

ODEs Modeling Chaotic Convection: Data vs. Physics

Ashley McKhann

April 25, 2011

Contents

Acknowledgements	3
Abstract	4
1 Introduction	5
2 Background	5
3 Experimental Analysis	10
4 Numerical Analysis	14
4.1 Eureka	14
4.2 Computational Fluid Dynamics (CFD) Experiment	15
4.3 Analysis and Results	15
5 Conclusions	23

Acknowledgements

I would first and foremost like to thank Chris Danforth for all his hard work and help in this process. Of course, I would also like to thank my parents for all the opportunities they have given me and for being there for me whenever I need them. Many thanks to Floyd Vilmont and Darren Hitt for their help with building the thermosyphon and making it functional. Thanks to Kameron Harris: without all the work he had done, none of this would have been possible. Thanks to UVM Undergraduate Research and Vermont Space Grant Consortium for the URECA grant. I would also like to acknowledge Peter Dodds and Kelvin Chu for being members of my thesis committee.

Abstract

This research investigates the difficulties associated with climate and weather prediction by forecasting the future state of a toy climate, analogous to the Lorenz model of natural convection. Temperature measurements, collected from a computational fluid dynamics (CFD) simulated thermal convection loop, are fit with a set of three differential equations using Eureka, a recently developed software tool for inferring natural laws from freeform data. Several forecasts, made using the optimal differential equations, are compared against a verifying set of temperature data from the CFD system to evaluate the accuracy of the forecast model. The results are interpreted with respect to a physical model derived from first principles. The goal is to discover reasonably accurate, low-dimensional prediction equations which model the physical laws governing this system.

1 Introduction

The ability to forecast the future state of chaotic physical systems such as Earth’s climate and weather has become increasingly vital [5]. Natural systems such as these are nonlinear, making them difficult to predict for reasons we will discuss. Attempts can be made to improve the forecasting models for these systems by observing simplified versions with reduced structure and fewer parameters, but which exhibit qualitatively similar dynamics. We utilize a toy climate in the form of a thermal convection loop, or thermosyphon, to produce a prediction model. By improving the prediction capabilities for our toy climate, we are working toward the long term objective of making improvements in climate forecasting.

Our goal for this research is two-fold: (1) construct a thermal convection loop physical experiment and (2) find low-dimensional equations which forecast future states of the system. In addressing this second objective, we compare a model derived from first principles to a model generated by a genetic algorithm. Eureka is a recently developed software program which determines governing equations based on a series of data points, in our case temperature data. We intend to input data collected from a thermosyphon to generate governing equations for the system, and then test the ability of these equations to predict future states of the system.

The thermosyphon is a toroidal loop which is filled with fluid and oriented vertically. It is encircled by two baths, one enclosing the top half and one surrounding the bottom half, both of which are filled with fluids of fixed temperature. The lower bath is heated, while the upper bath is cooled, creating a temperature difference between the top and bottom of the loop. For a moderate temperature difference, the fluid will rotate either clockwise or counterclockwise. As the temperature difference between the top and bottom is increased, a bifurcation leads the fluid to switch the direction in which it is flowing nonperiodically.

The aperiodicity in the system gives rise to the difficulties associated with forecasting future states of the thermosyphon. This aperiodicity causes the system to be chaotic and therefore experience sensitive dependence on initial conditions. In other words, the trajectories of two initial states which differ by a small amount will diverge exponentially in time on average. Sensitive dependence on initial conditions causes forecasting to be very difficult, since any prediction will diverge rapidly from the truth, no matter how accurate the initial measurement. It is this difficulty we seek to mediate in creating our prediction model.

2 Background

In the early 1960s, the National Weather Service was using linear differential equations to predict the weather. They knew their model was not optimal, but they did not discover how inaccurate the linear assumption was until Edward Lorenz presented his work. Lorenz,

a mathematician and meteorologist, decided to test a nonlinear model of convection in his 1963 paper “Deterministic Nonperiodic Flow” [10]. In this paper, Lorenz looked at a system of three ordinary nonlinear differential equations, now called the Lorenz equations. The equations attempt to describe the fluid flow in a Rayleigh-Bénard cell, where convection cells form in a fluid which is heated from below and cooled above. This system of deterministic equations represents an idealization of a forced dissipative hydrodynamical system (e.g. Earth’s weather). The equations are:

$$\frac{dx}{dt} = \sigma(y - x) \tag{1a}$$

$$\frac{dy}{dt} = \rho x - xz - y \tag{1b}$$

$$\frac{dz}{dt} = xy - \beta z \tag{1c}$$

Where x is proportional to flow velocity, y is proportional to the temperature difference across the convection cell, z is proportional to the deviation of the vertical profile of convecting temperature from that of conduction, and σ (the Prandtl number), ρ (the Rayleigh number), and β are parameters related to the fluid flow. The Rayleigh number is proportional to the temperature difference between the heat source and the heat sink and is typically the parameter adjusted to observe flow instability. The Prandtl number is the ratio of momentum diffusivity to thermal diffusivity, and β is proportional to tube geometry.

By numerically solving these equations, Lorenz found that the nonperiodic solutions for the system of equations - which make up almost all of the solutions - must be unstable. Since Earth’s atmosphere is nonperiodic, Lorenz concluded that long-range weather prediction will be impossible, since the instability causes sensitive dependence on initial conditions. In fact, even with sensors hanging every square foot throughout the atmosphere, feeding perfect observations into an infinite computer with exact knowledge of the atmosphere’s governing equations, the limit of predictability is only two weeks - the time required for the unresolved behavior to become relevant [11].

In an attempt to simplify the Lorenz system, later researchers used a thermosyphon to study the chaotic behavior of a convection cell [4, 6, 7]. The smaller radius of the loop, approximately 3 cm, to an outer radius of 50 cm, allows only a single major convection cell to form ¹. There are three equilibrium states, or regimes, which the thermosyphon experiences: conduction, clockwise (CW) rotation, or counter-clockwise (CCW) rotation. Thermocouples, the temperature measuring devices on the thermosyphon, placed at the 3 o’clock and 9 o’clock positions (see Figure 1) measure the temperature T inside the loop

¹Though the thermosyphon is used for the simplification provided by creating a single convection cell, it has been shown that minor disturbances can occur inside the loop where small pockets of the fluid do not flow in the predominant direction [12].

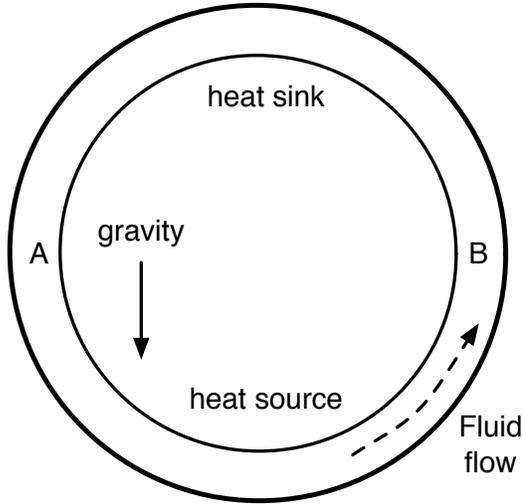


Figure 1: The heating bath surrounds the bottom half, heated to temperature T_{bottom} . The cooling bath surrounds the top half, chilled to temperature T_{top} . The direction of the fluid flow can be determined by measuring the temperature difference between points A and B. If $T_A - T_B$ is positive (negative) the flow is CW (CCW).

