Developing accurate forecasts for wildfire smoke dispersion

The University of Maryland, Baltimore County (UMBC) maintains a high standing in national rankings and a notable reputation for research in many areas. Since poor air quality can greatly impact health and life, timely observations and reliable forecasts are important for hazard control and protecting public health. In 2008, two professors and a Ph.D. graduate student in the Department of Computer Science and Electrical Engineering at UMBC initiated a research project to analyze wildfire smoke patterns. The ability to provide fire and public safety officials with a real-time assessment of smoke patterns during a fire would support more informed decisions on public evacuations and health warnings.

Dr. Kostas Kalpakis, Associate Professor in the Department of Computer Science and Electrical Engineering at UMBC, and principal researcher on this project, explains: “We saw the availability of new data and the demand for better forecasts, but we had limited means to integrate the new data to improve dynamic model forecasts.”

The UMBC research team began a search for technology that would help process large volumes of real-time data streams from various atmospheric sensors and weather reporting agencies and provide input for forecasting air quality and improving the accuracy of predictions.
Business benefits

- Fuses air quality data from multiple sources to provide public officials with real-time data about fire and smoke status
- Enables an improvement of about 16 percent in forecast accuracy in the first three days of an emission event based on prototype results
- Accelerates expected times for wildfire analysis from hours to minutes to promote timely decisions concerning firefighting and public safety
- Delivers perpetual analytics to reconcile differences between real-time data and associated predictions to continuously refine forecasts

“Our project required collecting and analyzing massive amounts of air quality data from a variety of sources in real time or near real time,” says Dr. Yaacov Yesha, Professor at UMBC’s Department of Computer Science and Electrical Engineering. “Our immediate goal was to demonstrate a prototype system that could fuse near real-time observations from multiple sensors and improve the accuracy of near-term forecasts for pollutant dispersion. The longer-term goal is to develop an operational system for air quality forecasting. We knew that the idea for our Data Assimilation research project would align nicely with IBM’s Smarter Planet initiatives and saw that we could benefit from IBM InfoSphere Streams to overcome problems in analyzing and correlating the data.”

Developing a prototype that improves forecasting accuracy

Developing a prototype that improves forecasting accuracy

The Data Assimilation research project comprises techniques used to fuse real-time data obtained through observations with results from mathematical models, while considering the uncertainties of both, as well as any constraints on consistency resulting from the underlying physical system. A complete data assimilation process requires observational data obtained from a variety of air quality sources, a processing model and an assimilation algorithm.

Smarter Research: Bridging the gap between predictions and real-time observations

Instrumented
Obtains real-time air quality data available from government environmental agencies and weather services for analysis.

Interconnected
Analyses real-time air quality data using a scientific algorithm and processing model designed to use atmospheric conditions to generate air quality forecasts.

Intelligent
Reconciles the differences between the actual measurements and the associated predictions, and recalibrates model algorithms to continually improve forecast accuracy.
First, real-time air quality data is collected from multiple observation sources to more accurately predict the concentration of particulate matter in the air. Next, UMBC researchers feed this data into the HYSPLIT (Hybrid Single Particle Lagrangian Integrated Trajectory) model developed by the NOAA (National Oceanic and Atmospheric Administration) Air Resources Lab. This model provides a complete system for computing particle dispersion patterns to indicate how particles travel through the atmosphere. It also allows for performing complex dispersion and deposition simulations.

Lastly, a data assimilation algorithm (known as LETKF or Local Ensemble Transform Kalman Filter) provides an ensemble analysis of the data to deliver more accurate initial conditions used to create forecasts. The algorithm runs on IBM® InfoSphere® Streams, which provides parallel and high-performance stream processing and continually receives and analyzes data from the variety of sources. InfoSphere Streams is part of IBM’s big data platform, and UMBC is using InfoSphere Streams version 1.2, with plans to upgrade to version 2.0.

The new system enables researchers to use streaming data in real time from sources, such as geostationary and orbiting satellites and drone aircraft, to help improve forecasting accuracy.

Dr. Kalpakis explains: “To generate a forecast, first we have historical weather and air quality data from prior experiments conducted by NOAA while developing the HYSPLIT model for pollutant dispersion. It is critical to note that HYSPLIT is the operational forecasting model that NOAA uses to advise government agencies. The same types of data used to validate the HYSPLIT model are being used to assess the quality of our fusion architecture and the algorithms that we have to fuse measurements with the forecast. And we have seen quite an improvement in terms of the forecasting accuracy.”
“We knew that the idea for our Data Assimilation research project would align nicely with IBM’s Smarter Planet initiatives and saw that we could benefit from IBM InfoSphere Streams to overcome problems in analyzing and correlating the data.”

—Dr. Yaacov Yesha, Professor, Department of Computer Science and Electrical Engineering, UMBC

Stationary satellites provide measurements approximately every half-hour, based on goals set by NOAA. The orbiting satellites, operated by NASA (National Aeronautics and Space Administration) provide measurements twice a day. “Going forward, we want to obtain data from UAVs (Unmanned Aerial Vehicles) that can be flown over wildfire areas to obtain higher spatial and temporal resolution data than is available today with the satellite instruments,” says Dr. Kalpakis.

