense, nature can perhaps be more predictable than we think, and at worst, we can expect that it remain predictable in its unpredictability, and thus we can make progress by applying a stochastic characterization to the empirical observations.

Maximum Entropy Principle Modeling

To apply context modeling effectively one has to approach the environmental domains pragmatically. We want to find solutions with the minimal amount of complexity that conversely generates the most general benefit. A model with a very detailed representation will typically apply only in specific cases and be of little use. So we first seek the simplest possible approach and find out if that has general applicability.

Fortunately, nature helps us out with the bookkeeping, through its implicit use of information theory, or as it is known in the physical sciences world – statistical mechanics. Thermodynamics and the essential notion of entropy both derive from statistical physics. The essential idea behind information theory is to try to describe an observed phenomenon with the least amount of words as possible.

A simple example of connectivity patterns drawn from Gell-Mann[11] demonstrates this point. Consider Figure 1 below, a series of networks showing increasing levels of seeming complexity. Graph **A** seems the least complex as it has no connections, while Graph **F** seemingly shows the most amount of complexity as it has all nodes interconnected.

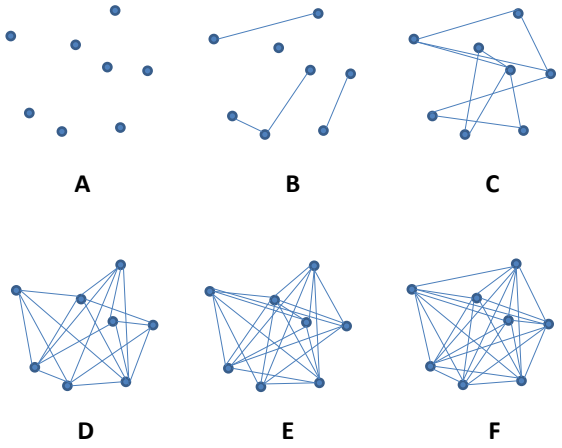


Figure 1: Gell-Mann's connectivity patterns. Increasing levels of complexity shown as an interconnected graph.

Using information theory, we can estimate the actual complexity by algorithmically describing each of the connectivity arrangements. Graph **A** can be described as *"no nodes connected"*, while Graph **F** can be described as *"all nodes connected"*. These therefore have very close to the same information content, because they can be reduced to an equally concise description. The graphs in between these end-points have higher complexity because they will require a more elaborate description.

In physics, this approach has long been applied to describe topics such as the ideal gas law, which stems from seminal ideas of statistical mechanics. Ensembles of particles, though complex at a micro-scale, can be described more concisely at the macro-scale, with the connecting thread elegantly applied from entropy and information theory.

These ideas can be applied to other natural phenomena. The key idea that we propose to apply is that of the Maximum Entropy Principle (MaxEnt)[2] . The principle of maximum entropy states that those configurations described by the least amount of constraints and moments (such as mean, variance, and higher-order moments) will tend toward probability density functions that maximize entropy. The MaxEnt solution is essentially to maximize **S** = entropy subject to constraints of probability, $f(p_i)$.

$$\max_{f(p_i)} S = -\sum p_i \cdot \ln p_i \tag{1}$$

In practice, MaxEnt gets routinely applied, even without the knowledge of the practitioners. For example, with the assumption of only a mean and a standard deviation in a set of empirical observations, the conventional approach is to apply the standard Gaussian or normal statistical distribution. That indeed coincides with the MaxEnt derived probability distribution for that particular set of moments. The canonical example of this is a manufactured component with a tight tolerance; this would routinely show a Normal distribution in the measurements of the produced parts. (A description of this example for an RLC[1] circuit using ratio distributions can be found in a report from DARPA's Adaptive Vehicle Make META program[16]).

Another simple example is that of a dye mixing in a glass of water. The eventual spatial distribution of the dye in the volume will approach a uniform level over time. This is straightforwardly modeled as a MaxEnt calculation given the boundary conditions of the volume. There is no preference for the dye occupying one subvolume over another so that we can only apply the known physical constraints. The barometric pressure law with altitude is one of the simplest examples of adding an energy constraint to Jaynes' formulation; and this gives the characteristic damped exponential decrease of pressure with altitude.

In addition, countless other phenomena, especially those showing great amounts of disorder, do not follow either a Normal distribution, or the uniform distribution or exponential just described. This includes many of the so-called "fat-tail" distributions, popularized by recent events[9]. Further, some of the fat-tails come about not from any intrinsic physical property, but from a derived property of the measure space (such as the ratio distribution mentioned parenthetically above).

Uncertainty Quantification

In addition to the natural random variation in the values of a particular environmental observable (**Type 1**: aleatoric uncertainties), our limited abilities to quantify the numbers can lead to a further uncertainty spread (**Type 2**: epistemic or systemic uncertainties). Whether the epistemic uncertainties allow us to capture the underlying randomness depends on the strength of the characterization and modeling. For example, we can have uncertainty in the actual measurements due to calibration or precision errors, uncertainty due to counting statistics, and uncertainty in the applicability of a model.

If the observable has a wide dynamic range and is readily quantified, the uncertainty in measurement is often absorbed in the natural randomness and has minimal impact on the data characterization. For counting statistics, uncertainty is minimized by drawing from a sufficient sample size that represents the complete dynamic range of the observable. By accumulating data over time, using approaches such as Bayes rule via additional prior knowledge, the additional evidence will reduce this uncertainty.

The applicability of the model for describing a natural phenomenon is referred to as the *likelihood* of the model and of its parameters. The certainty or confidence we have in a particular model is ultimately best gauged by comparison to an alternative model in which we can apply standard inference metrics such as conditional entropy, maximum likelihood, log-likelihood, and information entropy criteria such as Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC) [17]. The information criteria techniques are valuable because they penalize models that contain too many fitting parameters. Models based on first-principles with minimal parameterization (such as those derived from the Maximum

[1] RLC = resistive-inductive-capacitive. A configuration of RLC components will show a resonant frequency sensitive to the choice of parameter values.

Entropy Principle) will always score higher than, for example, a naïve high-order polynomial fit that contains many adjustable parameters.

The goal of context modeling is to converge to only **Type 1** uncertainty, where we can apply what we consider the most likely model and use its probability distribution function to provide Monte Carlo or importance sampling for design verification. In previous work[16] we applied propagation of uncertainty in combination with various physical and information-entropy-based models to arrive at estimates of a probabilistic certificate of correctness (PCC) for a given design. This included exogenous artificial effects and sets of metrics dealing specifically with exogenous variables, those variables whose value is determined outside the model in which it is used:

- Manufacturing variance (the RLC example)
- Semantic network links
- Travel dispersion
- Wireless signal latency
- Human reaction times

This approach is based upon application of the maximum entropy formulation and a careful consideration of the measure space[2]. By applying minimal information to stochastic models of various behaviors, that we can infer the essential probability distributions, and therefore concisely model contextual behaviors suitable for conversion (in reverse) to uncertainty bounds — necessary for tasks such as verifying vehicle or system environmental suitability. This choice results in a modeling approach for disordered systems and data containing uncertainty, using techniques such as the maximum entropy principle and superstatistics, has a more formal basis than the heuristic or fractal models conventionally used to empirically fit data.

In the following sections, we describe the basics of this approach. More sophisticated analyses can be developed on this foundation, for example:

- Growth curves
- Generalized correlation functions in the real-space domain
- Combining correlation functions with spectral representations, both in the spatial and temporal domains of system context.

These analyses will be covered in subsequent volumes (see Appendices B, C, D).

Characterization to Modeling

Fundamentally, context modeling involves applying the scientific method to describe the physical world. Figure 2 below shows the process going from (1) initial observations, to (2) characterization, and then to (3) modeling.

[2] See .This approach is not as common as one would think, considering the amount of detailed analysis undertaken with conventional statistical techniques (by non-domain experts who have a knowledge of statistics). We want to occupy the pragmatic middle ground and borrow insight from both physics and statistical camps.

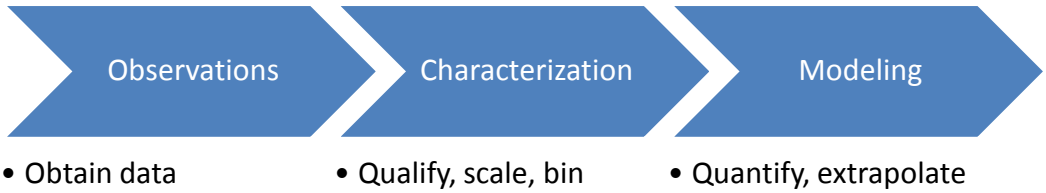


Figure 2: Process of Modeling

Stage 1 represents gathering the initial data from observations and measurements.

Stage 2 transforms the raw data to usable information —characterization imparts deeper categorical qualities and meaning to the data. In characterizing the initial data, we may not yet know the process or theory explaining how the data was generated, but inferences can be made from information made available, through for example a graphical form or as a lookup table. Such characterization of categorized data will allow the differentiation from other sets of data. For example, by charting the frequency of slopes in an environmental terrain data set, we can tell if one region is on average steeper than another region.

Stage 3 transforms the information to real knowledge via scientific modeling. By modeling we can bring a much deeper understanding to the information at hand than someone who only has lookup tables. To provide one benefit, we can extrapolate to regions outside the scope of our data ranges. Then, we can not only say how much steeper one area is than another but we can explain the reason and thus have additional knowledge to infer from. In other words, we can extend our reasoning capability.

We will treat these models in as concise a formulation as possible to avoid undue complexity, and we apply physical first principles to justify their description

Context Modeling of Environmental Domains

To better convey how this final modeling step is performed, we now provide examples applying maximum entropy principles and superstatistics to several practical examples: (1) wind speed, (2) terrain slopes, (3) rainfall intensity, (4) wave crest heights, and (5) electro-magnetic signal clutter. These all share a common foundation, as each draws from energy from the environment to give a range in intensity of some measure. In each case, the intensity is a readily measured observable, with the general trend of higher intensity values occurring less frequently than the low intensity values. The zero energy state is indeed the most common in all these measures.

An initial set of questions to consider from a context modeling perspective is what general characterization can we make from these observables, and then what universal modeling approaches can we apply?

Drawing from the perspective of maximum entropy, we first consider that the probability of a high energy configuration is typically treated as a variation of a Boltzmann factor:

$$e^{-E/E_A}$$

where E_A is the activation energy for occurrence. In this approach, the likelihood of higher energies becomes exponentially damped, scaled by the activation energy constant. The connection to maximum entropy is that we can treat this factor as a probability and then apply it as a probability density function of the measure of interest:

$$p(E)dE = \frac{1}{E_A}\ e^{-E/E_A}\ dE$$

The choice of the exponential in terms of MaxEnt is that it is the least biased estimator considering that E_A is the **average** energy of the configuration. If we happened to know the variance of the ensemble configuration, this would lead to a normal, Gaussian PDF. Yet, since we in general lack this knowledge, we need to rely on the least amount of information available, and this is the exponential, or Boltzmann factor.

Next we consider the application of the activation energy for the individual cases. For wind distributions, the kinetic energy is related to the square of the wind speed, *v*. This turns into the Rayleigh distribution.

$$E \sim kv^2$$
$$dE = 2\ k\ dv$$
$$p(v) \sim 2kv\ e^{-kv^2}$$

Electromagnetic signal clutter follows a similar derivation as the energy is the square of the amplitude of the electric signal.

For terrain distributions, we can to first order suggest that a potential energy is directly proportional to the terrain slope.

$$p(s) \sim\ e^{-ks}$$

For rainfall intensity, we consider the potential energy associated with a volume of water under gravitational forces. The larger the volume, the greater the encapsulated energy, which gets released scaled to the rate intensity of the rainfall.

Finally, for aquatic wave crests, the energy of the waves is proportional to square of the crest height. This works only to some level, as shoaling and non-linear fluid mechanics can prevent or attenuate taller waves.

Now consider that in each of these cases, the value of the activation energy can vary depending on regional or environmental conditions. For the case of rainfall, the intensity of the rainfall can be predicated on other conditions besides the volume of the water vapor alone. This leads to the idea of a super-statistical distribution. Here we not only apply the exponential PDF to the measures of interest, but we grant that probability an extra layer of uncertainty. That uncertainty would commonly apply to the value of E_A or to some other constant of proportionality.

Based upon activation energy proportionality or some other variant measure, different kinds of statistical distributions can be derived. For example, by considering variations in capacity and growth time, such phenomena as cloud sizes, lake sizes, and particulate sizes can be modeled. These become fat-tail distributions due to the weighting of the rate calculation, as a strong variant situated in the denominator of a stochastic ratio turns into a heavy weighting in the tail of a distribution.

The foregoing are exemplars of the general superstatistical approach we apply to natural context domains. The table below describes several of the practical examples. For many of these natural phenomena, empirical data sets can be used as samples from the observational space. By inferring information from a model that more accurately represents the underlying behavior than can a generic statistical distribution

more fundamental insights may be inferred. This has a number of benefits, including conciseness of representation and potentially better estimation of rare events.