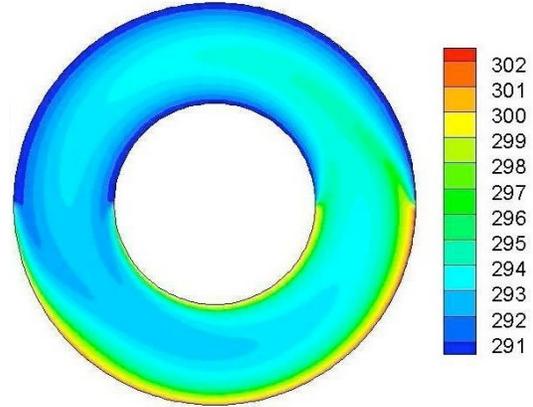


Figure 2: Snapshot of a 2-D CFD simulation simulating the flow inside the thermosyphon where the aspect ratio has been exaggerated into a donut for the purpose of illustration. Note that the fluid is rotating CCW since the heated fluid (light blue and green) is flowing past point B while the chilled fluid (dark blue) is flowing past point A. The colorbar units are in Kelvin. Picture from Harris et al. (2011) [8].

and can be used to identify which state the fluid is in. When T_A is greater than T_B the fluid inside the loop is rotating clockwise - the heated fluid from the bottom half of the loop is flowing past the point A, while the cooled fluid from the top is flowing past point B. Therefore, if T_A is less than T_B it is rotating counter-clockwise, and if T_A equals T_B then the fluid is not moving (see Figure 2).

Creveling et al. (1975) explored the different states produced by varying the temperature differences between the top and bottom halves of the thermosyphon using water for the fluid inside the loop. For small temperature differences between the top and bottom, ΔT , the fluid does not rotate and is in a stable conducting state. Once ΔT reaches a certain critical point, the fluid will begin to rotate either clockwise or counter-clockwise and will continue to rotate in that one direction stably as long as ΔT remains fixed. In this case, $T_A - T_B$ will not vary and the stable fixed state is now rotation in either the CW or CCW direction. As ΔT is increased further, the CW (CCW) rotation becomes unstable, seen as oscillations in the temperature reading of $T_A - T_B$. Eventually a second critical temperature difference is reached and the rotation becomes unstable. In this state, the fluid will repeatedly switch the direction in which it is flowing. These flow reversals occur nonperiodically, displaying chaotic behavior [4].

A flow reversal begins when a ‘pocket’ of fluid becomes hotter than the fluid around it. As the hot pocket passes point B (for fluid rotating CCW), it exerts a buoyancy force on the fluid, causing a positive acceleration in the counterclockwise direction and speeding up the rotation. As the pocket travels around the top of the thermosyphon, it does not cool as much and arrives at point A hotter than stably rotating fluid. Consequently, there is a buoyancy force in the reverse (clockwise) direction acting to slow down the speed of rotation. This deceleration allows the hot pocket to spend more time in the heated section causing it to become hotter. The pocket now arrives at point B hotter than it was on the previous rotation. Therefore, it has greater acceleration going through the top half and greater deceleration as it goes past point A. As the instability grows, this amplification process continues until the buoyancy force generated by the pocket at point A grows large enough, causing the flow to stop. With no rotation in the tube, the temperature difference between the top and bottom portions of the fluid grows undisturbed. The fluid will then ‘choose’ a direction of rotation, and if it begins to rotate clockwise, we call the rotation change a flow reversal, or regime change.

Gorman et al. (1986) categorized chaotic flow into three different regimes which depend on the parameters chosen for the system: transient chaos, subcritical chaos, and globally attracting chaos. Transient chaos occurs when the flow inside the loop at first displays chaotic behavior in the form of increasing amplitudes and flow reversals, but this behavior eventually decays and is replaced by a stable rotating flow. If all initial conditions display chaotic behavior (except the unstable equilibrium solutions), the behavior is defined as globally attracting chaos. Subcritical chaos is the in-between state where some initial conditions result in transient chaos while other initial conditions result in globally attracting chaos [7].

Several researchers have compared the behaviors found in the thermosyphon to the theoretical equations which describe Rayleigh-Bénard convection (i.e. the Lorenz model). Depending on the assumptions made and the data collected, most found that the Lorenz equations generally describe the fluid flow inside the thermosyphon. Ehrhard and Müller worked toward refining the model in their 1990 paper “Dynamical behavior of natural convection in a single-phase loop” [6]. They developed a transformed set of equations which, given certain parameter values, are equivalent to the Lorenz equations. Ehrhard and Müller’s equations include an expression using an experimentally determined constant K , which characterizes the nonlinear condition for the heat transfer coefficient. Ehrhard and Müller found that their experimental data agreed with values found using their theoretical model for conduction and steady convection. The values were not in agreement for time-dependent chaos and subcritical instability. They explain that the discrepancies are most likely due to an inaccurate measurement for K .

More recently, Harris et al. (2011) have looked at forecasting the chaotic flow reversals

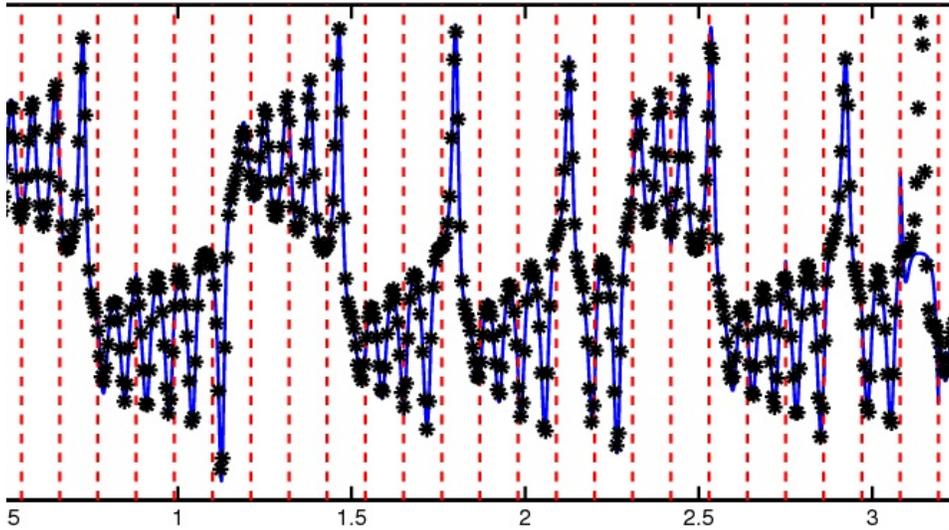


Figure 3: Graph of forecast (blue line) and data (black points). The forecast is the result of Equation 2 using the parameters defined below. The parameters have been tuned to minimize residuals over a set of windows (dashed red). Picture from Harris et al. [8]

in a CFD simulated thermosyphon. To do so, they derive a model from first principles using a variation of the Ehrhard-Müller (EM) derivations. The equations are:

$$\frac{dx}{dt} = \alpha(y - x) \quad (2a)$$

$$\frac{dy}{dt} = \beta x - y(1 + KH(|x|)) - xz \quad (2b)$$

$$\frac{dz}{dt} = xy - z(1 + KH(|x|)) \quad (2c)$$

These equations are equivalent to the Lorenz equations for $K = 0$. The main difference is due to $H(|x|)$ which represents the Heaviside step function. In Equation 2, the Heaviside step function causes $H(x)$ to vary as $p(x) = \frac{44}{9}x^2 - \frac{55}{9}x^3 + \frac{20}{9}x^4$ for $x \leq 1$ and $x^{1/3}$ for $x > 1$, where the piecewise function is designed to keep the model differentiable when $x = 0$ [8].

