Data-Driven Anomaly Detection Performance for the Ares I-X Ground Diagnostic Prototype

Rodney A. Martin [1], Mark A. Schwabacher [1], and Bryan L. Matthews [2]

[1] *NASA Ames Research Center, Moffett Field, CA, 94035, USA*
rodney.martin@nasa.gov
mark.a.schwabacher@nasa.gov
[2] *Stinger Ghaffarian Technologies Inc. at NASA Ames Research Center, Moffett Field, CA, 94035, USA*
bryan.l.matthews@nasa.gov

ABSTRACT

In this paper, we will assess the performance of a data-driven anomaly detection algorithm, the Inductive Monitoring System (IMS), which can be used to detect simulated Thrust Vector Control (TVC) system failures. However, the ability of IMS to detect these failures in a true operational setting may be related to the realistic nature of how they are simulated. As such, we will investigate both a low fidelity and high fidelity approach to simulating such failures, with the latter based upon the underlying physics. Furthermore, the ability of IMS to detect anomalies that were previously unknown and not previously simulated will be studied in earnest, as well as apparent deficiencies or misapplications that result from using the data-driven paradigm. Our conclusions indicate that robust detection performance of simulated failures using IMS is not appreciably affected by the use of a high fidelity simulation. However, we have found that the inclusion of a data-driven algorithm such as IMS into a suite of deployable health management technologies does add significant value.

1 INTRODUCTION

In preparation for the launch of Ares I-X, a data-driven anomaly detection algorithm was deployed as part of a suite of several software tools for inclusion in a ground diagnostics prototype to support detection and diagnosis of potential anomalies or failures during the prelaunch phase. The selected data-driven anomaly detection algorithm, IMS (Inductive Monitoring System), is based on incremental clustering, and operates with a semi-supervised anomaly detection paradigm, as defined in previous work (Chandola *et al.*, 2009). This implies complete reliance on training data of only the nominal class. As such the training data is implicitly labeled, and there are no labels for the anomalous class. The clustering is performed in an unsupervised manner, and any monitored data points falling outside of the clusters are flagged as anomalous. Detailed descriptions of how IMS performs anomaly detection are provided in previous work (Iverson, 2004), (Iverson *et al.*, 2009), (Martin, 2010).

Due to the lack of available nominal and fault data with which to validate and test the algorithm for Ares I-X, data from the Thrust Vector Control (TVC) System from previous Space Shuttle missions was used. The data collected served two purposes: as nominal data, and fault data which was constructed by seeding nominal data with failures of various types, severity, and fidelity for subsequent validation and testing. However, the ability of IMS to detect true failures may possibly be influenced by the realism of how they are simulated and subsequently tested. As such, a significant portion of this paper will be dedicated to investigating a computationally efficient approach to simulating such failures, and observing the effect of the increased fidelity on detection performance, extending what was presented in previous work (Martin *et al.*, 2010).

IMS was one of several data-driven anomaly detection tools that were evaluated for inclusion as part of the suite of technologies to be demonstrated during the Ares I-X test launch, which included both model-based and rule-based technologies. Data-driven algorithms are just one of three different types of algorithms that were deployed, the details of which were presented in previous work (Iverson *et al.*, 2009), (Schwabacher and Waterman, 2008) and (Schwabacher *et al.*, 2010a). The other two types of algorithms that were deployed include a "rule-based" expert system, and a "model-based" system. Within these two categories, the deployable candidates were selected based upon their flight heritage and system certifiability. For the rule-based system, SHINE (Spacecraft Health Inference Engine) (James and Atkinson, 1990) was selected for deployment, which is used within two components of BEAM (Beacon-based Exception Analysis for Multimissions)

(Mackey *et al.*, 2001). Other components of BEAM include various data-driven algorithms. BEAM is a patented technology developed at NASA's JPL (Jet Propulsion Laboratory). SHINE serves to aid in the management and identification of operational modes. For the "model-based" system, a commercially available package developed by QSI (Qualtech Systems, Inc.), TEAMS (Testability Engineering and Maintenance System) was highlighted in work subsequent to its debut (Cavanaugh, 2001), and was selected for deployment to aid in diagnosis. In the context of this particular deployment, distinctions among the use of the terms "data-driven," "rule-based," and "model-based," can be found in the previously cited paper (Schwabacher and Waterman, 2008). In the final deployed software package, we integrated TEAMS with IMS in the Ares I-X Ground Diagnostics Prototype (GDP) by running the two in parallel and displaying the outputs of both tools on the same console, and we used SHINE to provide the inputs to TEAMS.

In this effort, it is of great importance to provide for a robust and accurate detection of a variety of known fault modes that span a number of different rates of progression and severity. However, this capability is already well-provided for by other model-based tools (*i.e.* TEAMS) within the suite of deployed tools. IMS should be able to detect these, as well as unknown faults or anomalies that otherwise may have not been modeled from a top-down, or data-driven perspective, rather than a bottom-up, or model-based perspective. A review of the resulting performance of the entire deployed package has also been provided (Schwabacher *et al.*, 2010a), (Schwabacher *et al.*, 2010b). Other related work covering similar topics is also available in the literature (Iverson *et al.*, 2009), (Park *et al.*, 2002), (Pisanich *et al.*, 2006), (Rao *et al.*, 2009).

Some advantages that IMS has over the model-based and rule-based algorithms include the fact that:

1. It has the ability to detect anomalies that were previously unknown and not previously simulated or accounted for.
2. It has the potential to detect anomalies that are precursors of faults before a model-based system detects the fault.
3. It does not require a labor-intensive modeling process.

The disadvantages of IMS compared with model-based tools are:

1. It does anomaly detection only, not diagnosis, so that additional analysis is necessary to determine whether a detected anomaly is significant or not.
2. It only provides an acceptable level of accuracy if it is trained using a sufficient quantity of historical and/or simulated training data.

In previous work (Martin, 2010), we studied three candidates to provide the primary role of data-driven anomaly detection, which included IMS. Of the three algorithms tested, it was found that IMS was the best performing algorithm when considering both overall accuracy as quantified by the area under the Receiver Operating Characteristic (ROC) curve (AUC), and computational complexity. In this paper we aim to follow up with more detail on the performance of IMS in its designated primary roles as specified above, exploring both its advantages and disadvantages. In doing so, we will demonstrate that the other model-based and rule-based technologies with which IMS was deployed provided certain capabilities which IMS complemented well in some cases, while in other cases, the performance of IMS was less than desirable due to inappropriate use.