Table 2: List of stochastic models

Stochastic Metric	Elements	Description	Data
Wind	PDF, ME, SS	Model of wind speeds	Bonneville Power Authority[18]
Rainfall	PDF, ME, SS	Model of rainfall amount	Hydrometeorology Lab University of Iowa[19]
Clutter	PDF, ME, SS	Model of EMI	
Clouds	PDF, ME	Model of cloud sizes	NASA Goddard[20]
Lakes	PDF, ME	Model of lake sizes	Global Lakes and Wetlands Database[21]
Particles	PDF, ME	Model of particle sizes	NASA JPL [22]
Waves	PDF, ME	Model of crest heights	CDIP[23] and US Army Corps of Engineers[24]
Terrain slopes	PDF, ME, SS	Model of inclination	USGS DEM [25]

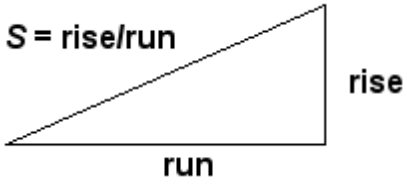
Examples

We now provide examples of almost predictable unpredictability that can arise in many natural environmental contexts, such as variability in terrain slopes, as well as in artificially man-made situations, such as a large highly-interconnected network [26]. These models can be applied in either analytic form or in a form suitable for Monte Carlo-type simulations.

Slope modeling and Topography

The first example is of a longitudinal slope context distribution. The macroscopic slope model of USA terrain follows closely that described by the Maximum Entropy Principle (MaxEnt).

How do we model and thus characterize disorder in the earth's terrain? Can we actually understand the extreme variability we see? If we consider that immense forces cause upheaval in the crust then we can reason that the energy can also vary slope topography widely.
The process that transfers potential energy into kinetic energy to first order has to contain elements of randomness. To the huge internal forces within the earth, generating relief textures equates to a kind of Brownian motion in relative terms — over geological time, the terrain amounts to nothing more than inconsequential particles to the earth's powerful internal engine.



We take the terrain slope *S* as our random variable (defined as rise/run).

The initial premise is the higher the slope, the more energetic the terrain. Applying the Maximum Entropy Principle to a section of terrain, we can approximate the local variations as a MaxEnt conditional probability density function, where E is the local mean energy and c is a constant of proportionality. But we also assume that the mean E varies over a larger area that we are interested in, as in the superstatistical sense of applying a prior distribution, where k is another MaxEnt measure of our uncertainty in the energy spread over a larger area.

The final probability is an integral over the marginal distribution consisting of the conditional multiplied by the prior. This integrates as a modified `BesselK` function of the zero order, K_0.

$$p(S) = \frac{2}{S_0} \cdot K_0(2\sqrt{S/S_0})$$

The average value of the terrain slope for this distribution is simply the value S_0.

The validity of the model thus derived can be assessed by comparing to a large set of data. The digital elevation model (DEM) data for the 1 degree quadrangles (aka blocks/tiles) in the USA from the USGS web site was characterized. This consists of post data at approximately 90 meter intervals (i.e. a fixed value of run) at 1:250,000 scale for the lower 48 USA and some spillover into Canada. From individual DEM files, we calculate the slopes between adjacent posts yielding an average slope (rise/run) of 0.039, approximately a 4% grade or 2.2 degrees pitch. The characterization takes the absolute values of all slopes so that the average is not zero.

The cumulative plot of terrain slopes for all 5 billion calculated slope points appears on the following chart[26]. The cumulative probability distribution of the `BesselK` model is plotted with the calculated average slope as the single adjustable parameter.

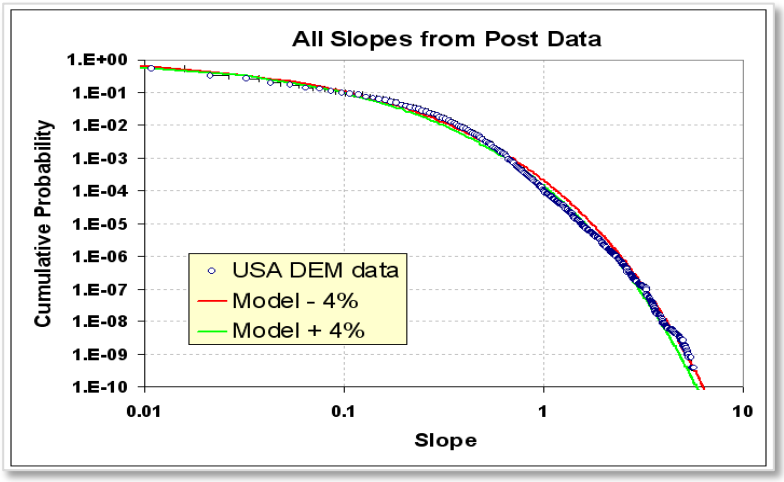


Figure 3: Longitudinal slope cumulative distribution function for the USA. The BesselK model with a small variation in S0 (+/-4% about the average 0.037 rise/run) demonstrates sensitivity to the fit.

The good agreement occurs because random forces contribute to maximizing the entropy of the topography. Enough variability exists for the terrain to reach an ergodic limit in filling the energy-constrained state space. As supporting evidence, we can generate a distribution that maps well to the prior by estimating the average slope from the conditional PDF of each of the 922 quadrangle blocks and then plotting this aggregate data set as another histogram.

For context modeling, a library function is used to generate Monte Carlo sample draws for the BesselK model without requiring a probability inversion. The resulting algorithm turns out surprisingly simple. First, draw two independent random samples from a uniform [0.0 .. 1.0] interval, then apply the natural log to each, multiply them together, and then multiply by the BesselK S_0 scaling constant.

This random draw algorithm will give the following cumulative if done 5 billion times, which is the same sample size as the real USA DEM data sample (see Figure 4). The only statistical noise is at the 10^{-9} level, which is roughly the same as in the DEM data set.

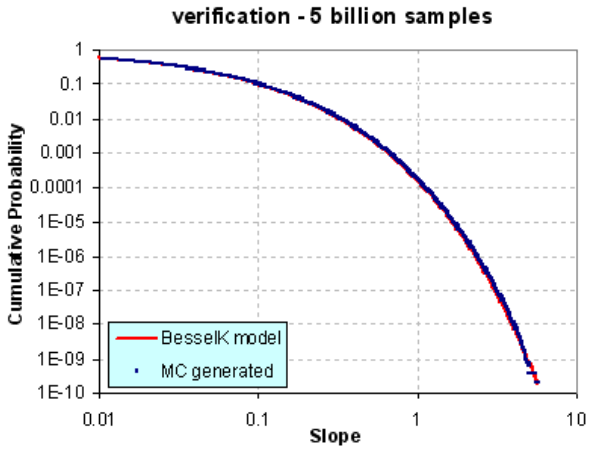


Figure 4: Drawing random samples to approximate slope distribution

Practically speaking, we see the variability in slopes expressed at the two different levels: the entire USA at the integrated (BesselK model) level and the aggregated regions at the localized (exponential prior) level. These remain consistent as they agree on the single adjustable parameter S_0.

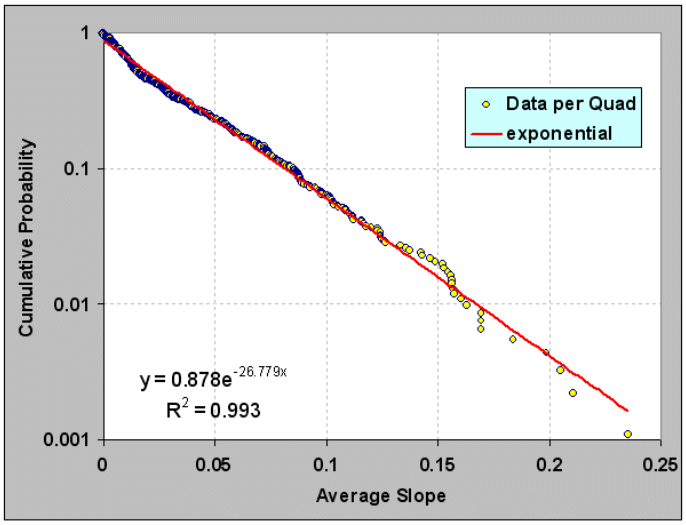


Figure 5: Generation of the prior distribution by taking the average slope of each of nearly 1000 quadrangles. The best fit generates a value of S_0 (1/ 27=0.037) close to that used in the previous figure.

The modeled distribution has many practical uses for analysis, including mobility studies and fuel efficiency planning (see Appendix F). Obviously, vehicles traveling up slopes use a significant amount of energy and a context modeler would need a model to base an analysis on without having to rely on the raw data by itself. Further, spatial correlations also exist and will prove useful as well. A survey of recent research gives no indication that others have discovered this rather simple model (See further Vico[27] Gagnon[28], Gonçalves[29], Guarnieri[30]

Wind Distribution

Wind velocities demonstrate a wide dynamic variability, ranging from calm to gale force. However intuitive the concept of "windiness", we may often miss the underlying mathematical simplicity behind wind speed variability. The complexity of the earth's climate and environment actually contributes to this simplicity as it generates more states for the system to exist within (see the Gell-Mann argument), which can also increase the likelihood of variability. With minimal knowledge as to the origin of the wind variance, we can apply the maximum entropy principle to its energy content.

The derivation of wind dispersion follows a few straightforward steps. We start with the premise that every location on Earth has a mean or average wind speed. This speed has a prevailing direction but assume that it can blow in any direction. Next we safely assume that the kinetic energy contained in the aggregate speed goes as the square of its velocity. If we assume only temporally-averaged mean wind energy and then relate the energy, $\boldsymbol{E}$, as the square of the wind speed, $\boldsymbol{v}^2$, the resultant maximum entropy probability distribution matches the Rayleigh distribution.

$$p(v) = p(E) \cdot \frac{dE}{dv} = 2cv \cdot e^{-cv^2}$$

This comes about from the Newtonian kinetic energy law $\frac{1}{2}mv^2$ and it shows up empirically as the aeronautical drag law (i.e. wind resistance) which also goes as the square of the speed.[3] Then apply the principle of maximum entropy to the possible states of energy that exist and come up with a probability density function that has no constraints other than a mean value (with negative speeds forbidden). In the equation above c is a constant and $1/c$ defines the mean energy (i.e. essentially acting as the Boltzmann activation energy). This describes a declining probability profile, with low energies much more probable than high energies. To convert to a wind dispersion PDF we substitute velocity for energy and simplify.

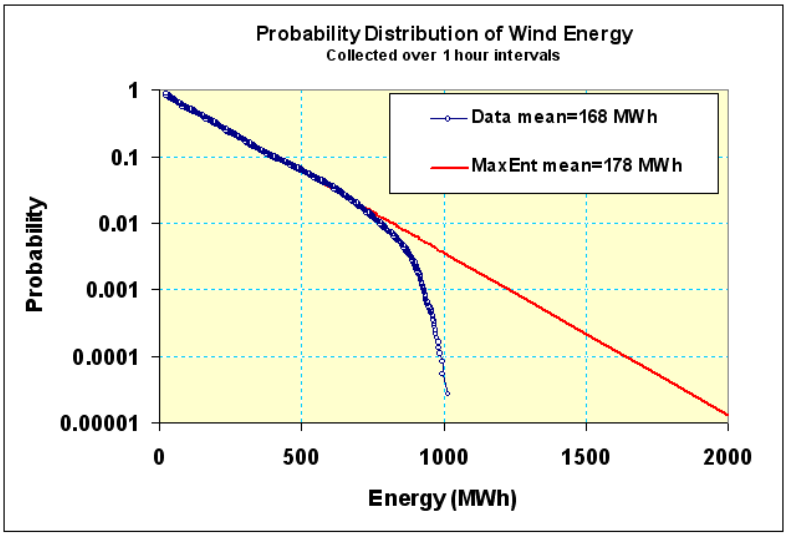


Figure 6: Wind distribution follows Rayleigh distribution closely. This data was collected from an archive of wind energy statistics collected in Ontario. The upper level cut-out is due to turbine governor regulation at high wind speeds.

[3] Note that we can consider E as an energy or modified slightly as a power, since the energy is sustained over time

Figure 6 shows an empirically observed wind speed distribution, showing a peak away from zero wind speeds and a rapid decline of frequency at higher velocity. Heuristically, many scientists refer to the model as following a Rayleigh or Weibull distribution. The Rayleigh comes out as the simpler model because it derives from first principles and any deviation from the quadratic exponent works as a refinement. The first data set shown consisted of about 36,000 sequential hourly measurements in terms of energy (kilowatt-hours) for Ontario.

By adding more data to our knowledge on wind dispersion, we can observe how dispersion in wind speeds has a universal character. The second data set (Figure 7) comes from northwest Germany and consists of wind power collected at 15 minute intervals.