Harris et al. compare various methods of data assimilation to predict the nonlinear phenomena a thermosyphon would experience. In particular, they look at 3D-Var and several forms of the Kalman filter to develop an accurate method for predicting the duration of a regime. They find that the amplitude of x (proportional to the mass flow rate), preceding a regime change is correlated to the duration of the following regime. In particular, larger amplitudes result in longer duration of following regimes with a threshold at approximately $x = 15$, after which the following regimes are not always longer. In addition, more unstable system states precede longer duration regimes.

Using a multiple-shooting method, Harris et al. solved for the optimal parameters to use for Equation 2 [2]. These values are: $\alpha = 7.99$, $\beta = 27.3$ and $K = 0.148$. Figure 3 displays

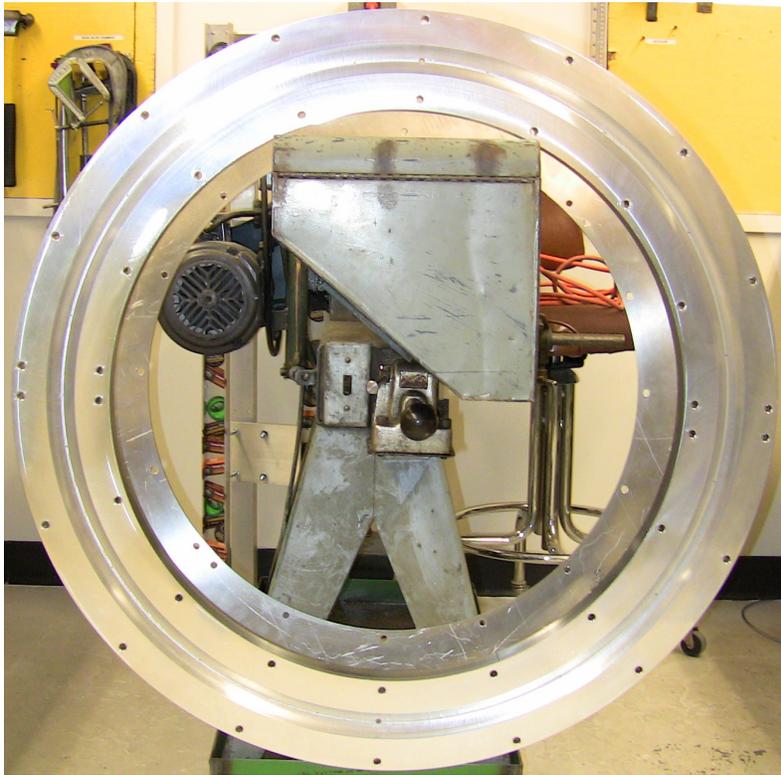


Figure 4: Picture of physical thermosyphon in process of being built.

a graph which applies these parameters to Equation 2 to fit a set of data points. These values were also used in our analysis to compare Equation 2 to the equations generated by Eureka.

3 Experimental Analysis

Our work began with developing a physical thermosyphon from which to collect data (see Figure 4). The thermosyphon was partially built when we began our work; the inner loop and surrounding baths were built and the thermocouples were in place. Sixteen thermocouples were placed evenly around the thermosyphon to measure the temperature data. Our first goal was to gather the data from the thermosyphon. For this purpose, the software TracerDAQ Pro was utilized. TracerDAQ Pro reads the temperature data from the thermocouples and creates a real-time graph of temperature versus time. To test the thermocouples, a heat gun was used to heat the water inside the inner loop. Using TracerDAQ Pro, data was generated that accurately portrayed the increasing temperature inside the thermosyphon.

Once the thermocouples and the temperature software were set up, we began working toward completing the heating and cooling baths. The cavities in which the fluid would run

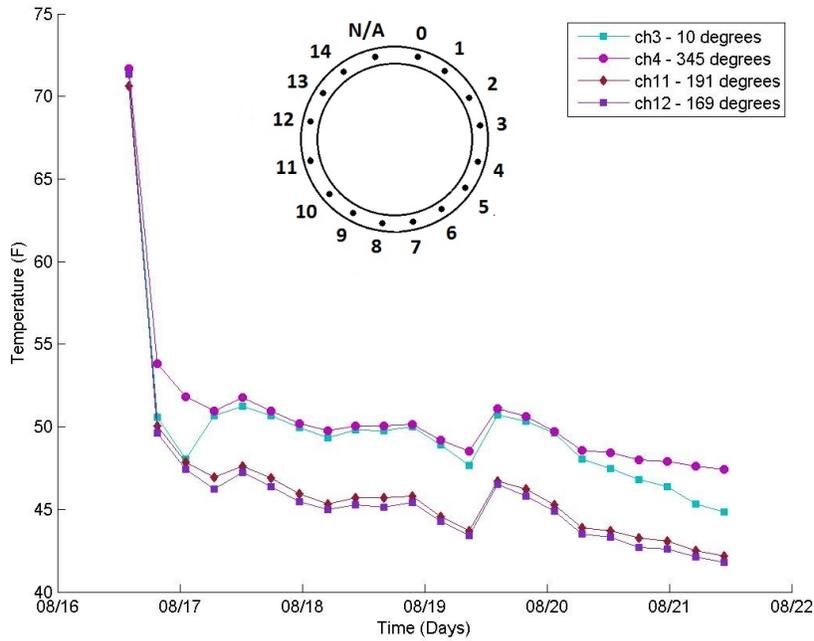


Figure 5: Temperature versus time plot of channels 3 and 4 (surrounding the 3 o'clock position of the thermosyphon) and channels 11 and 12 (surrounding the 9 o'clock position of the thermosyphon). Data taken with chiller on. Inset: Picture of where the channels are on the thermosyphon. There was no thermocouple in the N/A position.

through were in place, but both a heater and a chiller were needed to bring the fluid to the appropriate temperatures. First a small chiller was hooked up. The chiller was not fully functional in that we did not have control over setting the temperature. Instead, it would continuously drop in temperature until it reached an equilibrium point in the range of 41°F to 46°F. With the chiller running, we were able to collect data which revealed that the fluid in the inner loop was rotating counter-clockwise. Figure 5 shows that channels 3 and 4 (at the 3 o'clock position) were reading at a higher temperature than channels 11 and 12 (at the 9 o'clock position) leading us to this conclusion.

Next, a small hot water heater which could reach temperatures of 160°F was hooked up to the bottom half of the loop. With just the heater on, we again collected data which read that the fluid was rotating counter-clockwise (see Figure 6). A difficult issue arose when we tried to run the heater and the chiller at the same time to produce a large temperature difference between the top and bottom halves of the thermosyphon. The heater was running at a much higher wattage than the chiller, leading it to overpower the chiller. While the heater was running, the chiller read approximately 104°F to 114°F. These temperatures, 160°F for the heater and 110°F for the chiller, did not create a large enough temperature difference to produce chaotic results.