The remainder of this paper will be organized as follows: Section 2 will provide a detailed description of all simulated failures to be tested, including the higher fidelity version based upon physics. Section 3 provides a comparative discussion of the performance of IMS as it relates to the ability to robustly detecting simulated failures of varying fidelities. Section 4 will provide a general discussion of the selection of IMS as the data-driven anomaly detection algorithm, selection of parameters, training, validation & testing procedures. Both quantitative and qualitative performance results for Shuttle and Ares I-X data at the pad and at the Vehicle Assembly Building (VAB) will also be discussed. Section 5 provides a comparative discussion of the performance of IMS, and the model-based detection and diagnostic tool, TEAMS. The final concluding section will provide an overall summary and epilogue.

2 SIMULATED FAILURES

Historical Space Shuttle data was used to test the entire Ares I-X ground diagnostic prototype. The Space Shuttle Solid Rocket Booster (SRB) TVC is virtually identical to the Ares I-X first-stage TVC, so the SRB TVC data was expected to be very similar to the Ares I-X TVC data. Similarly, the ground hydraulic system used with the SRB TVC is virtually identical to the ground hydraulic system used with the Ares I-X TVC. These assumptions held up modestly well after our post-flight analysis, in consideration of all the tools that were deployed to support failure and anomaly detection. The differences that we found in the data were caused by differences in operations between Shuttle and Ares I-X, rather than by differences in the TVC or HSS hardware.

The SRB TVC and the associated ground hydraulic system have had very few failures. We thus had available to us an abundance of nominal data, but very little failure data. We therefore decided to develop a set of failure simulations that could be used to test the ability of the prototype to detect and diagnose failures. We inserted simulated failures into the historical Shuttle data, and used the resulting data sets to test the prototype before the Ares I-X launch.

Table 1 provides a summary of the failure modes that we simulated for each vehicle location. In order to test the integration of the TVC and HSS TEAMS models, we decided to select one failure mode that can be isolated to the TVC (Failure Mode 1a, FSM Leak)[1], one that can be isolated to the HSS (Failure

[1]The Fuel Supply Module (FSM) leak is a N_2H_4 (hydrazine) leak resulting in a pressure drop, and is simulated within 1 min prior to launch at the pad and within the 34 minute period after the calibration test in the VAB.

Table 1: Failure Mode Summary

Failure Mode Label	Vehicle Location	Failure Mode
1a	Pad	FSM Leak
	VAB	
2	VAB	HPU overheat
3	Pad	Hydraulic Leak
	VAB	
4	VAB	Stuck actuator

Mode 2, HPU overheat)[2], and one that would produce a TEAMS ambiguity group including both TVC and HSS candidates (Failure Mode 3, Hydraulic Leak).[3] In addition, because the actuator positioning test was considered to be the most important pre-launch test of the TVC, we decided to simulate a failure during this test (Failure Mode 4, Stuck actuator).[4] We will only describe failure mode 1a in the remainder of this section, as it includes examples of simulations that span the range of fidelity used for all of the failure modes. For the remaining failure modes, low fidelity linear simulations were used and simulated in a similar fashion as the low fidelity version of failure mode 1a. Furthermore, although the motivation for selecting these specific failure modes were based upon support for testing and integration of TEAMS models, they also serve as proving grounds for testing the anomaly detection capability of IMS.

As shown in Table 1, a leak in the fuel supply module can be simulated either at the pad or at the VAB. The leak at the pad was simulated to occur between Go for GLS Start (at approximately T-31 sec) and Go for SSME Start (at approximately T-10 sec). The FSM pressure is simulated to drop to an off-nominal value instead of nominally staying above a specified threshold.

Similar to the other simulated failure scenarios, an initial attempt at the construction of the FSM failure simulation involved the simple use of a linearly decreasing ramp, given a predefined rate of degradation from the nominal operating pressure to an off-nominal value. This linear simulation was used to support the ROC analysis performed in a previous study (Martin, 2010). However, it is possible to use a higher fidelity physics-based simulation for this scenario because all of the relevant data is available for its construction. A higher fidelity failure scenario may provide a more realistic test of our algorithm's ability to detect the failure in reality. The method used for the same simulated failure occurring at the VAB spans the period of time during which APU (Auxiliary Power Unit) system checks are conducted. Both low fidelity (linearly decreasing ramps) and high fidelity (physics-based) failure simulations for the FSM leak will be used for analysis of data at both the pad and the VAB to offer a fair basis for comparison in how fidelity affects final performance. This is primarily due to the fact that differences in detection performance between the VAB and pad may be due to differences in operational procedures regardless of simulation fidelity.

The FSM pressure will begin dropping from a nominal value to venting at atmospheric pressure over the course of a few minutes. As the FSM pressure drops, the FSM pressure sensor will redline on a low value. To simulate this failure, we must account for both fluid phases contained in the FSM, the liquid hydrazine and the gaseous nitrogen used to pressurize the spherical tank, such that it is completely voided. The leak in the FSM will be simulated to evolve according to the following assumptions:

1. Assume that the geometry of the FSM is established according to available documentation.
2. Assume that the liquid hydrazine (N_2H_4) is filled only to midpoint of the spherical tank.
3. Assume that the leak is below the surface of the liquid.

In order to simulate the FSM leak according to physics, we will also implicitly use all of the assumptions that result from applying the unsteady form of Bernoulli's equation as presented in (Munson *et al.*, 1998) to solve the differential equation shown as Eqn. 1 associated with the initial leak of the liquid hydrazine. Fig. 1 depicts the leak along with some of the geometrical constants and subscripted reference points used in Eqn. 1.

$$\int_{s_1}^{s_2} \frac{\partial V}{\partial t} ds = \frac{p_g}{\rho} + \frac{1}{2}\left[v_1^2 - \left(\frac{v_2}{C_d}\right)^2\right] + g(h_1 - h_2) \tag{1}$$

$p_g \triangleq p_0 - p_a$ is the gage pressure in the tank, where p_0 is the pressure to which the tank is pressurized with GN_2, and p_a is atmospheric pressure. ρ is are the density of liquid hydrazine, and g represents the gravitational constant. C_d is the coefficient of discharge at the leak point, and s defines the fluid streamline along which Bernoulli's equation is being applied. v_1 and h_1 define the velocity and height from the ground to the top of the liquid hydrazine, respectively. Similarly, v_2 and h_2 define the exit velocity and height from the ground to the site of the leak, respectively.

We assume the sphere has radius r, and the cross-sectional disk representing the top surface of the liquid

[2]The Hydraulic Pumping Unit (HPU) overheat failure is an over-temperature failure simulated within a 25 min period during tests in the VAB.

