

Available online at www.sciencedirect.com**ScienceDirect**

Procedia Engineering 99 (2015) 1082 – 1088

**Procedia
Engineering**www.elsevier.com/locate/procedia

“APISAT2014”, 2014 Asia-Pacific International Symposium on Aerospace Technology,
APISAT2014

Monte Carlo Analysis for Significant Parameters Ranking in RLV Flight Evaluation

Jie Gu, Shuguang Zhang*, Baoyin Wang

School of Transportation Science and Engineering, Beihang University, 37 Xueyuan Road, Beijing 100191, P.R. China

Abstract

Monte Carlo simulation is an effective method for evaluating complex systems. Besides estimating the performance level of the system through Monte Carlo method, it is more wanted to identify key factors in system operation so as to improve or redesign the system. When estimating the performance level, in order to obtain sufficient evaluation accuracy while keeping time cost as low as possible, its relation with confidence level and number of simulation runs is explained according to probability and statistics theory. To identify key factors, a method ranking the significant influencing parameters automatically for complex systems based on naive Bayes classifier (NBC) and kernel density estimator (KDE) is developed. NBC used for classification makes the method valid for all kinds of linear and nonlinear complex systems, and KDE contributes greatly to identifying significant influencing parameters in automated manner. The method above is applied to a reusable launch vehicle (RLV) flight evaluation. Through the evaluation, bias of atmosphere density is identified as the most significant parameter which relies on the flight control mode in the terminal flight phase.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of Chinese Society of Aeronautics and Astronautics (CSAA)

Keywords: number of simulation runs; naive Bayes classifier; kernel density estimator; posterior probability; TAEM;

Nomenclature

$P(A)$ probability of event A occurring

$P(A|B)$ conditional probability of event A occurring given event B occurred, or posterior probability of A given B

$p(x)$ probability density function (PDF) of the random variable x

* Corresponding author. Tel.: +86-10-82315237; fax: +86-10-82315237.

E-mail address: gnahz@buaa.edu.cn

$\Phi(\cdot)$ cumulative distribution function (CDF) of the standard normal distribution

1. Introduction

Identification of key factors in system operation is critical for any system design. In modern aerospace systems, the increasing complexity has made it impossible to explicitly tell the key influencing factors based on intuition; and thus methods like Monte Carlo have been extensively adopted [1-3]. The key philosophy of Monte Carlo analysis is to find the relations between input and output parameters of a system, from data produced in sufficient simulations [4, 5]. However, two problems were encountered when applying the Monte Carlo analysis to complex guidance and control systems.

The first problem is to determine how many simulation runs are sufficient for Monte Carlo analysis. Traditionally, it was determined by experience or on-line monitoring output variance. However, to obtain results with sufficient accuracy while keeping time cost as low as possible, it is preferable to predetermine the number of simulation runs needed for assessing resources and time required. Recently, Hanson and Beard explain a method (order statistics) for determining the runs required. But it is not familiar to the GN&C community and somewhat complex [1].

In Monte Carlo analysis, Gardner et al. recommended using simple correlation coefficients to rank model parameters [6]. But its use is limited by the inherent assumption that the input/output relationship is linear [7], for the world of complex systems can seldom be fully linearized [8]. To analyze complex systems with nonlinearity, some graphical methods such as scatter plots are applied, which need engineers to analyze all the graphs manually [2, 9]. However, for complex systems, there would be considerable number of input and output parameters and a great amount of Monte Carlo data to be analyzed. So how to select the most significant parameters of complex systems with nonlinearity in automated manner, is the second problem arising.

To solve the problem above, this paper derives two formulas to determine appropriate number of simulations according to probability and statistics theory, and a method based on NBC and KDE is applied to Monte Carlo analysis for significant parameters ranking. By applying the method above to a reusable launch vehicle (RLV) flight evaluation, significant uncertain parameters are pointed out and further analyzed.

2. Number of Monte Carlo runs

When applying Monte Carlo method to evaluate a system, the system is simulated a number of times under input parameters with a desired distribution generated by a high-quality random number generator (RNG). Then the property of the system can be estimated by comparing the statistics of the output with a criterion. In order to arrange appropriate resources and time for the evaluation, this section will explain how to predetermine the number of Monte Carlo runs that influences the evaluation accuracy.