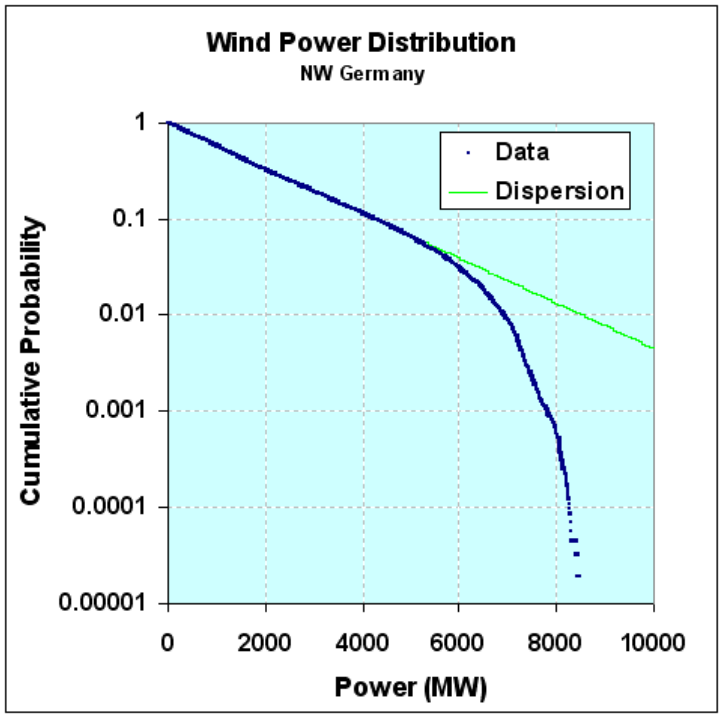


Figure 7: Wind variations for Germany. The curve has all the same characteristics as that for Ontario, demonstrating the universal behavior in wind speed variability.

Note that the same wind dispersion holds as for Ontario. Both curves display the same damped exponential probability distribution function for frequency of wind power (derived from wind speed). We also see the same qualitative cut-out above a certain power or wind energy level.

Adding More Variability. Since E_A can vary from region to region, we leave it as a conditional, and then set that as a maximal entropy estimator as well

$$p(E_i) = \alpha \cdot e^{-\alpha E_i}$$

then we integrate over the conditional's range according to standard practice and arrive at a cumulative.

$$P(E) = \int_0^\infty P(E|E_i)p(E_i)dE_i$$

This results in a simple lookup in your favorite comprehensive table of cataloged integration formulas, which leads to the following solution:

$$P(E) = 2\sqrt{\frac{E}{\overline{E}}}\,K_1\left(2\sqrt{\frac{E}{\overline{E}}}\right)$$

where K_1 is the modified BesselK function of the second kind, in this case of order 1, which is found in any spreadsheet program (such as Excel). Note that this is the same function that we used for the distribution of terrain slopes. The order 1 is the variant used for the *cumulative* distribution function

We tested this formulation against wind data from Bonneville Power Administration, which has over 20 meteorological stations set up around northern Oregon. The download consisted of over 2.5 million data points collected at 5 minute intervals, archived over the span of a little less than 2 years.

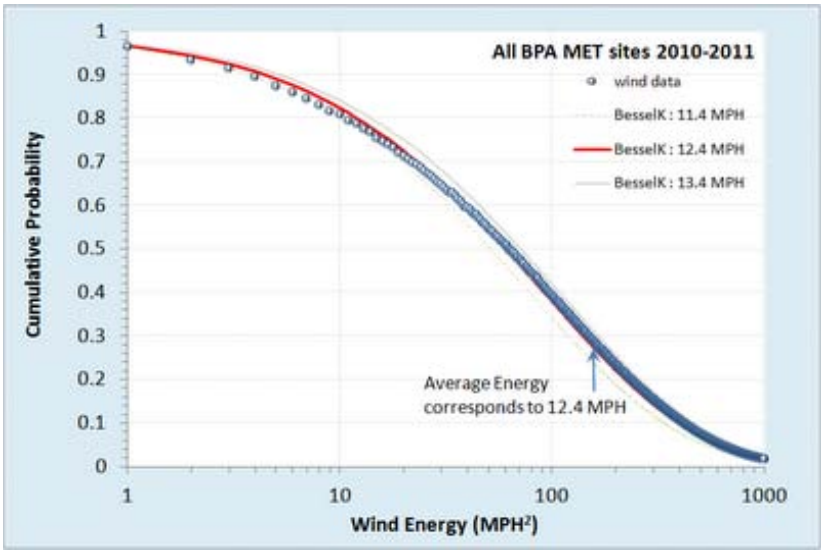


Figure 8: Cumulative distribution function of wind energies from Bonneville with model fit.

For the fit of this curve, the average energy was derived directly by computing the mean over the entire set of data separately. This corresponded to a value of 12.4 MPH, and placed a pair of positive and negative tolerances to give an idea of the sensitivity of the fit.

As this is a single parameter model, the only leeway we have is in shifting the curve horizontally along the energy axis, and since this is locked by an average, the fit becomes essentially automatic with no room for tweaking and little for argument. The probabilities are automatically normalized.

Figure 9 shows the log-log plot, which reveals a departure at high wind speeds. This shows that excessive gale force winds (greater than 60 MPH) did not occur over the extended region during the span of two years data collection.

Wind dispersion analysis has obvious applications for context modeling. Fuel efficiency is impacted by aerodynamics and drag goes up as the square of the wind speed. Vehicle cooling also is impacted by convection due to local winds. Applying this approach for context modeling has the benefit of allowing simple sampled data generation for verification and PCC bounding, similar to that applied for terrain slopes.

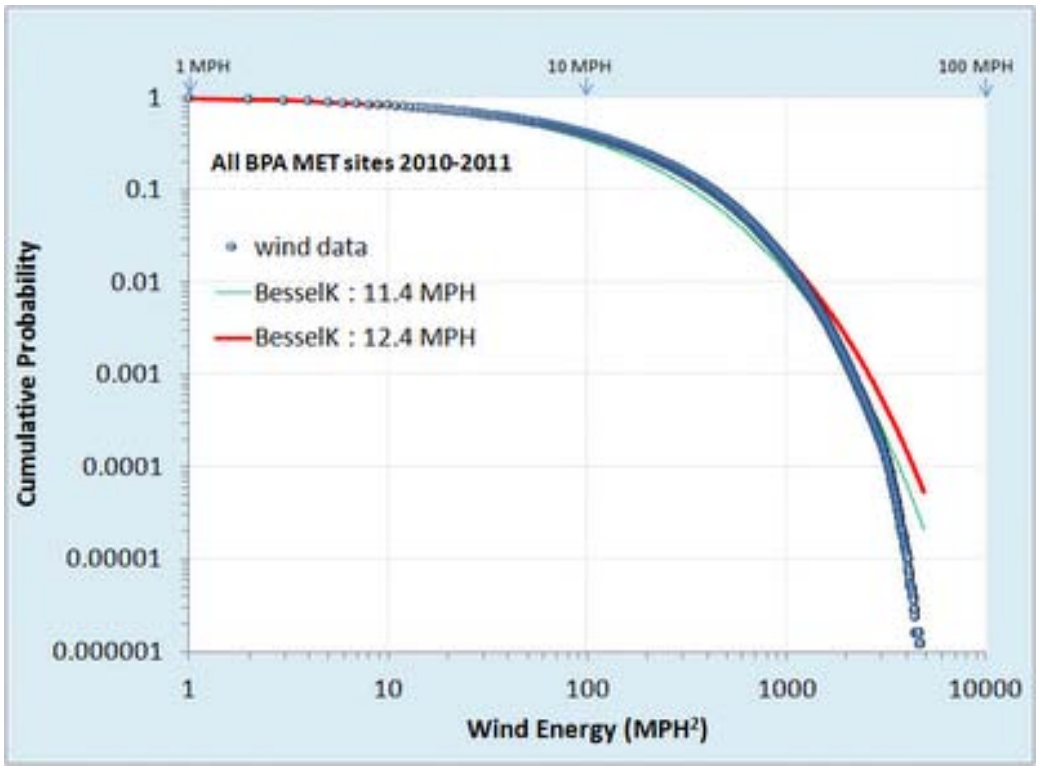


Figure 9: Cumulative distribution function of wind energies on a log-log plot.

The previous maximum entropy derivation assumed only a known mean of wind energy levels (measured as power integrated over a fixed time period). From this simple formulation, one can estimate or extrapolate a wind speed probability. Knowing the probability of wind speed, one can also perform all kinds of interesting extrapolations — for example, we can project the likelihood of how long it would take to accumulate a certain level of energy[31].

Aquatic Waves

Ocean waves exist in as disordered and unpredictable state as the wind. We may not always notice this as the scale of waves is smaller and often takes the form of a regular lapping of swells. In practice, the wind and wave energy distributions relate via similar maximum entropy disorder considerations. The following derivation assumes a deep enough water such that the wave troughs do not touch bottom

First, we make a maximum entropy estimation of the energy of a one-dimensional propagating wave driven by a prevailing wind direction. The mean energy of the wave is related to the wave height by the square of the height, ***H***. This makes sense because a taller wave needs a broader base to support that height, leading to a scaled pseudo-triangular shape of a *gravity* wave, as shown in Figure 10 below.

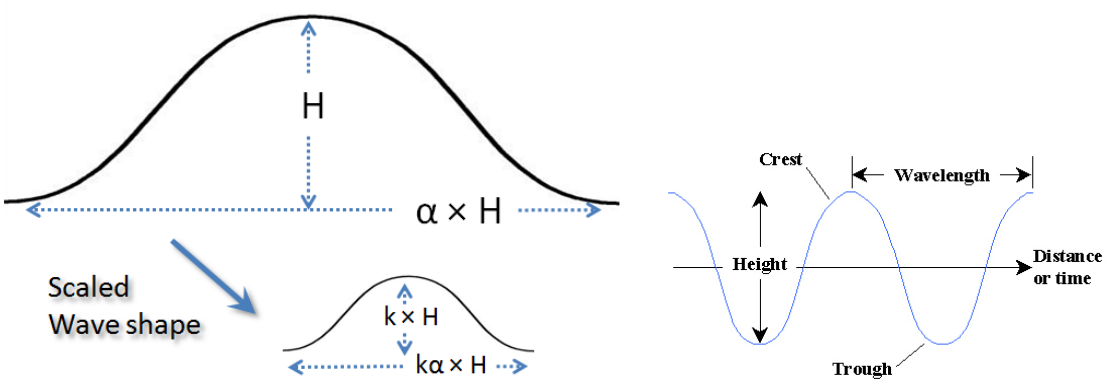


Figure 10: Total energy in a directed wave goes as the square of the height, and the macroscopic fluid properties suggest that it scales to size. This leads to a dispersive form for the wave size distribution

Since the area of such a scaled triangle goes as H^2, the MaxEnt cumulative probability is:

$$P(H) = e^{-\alpha H^2}$$

where α is related to the mean energy of an ensemble of waves. This relationship is empirically observed from measurements of ocean wave heights over a sufficient time period. This looks at height alone.

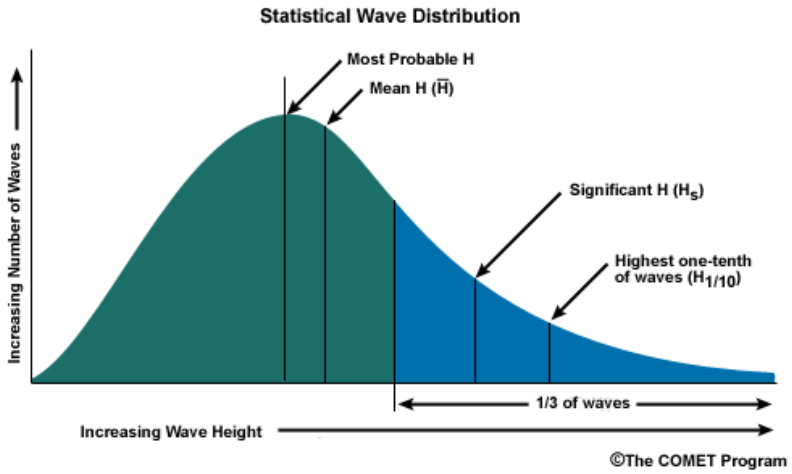


Figure 11: Statistical Wave Distribution (source: NOAA UCAR COMET Program)

Long-lived ocean and lake measuring stations have recorded historical records of wave crest data over the span of decades. From the US Army Corps of Engineer's Wave Information Studies project[24], the following figure collects chop and swell data from over several hundred million data points on Lake Michigan:

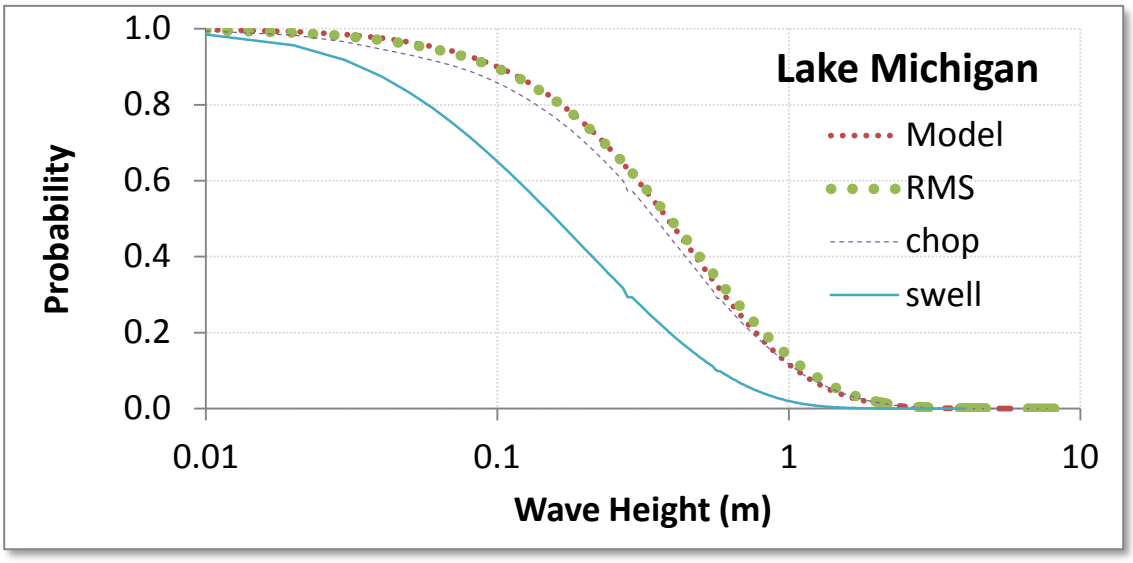


Figure 12: Lake Michigan wave height PDF on log-linear scale.