[3]A hydraulic fluid leak will result in a hydraulic fluid reservoir level drop that is simulated within 1 min prior to launch at the pad and within the 10 minute period after the calibration test in VAB.

[4]The actuator is simulated to be stuck during the actuator positioning test during a 2.5 min test in VAB.

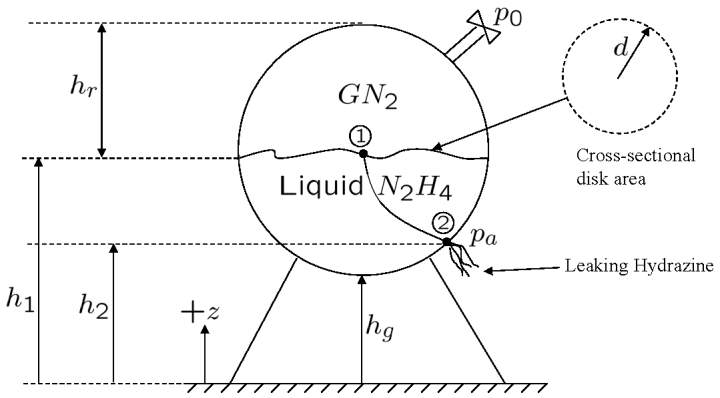

Figure 1: Fuel Supply Module Schematic and Geometry

hydrazine shown in Fig. 1 has a radius of d. Since we are interested in unanticipated decreases in the height of the liquid hydrazine in the tank, let us define $h \triangleq h_1$ as our independent variable to simplify Eqn. 1 for the one-dimensional case, defined with respect to the reference $+z$ shown in Fig. 1. Furthermore, we may apply Eqn. 2 for conservation of mass, and Eqn. 3 defines the velocity v_1 as a function of the height h_1. The ideal gas laws Eqns. 11-12 are defined for constant temperature (de)pressurizaton, and we assume constant acceleration via Eqn. 4. The geometry defined in Fig. 1 and auxiliary Eqns. 5-10 involve h_g, the distance from the ground to the bottom of the tank, and h_r, the distance from the top surface of the liquid hydrazine in the tank to the top of the tank. Thus, the simplified version of Eqn. 1 results in the differential equation shown as Eqn. 13.

$$\dot{m} = \rho A_1 v_1 = \rho A_2 v_2 \quad (2)$$

$$v_1 \triangleq \frac{dh_1}{dt} = \frac{dh}{dt} \quad (3)$$

$$\int_{s_1}^{s_2} \frac{\partial V}{\partial t} ds = \frac{dv_1}{dt}(h_2 - h_1) = \frac{d^2h}{dt^2}(h_2 - h) \quad (4)$$

$$A_1(h) = \pi d(h)^2 \quad (5)$$

$$A_2 = \pi d_0^2 \quad (6)$$

$$V_g(h) = \frac{\pi}{3} h_r^2 (3r - h_r) \quad (7)$$

$$V_0 = \frac{2\pi}{3} r^3 \quad (8)$$

$$d(h) = \sqrt{h_r(2r - h_r)} \quad (9)$$

$$h_r = 2r + h_g - h \quad (10)$$

$$p_0 V_0 = m_g R T \quad (11)$$

$$p(t) = \frac{p_0 V_0}{V_g} \quad (12)$$

$$(h_2 - h)\left(\frac{d^2h}{dt^2} + g\right) = \frac{1}{\rho}\left(\frac{p_0 V_0}{V_g(h)} - p_a\right) + \frac{1}{2}\left(\frac{dh}{dt}\right)^2 \left[\frac{(C_d A_2)^2 - A_1^2(h)}{(C_d A_2)^2}\right] \quad (13)$$

where d_0 represents the radius of the leak area assumed to be a round hole, and R represents the ideal gas constant for GN_2. $p(t)$ and T represent the absolute pressure as the leak evolves as a function of time, and absolute temperature of the GN_2, respectively. A_1 and A_2 represent the surface areas of the N_2H_4/GN_2 fluid interface and the round hole through which liquid hydrazine is leaking, respectively. V_0 and V_g are the initial volume of GN_2 and the volume of GN_2 as the leak evolves, respectively. Finally, $\dot{m}$ represents the mass flow rate of the liquid hydrazine (N_2H_4), and m_g represents the total mass of the GN_2 in the tank.

An approximation to the resulting differential equation can be used to yield a separable nonlinear differential equation that can be solved in closed form, shown as Eqn. 14. This approximation is applied by recognizing that the left hand side of Eqn. 13 (quantifying the gravitational and acceleration terms) is negligible relative to the right hand side. The gravitational term is always negligible, and the acceleration term is important only for quantification of a negligibly small transient at the very beginning of the leak. Furthermore, constants characterizing the FSM geometry can be simplified due to the relative sizes of the leak radius and the radius of the N_2H_4/GN_2 fluid interface (*i.e.* $d_0 \ll d(h)$). The last assumption is that $p_a \ll p(t)$, which may contribute most to the approximation error since the tank pressure evolves over time and will not necessarily always be much greater than atmospheric pressure. Thus the error may potentially grow over time as the tank pressure decreases due to evolution of the leak. However, in general the resulting closed-form representation will help to relieve the computational burden associated with numerical methods otherwise required to solve the differential equation (*i.e.* a stiff solver).

$$\frac{dh}{dt} \approx -\frac{C_d A_2}{A_1(h)} \sqrt{\frac{2 p_0 V_0}{\rho V_g(h)}} \quad (14)$$

Note that the *negative* square root of $\left(\frac{dh}{dt}\right)^2$ must be used in Eqn. 14 in order to yield a real solution.