The output of a Monte Carlo run is denoted by a random variable Y_i ($i = 1, 2, \dots$), whose expectation and variance are $E(Y_i) = \mu$ and $D(Y_i) = \sigma^2$ ($\sigma \geq 0$) according to the law of large numbers [5]. For N Monte Carlo runs, random variables Y_1, Y_2, \dots, Y_N are independent and identically distributed. According to central limit theorem [5], the statistic \bar{Y} , average of Y_1, Y_2, \dots, Y_N , used to estimate the system output expectation, will have a distribution that is approximately normal with mean μ and variance σ^2/N , then for each fixed positive number z ,

$$\lim_{N \rightarrow \infty} P\left(\left| \frac{\bar{Y} - \mu}{\sigma/\sqrt{N}} \right| \leq z\right) = 2\Phi(z) - 1 \quad (1)$$

For the statistic \bar{Y} with expectation $E(\bar{Y}) = \mu$ and variance $D(\bar{Y}) = \sigma^2/N$, according to the Chebyshev's inequality [5], there will be a relation that for any positive number ε ,

$$P\left(\left| \bar{Y} - \mu \right| < \varepsilon\right) \geq 1 - \sigma^2/N/\varepsilon^2 \quad (2)$$

Setting $\varepsilon = z \sigma/\sqrt{N}$, and considering N is finite, a relation is derived from Eq.(1) and Eq.(2),

$$1 - 1/z^2 \leq P(|\bar{Y} - \mu| < \varepsilon) \leq 2\Phi(z) - 1 \tag{3}$$

When evaluating the output of the system with a criterion, two results are set based on whether the criterion is satisfied. So the result would have the Bernoulli distribution with parameter $p(0 \leq p \leq 1)$. And the evaluation result can be denoted by the random variable, $Y_i (i = 1, 2, \dots)$, with expectation $E(Y_i) = p$ and variance $D(Y_i) = p(1 - p) \leq (p + 1 - p)^2/4 = (1/2)^2$. So, $\sigma^2 = D(Y_i) \leq (1/2)^2$. Hence,

$$N = (z\sigma/\varepsilon)^2 \leq (z/\varepsilon/2)^2 \tag{4}$$

where z represents the confidence level, and ε represents the evaluation error. So the relation among Monte Carlo analysis accuracy, confidence level, and the number of simulation runs can be represented by Eq.(3) and Eq.(4). For example, if $z = 3, \varepsilon = 0.05$, there will be $N \leq 900$, and $0.8889 \leq P(|\bar{Y} - \mu| < 0.05) \leq 0.9974$, which means that the statistics of 900 Monte Carlo simulations will have a maximal error of 5% with a confidence level between 0.8889 and 0.9974.

3. Method of significant analysis based on Monte Carlo data

Evaluating the system by Monte Carlo method is not only to estimate the performance, but also to identify the significant input parameters influencing the performance so as to improve or redesign the system. This section introduces naive Bayes classifier (NBC), a machine learning method, to evaluate the system by computing posterior probability of input parameters based on Monte Carlo data. And the probability is estimated by kernel density estimator (KDE). The difference of the posterior probability of the same input parameter given different types of output would represent the influence of the input parameter on the system.

3.1. Naive Bayes classifier

Naive Bayes classifier is a classification algorithm by learning from the training data set based on Bayes' theorem and the conditional independence assumption between the features, according to the maximum a posterior probability (MAP) decision rule derived from the principle of minimization risk [10].

For a system, input space $\mathbb{X} \subseteq \mathbb{R}^n$ is a set of n -dimensional vectors, and output space $\mathbb{Y} = \{c_1, c_2, \dots, c_k, \dots, c_K\}$ is a set of class labels; thus, input $x \in \mathbb{X}$ is an n -dimensional feature vector, and output $y \in \mathbb{Y}$ is a class label. Let X be the random vector defined in the input space \mathbb{X} , and let Y be the random variable defined in the output space \mathbb{Y} .

Every specific input is an instance, shown as $x = (x^1, x^2, \dots, x^j, \dots, x^n)^T$, where x^j denotes the j th feature. While $x_i = (x_i^1, x_i^2, \dots, x_i^n)^T$ denotes the i th instance, and y_i is the output corresponding to x_i . Thus, the training data set can be denoted by $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$.

