

**NUCLEATION AND VAPOR COMPOSITION: STATISTICAL ANALYSIS AND  
INTERPRETATION OF RATE MEASUREMENTS USING PCA**

R. McGraw  
Environmental Sciences Dept./Atmospheric Sciences Div.  
Brookhaven National Laboratory  
Upton, NY 11973-5000

Extended Abstract

Presented at the Hyytiälä Conference,  
"Formation and Growth of Secondary Atmospheric Aerosols",  
Hyytiälä, Finland  
August 14-17, 2005

September 2005

# Nucleation and vapor composition: Statistical analysis and interpretation of rate measurements using PCA

Robert McGraw  
Environmental Sciences Department  
Atmospheric Sciences Division  
Brookhaven National Laboratory, Upton, NY 11973

## Extended Abstract

**Overview:** Recent kinetic extensions of the nucleation theorem (KNT) [1] suggest that the logarithm of the steady-state nucleation rate has strong multi-linear dependence on the log concentrations of condensable species present in the vapor phase. A further remarkable result from the KNT is that the coefficients of this linear dependence provide a direct (model-free) determination of the molecular content of the critical nucleus itself. Building on these results, we demonstrate here the power of multi-linear regression methods, with emphasis on principal components analysis (PCA), for least squares parameterization, and molecular-level interpretation of nucleation rate measurements. The new approach is illustrated using laboratory data on the ternary vapor system p-toluic acid/ sulfuric acid/ water obtained by Renyi Zhang and co-workers [2]. The present analysis, part of which was carried out in collaboration with the Zhang group [3], yields direct information on the molecular content of the critical nucleus as well as a convenient, least-squares, parameterization for the vapor composition dependence of the nucleation rate. New results that bring in temperature as well as vapor composition dependence are described. Future extensions of these methods beyond laboratory nucleation rate measurements to the analysis of new particle formation in the atmosphere are discussed.

**Introduction:** The nucleation theorem (NT) is a thermodynamic result relating the sensitivity of the nucleation barrier height to changes in log vapor species concentration. Extensions to non-isothermal conditions, yielding the sensitivity of the barrier height to changes in temperature (the so-called ‘second nucleation theorem’) have also been developed (Ref. 1 includes many citations to this early work). A clear limitation of the NT is that the barrier height cannot be measured directly – unlike nucleation rate itself. NT-based estimates of the relative rate sensitivity are thus, at best, indirect. The early estimates generally assumed a rate expression in prefactor-exponent form,  $J = K \exp(-W^*/kT)$ , where  $K$  is the kinetic prefactor and  $W^*$  is the barrier height, and required additional arguments less fundamental than the NT itself. For example, the sensitivity of the kinetic prefactor  $K$  is now required, and this depends on which nucleation theory is used. Additionally, the prefactor-exponent form for  $J$  requires an activated process and is thus inapplicable under kinetically-controlled conditions where no barrier is present.

This situation was improved significantly through the development of kinetic nucleation theorems (KNTs) beginning with the work of Ford (see Ref. 1). KNTs provide for direct calculation of the rate sensitivity by incorporating the full Becker-Döring expression for the nucleation rate, thus including contributions from the multistate kinetics as well as thermodynamics to the nucleation rate. Under isothermal conditions:

$$\left( \frac{\partial \ln J}{\partial \ln S} \right)_T = \left( \frac{\partial \ln J}{\partial \ln n_1} \right)_T = \bar{g} + 1 \approx g^* + 1 \quad (1a)$$

where  $S$  is saturation ratio and  $n_1$  is monomer concentration. The overbar denotes the following average, defined over a molecular distribution of reciprocal equilibrium cluster growth rates:

$$\bar{y} = \frac{\sum_g \frac{1}{\beta_g n_g} y(g)}{\sum_g \frac{1}{\beta_g n_g}} = \sum_g P(g) y(g) \quad (1b)$$