Furthermore, by recognizing that $\frac{dV_g(h)}{dh} = -A_1(h)$, Eqn. 14 can be simplified to Eqn. 15.

$$V_g^{\frac{1}{2}} \frac{dV_g}{dt} \approx C_d A_2 \sqrt{\frac{2p_0 V_0}{\rho}} \tag{15}$$

Integrating both sides of Eqn. 15 and combining the result with Eqn. 12, we may now write the resulting closed-form expression for the tank pressure as a function of time, $p(t)$, shown as Eqn. 16.

$$p(t) \approx \frac{p_0 V_0}{\left(V_0^{\frac{3}{2}} + \frac{3}{2} C_d A_2 \sqrt{\frac{2p_0 V_0}{\rho}} t\right)^{\frac{2}{3}}} \tag{16}$$

Simulation of the voiding of the remaining gaseous nitrogen (GN_2) in the FSM is performed by use of a linear 1^{st} order approximation of a differential equation governing the release of an ideal gas as used in (Tchouvelev *et al.*, 2007). The solution of the differential equation is shown as Eqn. 17. The mass of the GN_2 was obtained by use of the design condition (1.1 lbs of gaseous nitrogen at 400 psig as a baseline), obtained from the seminal paper on the introduction of the FSM (McCool *et al.*, 1980). It was also assumed that the GN_2 underwent a constant temperature and constant volume ideal depressurization (bleeding off tank pressure by operating a GN_2 pressurization valve) from this design condition to the nominal value that existed at the time the leak was simulated when in the VAB. The constant temperature assumption also holds for evolution of the leak from the nominal pressure value to $p(t) = p_a$.

$$p(t) \approx p_v e^{-\left(\frac{C_d A_2}{V_v}\right) t \sqrt{\gamma \left(\frac{2}{\gamma+1}\right)^{\frac{\gamma+1}{\gamma-1}} RT}} \tag{17}$$

Of course, when $h = h_2$, the liquid hydrazine will have emptied out to the point that it can no longer escape from the hole, and only the gaseous nitrogen is left to escape. We call the pressure at which this occurs the *vent pressure*, p_v, which can easily be computed using Eqns. 7, 10, and 12. The time of this event can be approximated by using Eqn. 16. The corresponding volume of gas left to be evacuated from the tank is V_v, and γ is the ratio of specific heats for GN_2. Therefore, Eqn. 16 governs the release of liquid hydrazine until the time of the vent pressure. At this point, Eqn. 17 governs the subsequent release of gaseous nitrogen and complete voiding of the tank at which point $p(t) = p_a$.

3 COMPARATIVE ANALYSIS

In this section we aim to investigate and observe the effect of increased simulation fidelity on detection performance. In doing so we hope to gain a better understanding for and develop an appreciation of possible improved ability of IMS to detect simulated failures that may be more realistic. In the previous section, we have provided the details for how a high fidelity, physics-based simulation of a fuel supply module leak is to evolve, according to Eqns. 16 and 17. Using these equations, the time at which the pressure in the FSM approximately reaches atmospheric pressure associated with the high fidelity simulation can be used to construct the slope of the line associated with the low fidelity simulation, for a fixed leak radius. Thus, implicit linearized versions of Eqns. 16 and 17 represent low fidelity simulations. The slope of the resulting line will determine the rate of degradation, to be used as a fair basis for comparison to the nonlinear rate of degradation which evolves according to physics. Fig. 2 illustrates the times to be used to construct the slopes of low fidelity linear simulations, based upon various leak radii that were simulated with the high fidelity physics based simulations.

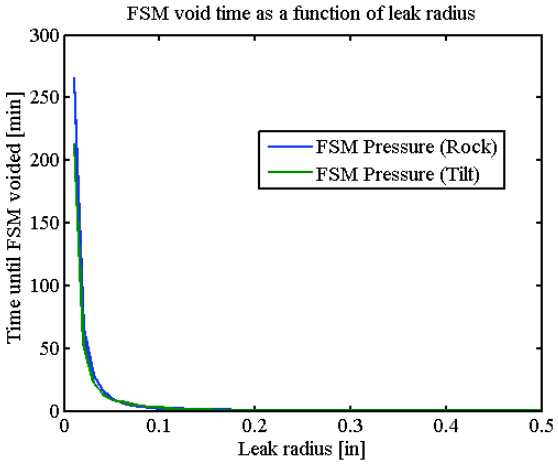

Figure 2: Time to FSM Voiding for Various Leak Radii

The detection performance can be quantified by the Area under the ROC (Receiver Operating Characteristic) curve (AUC). The ROC curve is a plot of the true positive rate against the false positive rate, and can be used to help make the tradeoff between these two rates. The curve is constructed by treating time points as representative samples, all of which are implicitly used to compute the true and false positive rates. The AUC is loosely a measure of accuracy over all possible tradeoffs between the true positive rate and the false positive rate, computed by numerically integrating the area under the ROC curve. More formally, the AUC represents the probability that a randomly chosen failure data point is more suspect than a randomly chosen nominal data point (Rosset, 2004). An AUC of one thus indicates perfect ranking of these two randomly selected data points.

As such, Fig. 3 demonstrates how detection performance varies across a range of leak radii for both the high and low-fidelity simulations of FSM leaks, using Shuttle data at the pad as the sole exemplar. Detection performance using Shuttle data from the VAB is poorer than that at the pad due to reasons to be described in Sec. 4. These reasons are also specific to the FSM leak failure mode 1a, but otherwise performance using the VAB Shuttle data exhibit the same tendencies as performance based upon using data from the pad.

Two main observations can be made regarding Fig. 3. First, it is evident that robust detection performance improves as the leak radius increases, as quantified by the AUC, regardless of the simulation fidelity. This meets with intuition, since a faster leak should be more

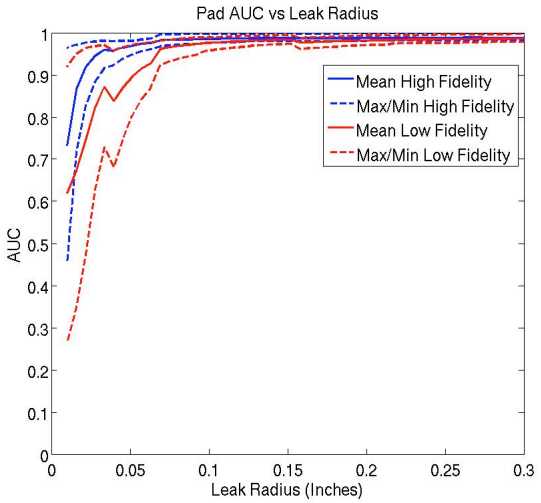

Figure 3: High vs. Low Fidelity Simulation Detection Performance

easily and quickly detectable. The second observation relates to the fact that the detection performance of both the high and low fidelity simulations converge as the leak radius increases. This also meets with intuition, since a faster leak can be more easily approximated in a linear fashion. However, as we can tell from the error bars, there is quite a bit of overlap between the high and low fidelity simulation methods for fast leaks, and even for slow leaks. Thus, there is no appreciable difference between the detection performance results for the low and high fidelity simulations, and as such we will not make the distinction between the two for the remainder of the paper.