Classification is to label a new instance x with a label y , namely, to find a function $y = f(x)$, making the expectation of the difference between predicted value $f(x)$ with actual value minimum, which is the principle of minimization risk. While NBC uses the MAP decision rule to find the classification function, which is equivalent with the principle of minimization risk, shown as below,

$$f(x) = \arg \max_{c_k} P(Y = c_k | X = x) \quad (k = 1, 2, \dots, K) \tag{5}$$

The condition probability of $Y = c_k$ given $X = x$ in Eq.(5) is calculated based on Bayes' theorem and conditional independence assumption between the features, as follow,

$$P(Y = c_k | X = x) = \frac{P(X = x | Y = c_k)P(Y = c_k)}{\sum_{i=1}^K P(X = x | Y = c_i)P(Y = c_i)} = \frac{P(Y = c_k) \prod_{j=1}^n P(X^j = x^j | Y = c_k)}{\sum_{i=1}^K P(X = x | Y = c_i)P(Y = c_i)} \tag{6}$$

Considering denominator values in Eq.(6) are same for all the $c_i (i = 1, 2, \dots, K)$, NBC will be,

$$f(x) = \arg \max_{c_k} P(Y = c_k) \prod_{j=1}^n P(X^j = x^j | Y = c_k) \quad (k = 1, 2, \dots, K) \tag{7}$$

3.2. Kernel density estimator

Kernel density estimator method is a non-parametric way to estimate the PDF of a random variable without any prior distribution knowledge, which is also called Parzen window method [11]. Different from the simple histogram density estimator, KDE uses a kernel function to construct a smooth PDF, which will in general more accurately reflect the underlying variable.

For a random variable x with a sample data set $\{x_1, x_2, \dots, x_N\}$, the PDF estimated by KDE is [12],

$$p(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{h^D} k\left(\frac{x-x_i}{h}\right) \tag{8}$$

where $k(\cdot)$ is the kernel function, a given PDF, h is called bandwidth playing the role of a smoothing parameter, D is the variable dimension, and N is the sample size. Generally, Gaussian kernel is taken as the kernel function to make the PDF smooth, thus Gaussian kernel PDF of a one-dimensional random variable is,

$$p(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{\sqrt{2\pi}h} e^{-(x-x_i)^2/2h^2} \tag{9}$$

where the bandwidth h also represents the standard deviation of the Gaussian components, the optimal choice of which is proposed by Silverman [12], shown as $h_s = \sigma[4/(3N)]^{-1/5}$, where σ is the standard deviation of the sample data set. Thus the PDF estimated by KDE is the sum of contributing Gaussian kernels located at each i th sample, as shown in Fig. 1. (a).

3.3. Significant parameters ranking

Based on Monte Carlo data, the system output type corresponding to a new input can be predicted by NBC, shown as Eq.(7), which is valid for any system including the system with nonlinearity. Since the probability of different output types ($Y = c_k$) based on the Monte Carlo data are fixed whatever the input is, the output types would be influenced by the difference of the posterior probability of the same input parameter given different output types, which could be obtained according to the posterior PDF of the input parameter estimated by KDE.

For a system evaluation with two types of results according to whether the criterion is satisfied, define the output space as $\mathbb{Y} = \{c_1 = \text{'success'}, c_2 = \text{'failure'}\}$. Thus the posterior PDF of the input parameter will be denoted as $p(x|c_k) (k = 1, 2)$. Fig. 1. (b) shows the two curves of the posterior PDF of an input parameter given two types of results. The non-overlapping area underlying the curves could represent the difference of the posterior probability of the input parameter, which can be as the index of significant input parameters for ranking (JR), expressed as,

$$JR = \int_{-\infty}^{+\infty} |p(x|c_1) - p(x|c_2)| dx \tag{10}$$

In order to compute the index automatically, the form of numerical calculation is given as below,

$$JR = \sum_{j=1}^M \left| \frac{1}{N_1} \sum_{i=1}^{N_1} \frac{1}{\sqrt{2\pi}h_1} e^{-(x_0+j|\Delta-x_i)^2/2h_1^2} - \frac{1}{N_2} \sum_{i=1}^{N_2} \frac{1}{\sqrt{2\pi}h_2} e^{-(x_0+j|\Delta-x_i)^2/2h_2^2} \right| \Delta \tag{11}$$

where $\Delta = [\max(x_i) - \min(x_i) + 6 \max\{h_1, h_2\}] / M$, $x_0 = \min(x_i) - 3 \max\{h_1, h_2\} - \Delta$, and M is the number of discrete intervals which are added up to approximate the integral value.

