Climate Change Data Modeling in Mountainous Regions

Justin Olson

Mountain Geography Term Paper

Background

Climate change is, arguably, the biggest challenge the human race will face in the 21st century. Nations across the planet dedicate billions of dollars towards research, analyzing the source of the problem, and developing potential solutions. The effects of climate change can be noticed all over the world, in all aspects of life. In the largest picture of the subject, the rising of global temperatures knocks over one domino eventually causing an uncountable number of changes to our planet, but most departments are dedicated towards looking at a single change. These changes are tracked over time, and used to predict future trends. This sort of predictive analysis is done in the context of modeling animal migration and population patterns, glacial monitoring and sea level estimation, and even weather patterns and fire risk estimations. This is what brings up the largest software challenge of our century.

Software can be the key tool to cracking the problem. A paper out of Tsinghua University in Beijing dives into engineering software for climate change. The paper states “...just like all other complicated problems, climate change problem also needs powerful software systems to enable researchers better understanding the issue, and to provide systematic, intelligent, effective information services and decision supports for

the governments, enterprises and individual citizens to better manage their daily operation, behavior and working processes." [1] This accurately states softwares role in the solution, as a tool in understanding the problem and finding the best solution. Another fantastic paper overlaying the role of software in the issue, is written by Steve Easterbrook from the University of Toronto. In this paper he discusses software engineering in the field of climate change. There is an important definition between the researchers, the software engineers, and the entities that attempt promote change. He outlines how everyone has a critical role, and promotes the idea of "open notebook"[2] science. This sort of data sharing is crucial into designing our system.

The massive amount of data collected and used to model a single effect of climate change is a technical challenge, and in the industry, this challenge is known as big data. Systems are designed to be highly efficient at computing data at unthinkable rates. Data storage systems are built to manage exabytes of data, and be able to access it in a meaningful way. Billions of dollars of funding are pumped into research and development of these systems, and they are able to do a decent job at what they are designed for; modeling a change. This is only a single change, a single puzzle piece, in a much larger puzzle. If these systems were able to assemble the pieces of the puzzle, rather than render a single piece, we would reach a solution much faster. If there was no computational limitation, we could combine each one of these big data models into one much larger model for our global climate. Every day there are new breakthroughs in the field of computation, and someday the technology will be ready. When that day comes, and there is hardware with the necessary speed capabilities, the theory and software behind building a massive scale climate model needs to be ready. This is the challenge we are currently able to work on.

Introduction

A similar large scale model was described in a paper co-authored by the same author from Toronto. Here Easterbrook depicts exactly how large of a computing challenge a model of this size can be. “To simulate climate change over a century, a fully coupled model can take several months to run on a supercomputer.”[3] While several month compute times are technically possible, we would not be able to do fully utilize the model. It wouldn’t be as simple as adjusting a parameter and watching how that changed the model, as it could take months to rebuild. To use a model for the functionality needed, it would have to be fast enough to calculate predictions based off of real time changes in any feature. While we are at a current technological limitation, there is still much that can be done. Until researchers discover computational methods capable of processing data, magnitudes faster, or hardware capable of running current implementations even faster; it is important to sort through the data we are currently collecting. If we are only able to effectively use a small percentage, we need to make sure that we are looking at the subsection that gives us the most information possible. If we are looking at climate change, different climate zones are going to respond differently. Some of the effects are much more noticeable and predictable in specific places. But, in another climate zone, the changes could seem much more sporadic or random. It is important that the selected zone yielded data with statistically consistent changes, and is as self contained as possible. The climate surrounding a mountainous region can be an ideal scenario for data-scientist’s to apply more global models. The distinct characteristics of mountain climates should create clean data to support the creation of a regional climate model that can be used to greatly enhance environmental protection efforts.