4 ANOMALY DETECTION

As mentioned previously, IMS works under the principle of semi-supervised anomaly detection by building a model of the nominal historical data on which it is trained. Because IMS only models the nominal data, and does not model any failure modes, it can potentially detect unknown failure modes. The model takes the form of a knowledge base (KB) of clusters. Once the KB has been learned, unseen data points are evaluated against the KB and assigned anomaly scores based on how anomalous the data points are with respect to the training data. If a new point falls within an existing cluster, then it is assigned an IMS score of zero. If it does not fall within an existing cluster, then the distance to the nearest cluster is used as the IMS score. When an anomalous period of the testing data is localized, the contributing IMS scores can be identified, helping to diagnose the issue. Prior to the Ares I-X launch, we trained IMS on historical Space Shuttle data, and tested it using historical Shuttle data into which we had inserted simulated failures. During the Ares I-X pre-launch period, IMS processed live Ares I-X data, using a knowledge base that was the result of training IMS on historical Shuttle data. The remainder of this section describes the selection of measurements for use with IMS, the training and testing procedures used, and the results obtained both on Shuttle data and on Ares I-X data. The section concludes with a summary of the results.

4.1 Parameter Selection

For Ares I-X data to be compatible with historical Shuttle data a common set of parameters needed to be chosen. During the Shuttle analysis on chosen simulated faults, all continuous-valued parameters were selected along with one discrete parameter that was known to be critical in detecting one of the three failure simulations, for a total of 137 parameters. The choice of mostly using continuous parameters was made because historically IMS has performed better when operating on mostly continuous sensor values. After running an analysis on the failure simulations, some false alarms were detected and an additional set of parameters were eliminated, leaving 102. For the purposes of feature selection (parameter elimination), a false alarm is defined qualitatively as a large excursion above an apparent “baseline” in the composite score produced by IMS, which characterizes the anomalousness of a specific point in the time series. With the elimination of these parameters the false alarms were significantly reduced. When the first set of Ares I-X VAB data was recorded a common subset was selected between the Ares I-X parameter set and the 102 parameters from the Shuttle resulting in the 33 parameters used for analysis on the Ares I-X data.

4.2 Training and Testing Procedures

For the purpose of training and testing IMS, we used historical Space Shuttle data into which we inserted simulated failures, with varying rates of degradation, and spanning fixed time periods in a random fashion. Although the main purpose of using IMS in the Ground Diagnostics Prototype is to detect unknown failures, we tested it by using simulations of known failures. (For obvious reasons, we were unable to simulate unknown failures.) IMS has a number of tunable input parameters, however one key parameter that was very important to tune was the maximum interpretation (max interp) parameter. This parameter governs the threshold in the learning phase that determines if a new data point should be placed in the current cluster or used to generate a new cluster. The parameter directly influences the number of clusters created in the learning phase and therefore has a major influence in the final anomaly score calculated by IMS. As the max interp value increases the total number of clusters formed becomes smaller.

To determine the optimal max interp value and corresponding number of clusters a set of cross validation runs was performed on a set of Shuttle VAB and pad data, using the AUC as the governing metric for optimization. Cross validation is a technique for estimating the accuracy of a machine learning algorithm, by training and testing the algorithm multiple times, each time using different subsets of the available data for training and testing, and then averaging the results.

4.3 Results on Shuttle Simulations

Once the cross validation runs were complete, the areas under the ROC curves were calculated using data that spans the time that the shuttle was still in the VAB.

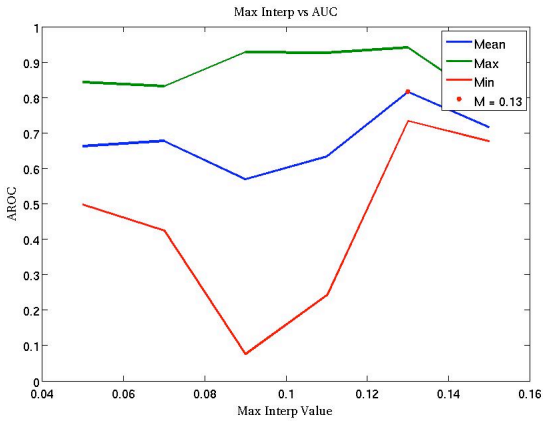

Figure 4: AUC as a function of IMS Parameter Max Interp for Shuttle data from the VAB

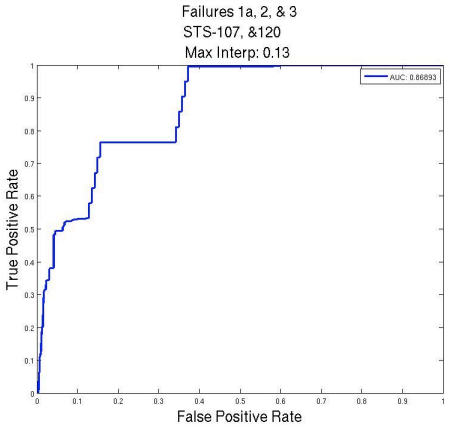

Figure 5: ROC Curve for Optimal Max Interp for Shuttle data from the VAB

Figure 4 shows the maximum, minimum, and average AUC over the three-fold cross validations and three fault scenarios (listed as failure modes 1a, 2, and 3 in Table 1) for each max interp value. The optimal max interp value that was chosen is marked in the plots. The mean AUC with the highest value is 0.86893, and corresponds to the optimal max interp value of 0.13, which can be seen in Figure 4. The ROC curve associated with this optimized max interp value can be seen in Figure 5. The relatively modest detection performance at the VAB can be attributed to the fact that IMS had difficulty detecting simulated failure 1a. This difficulty stemmed from the fact that the increase in IMS score resulting from this simulated failure was not much larger than the nominal variation in the IMS score, so it was not possible to select a threshold that would allow IMS to detect all of the simulated failures without increasing the number of false alarms. Thus, some failure modes are easily detected using IMS' distance-based approach with clustering, while others are not. When IMS is used in parallel with TEAMS-RT, TEAMS-RT should detect all of the failures that

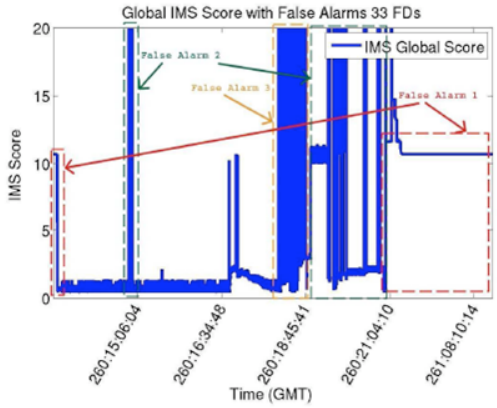

Figure 6: Ares I-X VAB Global IMS Score with False Alarms

are modeled in the TEAMS model; the advantage of using IMS in addition is that it has the potential to detect failures that were not modeled, as well as anomalies that are not yet failures. For the pad, the AUC is 0.99919, indicating that IMS does an excellent job of detecting the two simulated failure modes 1a and 3 at the pad, and performs much better than at the VAB. An intuitive explanation for this discrepancy relates to the fact that at the pad only a small portion of the data has high "activity," during the last minute before launch. However, data from quiescent periods previous to the last minute before launch are also used for analysis. As such, this translates to a lower signal to noise ratio, which directly influences the AUC, resulting in a higher value and thus fewer false positives.