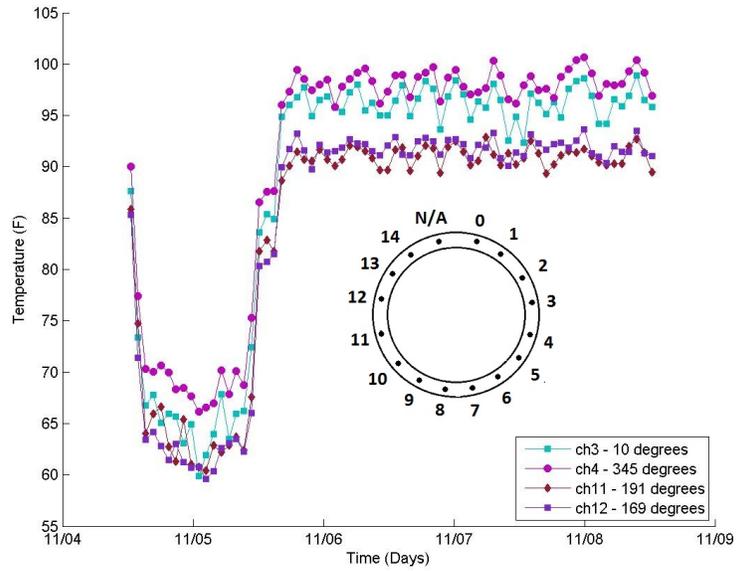


Figure 6: Temperature versus time plot of channels 3 and 4 (surrounding the 3 o'clock position of the thermosyphon) and channels 11 and 12 (surrounding the 9 o'clock position of the thermosyphon). Data taken with heater on.

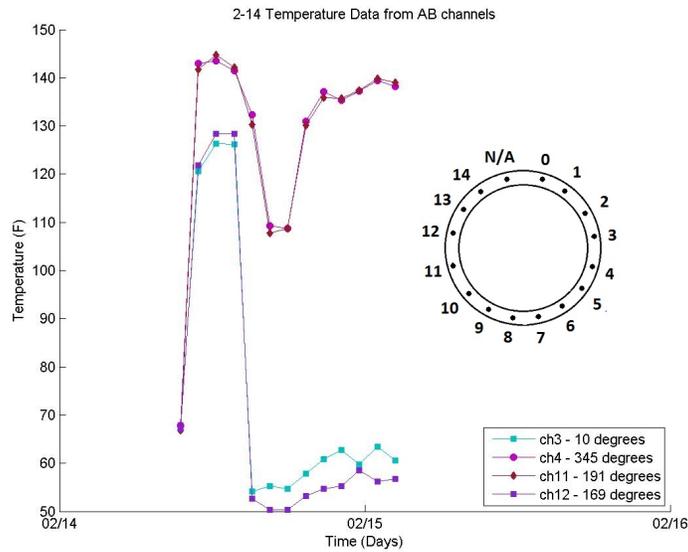


Figure 7: Data taken with heater on and tap water running through top half. Channels 3 and 12 are reading much colder than channels 4 and 11. This signifies that the fluid inside the thermosyphon is not moving since channels 3 and 12 are in the chilled section, while channels 4 and 11 are in the heated section. The large dip in channels 4 and 11 was due to decreasing the heater temperature and then raising it again.

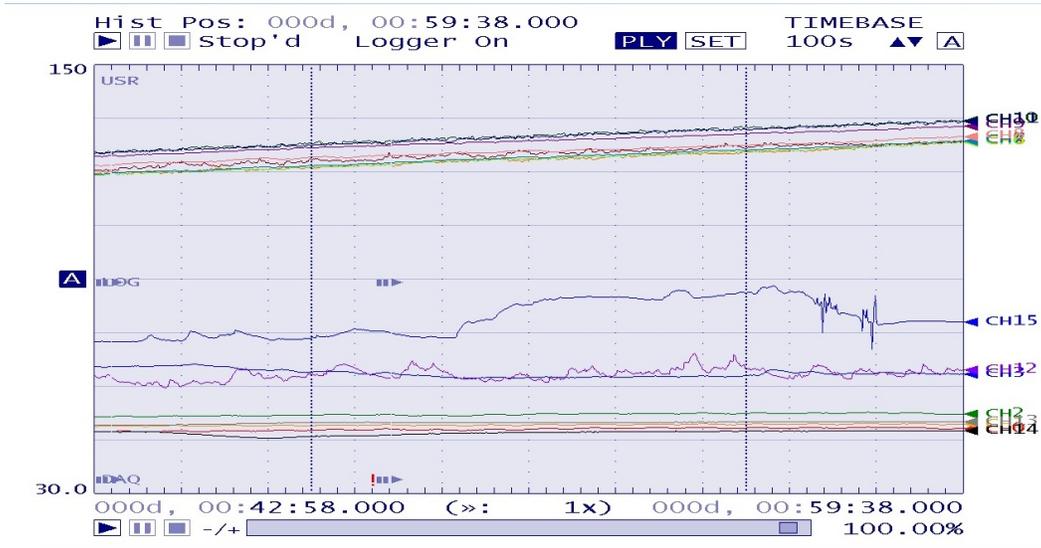


Figure 8: Temperature versus time plot from TracerDAQ Pro showing the increase in channel 15.

In an attempt to reach lower temperatures, tap water was run through the upper bath. The tap water temperature was approximately 45°F, and the heater was running at approximately 145°F. This temperature difference, 100°F, should have been enough to produce chaotic results. Instead, the temperature readings were the exact opposite. The data showed that the fluid in the inner loop was not moving, i.e. it was in a state of conduction (see Figure 7). This state is very unstable and is analogous to a pendulum holding steady while pointing straight up. Knowing that this situation could not be correct, we realized that there must be another issue at hand.

After coming up with several hypotheses to explain the data we were receiving, a test was run to see if the thermocouples were the issue. Thermocouples only read temperature at their tip. We thought that the sides of the thermocouples were touching the aluminum casing which encloses the thermosyphon. The aluminum heats up (or cools) due to the bath water running through the system; we believed that the thermocouples were therefore heating up (or cooling) along their length, causing the temperature reading to be inaccurate. To test the theory, a thermocouple was held against the heated section of the thermosyphon while its tip was in open air. As seen in Figure 8, channel 15 (the tested thermocouple) did read at an increased temperature even though its tip should have read room temperature.

We were not able to resolve this issue in time to collect physical data for this paper. Instead synthetic data from a CFD experiment was used.

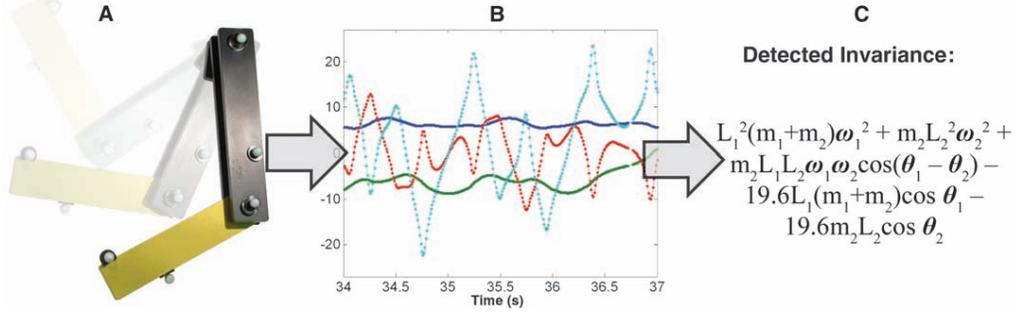


Fig. 1. Mining physical systems. We captured the angles and angular velocities of a chaotic double-pendulum (A) over time using motion tracking (B), then we automatically searched for equations that describe a single natural law relating these variables. Without any prior knowledge about physics or geometry, the algorithm found the conservation law (C), which turns out to be the double pendulum’s Hamiltonian. Actual pendulum, data, and results are shown.

Figure 9: Picture taken from Schmidt and Lipson paper representing how Eureka forms equations from data [13].