Specific Interest in Mountain Climate

The statistical analysis being done looks at the link between causes and effects in order to try to predict effects, or determine what causes need to happen to reach a desired effect. Changes in animal populations across the globe are the effect of multiple causes. With our current technology, we are unable to break down the change in animal

populations to each individual cause, that is, to what magnitude each cause attributed to the total effect, let alone use a model with that cause to work backward and apply it to the change in stream patterns, the change in precipitation, and the change in glacial levels to show the effects of glacial change on animal populations. Mountainous climates are very well self contained. By focusing on the climate of a specific mountain region, we can analyze data that has far fewer variables than more complex climate regions. If scientists were able to create a completely closed-system model of an entire region, we could use predictive analytics to make a window into the future of our planet. Looking at our data model, we could change individual features, and see how the larger picture changes.

The region already presents a difficult challenge to any life that attempts to thrive at elevation. Most of the plants and animals have evolutionarily adapted traits that give them the best chance possible of surviving in the difficult climate. These traits make them particularly subjective to a larger impact from a change in the local climate. This makes monitoring that change much easier. From a data perspective, it produces a few ideal conditions that give it a statistical advantage. There is a low variance in consistent data, and will be larger and more noticeable. The magnitude of these changes will be more consistent which provides a very clean dataset for statistical analysis. The effectiveness and accuracy of a model greatly depends on the shape of the data it can be trained with. The unique traits of the mountain climate allow us to produce very clean data ideal for analytics. While we cannot get where we need to go without theoretical or hardware improvements, this is a great place to apply current efforts. The data collected from the region has the advantage of a low number of outlying observations, and when there is change it is consistent.

Advantages of Studying Mountain Climate

While having an abstract conceptual understanding of the process and goal, data scientists still need to figure out exactly how to apply the concepts. A basic understanding of how the data is used is necessary to understand exactly how we can

use the statistical advantage that data collected from mountainous regions provides us. As a brief introduction to a single method of data modeling, below is a simplified abstract explanation of how we can look at the data.

Consider a theme park and a carwash, who are either open or closed depending on the status of the weather. Now consider a set of collected data that has three columns: the current weather, the status of the theme park, and the status of the car wash (see fig 1). These are also referred to as features in a model. While the status of the theme park and car wash are dependent on the weather, the weather is not determined by the status of the theme park or the car wash. That is, the owner of the car wash cannot turn the light on his open sign to make the rain stop. The owners of the park and of the car wash determine whether or not they will be open or closed in the morning when they check different weather sources. Just like all weather predictions, they are not 100% accurate. Sometimes the owner of the theme park will send out the email to his employees that they may have the day off, and the snowstorm will blow over. At the same time the car wash owner decides the chance is low and stays open. This means that sometimes an observation may have the theme park closed, the car wash open, and the weather is clear. However the majority of the data for clear weather will have them both open. A quick glance at the following set of data and we can come up with three rules of operation for the car wash and theme park.

Weather	Theme park Open?	Car Wash Open?
Sunny	y	y
Rainy	y	n
Snowy	n	n

Fig. 1

- The car wash is open only if it is sunny
- The theme park is open if it is sunny or raining
- The theme park is closed if it is snowing

Now while that was an easy task for a human to generate the above list, when adding the inconsistency of a mistaken weather report into the dataset, it becomes a difficult task, even for a machine to complete. If there are only three days of observations and they represented the correct conditions, the machine can offer predictions at 100% accuracy. If the machine is provided a ‘yes and no pair’ for the status of the businesses, it can guarantee you the weather on that day. As the inconsistencies are added it is no longer a guarantee, but the machine can still return a prediction with a calculated percentage of accuracy. If the machine is trained with four data observations and one is an outlier, the percent accuracy will very low. The machine would have a 50/50 chance at guessing between the two options it narrowed down to. Add one more correct observation and we can give a 66% accuracy. It would need to be trained with many more correct observations before it could ensure a high accuracy. So for every statistical anomaly in the data, the necessary size of the dataset to ensure an accurate prediction increases. The more data used to train the machine, the more it can find the statistical consistencies, and the more accurate it is.