Based on Monte Carlo data, JR s of all the input parameters are computed according to Eq.(11) automatically. And the larger JR is, the more significant the input parameter will be, which should be further analyzed.

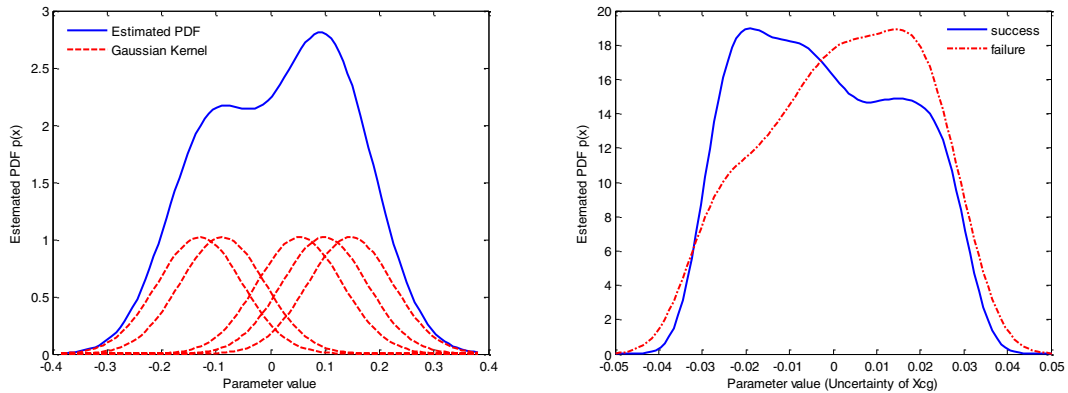


Fig. 1. (a) estimated PDF by KDE; (b) estimated PDF of parameter Xcg given success/failure.

4. Application to a reusable launch vehicle flight evaluation

The method above is applied to evaluate the robustness of the guidance and control system, and to identify the significant uncertain parameters for a RLV in the flight phase of terminal area energy management (TAEM) [13], which is an unpowered gliding flight. Firstly, Monte Carlo data is generated by running the TAEM flight simulation, the statistics of which would show the robustness of the system. Then, rank the uncertain parameters by the method above based on the Monte Carlo data. At last, top part of significant uncertain parameters are further analyzed to understand the system better, so as to improve or redesign the system.

4.1. Monte Carlo simulation

According to section 2, 1000 simulation runs are chosen to make the statistics of the Monte Carlo data having smaller estimation error at a higher confidence level. There are 22 uncertain parameters as the system input, including parameters of aerodynamics, vehicle configuration, mass property, and environmental condition, which have a normal or uniform distribution. Thus 1000 groups of input parameters are randomly generated according to their distribution types and distribution parameters. Then, run the TAEM flight simulation, and record the corresponding output, which is compared with the criterion that whether the dispersions of flight parameters (downrange, cross-range, Mach, heading angle, etc.) at the end of the TAEM flight exceed the acceptable range.

4.2. Ranking of uncertain parameters

The statistics of Monte Carlo data obtained in last subsection show that the guidance and control system of TAEM flight is robust with 88% mission success ratio, and lots of mission are labeled as failure for Mach exceeding the acceptable range. In order to find the mission failure reason, the uncertain parameters are ranked according to their JR s computed based on Eq.(11) automatically. The result is shown in Table 1 and Fig. 2. (a). Thus, the top parameter Rho, JR of which is far larger than the others, has main contribution to the mission failure, and should be further analyzed. What's more, the shape of PDF of the uncertain parameters can reflect the trend of its effect on the system visually, which could be useful when making a detailed analysis. PDFs of the top parameter Rho and the last parameter CY are shown in Fig. 3, and PDFs of the second parameter Xcg are shown in Fig. 1. (b).