The probability density function for significant wave height empirically follows a modified Bessel function of the second kind. This essentially derives from a maximum entropy distribution for a wave height with a mean energy value, which is then dispersed again by maximizing the uncertainty in the energy. With precisely the same pattern that we derive the Bessel function from a range in wind speed values, we can derive the PDF for wave energy.

The Bessel fit works well for small wave heights but then starts to diverge when the wave height starts to exceed a critical level. This critical level is essentially the cresting limit of a wave given the average depth of the water. At this point we can apply an empirical correction factor first proposed by Jahns and Wheeler[32] and further analyzed by Haring [33]. This factor is essentially a 2nd-order polynomial which gradually suppresses the wave height from exceeding the critical cresting value.

$$1 - 4.37\frac{H}{d}(0.57 - \frac{H}{d})$$

The rationale for the factor arises from the remote likelihood of a wave height from exceeding the average water depth (d) in a region.

Based on data collected from coastal waters of two large lakes, Superior and Michigan, and that along the eastern USA seaboard of the Atlantic Ocean, we can see the characteristic bend on the PDF at approximately 10 meter height.[4]

According to Figure 13, the same basic Jahns/Wheeler correction is applied across the bodies of water. Both Michigan and Superior use a Bessel function of order 1, while the Atlantic uses a Bessel of order 2, which is generated by assuming an uncertainty that is not a maximum entropy exponential in the mean, but a MaxEnt that is a gamma of order 2 (i.e. two exponentials convolved which reduces the variance in ½)

[4] The data from Lake Superior was not as extensive as the other two bodies of water

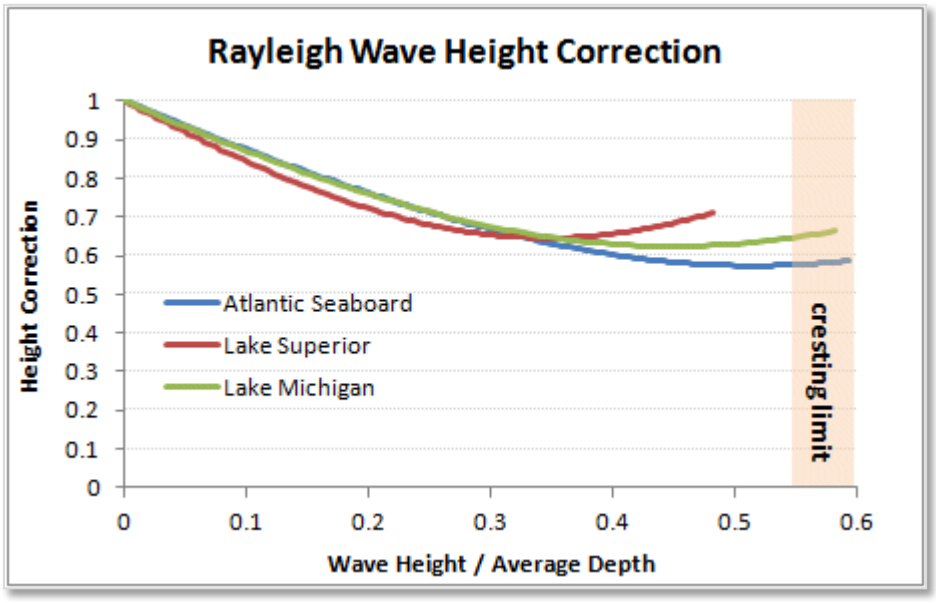


Figure 13: The Jahns/Wheeler correction applied to various bodies of water. The Lake Superior used coefficients closer to the heuristic J/W values (see Table 3) but Atlantic and Michigan appeared to asymptotically merge at the Nelson cresting limit of 0.55 for the wave height / average depth ratio.[34] [35].

Table 3: Heuristic Jahns/Wheeler correction used different coefficients to achieve the best fit.

	Atlantic Seaboard	Lake Superior	Lake Michigan	Haring
Inferred Average Depth	25.13 m	22.82 m	13.75 m	-
Haring Coefficient 1	2.52	4.86	3.03	4.37
Haring Coefficient 2	1.03	0.69	0.90	0.57

To evaluate the fit across the extreme values, Figure 14 shows the model profile on a log-log scale. The long tails are important for evaluating the probabilities of high sea-state values.

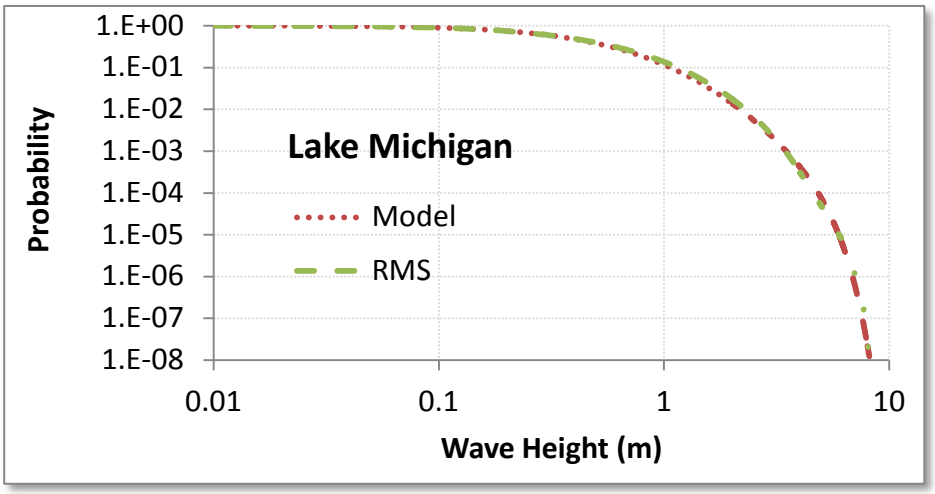


Figure 14: Lake Michigan log-log scaled version of wave height PDF

Lake Superior measurements (Figure 15) were sparse, yet the same profile is observed.

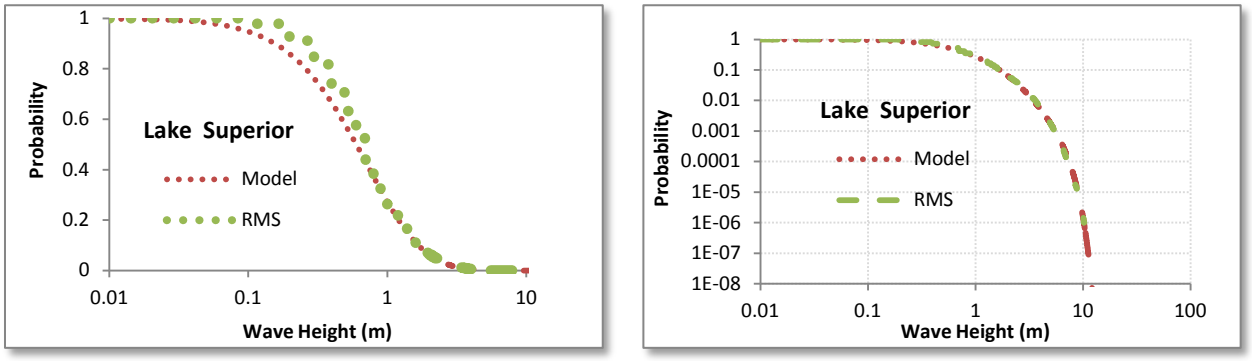


Figure 15: Lake Superior wave height PDF log-linear (left) and log-log (right)

Wave measurements from the Atlantic Coast along the length of the USA were less widely dispersed (see below). It is much more unlikely to find very calm waters in the data set. In this case, a higher order BesselK function was used to model the wave height distribution.

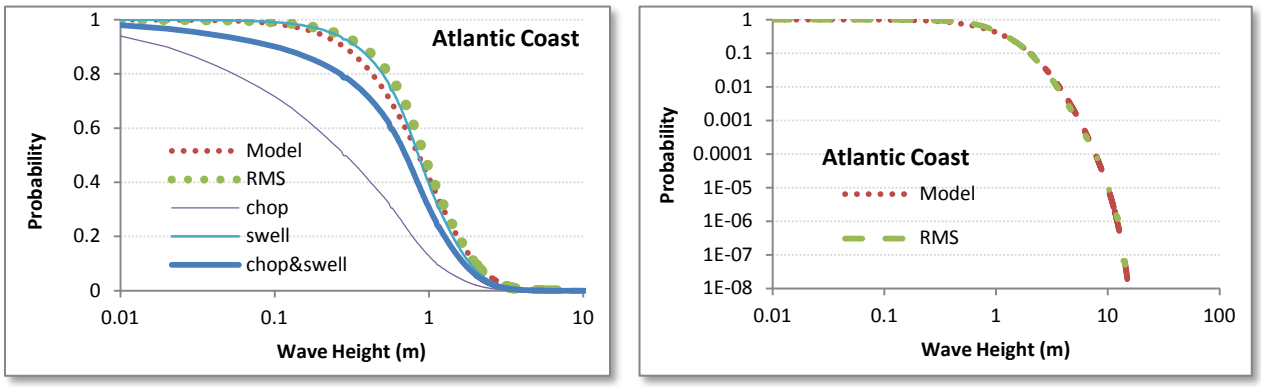


Figure 16: Atlantic Coast wave height PDF plotted as log-linear (left) and log-log (right)

In general, the divergence between the BesselK maximum entropy estimator for wind and that for waves is due to nonlinear effects at large wave-height. The heuristic Jahns/Wheeler correction factor is physically modeled by wave cusping, which generate larger heights than the triangular base wave predicts. This essentially rationalizes the sharpened crests and flattened troughs before it hits the cresting limit. We used a heuristic, but other corrections are available, such as derivations from the Rayleigh-Stokes (Tayfun model[36] [37]) process which is classified as a narrow-banded random process[38].

For use in simulations, other factors also play in such as, the probability of consecutive waves and the fact that phase velocity increases with the increase in wave steepness[39] .

Even with the complexity inherent in modeling turbulence, the modeling at the PDF level has some predictive power. For example, using the Atlantic model parameters we can estimate the wave height for the Mediterranean coast of Greece [40]. In this sense, waves have universal characteristics.

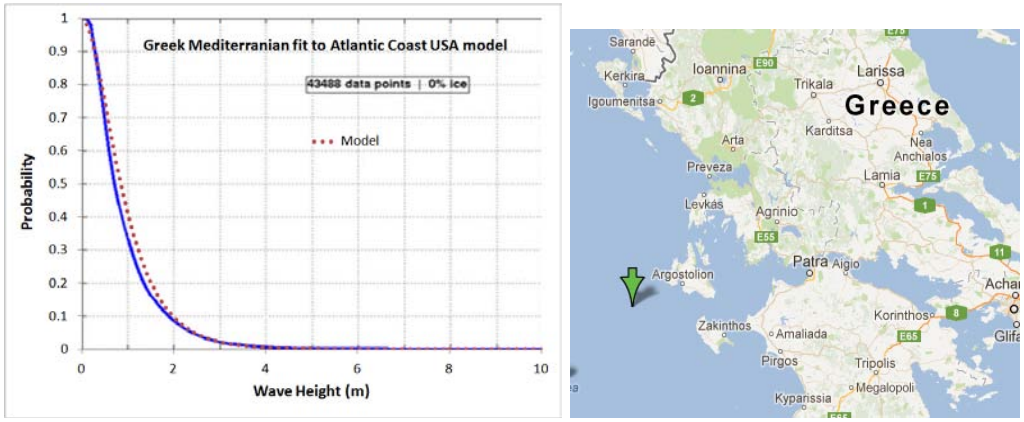


Figure 17: Mediterranean sea wave height assuming Atlantic coast model

Rainfall Intensity

As a premise, we want to consider whether a simple stochastic model can generate statistical patterns of rainfall events. We postulate that a critical point exists for rain to fall. The volume and density at which nature decides it reaches this critical point has much to do with the rate at which a cloud develops in its intensity and payload.