Now, in the above example, we saw how badly one outlier messed up a model with only three features. With a few changes this model becomes inconceivably complicated. This is a trivial example in comparison to the real application. The features in this climate change model wouldn't be status of a car wash or theme park, but one measurement of a tracked change, thought to be related to climate change. It is no longer a binary yes or no, but could be any continuous value. A slightly more complicated example would be estimating the value of a car based off of a dataset containing car sales with the following features: value, year, color, mileage, make. While the year and color are going to operate relatively similar to binary, milage and make

throw everything off. The value of a car is directly affected by its milage in a nonlinear rate that could be very difficult to compute. While the make of a car offers some classification to value, different manufacturers have different cars in all different price ranges. Without the model of the car as well, it is hard to tell exactly how this can be used, we just know it is relevant data. On top of all of that, we are asking the machine to compute a continuous value, not a discrete label. Just adding a few degrees of complexity, made this problem scale terribly.

Finally, with all of this in mind, let's think about the model we are actually trying to create, and why the properties of mountainous regions provide an advantage against these two scaling issues. This data in comparison to other regions, will have a lower percentage of outlying observations, so the predictions will be accurate. Having consistent changes when using a non binary scale, makes it as close to a binary problem as possible. The same reason why using year as a feature was better than milage. Year and milage have drastically different scales. With the incredible complexity of the model needed, these advantages are crucial.

Implementation of a Statistical Model

If we were to model an entire mountain region's climate, it could have some extremely powerful functionality. We could use it to change the value of any given feature and predict the resulting outcome. The first issue is the data itself. Even if we have the theory behind creating such a model, we would need to be able to represent an entire climate only by data. This would require the combined efforts of everyone doing similar research. All of the data collected is logged in time series form, and can be joined on the timestamp. This data would include every feature that could be argued has an indirect effect due to climate change. Then just the subsection of features you are interested in can be selected. All of a sudden, glacial observations are now in direct comparison with things such as, precipitation, stream volume, floral patterns, or even animal populations. Then the direct impact of glacial melting can be seen on the food

chain. Taking this even further, if the population of a specific animal species is selected, the other values in the model can be tweaked to predict endangerment and extinction timelines. Out of this master data set, a bunch of smaller models can be built to predict specific features, and these smaller models are useful for specific things. Some useful models could be: At risk animal population levels, forest fire danger levels, invasive species populations, short and long term environmental impact of road construction, tree density over time, and geologic activity. The list could be endless.

Coincidentally, most, if not all, of these suggestions are already statistical models being used by different environmental departments across the globe. If creating this huge model is so complex, and they are already doing most of it already, what is the point? Currently all of these models are acting as independent entities. By creating sub-models from a feature pool of a large model, they now have the ability to work together. This gives each individual model the enhanced power and accuracy from every other one created from the same data set. If the real world application of this is not quite clear yet, consider this example.

Before constructing a new road and tunnel, environmental studies are done. A short term study is done to determine how the effect of the actual construction will have on the surrounding environment. A long term study is also done to determine how the formation of the new road and tunnel will affect the environment over time. Similar statistical models are used for both of these. At the same time several other corporations are modeling animal populations, rockslide potential, and forest density. Now the rockslide model can predict the effects of the new tunnel with information from the forest density over time. Instead of using a rough estimated guess of how many trees will be in the area in 20 years to find rockslide potential after the construction, it can use a statistically trained model to include a much more accurate tree density estimation. Similar to the previous example with the theme park and car wash, increasing the accuracy of one feature dramatically increases the simplicity and accuracy of the computation. If the weather man was always right then you could always tell the weather based on if the car wash and theme park were open or closed.

Increasing the weatherman's accuracy percentage directly increases the accuracy of the determining the weather from the status of the store. Providing more accurate floral and fauna statistics, to the construction impact model only greatly increases its accuracy.

Application to Solving Climate Change

The goal of creating a universal closed-system mountain climate model was for it to be used as a research tool for climate change, so far I have only proposed improvements to more simple models. The original issue with a massive climate change model was the uncertainty of the endless amounts of possible changes tracked. If each of these changes are set up as a model, it is no longer an uncertain value. The machine calculated the most likely outcome with a given percentage likelihood that can be taken into account in the big picture.