4 Numerical Analysis

4.1 Eureka

In our research, we tested the predictive ability of equations derived from the Navier-Stokes equations, namely Equation 2, to those found using Eureka software [16]. Eureka seeks to find analytical expressions which explain symbolically precise conserved relations. To accomplish this goal, symbolic regression is used by piecing equations together using basic building blocks including algebraic operators, analytical functions, constants, etc. If the resulting equation models the data within a certain fitness the equation is subsequently refined or mutated to better fit the experimental data. To identify equations which are nontrivial and can model conservation laws, partial derivatives between pairs of variables are used. Since the desired equation needs to accurately model the data while not overfitting it, the algorithm produces a number of equations which are listed based on their accuracy and their complexity.

The complexity is measured as the number of nodes in the expression trees; more specifically, constants, algebraic operators (except division) and variables add one to complexity while division operators and trigonometric functions add two². The accuracy, or fitness, is a numerical measure of how well a candidate equation fits the data. Eureka uses the following equation to measure fitness: $-\frac{1}{N} \sum_{i=1}^N \log \left(1 + \text{abs} \left(\frac{\Delta x_i}{\Delta y_i} - \frac{\delta x_i}{\delta y_i} \right) \right)$. This equation measures how well the generated equation predicted the data by taking the difference between estimated partial derivatives - one from the data $\left(\frac{\Delta x_i}{\Delta y_i} \right)$ and one predicted by the candidate law equation $\left(\frac{\delta x_i}{\delta y_i} \right)$ - then uses a mean-log-error to combine the residuals. The mean-log-error was chosen due to its ability to reduce high-magnitude residuals which can occur when the

²Other nodes may add a different amount to complexity, but the above values were the only ones used in this work.

denominator approaches or crosses zero, without discarding them entirely. [13]

This algorithm was tested by solving for the laws governing motion data collected from an air-track oscillator and a double pendulum. The results of running the experimental data, consisting of position and velocity measurements, through Eureka came in the form of energy laws for each system i.e. Hamiltonian and Lagrangian equations. In particular, for chaotic data from the double pendulum, the algorithm produced a law of conservation of angular momentum. While the equation $F = ma$ took Newton years to formulate, it took Eureka a few hours to find this governing law from the inputted data.

4.2 Computational Fluid Dynamics (CFD) Experiment

Due to our inability to use physical data, synthetic data from a high resolution simulation CFD experiment was employed as input to produce solutions from Eureka. This CFD data was generated by observing the velocity, weighted by the mass of the fluid (mass flow rate), flowing through a single cross-section of the loop. This value, which is analogous to the x variable of the Lorenz equations, was related to Equation 2 in order to find the best x , y , and z variables to represent the synthetic observation. This was done using data assimilation, a process which takes observations of past states of the system and combines them with numerical forecasts made by Equation 2 to create the analysis - a best estimate of the system with a defined area of uncertainty. Examples of data assimilation methods include the Kalman filter, the ensemble square root filter, and 3D-Var.

The synthetic data was comprised of 250,000+ points in time for each of the three variables. We used the first 100,000 data points as training data for finding equations in Eureka and the subsequent 100,000 data points as validating data to test the predictive abilities of the generated equations.

4.3 Analysis and Results

The three variables x , y , and z from the synthetic data were input into Eureka along with a time variable using 100,000 data points each. The three equations for $\frac{dx}{dt}$, $\frac{dy}{dt}$, and $\frac{dz}{dt}$ were solved for separately using the following basic building blocks: constants, algebraic operators, and the sine and cosine trigonometric functions. Eureka produced approximately 20 equations for each expression. We chose two equations from the resulting list - the equation which most closely resembled the Lorenz equations and the equation with the best fitness.

$$\frac{dx}{dt} = 0.310112y - 0.306327x \quad (3a)$$

$$\frac{dy}{dt} = 1.54139x - 0.0738879y - 0.0459884xz \quad (3b)$$

$$\frac{dz}{dt} = 0.0473712xy - 0.0536667z \quad (3c)$$

$$\frac{dx}{dt} = \frac{1.46267x + x \cos(0.0598488 - 0.0775666x) - 0.169405 \cos(0.0598488 - 0.0775666x) - 2.29592y}{\cos(-0.256487x - 0.33668) - 7.39457} \quad (4a)$$

$$\frac{dy}{dt} = \frac{0.242576z}{46.1372 + xy + y^2} + \frac{432.735x}{65.3906 + z} - 0.0737205y - 4.39379x - 0.0954854 \quad (4b)$$

$$\frac{dz}{dt} = \frac{0.470141x^3 + 2.35432y^2 + 49.0364xy - 4.15505x^2 - 8.97079y}{1008.87 + 5.67322x} - 0.0485457z - 0.125211 \quad (4c)$$

For each set of equations, Eureka provided a fitness and complexity measurement, described above. These values were as follows:

Equation	Complexity	Fitness
3a	7	0.148
3b	13	0.118
3c	9	0.069
4a	38	0.093
4b	33	0.075
4c	42	0.037

Equations 3a and 3c, while resembling the Lorenz equations, were also chosen because they occur at the largest drop in fitness - the previous equations produced by Eureka had a significantly higher fitness than these equations, while all equations following had only a small reduction of the fitness (see Figure 10). The equations produced for $\frac{dy}{dt}$ did not display a significant drop in fitness; therefore Equation 3b was chosen based on its similarity to the Lorenz equation for $\frac{dy}{dt}$.

Next, we analyzed how well the generated Eureka equations could predict future states of the system by using the validating data. To begin, an initial condition (the 100,001st x , y , and z variables) from this data set was selected to evaluate the generated equations. This state was then integrated with each model using the Runge-Kutta (order 4) method in Matlab with a timestep of size 0.01. Figure 11 shows a comparison between the forecast and the validating data using this initial condition. From these graphs, we determined that the generated equations were able to predict the state of the system for approximately 150

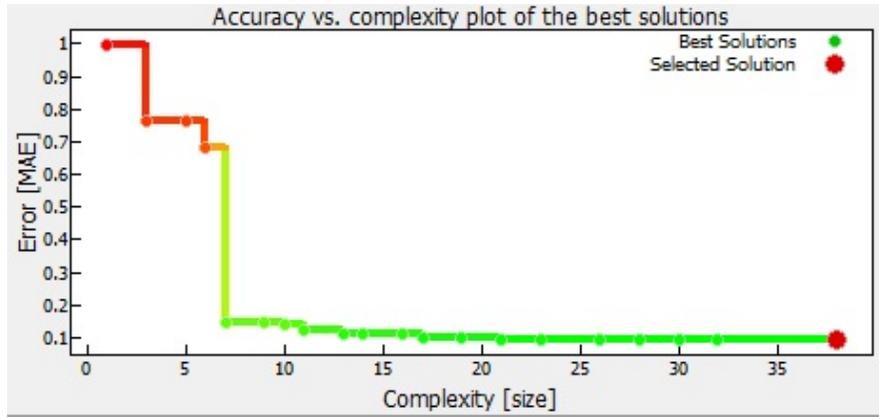


Figure 10: Graph produced by Eureqa for $\frac{dx}{dt}$ showing the fitness of the equation generated versus the complexity. The sharp drop at complexity = 7 lands on the point representing Equation 3a, which is similar in form to the Lorenz equation (Equation 1a). The large red dot (last point on the graph) represents Equation 4a.