Table 1. Ranked uncertain parameters

Rank	JR	Parameter	Rank	JR	Parameter
1	1.270136	Rho (atmosphere density)	4	0.201035	CNp (cross derivatives)
2	0.256878	Xcg (center-of-gravity in axial direction)	5	0.192071	lyz (product of inertia)
3	0.230256	CR (rolling moment coefficient)	:	:	:

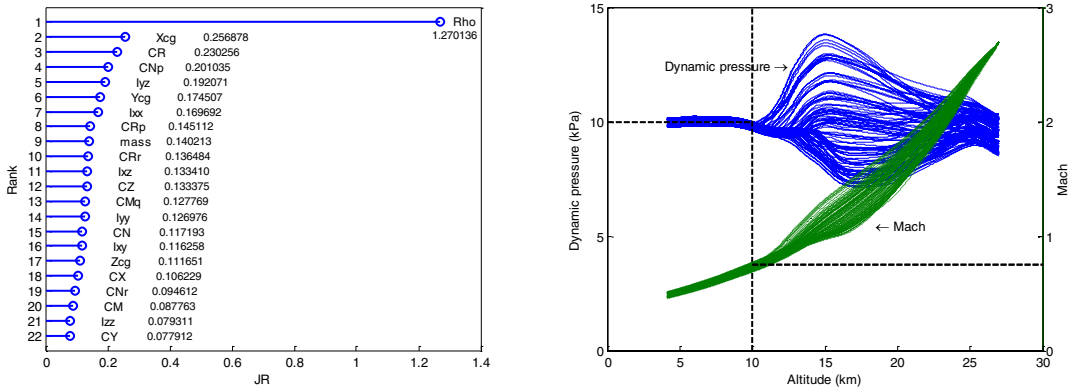


Fig. 2. (a) ranked uncertain parameters; (b) dynamic pressure and Mach profiles from Monte Carlo simulations.

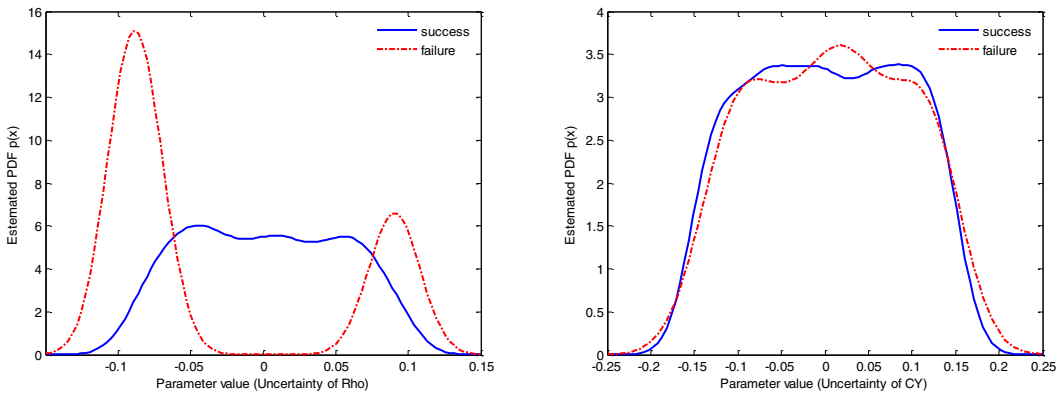


Fig. 3. (a) estimated PDF of parameter Rho given success/failure; (b) estimated PDF of parameter CY given success/failure.

4.3. Further analysis for the most significant uncertain parameters

From the analysis above, atmosphere density Rho is determined as the most significant uncertain parameter influencing the mission success, which has the main contribution to that Mach exceeding the acceptable range. Owing to Rho and Mach are associated only by dynamic pressure, expressed as $q = \rho a^2 Ma^2 / 2$, where q is dynamic pressure, ρ is atmosphere density, a is speed of sound, and Ma is Mach number, variation of dynamic pressure would determine how atmosphere density influence Mach. Since the flight guidance and control mode at the final phase of TAEM flight is using airspeed brakes for accurate control of dynamic pressure in the low-speed, low-altitude state, shown as in Fig. 2. (b), atmosphere density and Mach would be reciprocal relationship, which fits

the information shown in Fig. 3. (a), that the mission would more likely fail when the absolute value of uncertainty of Rho is large, and it would more likely succeed when the value is small. So in order to decrease the mission failure, uncertainty of atmosphere density should be lessened or the guidance and control system of TAEM flight should be improved.