where  $n_g$  is the (constrained) equilibrium concentration of  $g$ -clusters,  $\beta_g$  is the per-cluster collisional growth rate, and the last equality defines the normalized distribution  $P(g)$ . Eq. 1, originally due to Ford, was rederived for an ideal vapor, using only the deeply-rooted principles of mass action and detailed balance, and designated a ‘kinetic nucleation theorem’ in order to reflect this kinetic, as well as thermodynamic, foundation [1]. Multi-component extensions of the KNT, inclusion of temperature, chemically-induced nucleation, and expressions for higher-order derivatives based on the cumulants of  $P(g)$  were also developed [1]. For example, the second derivative is given by the negative variance of  $P(g)$ :

$$\left( \frac{\partial^2 J}{\partial (\ln S)^2} \right)_T = -[\bar{g}^2 - (\bar{g})^2] \equiv -\kappa_2 \quad (2)$$

where  $\kappa_2$  is the second cumulant and the averages are as defined in Eq. 1b. The  $n^{\text{th}}$  partial derivative is given by the remarkably simple result:  $(-1)^{n+1} \kappa_n$  [1].

A further advantage of the KNT is that kinetically-controlled cluster formation is also accommodated within its more general framework. (In this case the second equality of Eq. 1, and the equalities of Eq. 2, remain exact, with the averages from Eq. 1b still well defined. It is the third, approximate, equality of Eq. 1 that fails because  $g^*$  is no longer the well defined thermodynamic ‘bottleneck’ size that must be reached in order for still larger clusters to form.)

**PCA analysis of a ternary vapor studied by Zhang et al. - the p-toluic acid/ sulfuric acid/ water system [2]:** Experimental rate measurements tend to support the strong propensity to linear behavior when results are plotted in the  $\text{Log}(J)$  vs either  $\text{Log}(S)$  or  $\text{Log}(n_1)$  coordinates suggest by the NT/KNT. According to Eq. 2, this suggests that the distribution  $P(g)$  tends to be sharply peaked about the critical size – equivalently, that the change in  $\bar{g}$  is small over the range of the measurements. This suggest writing the nucleation rate, say for the above ternary system, in the multi-linear (using log coordinates) form, which on integration yields:

$$J = J_0 \left( \frac{[H_2SO_4]}{[H_2SO_4]_0} \right)^{a+\delta_a} \left( \frac{RH}{RH_0} \right)^{b+\delta_b} \left( \frac{[Org]}{[Org]_0} \right)^{c+\delta_c}, \quad (3)$$

where  $RH$  is relative humidity,  $[H_2SO_4]$  is the concentration of sulfuric acid ( $\text{cm}^{-3}$ ) and  $[Org]$  is the concentration of p-toluic acid (ppb) in the vapor phase. Consistency of Eq. 3 with the isothermal KNT follows where  $a$ ,  $b$ , and  $c$  are the number of molecules of sulfuric acid, water, and organic acid, respectively, in the critical nucleus.  $\delta_a$ ,  $\delta_b$ , and  $\delta_c$  are small quantities (between 0 and 1) that relate to the direction of the nucleation flux through the space of cluster size. Accordingly, the measurements from [2] are cast in terms of the coordinates  $\{x_i, y_i, z_i\}$  for

$i = 1, \dots, N$  where  $N$  is the number of measurements,  $x = \text{Log}_{10}[H_2SO_4, \text{ molecules } cm^{-3}]$ ,  $y = \text{Log}_{10}[\text{Organic}, \text{ ppb}]$ , and  $z = \text{Log}_{10}[J \text{ cm}^{-3} s^{-1}]$ . Temperature and  $RH$  are constant over the range of measurements.