4.4 Results on Ares I-X

Once the optimal max interp parameter was determined from the Shuttle data, IMS was trained on 33 measurements using Shuttle data from seven flights, which also represents the greatest common subset corresponding to equivalent Ares I-X measurements. After building the knowledge base, the Ares I-X data was evaluated against it, and ostensibly acts as hold out test data from a machine learning standpoint. The resulting IMS scores for the VAB are shown in Figure 6. With the initial set of 33 measurements, 3 periods of anomalous behavior were flagged by IMS; they are labeled as three "False Alarms" in Figure 6. We performed an analysis of each "false alarm"; here we present the analysis of False Alarm 1 as an example. We determined that False Alarm 1 was primarily caused by two measurements. The contributing IMS scores for these two measurements are plotted in Figure 7.

False Alarm 1 was caused by a difference between the Space Shuttle and Ares I-X data. In recent years, the TVC actuator tests performed in the VAB have all been "pinned" tests, meaning that the actuator is physically pinned to the nozzle during testing, so that the nozzle moves during the test. The first TVC actuator position test performed in the VAB for Ares I-X was an "unpinned" test, meaning that the actuator was detached from the nozzle, and the nozzle did not move during the test. Because the actuator was unpinned,

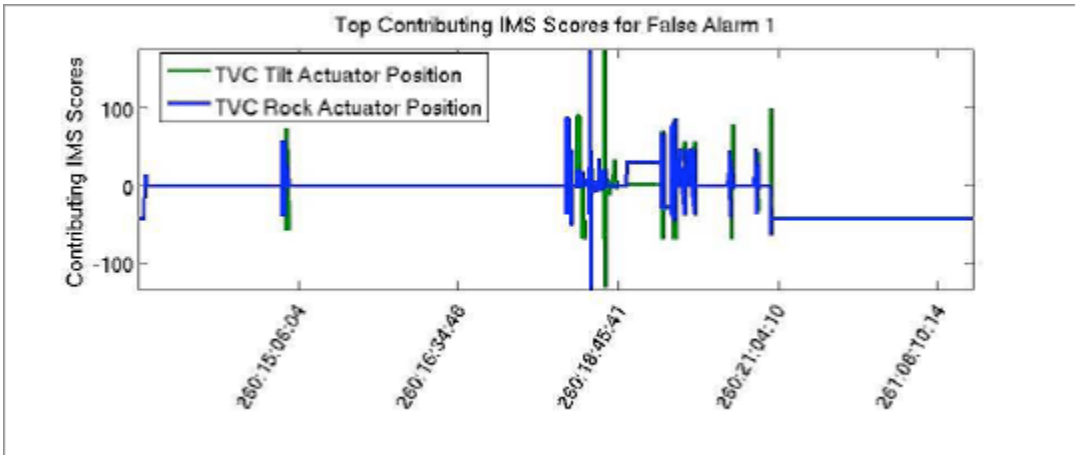

Figure 7: VAB Top Contributing IMS Scores For False Alarm 1

it was able to move through a larger range of motion that is not possible during pinned testing. IMS therefore saw rock and tilt position values that it had never seen in the Shuttle data, which it flagged as anomalies. These anomalies are "false alarms" in the sense that they are not failures, but they do illustrate the ability of IMS to detect new data that is different from what it has seen before. We performed a similar analysis for the pad, where there were fewer anomalies identified by IMS. Like the anomalies detected at the VAB, the anomalies detected at the launch pad were caused by operational differences between Shuttle and Ares I-X.

4.5 Summary of IMS Deployment Results

The experiments that we ran before the Ares I-X launch using historical Space Shuttle data with simulated failures demonstrated that IMS is able to detect most of the simulated failures, but not all of them. In particular, it had difficulty detecting the simulated failure mode 1a in the VAB due to its relatively small contribution to the overall IMS anomaly score compared to the other two simulated failure modes. IMS is not trained to detect specific failure modes; it detects data that is anomalous according to its cluster-based model. We expect that many known and unknown failure modes will be detected as anomalies by IMS, but it is not guaranteed to detect all possible failure modes. The advantage of using IMS together with a model-based diagnosis system such as TEAMS is that it adds the potential to detect unknown failure modes and to detect precursors of failures.

The results of running IMS on Ares I-X data, using a knowledge base that was trained on historical Space Shuttle data confirm our hypothesis that the Ares I-X TVC data is reasonably similar to the Space Shuttle SRB TVC data. Most of the time IMS produced small anomaly scores when run on the Ares I-X data. IMS did detect some "anomalies" in the Ares I-X data. These anomalies were "false alarms" in the sense that they were not failures but rather caused by operations performed differently for Ares I-X versus Shuttle; hence, they illustrate the ability of IMS to detect new data that is different from what has been seen in the past.

5 IMS/TEAMS PERFORMANCE COMPARISON

5.1 Anomalies Detected by IMS That Were Not Detected by TEAMS

We have seen that IMS detected some interesting anomalies that were not detected by TEAMS because they were not failures as defined in the FMEA (Failure Modes and Effects Analysis) and the other documents on which the TEAMS models were based. One such anomaly was the pinned/unpinned actuator anomaly mentioned previously. In the pinned/unpinned anomaly there were procedural differences between the TVC test for the Shuttle and Ares I-X, resulting in IMS signaling an anomaly in the TVC rock and tilt actuator positions. . This anomaly was not a failure; hence it was detected by IMS but not by TEAMS. Furthermore, it was found that there are other differences between Shuttle and Ares I-X actuator tests due to the sequence being changed slightly along with a greater max displacement. Ostensibly, this had an even greater effect than the pinned/unpinned variation alone for IMS.