After the Data Assimilation prototype generates a forecast, the objective is to validate it. “The most important contribution of our project is bridging the gap between air quality forecasts and real-time measurements to improve wildfire smoke prediction and monitoring,” says Dr. Kalpakis. “Reconciling the differences between the actual air quality measurements and the associated predictions enables us to recalibrate model algorithms to improve future forecasts. By making continuous improvements, we can bring predictions and observed measurements closer together.”

Building a platform for computationally intensive models

The process of revising forecasts based on actual measurements can be computationally intensive. For example, when the weather service runs its numerical models, it is actually using big super computers to do those calculations because the process is computationally demanding. As a result, selecting an appropriate platform for this research project was critical.

The IBM Systems and Technology Group, University Alliances team worked with the IBM University Relations team to provide UMBC researchers with an IBM BladeCenter® H and 13 IBM BladeCenter HS22 servers comprising 104 cores (8×13) and a few hundred gigabytes of random access memory (RAM). These servers support a broad range of workloads. The Data Assimilation prototype runs on a University cluster, called BlueGrid, which hosts eight IBM BladeCenter HS22 servers with Intel Xeon processors running Red Hat Linux (RHL) 5.5, a 64-bit operating system.
InfoSphere Streams is quite suitable for scientific calculations since it integrates complex data types for vector and matrix operations, as well as extended libraries, which we need to use extensively in the Data Assimilation system.”

—Dr. Kostas Kalpakis

InfoSphere Streams offers advantages in scaling across a wide range of hardware platforms and can be reconfigured automatically in response to changes in data and system resource availability. With InfoSphere Streams, UMBC can respond to events and trends immediately, while it is still possible to improve outcomes.

Dr. Kalpakis comments, “Using the LETKF algorithm and InfoSphere Streams, multiple localized and independent processes can happen simultaneously to generate results much faster.”

**Promising results from the Data Assimilation prototype**

“T o my knowledge, this is the first time anyone has done air quality forecasting in a systematic way using an operational model like HYSPLIT from NOAA and a diverse set of measurements to improve the accuracy of sensor forecasts,” says Dr. Kalpakis. “We managed to get this project off the ground in a little over a year and also provided the capability to improve forecast accuracy by fusing these observations into the system. The preliminary results on our Data Assimilation prototype indicate an improvement of about 16 percent in forecast accuracy in the first three days of an emission event. It’s preliminary, but it represents quite an improvement over traditional forecasts. That is the thing that makes our research interesting and exciting for us to continue.”

**Looking clearly toward the future**

If made available in a production environment, UMBC’s Data Assimilation system would help improve the way wildfires are tracked and extinguished and improve decision making for issuing health and safety alerts.
“We are fortunate that we started this project with IBM, and so far, our activity has been well received by some of the major groups, for example NOAA,” says Dr. Yesha. “Our long-term goal is to move the prototype design into an operational system for air quality forecasting. One factor of paramount importance is the accessibility and interoperability of the prototype system. A fully operational system will need to interface with many other systems.”

“InfoSphere Streams is quite suitable for scientific calculations since it integrates complex data types for vector and matrix operations, as well as extended libraries, which we need to use extensively in the Data Assimilation system,” Dr. Kalpakis adds. “InfoSphere Streams offers scalability and strong support for data stream manipulation that will make developing an extensible system and adding functionality much easier. Moreover its mixed-mode code design also helps us add generic code to handle different situations and hence makes the code easier to manage. In addition, the support InfoSphere Streams provides for the operator graph structure allows us to conceptualize and visualize the system design, reduce the development effort and facilitate the design of a clear extensible system.”

IBM continues to provide strong technical support in promoting InfoSphere Streams,” concludes Dr. Kalpakis. “We are excited to be part of demonstrating the Data Assimilation prototype. IBM’s participation in this effort was crucial in terms of resources and technical guidance. Having a close collaboration with the InfoSphere Streams team was extremely critical for us to reach this point. We are very grateful for their guidance, support and encouragement.”
The inside story: Getting there

According to Dr. Kalpakis, the Data Assimilation project started with an idea to align environmental research with IBM’s Smarter Planet initiatives.

“The inspiration for this project was driven by a few key factors,” Dr. Kalpakis explains. “First, we have the increasing availability of near real-time measurements and various sensors, including space borne sensors and those aboard unmanned aircraft vehicles (UAVs). Second, we have increasing demand for continuous, near real-time forecasts about the future state of the world and our environment. So we had the availability of new data and demand for better forecasts, but limited means to integrate observations with our dynamic mathematical system models to gain the actionable insight. These factors aligned closely with IBM’s Smarter Planet initiative. So we started talking to the IBM InfoSphere Streams group and we got the project started.”

The biggest incentive was to demonstrate the value of the data assimilation techniques. The researchers proposed starting with a simple prototype that exhibits most of the characteristics of the target application for analyzing pollutant dispersion.

“By working out the details using the prototype, we were able to gain support from the various University stakeholders for the project,” says Dr. Kalpakis.

For more information
To learn more about how IBM can help you transform your business, please contact your IBM sales representative or IBM Business Partner.

To learn more about IBM big data, visit: ibm.com/software/data/bigdata

To learn more about IBM InfoSphere Streams, visit: ibm.com/software/data/infosphere/streams

To get involved in the conversation:
www.smartercomputingblog.com/category/big-data

For more information about the University of Maryland, Baltimore County, visit: www.umbc.edu