We assume a maximum entropy probability distribution function which assumes an average energy of rainfall rate for a storm, ***i*** :

$$P(E|E_i) = e^{-E/E_i}$$

The rationale for this is that the rainfall's energy is proportional to the rate of the rainfall, since that amount of moisture had to be held aloft by gravity.

$$Rate_i \sim E_i$$

Yet we know that the E_i can vary from storm to storm, so we leave it as a conditional, and then set that as a maximal entropy estimator as well

$$P(E_i) = \alpha\, e^{-\alpha E_i}$$

then integrating over the conditional's range.

$$P(E) = \int_0^\infty P(E|E_i)p(E_i)dE_i$$

This leads to the following solution:

$$P(E) = 2\sqrt{\frac{E}{E}}\, K_1\left(2\sqrt{\frac{E}{E}}\right)$$

where K_1 is the modified BesselK function of the second kind, in this case of order 1. This is the same general derivation as we performed for wind speed.

This analysis was compared against this recent paper by Papalexio: "Can a simple stochastic model generate rich patterns of rainfall events?" [19], and graphed as shown below in Figure 18. The green points constitute the BesselK fit which lies right on top of the blue empirical data set.

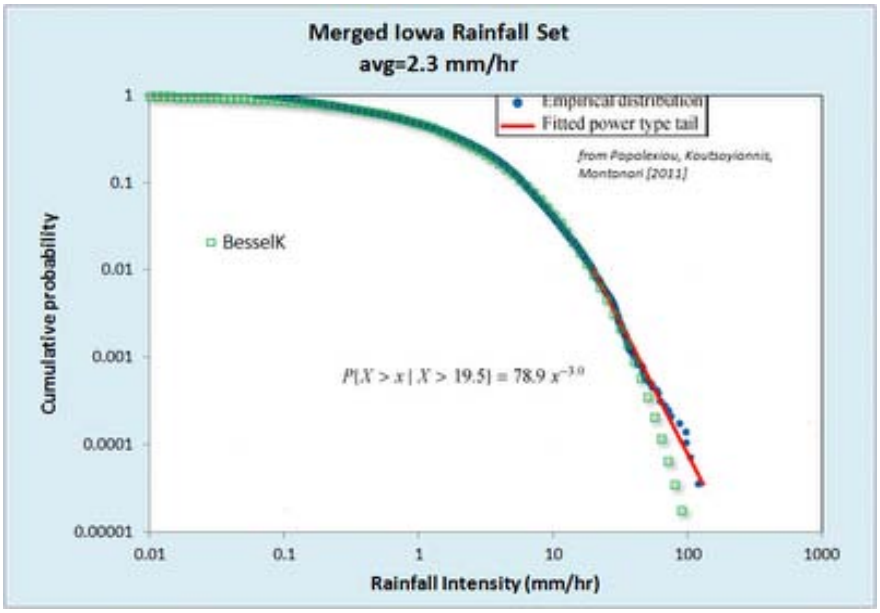


Figure 18: Cumulative rainfall distribution statistically gathered from several storm events.

From the table below reproduced from the paper, we can see that the mean value used in the BesselK distribution was *exactly the same* as that from the standard statistical moment calculation

Table 4: Rainfall moments

Event#	1	2	3	4	5	6	7	All
SampleSize	9697	4379	4211	3539	3345	3331	1034	29536
Mean(mm/h)	3.89	0.5	0.38	1.14	3.03	2.74	2.7	**2.29**
StandardDeviation(mm/h)	6.16	0.97	0.55	1.19	3.39	2.2	2	4.11
Skewness	4.84	9.23	5.01	2.07	3.95	1.47	0.52	6.54
Kurtosis	47.12	110.24	37.38	5.52	27.34	2.91	-0.59	91
HurstExponent	0.94	0.79	0.89	0.94	0.89	0.87	0.97	0.89

In contrast, Papalexio tried to apply Hurst-Kolmogorov statistics to the problem in search of a power law solution, claiming significance for the finding of a power law tail of -3. Instead we suggest that the slight deviation in the tail region is likely caused by insufficient sampling in the data set. A slight divergence starts to occur at the 1 part in 5,000 resolution level and since there are only 30,000 points in the merged data set, indicating that statistical fluctuations could account for the difference. See Figure 19 below which synthesizes a BesselK distribution from the same sample size, and can clearly duplicate the deviation in the tail.

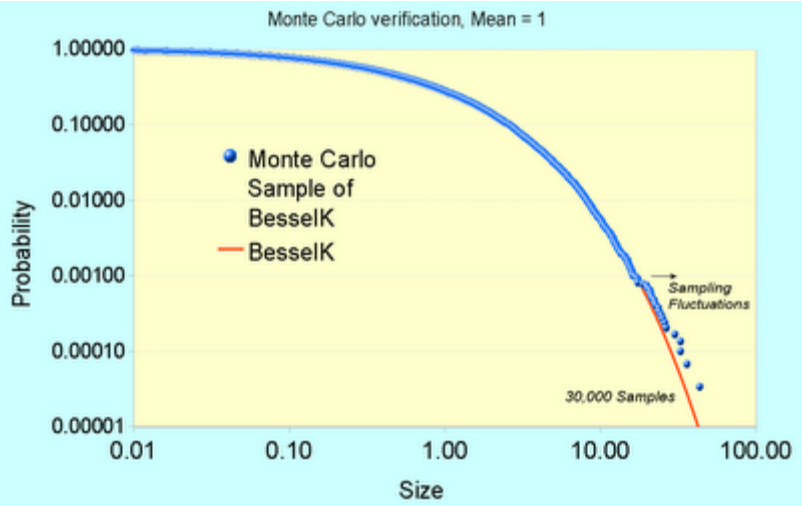
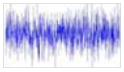


Figure 19: Monte Carlo sampling from a synthetic BesselK distribution reveals the same departure at low probability events.

Fluctuations of 1 in 30,000 are clearly seen at this level, so care must be taken to not read fat-tail behavior that may not exist.

The bottom-line is that this simple stochastic model outperforms Papalexio's fractal model, which doesn't seem as well-grounded with respect to first-principles once we apply an elementary uncertainty analysis to the data. The only physical premise needed to derive the physics grounded model specification is the intuition that rainfall buildup is a potential energy argument. This translates to a Boltzmann/Gibbs probability with a large spread in the activation energy. The large spread is modeled by maximum entropy — since the activation energy is unknown we assume a mean and let the fluctuations about that mean vary to the maximum amount possible. That is the maximum entropy activation which is proportional to a mean rainfall rate — the stronger the rate, the higher the energy.

That essentially becomes the premise for predicting the probability of a given rainfall rate within the larger ensemble defined by a mean. Note that Papalexio[19] seems to imply that the PDFs can go beyond the variability of a BesselK and into fatter tail territory (such as the MiejerG distribution, which is exponentially mixed prior with the Bessel). That is certainly acceptable as it will simply make it into a higher probability of outlier distribution, and if a fatter tail matches the empirical observations on a grander ensemble scale, then that might be a better characterization of the natural process [41].



EMI Clutter

The intermittent nature of wind power has a fundamental explanation based on entropy arguments. This same entropy-based approach explains some other related noisy and intermittent phenomena that modelers have to deal with. One case involves the use of mobile wireless gadgets such as WiFi devices, cell phones, and global positioning system (GPS) navigation aids in an imperfect (i.e. noisy) environment crowded out by electro-magnetic interference (EMI).

These wireless devices are often used in cluttered environments where ideal transmitted power mixed with EMI noise results in frustrating fadeouts that we need to patiently wait out. An example of Rayleigh fading appears in Figure 20 below. Signal interference-based explanations for why this happens can be

found, originating via the same intentional phase cancellations that occur in noise-cancelling headphones. The electronics in noise-cancelling headphones flip the phase so all interferences turn destructive, but for wireless devices, the interferences turn random, some positive and some negative, so the result gives the random signal shown. In the limit of a highly interfering environment the amplitude distribution of the signal shows a Rayleigh distribution, the same observed for wind speed.

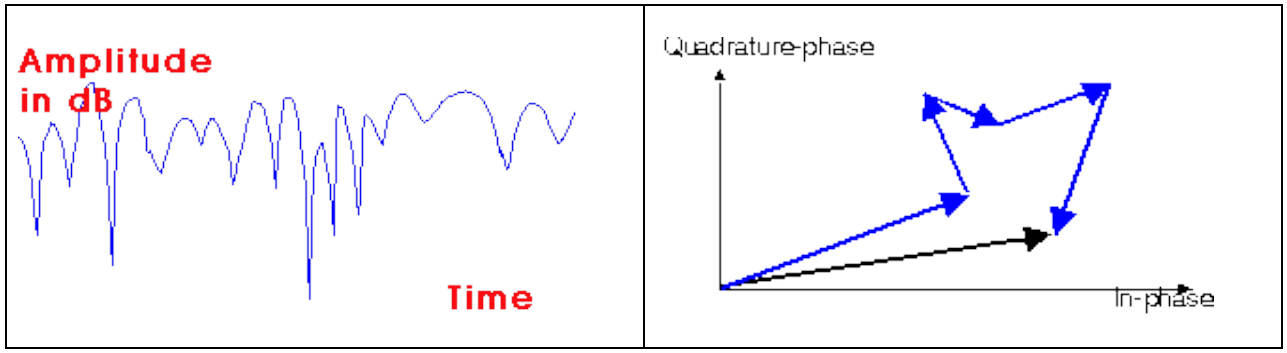


Figure 20: Noisy variation of signal amplitudes (left) is typically explained as a quadrature summation of randomly directed vector amplitudes (right). The variation in noise amplitude levels with time and the effect this has on a signal is known as Rayleigh fading.

Our knowledge of the situation reduces to that of knowing only the average power level of the signal. In that case, we can use Maximum Entropy Principles to estimate the amplitude from the energy stored in the signal, just like one can derive it for wind speed. So, as a starting premise, if we know the average power alone, then we can derive the Rayleigh distribution. The following figure shows the probability density function of the correlated power measured from a GPS signal. Since power in an electromagnetic signal relates to energy as a flow of constant energy per unit time, then we would expect the energy or power distribution to look like a damped exponential, in line with the maximum entropy interpretation. And it does exactly match a damped exponential.

$$p(E) = k \cdot e^{-kE}$$

This matches the observation of noise power level as shown in Figure 21.

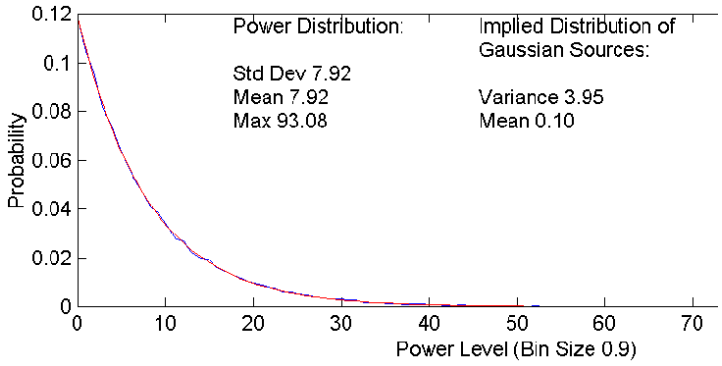


Figure 21: Note that the standard deviation equals the mean, a clear indication of a damped exponential. (from Watson[42])

Yet since power (E) is proportional to Amplitude squared, we can derive the probability density function by invoking the chain rule.

$$p(A) = p(E) \cdot \frac{dE}{dA} = e^{-kA^2} \frac{d}{dA}(A^2) = 2kr \cdot e^{-kA^2}$$

This precisely matches the Rayleigh distribution, implying that Rayleigh fits a Maximum Entropy (MaxEnt) distribution. So too does the uniformly random phase in the destructive interference process qualify as a MaxEnt distribution, which will range from 0 to 360 degrees (which gives an alternative derivation of Rayleigh). So all three of the distributions, the Exponential, Rayleigh, and Uniform, work together; and this provides a rather parsimonious application of the maximum entropy principle.

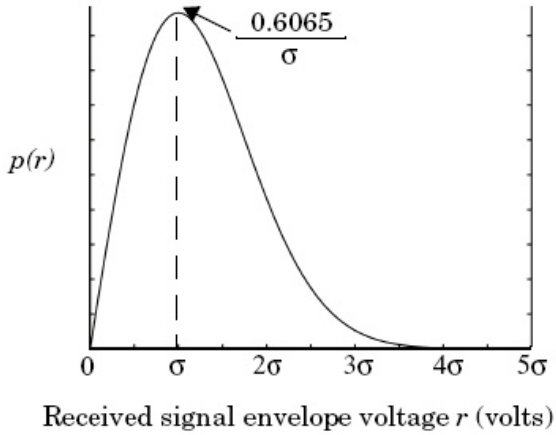


Figure 22: The Rayleigh Distribution of signal strength.

The most interesting implication of an entropic signal strength environment relates to how we deal with this power variation in our electronic devices. In a GPS, a vehicle's navigation system will experience this when trying to acquire a GPS signal from a cold-start. The amount of time it takes to acquire GPS satellites can range from seconds to minutes, and sometimes we don't get a signal at all, especially if we have tree cover with branches swaying in the wind.