5.2 Failures Detected Earlier by IMS Than by TEAMS

Table 2 summarizes the detection times for the simulated failures that were detected both by IMS and TEAMS in minutes after injection of the failure. A hypothesized advantage of IMS is that it may detect certain failures before TEAMS. However, on average the results show that TEAMS detected failures prior to the time that IMS did. On two occasions, IMS was able to detect a simulated failure prior to TEAMS, as shown in red in Table 2. In the case of failure 3, which was simulated with a simple bit flip at the VAB, the detection occurred at approximately the same time. The other two failures are more complicated, and are described by gradual ramps of continuous-valued parameters rather than instantaneous bit flips of discrete-valued parameters, owing to the notable differences in detection times. It can be seen from Table 2 that IMS sometimes detected failures earlier than TEAMS did, but more often it detected them later. There may be some advan-

Table 2: Summary of simulated failure detection times

Failure	Flight	Trial	IMS Detection Time	TEAMS Detection Time	Difference
1a	STS-107	1	8.77	2.24	6.53
		2	19.51	5.04	14.47
		3	13.48	1.39	12.09
	STS-112	1	217.37	2.2	215.14
		2	3.27	5.04	-1.77
		3	222.07	1.38	220.7
	STS-120	1	0.89	2.23	-1.34
		2	8.69	5.02	3.67
		3	3.45	1.38	2.07
2	STS-112	1	1.76	1.5	0.26
		2	4.01	2.33	1.68
		3	3.78	2.4	1.38
	STS-120	1	4.92	2.57	2.35
		2	3.62	2.32	1.3
		3	3.89	2.39	1.5
3	STS-112	1	0	0	0
		2	0	0	0
		3	0	0	0
	STS-120	1	0	0	0
		2	0	0	0
		3	0	0	0

tage to running IMS in parallel with TEAMS in order to provide earlier detection of some failures. Another observation worth noting is that there appears to be a wider variance for the IMS detection latencies for a given failure simulation spanning several flights. This lends credence to the fact that TEAMS detection times are based purely upon logic rather than statistics, the latter of which IMS incorporates in its detection capability.

5.3 Failures Detected By TEAMS That Were Not Detected By IMS

IMS occasionally misses simulated failures, usually as a function of the fine tuning required to mitigate specific instances of false alarms on test (Ares I-X) data. This fine tuning involves varying the number of clusters in the knowledgebase, the measurements (sensor values) represented in the knowledgebase, as well as the threshold or qualitative heuristic used following the application of ROC analysis. Typically, ROC curves span multiple failures, but are based only on a limited few Shuttle flights for training data. As such, when applying the resulting knowledge-base to unseen hold out test (*e.g.* Ares I-X) data, simulated failures may not be detected. In fact, great measures may need to be taken in order for such failures to be detected, often at the expense of false alarms, as is apparent in the examples of false alarms presented previously.

5.4 Failures More Appropriate For Modeling With TEAMS

Anomaly detection methods such as IMS are not well suited for detecting some types of failures. As mentioned previously, we used simulations of known failure modes to test IMS. For some of these simulated failures, we expended a lot of effort in tuning IMS to get IMS to detect the simulated failures. This tuning process included reducing the set of measurements that were used to train IMS. For failure mode 4, a simulated failure covering a stuck actuator during a simulated positioning test at the VAB, almost all measurements other than the one required to simulate the failure had to be excluded in order to provide adequate detection capability. For this same case, a linear regression was required in order to facilitate the construction of commanded position computed by proxy of a commanded current measurement due to the absence of the requisite electromechanical conversion data. The difference between the quasi-commanded position and the actual measured position was then used as the sole parameter with which to train and test IMS. Any additional measurements included in the knowledgebase resulted in a missed detection. This is a case in which IMS was clearly not a good choice for detecting the particular failure mode.

Cases such as these serve as evidence that each tool should be leveraged to promote its strengths rather than re-adapting the tool to solve a problem that is outside of its domain of relevance. With IMS, we know that its strengths lie in a great potential to detect faults that are unknown or that otherwise have not been modeled and to detect anomalies that are precursors of faults before a model-based system detects the fault. We believe that it would be better to rely on TEAMS to detect the known failure mode described above, rather than tuning IMS to detect it. Reducing the set of measurements that are used to train IMS did allow IMS to successfully detect the simulated failures, but it reduced IMS' potential to detect other unknown failures.

6 SUMMARY AND CONCLUSIONS

As mentioned previously, we believe including a semi-supervised data-driven anomaly detection algorithm such as IMS alongside a model-based diagnosis system such as TEAMS in a diagnostic system adds sig-

nificant value, when used appropriately. Doing so will allow the overall anomaly detection system to be endowed with the potential to detect anomalies that cannot be detected by the model-based diagnosis system in isolation, either because they are unknown failures and therefore unmodeled, or because they are not failures. Furthermore, IMS may detect known failures in advance of the time that TEAMS would detect them, and in general IMS requires less modeling effort than TEAMS (although it does require a sufficient quantity of historical and/or simulated training data).

It was also important to consider the ability of IMS to detect failures in a true operational setting, but there was a dearth of true failures resembling those that we simulated with which to conduct experiments. Therefore, we hoped to demonstrate an improved ability of IMS to detect simulated failures that may be more realistic by increasing their fidelity. We have found that for fast FSM leaks, robust detection performance improves for both the high and low fidelity simulations, and the performance for both types of simulations also converges as the leak rate increases. However, overall we have also observed that there is no appreciable difference between the effect of using a low or high fidelity simulation on detection performance.

ACKNOWLEDGMENTS

We would like to acknowledge funding by the NASA Constellation Program, the NASA Exploration Technology Development Program, and by the NASA KSC Ground Operations Project. We would also like to thank all of the current and past members of the Ares I-X GDP team, including Charles Lee, John Ossenfort, Rebecca Oostdyk, Vijay Baskaran, Robert Waterman, Barbara Brown, and John Wallerius. Furthermore, we graciously acknowledge the reviews of Dawn McIntosh and Dr. Ann Patterson-Hine.

REFERENCES

(Cavanaugh, 2001) Kevin Cavanaugh. An integrated diagnostics virtual test bench for life cycle support. In *Proceedings of IEEE Aerospace Conference*, volume 7, pages 7–3235–7–3246, Big Sky, Montana, March 2001.

(Chandola *et al.*, 2009) Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

(Iverson *et al.*, 2009) David L. Iverson, Rodney Martin, Mark Schwabacher, Lilly Spirkovska, William Taylor, Ryan Mackey, and J. Patrick Castle. General purpose data-driven system monitoring for space operations. In *Proceedings of the AIAA Infotech@Aerospace Conference*, Seattle, Washington, April 2009.

(Iverson, 2004) David L. Iverson. Inductive system health monitoring. In *Proceedings of The 2004 International Conference on Artificial Intelligence (IC-AI04)*, Las Vegas, Nevada, June 2004. CSREA Press.