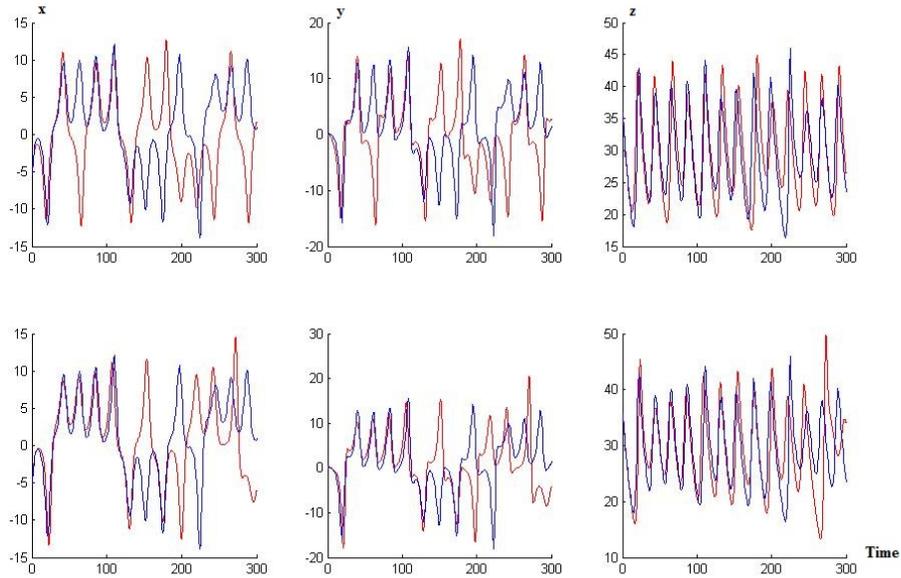


Figure 11: Plots of validating data (blue) and forecast (red) for variable x , y , and z respectively versus time. Above: Forecast using Equation 3. Below: Forecast using Equation 4

time units, where one time unit is equal to 30 seconds in real time³. Since these graphs only show the predictive ability for a single initial condition, we determined that 300 time units would be a sufficient amount of time in which to evaluate the ability of the model to predict the system's state.

³The CFD data has an increment of 5 timesteps of .01 between data points. A time step equals 6 seconds in real time, therefore 5 timesteps is equivalent to 30 seconds in real time.

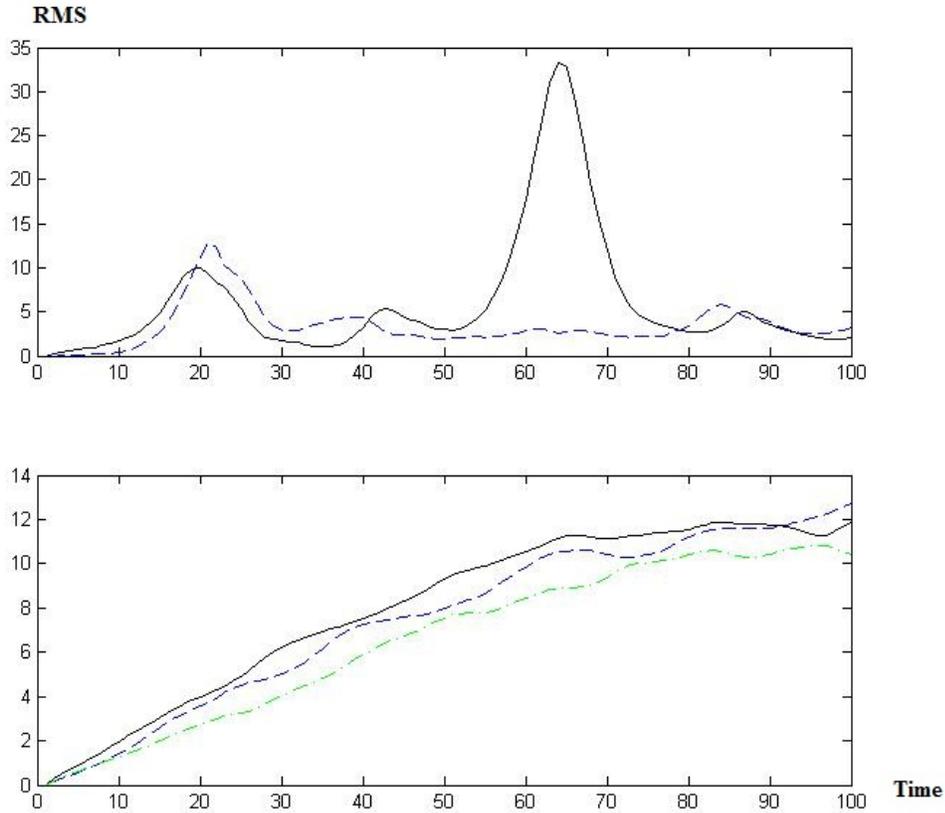


Figure 12: Above: RMSE for a single run using the 100,001st initial condition for Equations 3 (black) and 4 (dashed blue). Below: Average RMSE plots for 400 forecasts using Equation 3 (black), Equation 4 (dashed blue) and Equation 2 (dot dash green).

In order to more fully evaluate these equations, two tests were employed: the Root Mean Square Error (RMSE) and the Anomaly Correlation. The RMSE measures the 2-norm of the difference between the values predicted by the model and values taken to be the truth (i.e. direct observations of the analysis), in this case the validating data from the CFD experiment.

$$RMSE = \|\mathbf{z}^a - \mathbf{z}^*\| \quad (5)$$

where \mathbf{z}^a is the truth (validating data) and \mathbf{z}^* is the forecast (data from the generated equations).

We used 400 initial conditions to evaluate the RMSE. Each evaluation was compared to the truth for 300 time units. The average of the 400 runs was taken to see how accurately the forecast performed over the 300 time units. Figure 12 (top) shows the RMSE as a function of time for a single initial condition for both Equations 3 and 4. Figure 12 (bottom) shows the three models - Equations 2, 3, 4 - and their loss of predictive ability over time. The

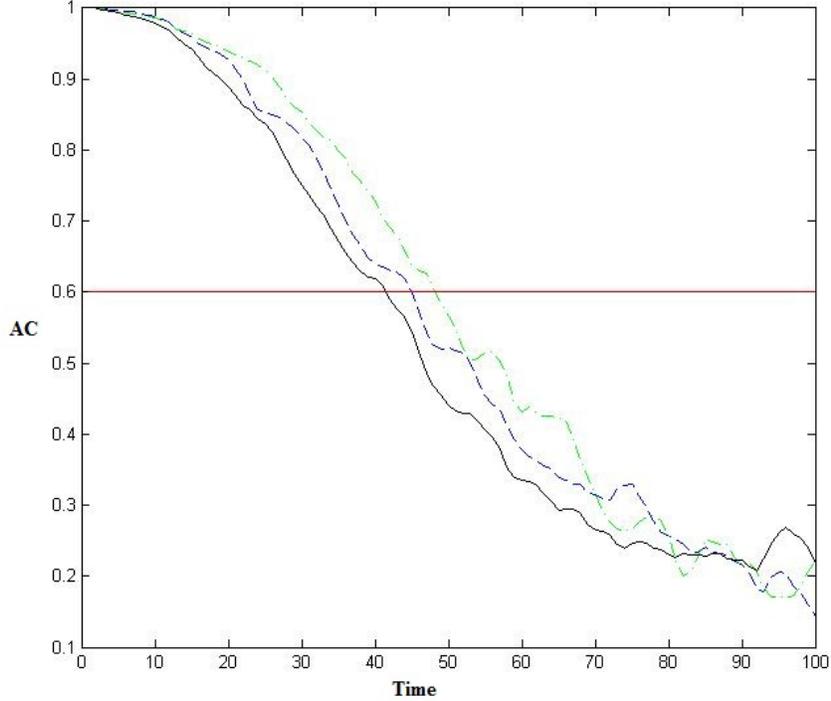


Figure 13: Anomaly Correlation plots for Equation 3 (black), Equation 4 (dashed blue), and Equation 2 (dot dash green). The threshold line $AC = 0.6$ (red) is crossed between $t = 41$ and $t = 42$ for Equation 3 and between $t = 44$ and $t = 45$ for Equation 4. The threshold line is crossed between $t = 48$ and $t = 49$ for Equation 2.

more complex inferred model performs better than its less complex counterpart, and Harris' version of the EM model slightly outperforms the Eureka equations.