Fig. 2. (a) shows that, the other uncertain parameters have far less effect on the mission than the top significant parameter Rho. But from the curves of their posterior PDFs, something about how they effect on the result can be inferred. Taking the second significant parameter Xcg as an example, the probability of mission failure is slightly greater than that of success, when Xcg is large; and the probability of success is slightly greater than that of failure, when Xcg is small, as shown in Fig. 1. (b). This information fits the flight dynamic characteristic that the longitudinal stability margin decreases when center-of-gravity of the vehicle shifts rearward. And the fact that uncertainty of center-of-gravity has little effect on the mission indicates that the longitudinal control system is robust.

5. Conclusions

This paper explains how many Monte Carlo samples are appropriate for systems evaluation, and introduces NBC and KDE to ranking significant parameters in RLV flight evaluation. The results indicate that:

(1) Comparing to the traditional way of determining simulation runs by experience or on-line monitoring output variance, the derived relations among analysis accuracy, confidence level, and number of simulation runs provides a guide to predetermine the number of Monte Carlo simulation runs needed.

(2) The NBC-KDE-based method provides a new approach to identify significant influencing factors for complex systems. It not only is valid for all kinds of linear and nonlinear systems, but also reduces manpower and time costs for running in automated manner.

(3) The uncertainty parameters influencing RLV flight are ranked automatically, and bias of atmosphere density has the largest effect on terminal performance of TAEM flight control, which lies in the TAEM flight control mode that using airspeed brakes for accurate control of dynamic pressure in the low-speed, low-altitude state.

Acknowledgements

The authors acknowledge Dr. Peng TANG in Beihang University for offering the guidance module of the TAEM simulation system. And the authors are also grateful to PhD Candidate Lei GONG in Beihang University for cooperating with the authors in designing the control module of the TAEM simulation system.

References

- [1] J.M. Hanson, B.B. Beard, Applying Monte Carlo simulation to launch vehicle design and requirements verification, *Journal of Spacecraft and Rockets*, 2012, 49(1): 136-144.
- [2] J.M. Hanson, C.E. Hall, Learning about Ares I from Monte Carlo simulation, AIAA paper 2008-6622, 2008.
- [3] Y.K. Chen, T. Squire, B. Laub, et al., Monte Carlo analysis for spacecraft thermal protection system design, AIAA paper 2006-2951, 2006.
- [4] J.M. Hammersley, D.C. Handscomb, Monte Carlo methods, Methuen & Co. Ltd., London, 1964.
- [5] Z. Sheng, S.Q. Xie, C.Y. Pan, Probability & statistics, third ed., Higher Education Press, Beijing, 2001. (in Chinese)
- [6] R.H. Gardner, R.V. O'Neill, J.B. Mankin, et al. A comparison of sensitivity analysis and error analysis based on a stream ecosystem model, *Ecol. Modelling*, 1981 (12): 173-190.
- [7] D.M. Hamby, A review of techniques for parameter sensitivity analysis of environmental models, *Environmental Monitoring and Assessment*, 1994, 32(2): 135-154.
- [8] P.H. Zipfel, Modeling and simulation of aerospace vehicle dynamics, second ed., AIAA Inc., Reston, 2007.
- [9] R.M. Cooke, J.M. van Noortwijk, Graphical methods for uncertainty and sensitivity analysis, In: G. Schueller, P. Kafka (Eds.), *Safety and Reliability, Proceedings of ESREL '99*, Balkema, Rotterdam, 1999, pp. 1187–1192.
- [10] H. Li, *Statistical learning method*, Tsinghua University Press, Beijing, 2012. (in Chinese)
- [11] E. Parzen, On estimation of a probability density function and mode, *Annals of Mathematical Statistics*, 1962, 33: 1065-1076.
- [12] B.W. Silverman, *Density estimation for statistics and data analysis*, Chapman and Hall, London, 1986.
- [13] T.E. Moore, Space shuttle entry terminal area energy management, NASA Technical Memorandum 104744, 1991.