PCA constitutes a general mathematical framework suitable for conventional least-squares regression as well as for providing a more ‘geometrically-flavored’ principal component analysis of the dataset. The procedure is carried out through the following sequence of steps: (1) compute the coordinate means,  $\{\mu_x, \mu_y, \mu_z\}$ , and the centered coordinates,  $\mathbf{w}(\mathbf{i}) = \{x_i - \mu_x, y_i - \mu_y, z_i - \mu_z\}$ , (2) form the 3x3 covariance matrix,  $\mathbf{C}$ , whose elements are the variances and covariances

$$\langle xx \rangle = N^{-1} \sum_{i=1}^N (x_i - \mu_x)(x_i - \mu_x); \quad \langle xy \rangle = N^{-1} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y); \text{ etc.},$$

and (3) solve the eigenvalue problem associated with  $\mathbf{C}$ . The normalized eigenvectors of  $\mathbf{C}$  are the principal component basis vectors ( $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ ) and the corresponding sorted eigenvalues ( $\lambda_1 \geq \lambda_2 \geq \lambda_3$ ) give the variances of the dataset along the directions of the principal components.

The transformed (principal) coordinates, shown in Fig. 1, are obtained as the scalar products:  $\eta_i = \mathbf{w}(\mathbf{i}) \cdot \mathbf{v}_i$ , etc. According to the KNT (Eq. 3) these should be nearly coplanar - lying in  $(\eta_1, \eta_2)$  plane (orthogonal to  $\eta_3$ ). Apart from a small amount of apparently random noise, this indeed appears to be the case. Transforming back to the original  $(x, y, z)$  coordinates, the equation of the plane takes the form:

$$z = -3.89 + 8.38x + 1.74y \quad (4)$$

thus providing an especially compact parameterization of the entire ternary data set.

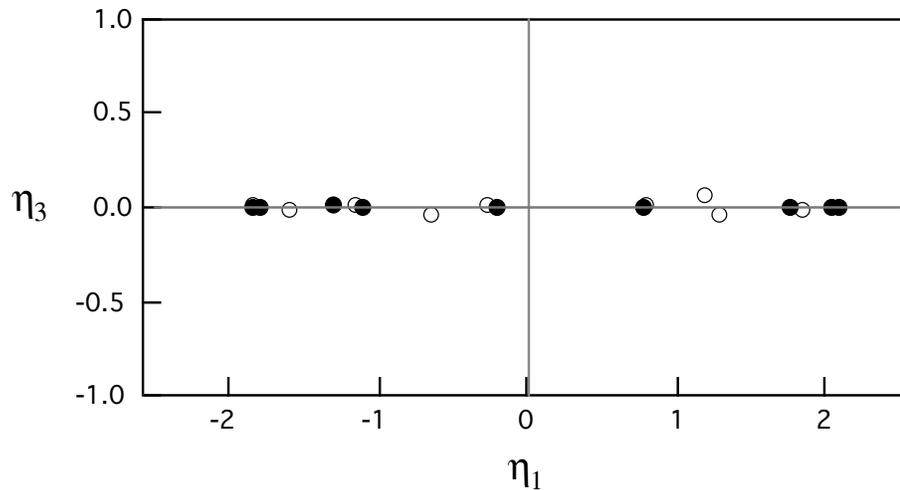


Figure 1. Ternary data set in principal coordinates.  $\eta_1$  is the axis along which the variance is greatest,  $\eta_3$  is the axis along which the variance is least. Filled circles, p-toluic acid concentration = 0.2ppb. Open circles, p-toluic acid concentration = 0.4ppb. Data shows no systematic departure from the  $(\eta_1, \eta_2)$  plane.