Explaining the variable delay in GPS comes out quite cleanly as a fat-tail statistic if we consider how the GPS locks into the set of satellite signals. The solution assumes the entropy variations of the signal strength and integrating this against the search space that the receiver needs to lock-in to the GPS satellites. Since the search space involves time on one axis and frequency in the other, it takes in the limit $\sim N^2$ steps to decode a solution that identifies a particular satellite signal sequence for a particular unknown starting position. This gets reduced because of the mean number of steps needed on average in the search space. Dynamic programming matrix methods and parallel processing (perhaps using an FFT) can be used to get this to order *N*, so the speed-up for a given rate is t^2. So this will take a stochastic amount of time according to the MaxEnt criteria.

However, due to the Rayleigh fading phenomenon, we don't know how long it will take to integrate our signal with regard to the rate *R*. This rate has a density function proportional to the power level distribution, then according to the rules for marginal distributions the conditionals line up to give the probability of acquiring a signal within time *t*.

$$P(t < T) = \frac{1}{1 + \left(\frac{T}{a}\right)^2}$$

This leads to the dispersion result where *a* is an empirically determined number derived from *k* and *c*. This is not considered an extremely fat tail because the acceleration of the search by quadrature tends to mitigate very long times.

Data Analysis. Data was collected from a GPS project that has a goal to speed up wild-fire response times by cleverly using remote transponders[43]. They published data for cold-start times as shown in the histogram below. Note that the data shows many times that approach 1000 seconds. The single parameter entropic dispersion fit (a=62 seconds) appears as the blue curve, and it fits the data quite well.

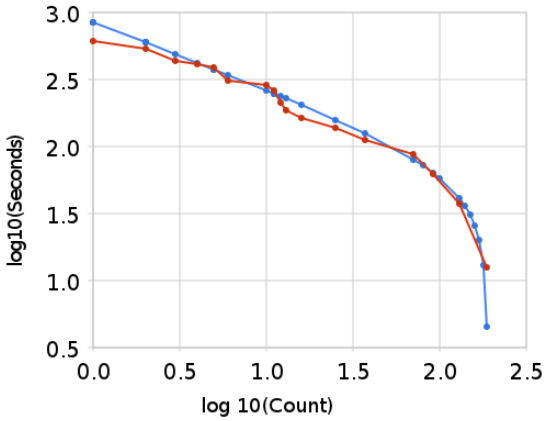


Figure 23: Fit of GPS acquisition times to a dispersive model

Beyond the Rayleigh, we will also find the BesselK PDF, as that is what describes the K-distributed clutter which appears in large regions collected by radar stations against random terrestrial and oceanographic features. The heterogeneous nature of that environment will generate clutter with longer tails than the Rayleigh [7]. The fatter tails of K-distributed clutter has implications for triggering false alarms when trying to distinguish signal from noise [44].

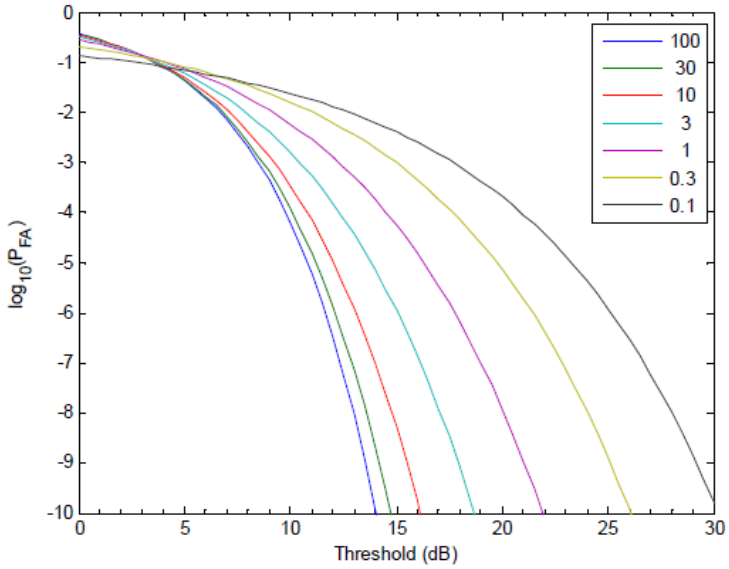


Figure 24: Probability of false alarm for a single pulse return from K-distributed clutter, for different values of the shape parameter v. Adapted from Bocquet [44]

Clouds

If we assume that clouds develop by some sort of preferential attachment, then the uncertainties at which the preferential attachment process increases with time balanced against the uncertainties in the critical point contribute to the dispersion. This is an analysis based on work by Yuan at NASA Goddard, "Cloud macroscopic organization: order emerging from randomness" [20]

The cloud size distribution follows a variant of the Zipf-Mandelbrot law, which also fits several other natural characterizations, such as oil field volume[26] and lake size distributions (see next section).

$$P(Size) = \frac{1}{1 + \frac{Median}{Size}}$$

We first assume that water vapor disperses through the atmosphere freely. The rate ***r*** at which it does this we treat as a stochastic variable with a probability density

$$p(r|g) = \frac{1}{g} \cdot e^{-\frac{r}{g}}$$

This introduces two concepts at once: the idea that we do not assume a single rate (i.e. assume instead *dispersion*) together with the idea that we can only assume at best a mean (as the growth rate ***g***) *and* to treat the standard deviation as equivalent to the mean. This type of assumption makes the least presuppositions as to what has happened — we know we have a mean value but beyond that, the rate can vary to the maximum entropy limit.

If we next assume that a collection of these rates can act to sweep out a selected uniform unit of volume, then over time we can imagine that a cloud will capture this moisture.

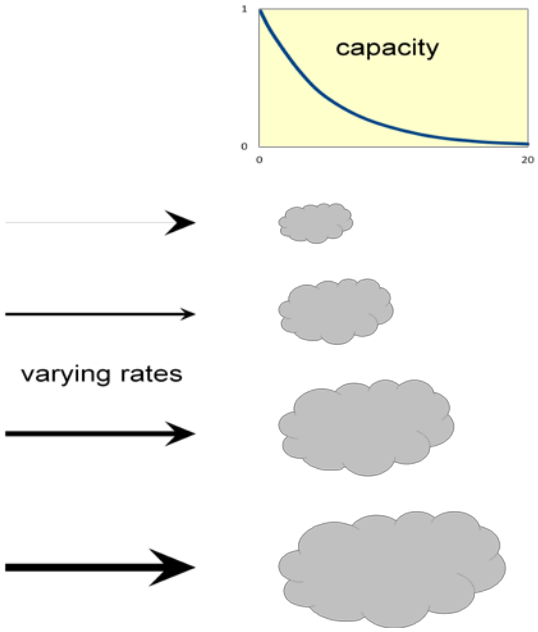


Figure 25: Cloud formation process leads to dispersion in sizes

Suppose that the water vapor diffuses outward, so for a given time period ***t***, the moisture will diffuse over a distance ***x*** = ***r t***, a simple variable change gives

$$p(r|g) = \frac{1}{gt} \cdot e^{-\frac{x}{gt}}$$

Over time, the probability that some moisture will migrate at least a length x_0 distance is:

$$p(x > x_0|g,t) = \int_0^\infty p(x|g,t)\,dx = e^{-\frac{x_0}{gt}}$$

Alternatively, the following relation tells us the cumulative probability of the distance covered by material after time ***t***. This again assumes a distance travelled ***x*** = ***r t***.

$$P(x_0|g) = \int_{r=x_0/t}^\infty p(r)\,dr = e^{-\frac{x_0}{gt}}$$

This relation also crops up in terms of the *population balance equation*. It basically relates a conservation of particles law, in that we do not lose track of any material due to a flow.

So next we have to accumulate this over a volume or depth at which a cloud develops. The simplest approximation assumes that the water droplets get distributed to a mean height (*H*) with a similar exponential distribution — this is like a capacity for the cloud formation (see Figure 25):

$$f(x|H) = \frac{1}{H} e^{-\frac{x}{H}}$$

Combining the two relations turns into an *a priori* probability for the expected cumulative transfer after time t through the volume. Integrating over the atmospheric layer over which clouds can form, gives the average water volume accumulated:

$$C(t|H) = \int_0^\infty f(x|H) \cdot P(x|g)dx = \int_0^\infty f(x|H) \cdot e^{-\frac{x}{gt}}\,dx$$

$$C(t|H) = \frac{1}{1 + \frac{H}{gt}}$$

For the last assumption, we note that if ***t*** gets evenly spread from the over time, then the value ***gt*** becomes the effective collected thickness ***W*** of a distribution of clouds, where we add a factor ***k*** to indicate collection efficiency. Alternatively, we can interpret the stochastic variable ***W*** as the maximum net cloud thickness that would develop over a diffusion time ***t***. The term ***kH*** sets the potential maximum net thickness achievable turning it into a hyperbolic discounting probability distribution.

$$C(W|H) = \frac{1}{1 + \frac{k\,H}{W}}$$

Or as we first surmised:

$$P(Size|Median) = \frac{1}{1 + \frac{Median}{Size}}$$

The agreement to the collected data is remarkable over several orders of magnitude..

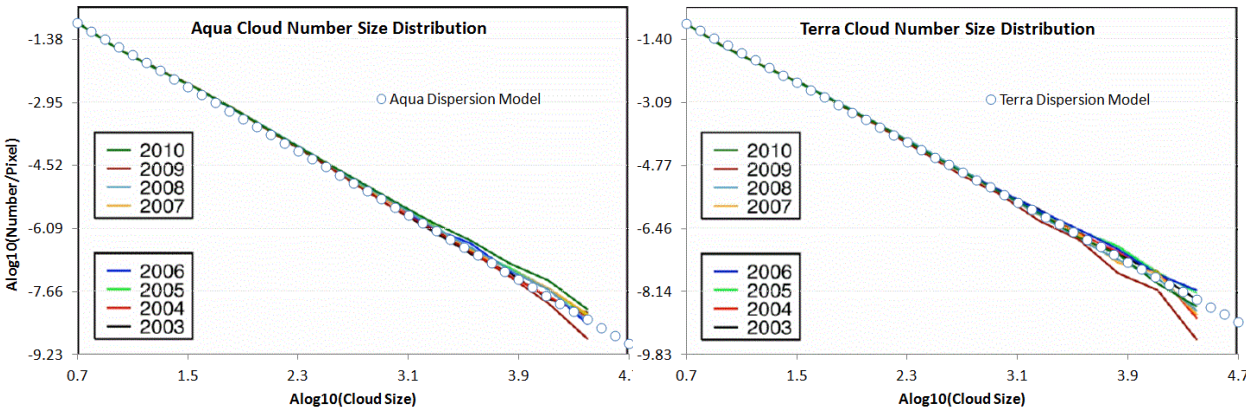


Figure 26: Cloud statistics using two different telemetry systems[20]. The same power law holds over several orders of magnitude.

Cloud size modeling may have some limited utility for nominal weather prediction (such as expected durations of shade and estimating the potential for lightning), but it does demonstrate well the wide aleatory uncertainty in a natural phenomena, and how widespread it is.

Lake sizes

Our environment shows great diversity in the size and abundance in natural structures. Freshwater lakes accumulate their volume in a dispersive manner. Over geologic time, water drifts into a basin at various rates and over a range in collecting regions. As lakes capture most of their volume through water drainage, one can imagine that the rate of influx plays a factor in how large a lake can become. The Maximum Entropy prediction of the size distribution leads to the following expression, exactly the same as for cloud formation:

$$P(Size) = \frac{1}{1 + \frac{Median}{Size}}$$

Surveys of lake-size show the same reciprocal power law dependence, with the exponent usually appearing arbitrarily close to one. In the Figure 27 below, the data plotted on a ranked plot clearly shows this dependence over several orders of magnitude.

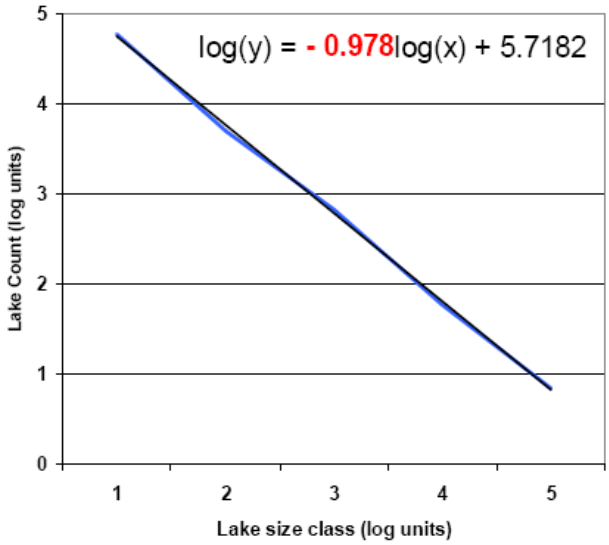


Figure 27: Northern Quebec lakes map and size distribution adapted from Telmer [45].

More revealing is that in Figure 28 below, we can observe the bend in the curve that limits the number of small lakes in exact accordance to the equation shown above. The agreement with such a simple model suggests that of a universal behavior.