Everything essentially boils down into a system that can do two very important tasks. The first being applying a current model and estimate the positive or negative effect on global climate. A reforestation act could include rigorously calculated estimates on the exact positive effects it will have. A construction job can model much more accurately the exact negative effect it will have. The second function is applying the climate change model, and know its impact on other models. We could know the exact contribution climate change has on population numbers, and calculate how much change would need to occur to save endangered species, what projects had to get cut or funded to meet that change, and estimate deadlines. Arguably the most important feature though is being able to quantify an exact population of people's contribution to climate change, model exactly what negative outcomes will occur at current values, and find what needed to be changed in order to not reach those negative outcomes.

The biggest factor in climate change is people, and the largest issue with applying green habits is the anonymity behind it. People think that them alone is not enough to make a difference, and that even when there are finally negative impacts they won't be around to see them. A highly accurate model would greatly improve

climate change education efforts. Once someone puts a quantifiable value to the negative impact your population is having, and a quantifiable value to the efforts needed to be done, there isn't as much anonymity behind it anymore. People are much more likely to change bad habits when they know the big picture. The application of a model like this would allow people to be well informed on their actions, and the effects of those actions.

Conclusion

For the immediate future there is a technological barrier limiting the size an effective data model can be, and how much data can be gathered and processed into that model. The enormous task of creating a global climate change model that consists of every possible direct and indirect effect, and every single direct and indirect cause, is not feasible in the foreseeable future. Until computation speeds catch up to what we need them to be, we can look at the climate of a specific region, and apply a large scale model. This target region needs to be as self contained as possible, and any outside interference needs to be predictable. Any unpredictability compounds in accuracy reduction, so selecting the ideal region is crucial. Mountainous regions have a very unique climate, with many properties that produce ideal conditions for such a data model. The elevated peaks provide a well isolated climate. The well tuned local species respond more noticeably to small changes in climate due to their evolutionary adaptations to the harsh environment. These factors make it produce very clean data, appealing to this kind of data modeling.

Many environmental protection agencies and corporations afraid of being sued by the environmental protection agencies, already create useful models that could be applicable to climate change. Almost all of these models are created from data that is observed over time. The fact that they all have a time field in common means they can be combined and used interchangeably. Aggregating data from many different sources, and using it all in an efficient manner, is the key to producing the most accurate model.

When creating a data model every small hiccup in the data, reduces the accuracy of the resulting predictions. If all of these models were created using features from a large model including every climate change cause and effect, then they can be used by one another to improve the accuracy of each other. Once the accuracy of each prediction is high enough then we can efficiently model climate change with the large model, using the smaller models to enhance the accuracy of features.

This mountain range climate change model can be used to predict many useful pieces of information, and is as accurate as you can adjust for influence from outside factors. The goal is to have the mountain climate model as close to a closed-system as possible, and then apply accurate corrections for the places the system cannot be closed. Using the model we can show the effect certain changes have on the regional climate. Scaling this out over time we can show the impact on features due to climate change. After this we can simply adjust the initial parameters to see what goals were needed to be met in order to prevent certain outcomes. This allows for greatly improved education, and could be used on a small enough scale to remove enough of the anonymity being climate change contribution. These two factors should be enough to spark change. The best way to make progress towards anything is to set a goal with a clear path. Isolating a mountain environment due to its statistical advantages, to create an accurate large scale climate computational model, could provide the tool necessary to create this path.

Related Work Referenced

[1] Liu, Lin, He Zhang, and Sheikh Iqbal Ahamed. Some Thoughts on Climate Change and Software Engineering Research. N.p.: n.p., n.d. PDF. http://www.cs.toronto.edu/wsrcc/WSRCC1/papers/Liu-WSRCC-1.pdf

[2] Easterbrook, Steve M. "Climate change: a grand software challenge." Proceedings of the FSE/SDP workshop on Future of software engineering research - FoSER '10 (2010): n. pag. Web.

[3] Easterbrook, Steve M., and Timothy C. Johns. "Engineering the Software for Understanding Climate Change." Computing in Science & Engineering 11.6 (2009): 65-74. Web.