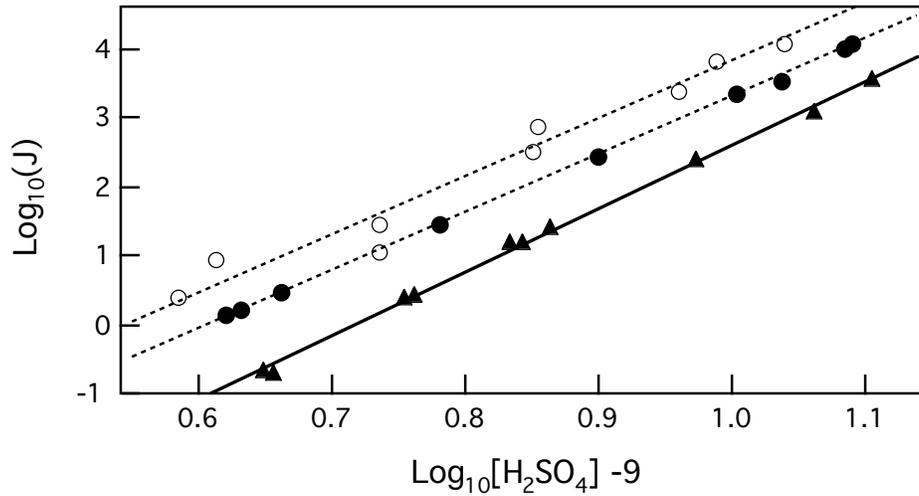


Figure 2. Zhang et al. ternary data set plotted in  $x$ - $y$ - $z$  coordinates and projected onto the  $x$ - $z$  coordinate plane. Filled circles, p-toluic acid concentration = 0.2ppb. Open circles, p-toluic acid concentration = 0.4ppb. Triangles, binary sulfuric acid-water measurements. Solid line, least-squares fit to the binary data. Dashed lines, projections from the PCA planar fit evaluated at the two p-toluic acid concentrations (these lines have the same slope).

Equation 4, with  $y$  held constant at each of the two measured p-toluic acid vapor concentrations, gives the parallel planar projections (dashed lines) shown in Fig. 2. The experimental data points are also shown on this log-log coordinate scale suggested by the KNT and are in excellent agreement with the planar fit (Eq. 2). Results for the binary sulfuric acid-water reference system are also shown.

The KNT, applied now to Eq. 4, takes the form:  $(\partial z / \partial x)_y = 8.38 = g^*_{H_2SO_4} + \delta_{H_2SO_4}$ ;  $(\partial z / \partial y)_x = 1.74 = g^*_{Org} + \delta_{Org}$ . Taking into account that the  $\delta$  values are positive and less than unity, we obtain the critical nucleus molecular composition: 8 sulfuric acid molecules and just 1 molecule of the organic acid!

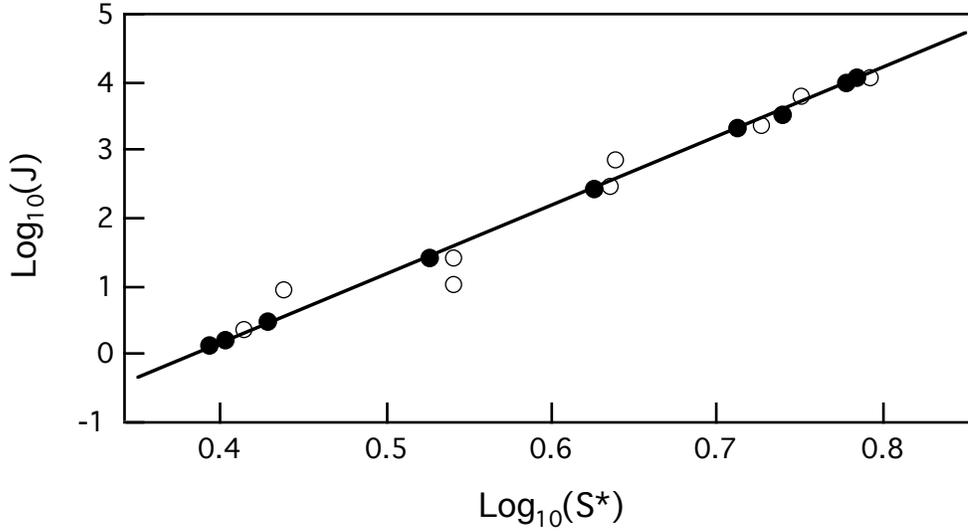


Figure 3. Re-plot of the ternary data from Fig. 2 using the effective homomolecular saturation  $S^*$  [4]. Filled circles, p-toluic acid concentration = 0.2ppb. Open circles, p-toluic acid concentration = 0.4ppb. Solid line is result from Eq. 5.