The Anomaly Correlation (AC) was also employed to measure the accuracy of the generated equations. It is used to analyze the difference between the forecast and the truth for climate data by measuring the cosine of the smallest angle between two vectors. The following equation is used for calculating the Anomaly Correlation:

$$AC = \frac{(\mathbf{z}^* - \bar{\mathbf{z}})^T (\mathbf{z}^a - \bar{\mathbf{z}})}{\|\mathbf{z}^* - \bar{\mathbf{z}}\| \|\mathbf{z}^a - \bar{\mathbf{z}}\|} \quad (6)$$

where \mathbf{z}^a is the truth (validating data), \mathbf{z}^* is the forecast (data from the generated equations), and $\bar{\mathbf{z}}$ is the average of the forecasted variables.

The Anomaly Correlation has a threshold at 0.6, above which the forecast is considered useful. This limit provides the ability to quantitatively measure the improvement of one generated equation over the other. For this experiment, we again used 400 initial conditions each of which ran for 300 time units. Averaging over the 400 runs resulted in a length 300 vector which was plotted against time. The threshold line $AC = 0.6$ was also plotted

for reference. As Figure 13 shows, Harris’ model does indeed provide an improvement over Eureka’s output. It was found that the Anomaly Correlation for Equation 3 crossed the line $AC = 0.6$ between $t = 41$ and $t = 42$ while the Anomaly Correlation for Equation 4 crossed between $t = 44$ and $t = 45$. This gave Equation 4 a 7.3% improvement over Equation 3. The improvement makes sense due to Eureka’s calculation of improved fitness for this equation with greater complexity.

Included in Figures 12 and 13 is the RMS and Anomaly Correlation for Equation 2, Harris’ model. By including this equation in our tests, we are able to see how well our Eureka-generated equations compare to an ODE derived from the physics governing the system. We see from both figures that Harris’ model does forecast better than the equations from Eureka, but the Eureka equations are not far behind. Looking at the Anomaly Correlation, Equation 2 crossed the line $AC = 0.6$ between $t = 48$ and $t = 49$, and therefore Equation 2 has a 6.7% improvement over Equation 4.

In order to obtain another perspective of how well the equations were performing based on the Anomaly Correlation, another graph was generated. Figure 14 is a 3D plot of each of the 100,000 validating points using the x , y , and z variables as coordinates. The Anomaly Correlation was calculated for each point in the validating data by evaluating Equation 3, the lower complexity Eureka equation, and using a run of 200 time units. The time at which the Anomaly Correlation falls below 0.6 was calculated for each point, and the point is colored based on this number.

From the plot in Figure 14, we can see that forecasts which begin near unstable connecting equilibrium states do well, since their Anomaly Correlation falls below 0.6 at a later time. The same is true for points further away from the foci. Points in dark blue in the middle area of the wings do not perform as well. There are two sections of note: one is the ‘spine’ of dark blue points between the two wings. These points have smaller z values. A small z value represents little to no rotation inside the thermosyphon. This is a very unstable state, and therefore it makes sense that the forecast might quickly deviate from observations in these situations. There is also a region of yellow to red points along the spine and below the wings. We believe that initial conditions in this region are well predicted by the model due to the trajectory they are most likely to follow. In Figure 15, one such trajectory is mapped for the point whose Anomaly Correlation remained above 0.6 for the longest amount of time. We note that the trajectory travels around the left wing once, immediately heads toward the center of the right wing, and then slowly winds its way out. This type of behavior allows the equation to forecast accurately for a long period of time.

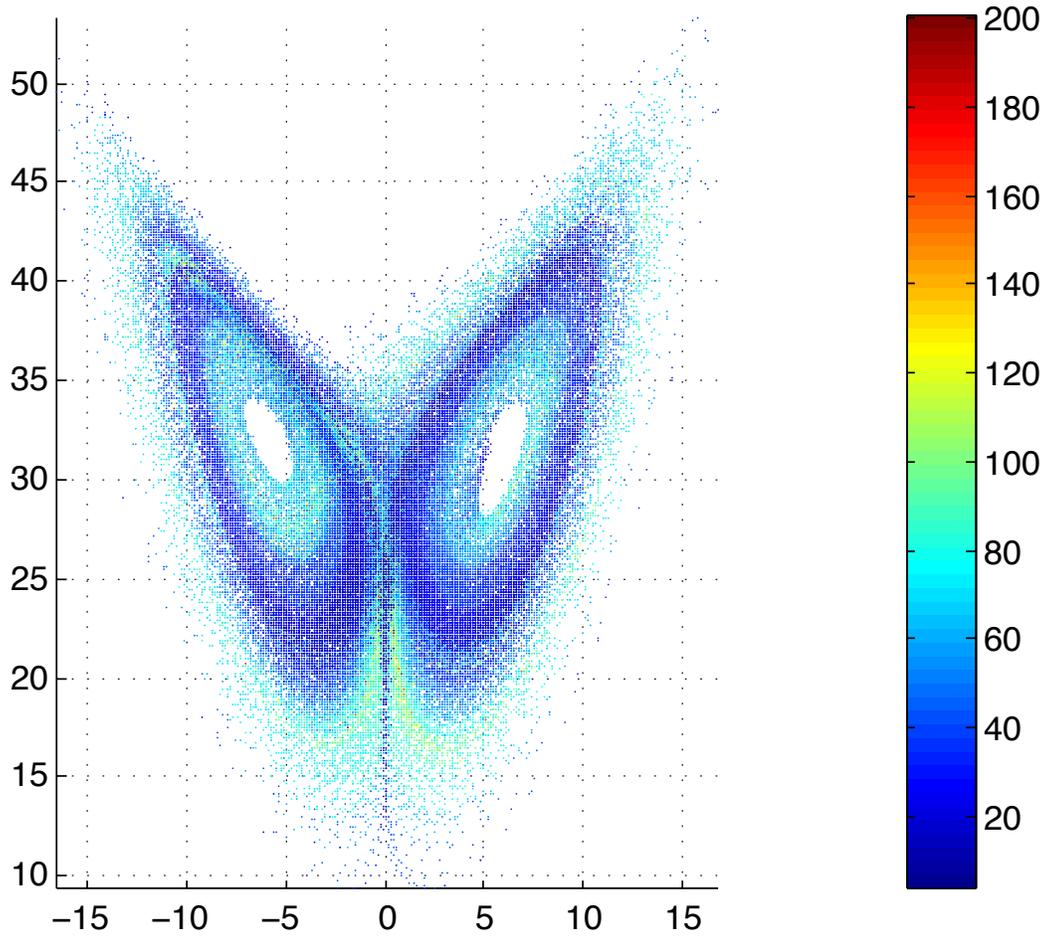


Figure 14: 3D plot of the 100,000 validating points. Each point is used as an initial condition and evaluated using Equation 3. The points are colored based on the duration until the Anomaly Correlation fell below 0.6. Points which are colored based on a lower number performed worse than points colored based on a higher number. The plot shows the X-Z view.