(James and Atkinson, 1990) Mark L. James and David J. Atkinson. Software for development of expert systems. *NASA Technology Briefs*, 14(6), 1990.

(Mackey *et al.*, 2001) Ryan Mackey, Mark James, Han Park, and Michail Zak. BEAM: Technology for autonomous self-analysis. In *Proceedings of the IEEE Aerospace Conference*, Big Sky, MT, 2001.

(Martin *et al.*, 2010) Rodney Martin, Mark Schwabacher, and Bryan Matthews. Investigation of data-driven anomaly detection performance for simulated thrust vector control system failures. In *Proceedings of the 57th Joint Army-Navy-NASA-Air Force Propulsion Meeting*, Colorado Springs, CO, May 2010.

(Martin, 2010) Rodney A. Martin. *Aerospace Technologies Advancements*, chapter Evaluation of Anomaly Detection Capability for Ground-Based Pre-Launch Shuttle Operations, pages 141–164. IN-TECH, January 2010.

(McCool *et al.*, 1980) Alexander A. McCool, Adas J. Verble Jr., and Jack H. Potter. Space transportation system solid rocket booster thrust vector control system. *AIAA Journal of Spacecraft*, 17(5):407 – 413, Sept. - Oct. 1980.

(Munson *et al.*, 1998) Bruce R. Munson, Donald F. Young, and Theodore H. Okiishi. *Fundamentals of fluid mechanics*. Wiley, 1998.

(Park *et al.*, 2002) Han Park, Ryan Mackey, Mark James, Michail Zak, Michael Zynard, John Sebghati, and William Greene. Analysis of Space Shuttle Main Engine Data Using Beacon-based Exception Analysis for Multi-Missions. In *Proceedings of the IEEE Aerospace Conference*, Big Sky, MT, March 2002.

(Pisanich *et al.*, 2006) Greg Pisanich, Ryan Mackey, David Iverson, and Dave Lawrence. Integrated system health management (ISHM) technology demonstration project final report. Technical report, NASA Ames Research Center, Dryden Flight Research Center, Jet Propulsion Laboratory, California Institute of Technology, January 2006.

(Rao *et al.*, 2009) Chinmay Rao, Asok Ray, Soumik Sarkar, and Murat Yasar. Review and comparative evaluation of symbolic dynamic filtering for detection of anomaly patterns. *Signal, Image and Video Processing*, 3(2):101–114, June 2009.

(Rosset, 2004) Saharon Rosset. Model selection via the AUC. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML'04)*, Banff, Alberta, Canada, July 2004.

(Schwabacher and Waterman, 2008) Mark Schwabacher and Robert Waterman. Pre-launch diagnostics for launch vehicles. In *Proceedings of the IEEE Aerospace Conference*, Big Sky, MT, March 2008.

(Schwabacher *et al.*, 2010a) Mark Schwabacher, Rodney Martin, Robert Waterman, Rebecca Oostdyk, John Ossenfort, and Bryan Matthews. Ares I-X ground diagnostic prototype. In *Proceedings of the 57th Joint Army-Navy-NASA-Air Force Propulsion Meeting*, Colorado Springs, CO, May 2010.

(Schwabacher *et al.*, 2010b) Mark Schwabacher, Rodney Martin, Robert Waterman, Rebecca Oostdyk, John Ossenfort, and Bryan Matthews. Ares I-X ground diagnostic prototype. In *AIAA Infotech@Aerospace Conference*. AIAA, April 2010.

(Tchouvelev *et al.*, 2007) Andrei V. Tchouvelev, Zhong Cheng, Vladimir M. Agranat, and Serguei V. Zhubrin. Effectiveness of small barriers as means to reduce clearance distances. *International Journal of Hydrogen Energy*, 32:1409 – 1415, 2007.

Rodney A. Martin was born in Poughkeepsie, NY, in 1970. He completed a B.S. in Mechanical Engineering at Carnegie-Mellon University, Pittsburgh, PA in 1992. Following 5 years of active duty as an officer in the U.S. Navy, he returned to graduate school, earning an M.S. degree in Mechanical Engineering from UC Berkeley in 2000, and a Ph.D. in Mechanical Engineering from UC Berkeley in 2004. He has worked at NASA Ames Research Center since completing his Ph.D. in 2004. His research interests span the areas of fault detection, estimation, prediction, control theory, and machine learning. Most recently he has worked on applying techniques from these research areas to projects at NASA that address aviation safety and space propulsion systems. He was a member of the team that won the 2007 JANNAF Outstanding Achievement in Liquid Propulsion Award in the Operational Systems category for work in the area of Data Driven Algorithms for SSME Fault Detection, bestowed by the JANNAF Liquid Propulsion Subcommittee. He is currently a member of IEEE and AIAA.

Mark A. Schwabacher was born in New York City in 1968. He completed a B.A. with a triple major in Computer Science, Mathematical Sciences, and Mathematical Economic Analysis at Rice University in Houston, TX, in 1990. He completed a Ph.D. in Computer Science at Rutgers University in New Brunswick, NJ in 1996. His thesis work applied artificial intelligence to engineering design. After that, he was an NRC Postdoctoral Research Associate at the National Institute of Standards and Technology in Gaithersburg, MD for two years. He has worked at NASA Ames Research Center since 1998, where he has worked on several systems health management activities. He served as the Software Lead of the NASA X-37 Integrated Vehicle Health Management (IVHM) Experiment, and led the development of the Ares I-X Ground Diagnostic Prototype. He is currently working on IVHM for commercial aviation. He has also applied anomaly detection algorithms to Earth science and to aviation security. He was a member of the team that won the Outstanding Achievement in Liquid Propulsion Award in the Operational Systems category at the JANNAF 3rd LPS Meeting in 2007. He was a co-author of the paper that won the Graduate of the Last Decade Best Paper Award at the International Conference on Prognostics and Health Management in 2008. He is a member of AIAA.

Bryan L. Matthews was born in Monterey CA in 1979. He completed his B.S. in Electrical Engineering at Santa Clara University, Santa Clara, CA, in 2002. Following an internship at NASA Ames Research Center in 2002 he worked in the field of human biometric sensing, which included brain computer interface and cognitive fatigue studies. More recently he has also been involved in various Integrated Vehicle Health Management projects (IVHM) including anomaly detection in commercial airline safety data as well as work on the Ares I-X Ground Diagnostics Prototype (GDP). He was presented ith the 2009 NASA ARC contractor award for work on Ares 1-X GDP as well as being a team member of the NASA software of the year award runner up for 2008.