The PCA plane of Eq. 4 can be further made to collapse to a single line by simply combining the terms in  $x$  and  $y$  into an effective homomolecular vapor saturation,  $S^* = [H_2SO_4]^{x_1}[Org]^{x_2}$ . Here  $x_1 = 8.38/(8.38 + 1.74)$  and  $x_2 = 1.74/(8.38 + 1.74)$  are the relative fractions of sulfuric acid and p-toluic acid in the critical nucleus (apart from the small  $\delta$  terms included here). Equation 4 thus becomes:

$$\text{Log}_{10}J = -3.89 + (8.38 + 1.74)\text{Log}_{10}S^*. \quad (5)$$

Transforming the data set to  $\text{Log}(J)$ - $\text{Log}(S^*)$  coordinates and comparing with Eq. 5 yields the results shown in Fig. 3. The concept of using an effective homomolecular saturation to reduce the dimensionality of the nucleation theory is described in Ref. 4. Here we have shown that such reduction is both readily understood and readily accommodated within the framework of PCA.

**Status of recent results:** Results were also presented at the Workshop for the non-isothermal case, thus bringing temperature as well as vapor species concentrations into the statistical analysis. For this purpose we employed a new version of the KNT in which the partial derivative is taken at constant monomer concentration,  $n_1$ , rather than at constant  $S$  – as this is the more convenient form when having to deal with species for which the saturated vapor pressure, needed to compute  $S$ , is not well known:

$$\ln J = \ln J_0 + \left( \frac{\bar{g}E_1 - \bar{E}_g}{k} \right) \left( \frac{1}{T} - \frac{1}{T_0} \right). \quad (6)$$

The subscript ‘0’ refers to a reference condition – e.g. the centroid of the dataset. Eqs. 1 and 6 were combined and a three-coordinate  $\{\log(n_1), 1/T, \log(J)\}$  version of PCA tested using a data set of water vapor measurements recently published by Wölk and Strey [5]. For this case the regression gives the coefficient in Eq. 6, which is the critical cluster energy relative to the monomeric vapor, as well as the molecularity of the critical nucleus. For best results we

included quadratic terms in the expansion for  $\text{Log}(J)$  (cf. Eq. 2) so as to account for mild curvature seen in the dataset relative to the planar fit. These results will be fully described in a future publication [5].

In their present form, these new methods already provide a powerful approach to parameterization and molecular-level interpretation of the nucleation processes that underlie atmospheric new particle formation. Future extensions should take into account the effect of background aerosol both on the nucleation rate itself [6] and on the scavenging loss of freshly nucleated particles during their subsequent growth to measurable and climate-significant size [7]. These two effects, each being characterized by the same non-dimensional parameter (a ratio of scavenging to growth rates [6,7]) are likely to be equally important and of comparable size.

### References:

1. R. McGraw and D. Wu, "Kinetic extensions of the nucleation theorem", *J. Chem. Phys.*, *118*, 9337-9347 (2003).
2. R. Zhang, I. Suh, J. Zhao, D. Zhang, E. C. Fortner, X. Tie, L. T. Molina, and M. J. Molina, "Atmospheric new particle formation enhanced by organic acids", *Science* *304*, 1487-1490 (2004).
3. R. Zhang et al., in preparation.
4. M. Kulmala and Y. Viisanen, "Homogeneous nucleation: reduction of binary nucleation to homomolecular nucleation", *J. Aerosol Sci.* *22*, Supplement 1, S97-100 (1991).
5. R. McGraw, in preparation.
6. R. McGraw and W. H. Marlow, "The multistate kinetics of nucleation in the presence of an aerosol", *J. Chem. Phys.* *78*, 2542-2548, 1983.
7. McMurry, P. H., M. Fink, H. Sakurai, M. R. Stolzenburg, L. Mauldin, K. Moore, J. Smith, F. Eisele, S. Sjostedt, D. Tanner, L. G. Huey, J.B. Nowak, E. Edgerton, D. Voisin, 2005, "A Criterion for New Particle Formation in the Sulfur-Rich Atlanta Atmosphere," *J. Geophys. Res.*, in press.