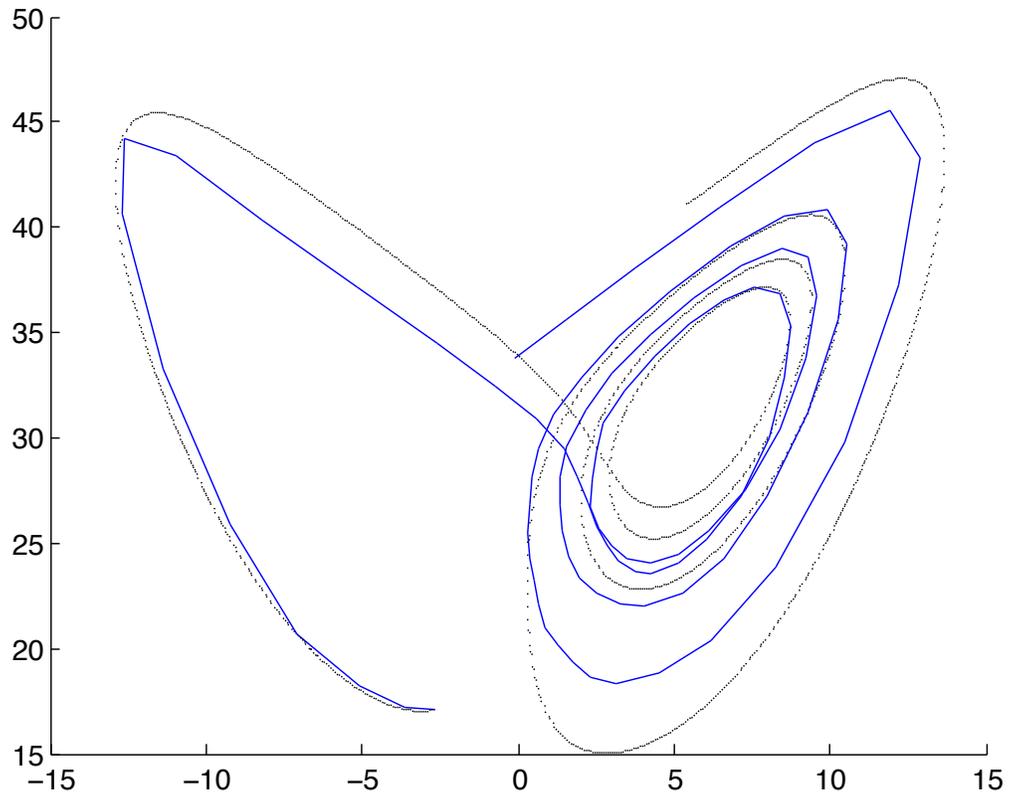


Figure 15: Trajectory for truth (blue) and forecast (dotted black) using the initial condition for which the Anomaly Correlation for Equation 3 remained above 0.6 the longest. The trajectory runs for 100 time units. The plot shows the X-Z view.

5 Conclusions

This work began with two objectives. Our first goal was not fully completed since we were unable to collect usable temperature data from the thermosyphon. However, the thermosyphon will soon be a functional toy climate which can produce data for further analysis as a sandbox for data assimilation algorithms. Regarding the second goal, we were able to utilize the Eureka software to solve for forecasting equations.

By analyzing these equations through forecasting tests, we were able to draw some conclusions about the ability of Eureka to infer governing laws from data. Though the Eureka generated equations did not outperform Harris' model with regards to forecasting, both Eureka equations came very close to matching the forecasting ability of an equation which was derived specifically for the purpose of describing the thermosyphon system. It is a remarkable feat that this software, just by inputting a data set, can output equations which describe the system so well. We are eager to see how further advances in equation-generating software can improve not only the prediction capabilities for climate forecasting, but work in other scientific studies as well. One potential mechanism for improving Eureka's performance is the involvement of scientists in the choice of mutations genetic algorithm implementation. Indeed, recent work on protein folding using online games has led to the realization that human-computer interacting systems are capable of producing far better results than either in isolation [3].

Another conclusion we were able to draw from this work was the intrinsic relationship between forecasting ability and initial conditions. From Figure 14, we saw that a data point's position in 3D space is a critical factor in determining how well the forecast model will perform.

The next steps in this process will be to complete work on the physical thermosyphon so it can produce temperature data which will be inputted into the Eureka software. It will be interesting to see if the physical data still produces equations with the same form as the Lorenz equations. From there, the generated equations can be used to work toward an improved forecasting model. Eventually, this work may lead to more accurate long-time forecasts for weather and climate prediction. However, the most interesting question raised by this work is: how long will physical principles be the bread and butter of science? When will software and computer systems, such as Watson, suggest laws we cannot interpret?

References

- [1] K. T. Alligood, T. D. Sauer, J. A. Yorke. *Chaos An Introduction to Dynamical Systems*. Springer, New York (1996).
- [2] E. Baake, M Baake, H. G. Bock, K. M. Briggs. “Fitting ordinary differential equations to chaotic data.” *Physical Review A* (1992)
- [3] Cooper et al. “Predicting protein structures with a multiplayer online game.” *Nature* 466, 756-760 (2010).
- [4] H. F. Creveling, J. F. De Paz, J. Y. Baladi, and R. J. Schoenhals. “Stability characteristics of a single-phase free convection loop,” *Journal of Fluid Mechanics* 67, 65 (1975).
- [5] C. M. Danforth, J. A. Yorke. “Making Forecasts for Chaotic Physical Processes,” *Physical Review Letters*, 96, 144102. (2006).
- [6] P. Ehrhard, U. Müller. “Dynamical behaviour of natural convection in a single phase loop,” *Journal of Fluid Mechanics* 217, 487 (1990).
- [7] M. Gorman, P.J. Widmann, K.A. Robbins. “Nonlinear Dynamics of a Convection Loop: A Quantitative Comparison of Experiment with Theory,” *Physica D* 19, 255-267 (1986).
- [8] K. D. Harris, E.-H. Ridouane, D. L. Hitt, C. M. Danforth. “Forecasting regime changes in a chaotic toy climate,” In Preparation (2011).
- [9] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. University Press, Cambridge (2003).
- [10] Edward Lorenz. “Deterministic Nonperiodic Flow,” *Journal of the Atmospheric Sciences*, 20 130141 (1963).
- [11] E. N. Lorenz, “A Study of the Predictability of a 28-Variable Atmospheric Model”, *Tellus* 17 321333 (1965).
- [12] E.-H. Ridouane, C. M. Danforth, D. L. Hitt. “A 2-D Numerical Study Of Chaotic Flow In A Natural Convection Loop,” *International Journal of Heat and Mass Transfer*. doi: 10.1016/j.ijheatmasstransfer.2009.10.003 (2009).
- [13] M. Schmidt, H. Lipson “Distilling Free-Form Natural Laws from Experimental Data,” *Science*, Vol. 324, no. 5923, pp. 81 - 85 (2009).
- [14] Z. Toth, E. Kalnay, Bull. “Ensemble Forecasting at NMC: The generation of perturbations,” *Amer. Meteor. Soc.*, 74, 2317-2330 (1993).
- [15] P. Welander. *Journal of Fluid Mechanics* 29, 17 (1967).

[16] <http://creativemachines.cornell.edu/eureka>