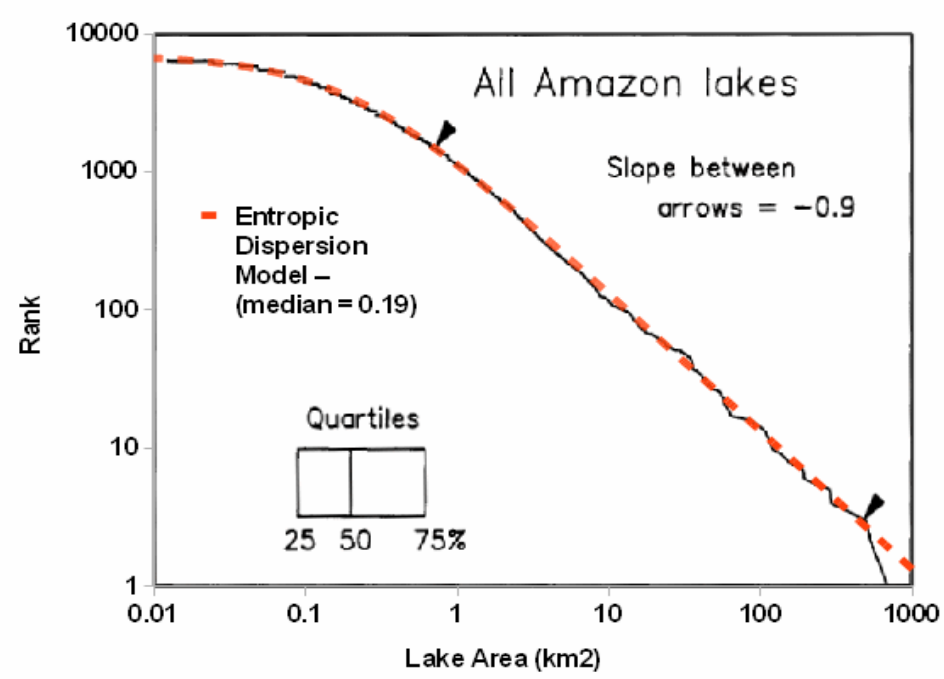


Figure 28: Amazon lake size distribution from "Estimation of the fractal dimension of terrain from Lake Size Distributions"[46]

This general trend is repeated for lakes around the world. All that is required is to have a median value for lake area and the rest of the distribution will rougly scale to this factor.

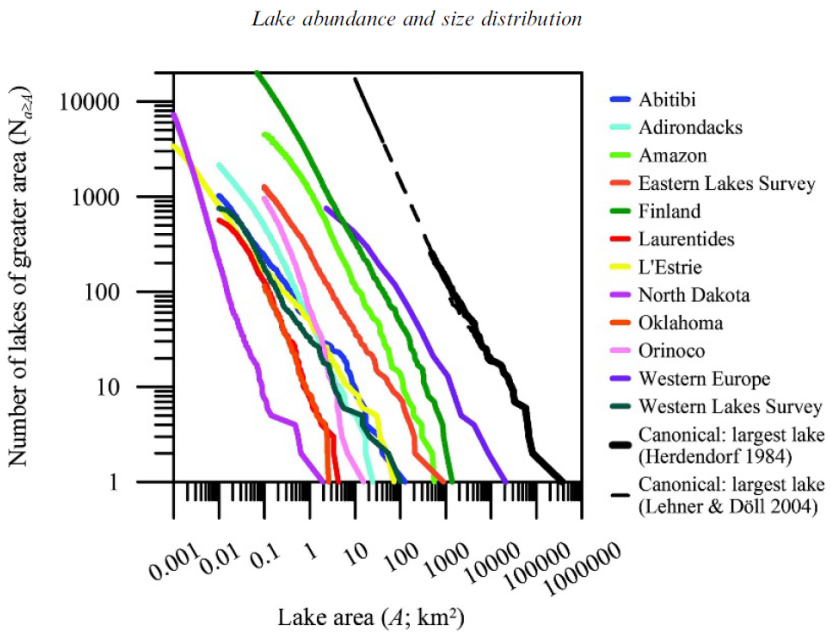


Figure 29: Near universal scaling of lake size distributions across the world [21]

Models of lake size distribution may have utility for vehicle fording and amphibious vehicle simulations.

Particulates

The role of entropy overrides other factors enough so that some simple dispersion arguments can explain the size distribution completely. Take as an example the formation of ice crystals in a cirrus cloud. Depending on the surrounding temperature, a crystal nucleates on some foreign particle and then starts growing. The atmospheric conditions have enough variety that the growth rate will disperse to the maximum entropy amount given a mean rate value. The end state for volumetric growth will also show the same amount of variation, where x is the size variate and S is the mean size.

The following particle size distribution (PSD) graph, Figure 30, shows measurements taken from high altitude cloud experiments [47][48][22]. The size gets measured along a single length dimension and the density of the particles takes the place of a probability.

The main profile follows a power-law in volumetric growth, where S is the median volume (see [26]):

$$p(x) = \frac{x}{(x+S)^2}$$

Crystal sizes get reported as a length and we have to convert that to a volume. This means the derivative has to include a chain rule to convert the volume x to a length parameter L, $x \sim L^3$ generates $dx/dL \sim L^2$.

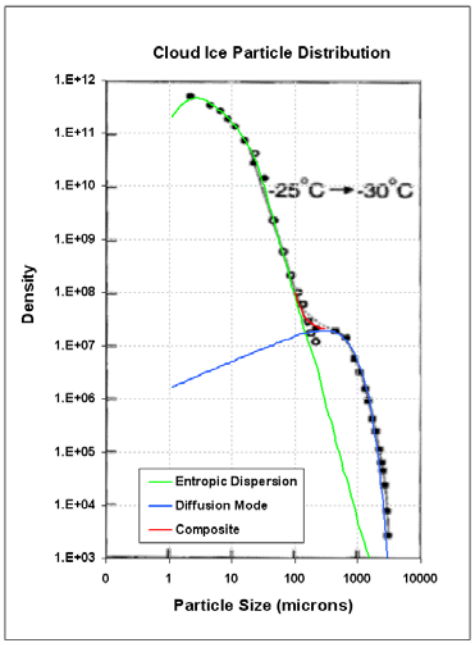


Figure 30: Cloud ice particle number density n(D) vs. the long dimension of particles as observed at the temperature range of - 25C and -30 C (from Platt[47]]). It shows a bimodal structure in the ice crystal distribution with the second peak at ~500 microns.

The data fits the dispersion model nicely (green line), but notice at low density that an extra mode shows up as the blue line. This clearly has a sharp exponential drop so likely has a non-dispersive origin. In terms of the higher density entropic model, this stands out as an ordered nucleation regime in the midst of a sea of disordered ice crystal growth modes. Applying a prior distribution is similar to integrating profiles of different weight (aka *superstatistics*) to come up with a power-law as shown below. By comparing this to the empirical sized distribution one can see how a distinct peak could occur.

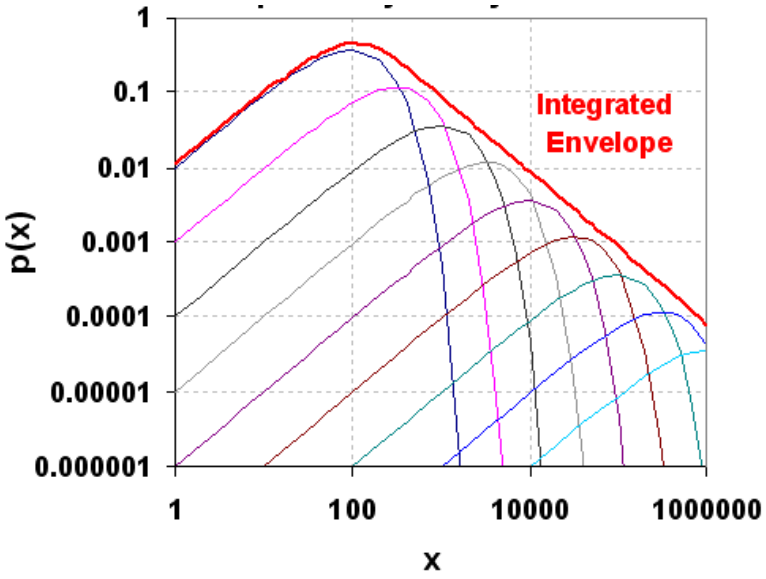


Figure 31: Superposition of exponential density profiles leads to a entropic dispersion envelope. The maximum entropy distribution as a prior distribution acts to smooth the envelope, but a strong component could stand out. Indeed this is typical in particle size distributions.

Physically why does this bump occur? Some unknown nucleation process has provided an optimal growth environment for these crystals to deviate from the entropic distribution. Hypothesizing, this could take the form of a catalyst or an accommodating growth substrate. With a power tail of -3/2 this might well have a strong diffusive growth component. However, the nuclei occur rarely enough so they do not drown out the much more common random or spontaneously occurring growth centers. It thus shows up as a clear non-dispersive growth mode in a sea of non-uniformity.

On a micro-level, we do have a population of reproducible structured shapes to bind against — as the airborne particulate world shows some uniformity in its density. This original analysis may prove of some help to those looking at particle size distributions of volcanic ash (see the figure below). Researchers routinely apply a log-normal fit to the data — yet one that uses a MaxEnt dispersion formulation with the appropriate volume/diameter exponent often can work just as well. Below, we use a root 1/2 dispersive growth rate on volume which may indicate a diffusion-controlled rate.

To account for wear/tear/corrosion modeling, one could apply particulate and ash models (Figure 32 and Figure 33, the latter featuring multi-modal size features) that feed into models for physical breakdown.

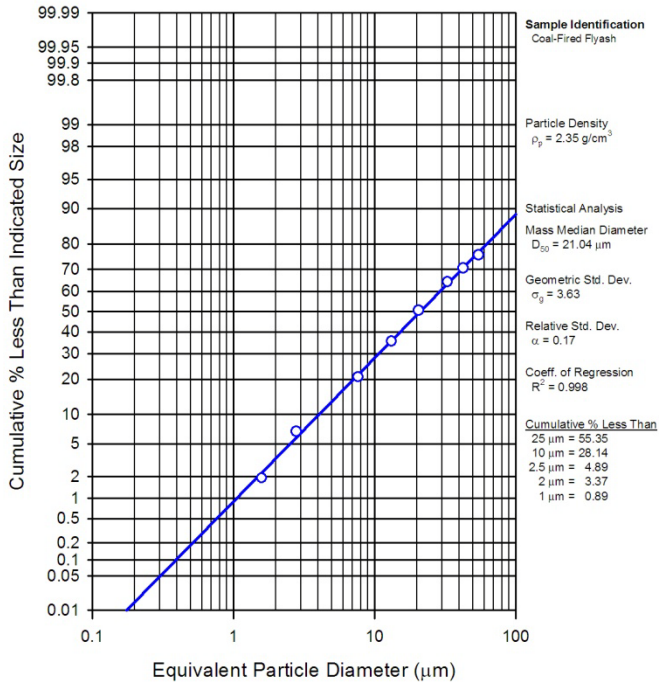


Figure 32: Wiki data for fly ash size distribution[49].

The signature feature of volcanic ash plumes is that they have dense populations of certain sized particulates (Figure 33), having derived from a homogeneous environment.

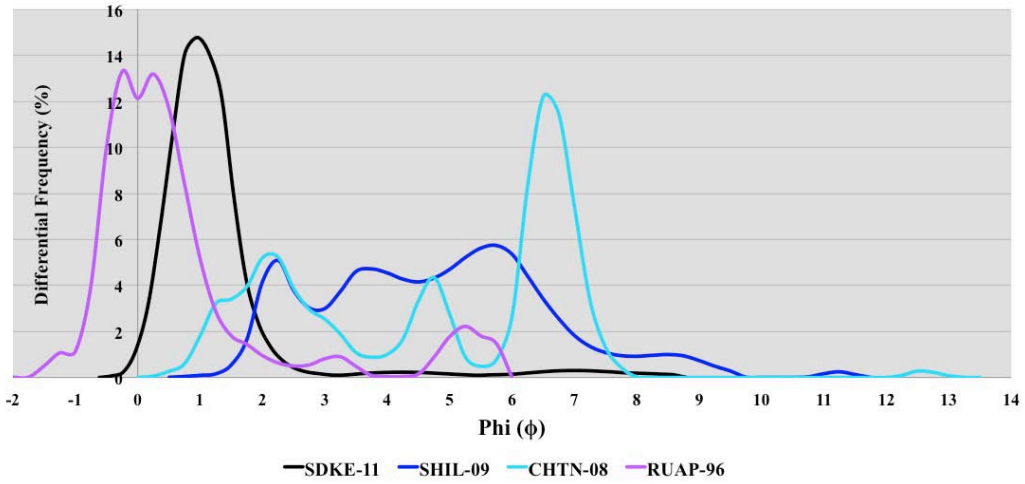


Figure 33: Volcanic Ash [50]. This is on the Krumbein phi *(φ) particle size scale [51] which is negative of the log base 2 of the particle diameter. That means smaller diameters are to the right side of the scale.*

Discussion

This study has described the characterization and modeling of natural phenomena in a form suitable for contextual use. We have provided examples of slope distribution, wind speed, rainfall rates, lake sizes, etc. Most of the models reduce to fairly simple analytical forms.

The probability measures derived from theory and applied to the data are often snapshots in time of a continuous growth process. Physical processes can accumulate or discard material following rates of material flow. While this may seem overly simplistic to consider, it does set the stage for revealing patterns in behavior.

For example, a statistically measured volume could be the result of various rates operating over varying time spans. Along with this variation, the concept of an average or mean value plays a role into modeling the behavior.

Generally speaking, two categories of models exist:

- *Models of variations around a mean energy*: Measures are bound by uncertainty in the mean energy value of a measured value, such as wind, rainfall rate, clutter, terrain slopes, wave crest height; these all generate the same universal curve, the ubiquitous modified Bessel of the second kind.
- *Models of variations around a rate and enclosed volume*: Several measures are bound by considerations of the rate of accumulation and the uncertainty of the eventual volume: such as clouds, particles, and lake sizes give the same entropic power law.[5]

We also described how the statistical spreads come about, via the concept of dispersion and maximum entropy. One of the outcomes of modeling is that a significant amount of data, 5 billion in the case of USA terrain and several hundred million for just the Lake Michigan wave data, can be expressed in a

[5] Plus oil reservoirs.

much more concise form.[6] This has a practical implication of allowing a terse model to summarized voluminous data to usefully assess candidate engineering designs. Further, as these models are parameterized by physical variables such as mean energy, the concise formulation allows exploration of extreme conditions extrapolated from but not exemplified by existing data sets.

Sampling approaches for simulation are often very straightforward for these distributions. We described several efficient applications of random sampling for transcendental functions such as the BesselK. We can further apply techniques such as importance sampling to reduce the sample size needed or to reduce the variance of the statistics.

In certain cases it may be possible to invert the stochastic models in order to support assessment of assume-guarantee conditions on component engineering models and to derive Probabilistic Certificates of Correctness (PCCs). The classical case is for normal distributions, but for fat-tail distributions such as the hyperbolic power law, this is also rather straightforward. To make sampling efficient, particularly for rare events, techniques such as importance sampling may be required. If a probability for a certain state is 10 orders of magnitude more rare than the most frequent state, the rule of thumb is that 10^{10} more samples may be required to catch this event. Importance sampling can cut the required number of samples down if the density function is well-characterized.

The fact that many of these models follow similar patterns makes it them useful for model automation. Abstract interfaces for common distributions such as the BesselK and hyperbolic power law can be generated and are of broad utility (see Volume 4).

Other characterization techniques such as the multiscale entropy measure[52] can be applied to temporal and spatial scales covering a wide dynamic range. This reveals the amount of disorder and uncertainty in the data, which is important as a quick characterization metric.

Summary and path forward

In general, support for context modeling and the general notion of Internet-based data collection allows us to leverage a vast amount available data and reduce its dimensionality and scope by careful characterization and modeling. Certain elementary characteristics related to probability density functions were covered in this paper.

Frequency characteristics and correlations will be described in the second paper in this series, “Terrain Characterization” (see Volume 2).

The third paper in this series, “Diffusive Growth” uses universal principles to apply the same pattern to characterize other disparate natural phenomenon, such as corrosion and thermal dispersion (see Volume 3).

Furthermore, by applying ontology-based approaches for organizing models and techniques we can set the stage for broader collections of such models discoverable by a general community of designers and analysts. Together with standard access protocols for context modeling these innovations provide the promise of making environmental context models generally available and reusable, significantly assisting the routine application of model based engineering. That is the scope of the fourth paper in this series (see Volume 4).

[6] Archived data often gets stale, and eventually gets deleted via some bureaucratic measure. Possessing a model that can potentially condense 4 billion data points into a single function containing a single parameter has clear benefits. It amounts to a 4 Billion to 1 reduction in storage size and encapsulated domain knowledge.

Acknowledgements

This work was performed under U.S. Department of Interior contract D12PC00241.

References

[1] F. E. Pritchard, C. C. Easterbrook, and G. E. McVehil, "Spectral and Exceedance Probability Models of Atmospheric Turbulence for Use in Aircraft Design and Operation," DTIC Document, 1965.
[2] E. T. Jaynes and G. L. Bretthorst, *Probability theory: the logic of science*. Cambridge Univ Pr, 2003.
[3] C. Beck and E. Cohen, "Superstatistics," *Physica A: Statistical Mechanics and its Applications*, vol. 322, pp. 267–275, 2003.
[4] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet mathematics*, vol. 1, no. 2, pp. 226–251, 2004.
[5] C. E. Shannon, W. Weaver, R. E. Blahut, and B. Hajek, *The mathematical theory of communication*, vol. 117. University of Illinois press Urbana, 1949.
[6] D. Mumford and A. Desolneux, *Pattern Theory: The Stochastic Analysis Of Real-World Signals*. A K Peters, Ltd., 2010.
[7] U. Grenander and A. Srivastava, "Probability models for clutter in natural images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 4, pp. 424–429, 2001.
[8] E. W. Montroll and M. F. Shlesinger, "Maximum entropy formalism, fractals, scaling phenomena, and 1/f noise: a tale of tails," *Journal of Statistical Physics*, vol. 32, no. 2, pp. 209–230, 1983.
[9] N. N. Taleb, *The black swan: The impact of the highly improbable*. Random House Inc, 2010.
[10] D. Sornette, *Critical phenomena in natural sciences: chaos, fractals, selforganization, and disorder: concepts and tools*. Springer Verlag, 2004.
[11] M. Gell-Mann, *The Quark and the Jaguar: Adventures in the Simple and the Complex*. St. Martin's Griffin, 1995.
[12] N. Taleb, *Fooled by randomness: The hidden role of chance in life and in the markets*. Random House Inc, 2005.
[13] E. J. Gumbel and J. Lieblein, "Some applications of extreme-value methods," *The American Statistician*, vol. 8, no. 5, pp. 14–17, 1954.
[14] R. Hanel, Thurner, and M. Gell-Mann, "Generalized entropies and the transformation group of superstatistics." [Online]. Available: http://www.pnas.org/content/108/16/6390.full. [Accessed: 15-Mar-2012].
[15] D. Mumford, "The dawning of the age of stochasticity," *Mathematics: Frontiers and Perspectives*, pp. 197–218, 2000.
[16] S. Bankes, D. Challou, T. Haynes, H. Holloway, P. Pukite, J. Tierno, and C. Wentland, "META Adaptive, Reflective, Robust Workflow (ARRoW)," BAE Systems, Final Report TR-2742, 2011.
[17] D. J. Noakes, A. I. McLeod, and K. W. Hipel, "Forecasting monthly riverflow time series," *International Journal of Forecasting*, vol. 1, no. 2, pp. 179–190, 1985.
[18] BPA, "BPA Meteorological Information," *Meteorological Information from BPA Weather Sites*. [Online]. Available: http://transmission.bpa.gov/business/operations/wind/MetData.aspx. [Accessed: 31-May-2012].
[19] S. M. Papalexiou, D. Koutsoyiannis, and A. Montanari, "Can a simple stochastic model generate rich patterns of rainfall events?," *Journal of Hydrology*, 2011.
[20] T. Yuan, "Cloud macroscopic organization: order emerging from randomness," *Atmos. Chem. Phys*, vol. 11, pp. 7483–7490, 2011.
[21] J. A. Downing, Y. T. Prairie, J. J. Cole, C. M. Duarte, L. J. Tranvik, R. G. Striegl, W. H. McDowell, P. Kortelainen, N. F. Caraco, J. M. Melack, and others, "The global abundance and size distribution of lakes, ponds, and impoundments," *Limnology and Oceanography*, pp. 2388–2397, 2006.
[22] D. L. Wu, "MLS Cloud Ice Measurements." [Online]. Available: http://mls.jpl.nasa.gov/dwu/cloud/index.html. [Accessed: 27-Apr-2012].
[23] CDIP, "CDIP Homepage." [Online]. Available: http://cdip.ucsd.edu/?units=metric&tz=UTC&pub=public&map_stati=1,2,3. [Accessed: 24-Apr-2012].
[24] C. of E. US Army, "Wave Information Studies." [Online]. Available: http://wis.usace.army.mil/hindcasts.shtml. [Accessed: 31-May-2012].

[25] USGS, “USGS/EROS Find Data/Products and Data Available/DEMs.” [Online]. Available: http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/DEMs. [Accessed: 24-Apr-2012].
[26] P. R. Pukite, *The Oil Conundrum: Vol. 1 Decline, Vol. 2 Renewal*, vol. 1,2, 2 vols. Daina, 2011.
[27] G. Vico and A. Porporato, “Probabilistic description of topographic slope and aspect,” *J. Geophys. Res*, vol. 114, p. F01011, 2009.
[28] J. S. Gagnon, S. Lovejoy, D. Schertzer, and others, “Multifractal earth topography,” *Nonlinear Processes in Geophysics*, vol. 13, no. 5, pp. 541–570, 2006.
[29] G. Gonçalves and J. Santos, “Propagation of Dem Uncertainty: An Interval Arithmetic Approach.,” in *XXII International Cartographic Conference, Spain*, 2005.
[30] A. M. Guarnieri, “SAR interferometry and statistical topography,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 40, no. 12, pp. 2567–2581, 2002.
[31] D. Carrier, “Renewable Energy Analysis for Afghanistan,” RDECOM AMSAA, 2010.
[32] H. Jahns and J. Wheeler, “Long-term wave probabilities based on hindcasting of severe storms,” *Journal of Petroleum Technology*, vol. 25, no. 4, pp. 473–486, 1973.
[33] G. Z. Forristall, “Wave crest distributions: Observations and second-order theory,” *Journal of Physical Oceanography*, vol. 30, no. 8, pp. 1931–1943, 2000.
[34] R. C. Nelson, “Depth limited design wave heights in very flat regions,” *Coastal Engineering*, vol. 23, no. 1, pp. 43–59, 1994.
[35] J. Fenton, “Nonlinear wave theories,” *The Sea*, vol. 9, no. 1, pp. 3–25, 1990.
[36] M. A. Tayfun, “Narrow-band nonlinear sea waves,” *Journal of Geophysical Research*, vol. 85, no. C3, pp. 1548–1552, 1980.
[37] H. Socquet-Juglard, K. Dysthe, K. Trulsen, H. E. Krogstad, and J. Liu, “Probability distributions of surface gravity waves during spectral changes,” *Journal of Fluid Mechanics*, vol. 542, no. -1, p. 195, Oct. 2005.
[38] A. H. Izadparast and J. M. Niedzwecki, “Empirical Moment-Based Estimation of Rayleigh-Stokes Distribution Parameters,” presented at the Proc. 21st International Ocean and Polar Engineering Conference, 2011.
[39] R. H. Stewart, *Introduction to physical oceanography*. A & M University, 2003.
[40] C. DMI, “DMI / COI [Wave statistics].” [Online]. Available: http://ocean.dmi.dk/wavestat/index.uk.php. [Accessed: 20-Sep-2012].
[41] D. Koutsoyiannis, “The scaling properties in the distribution of hydrological variables as a result of the maximum entropy principle,” presented at the Geophysical Research Abstracts, 2005, vol. 7, p. 03781.
[42] J. R. A. Watson, *High-sensitivity GPS L1 signal analysis for indoor channel modelling*. University of Calgary, Department of Geomatics Engineering, 2005.
[43] M. Chen, C. Majidi, D. Doolin, S. Glaser, and N. Sitar, “Design and construction of a wildfire instrumentation system using networked sensors,” *Network Embedded Systems Technology (NEST) Retreat, Oakland California. Retrieved April*, vol. 5, 2004.
[44] S. Bocquet, “Calculation of Radar Probability of Detection in K-Distributed Sea Clutter and Noise,” DTIC Document, 2011.
[45] K. Telmer, M. Costa, and J. MacGregor, “K&C Science Report–Phase 1 Global Lake Census.”
[46] S. K. HAMILTON, J. M. MELACK, M. F. GOODCHILD, and W. Lewis, “Estimation of the fractal dimension of terrain from lake size distributions,” *Lowland floodplain rivers: Geomorphological perspectives. Wiley*, pp. 145–163, 1992.
[47] C. M. R. Platt, “A parameterization of the visible extinction coefficient of ice clouds in terms of the ice/water content,” *Journal of the atmospheric sciences*, vol. 54, no. 16, pp. 2083–2098, 1997.
[48] D. L. Wu, J. H. Jiang, and C. P. Davis, “EOS MLS cloud ice measurements and cloudy-sky radiative transfer model,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 5, pp. 1156–1165, 2006.
[49] Wikipedia, “Particle-size distribution - Wikipedia, the free encyclopedia.” [Online]. Available: http://en.wikipedia.org/wiki/Particle-size_distribution. [Accessed: 31-May-2012].
[50] Wikipedia, “Volcanic ash - Wikipedia, the free encyclopedia.” [Online]. Available: http://en.wikipedia.org/wiki/Volcanic_ash. [Accessed: 31-May-2012].
[51] Wikipedia, “Particle size (grain size) - Wikipedia, the free encyclopedia.” [Online]. Available: http://en.wikipedia.org/wiki/Particle_size_(grain_size). [Accessed: 31-May-2012].
[52] P. Pukite and S. Bankes, “Entropic Complexity Measured in Context Switching,” in *Applications of Digital Siignal Processing*, vol. 17, InTech, 2011.

