Air Quality Modeling on the Cloud: **Pairing Data Availability and High-Performance Clusters**

by Elizabeth Adams, Christos I. Efstathiou, Zac Adelman, Kristen Foley, Fahim Sidi, Winston Hao, and Saravanan Arunachalam

This article reflects on how cloud computing makes state-of-the-science air quality models, including CMAQ, more accessible and shareable among non-profits, private entities, universities, and government organizations around the world.

Modelers using Air Quality Models (AQMs) on the cloud seek to accelerate the solution of global-to-local air quality and health problems by making them easier to run, lowering costs, and expanding access. The "cloud" refers to on-demand, pay-as-you-go computing services. Major cloud vendors provide user-defined high-performance computing (HPC) clusters (with multi-core central processing units (CPUs), memory, networking, pre-installed software, and data storage), using cutting-edge technology beyond what is available in-house at universities or government agencies.[1] The cloud enables running state-of-the-science AQMs on-demand by anyone in the world with access to a web browser and the internet. This paradigm shift could help modeling groups with either no access to HPC systems, or who compete for resources from a multi-user system.

Seeing a growing opportunity for cloud-based solutions for environmental modeling, the U.S. Environmental Protection Agency (EPA) and the Lake Michigan Air Directors Consortium (LADCO) with the help of the Community Modeling and Analysis System (CMAS) Center established the Modeling in the Cloud Workgroup in 2021.[2] The workgroup serves as a forum for the AQM community to discuss developing and executing modeling projects using different cloud service providers (CSPs) in support of regulatory and research applications. Presentations by CSPs; including Amazon Web Services (AWS) and Microsoft Azure, and by workgroup members about their experiences porting existing modeling capabilities to the cloud, help to identify knowledge gaps, and find cloud-based solutions to technical issues.

Scientists from LADCO, EPA, and the CMAS Center shared their experiences developing and testing modeling platforms on the cloud. LADCO is using cloud computers to supplement and replace their in-house compute cluster, access scalable HPC resources, circumvent information technology (IT) barriers at state air agencies, and enable collaboration in modeling and data analysis across their membership. The modeling platform (MP) developed by LADCO and available for the community includes software for meteorological and AQMs, including visualization and analysis of model inputs and outputs for a specific year, spatial domain, and grid resolution.

Cloud-Based Solutions for CMAQ

Recognizing the direction of LADCO and the needs of the larger AQM community, EPA, and the CMAS Center are investigating cloud-based solutions for running the Community Multiscale Air Quality Modeling (CMAQ) system.[3] CMAQ, a Linux-based suite of software used to model ozone, particulates, toxics, and acid deposition, requires significant resources to run that may be out of reach for some modeling groups. EPA and the CMAS Center worked with cloud engineers from the University of North Carolina (UNC), and CSPs to identify the best-fit cloud resources by optimizing the CMAQ MP for cost and run-time.

The CMAS Center installed and tested the MP and created the tutorials. On the cloud, charges accrue for as long as resources are in-use. The CMAQ on AWS Tutorial[4] reduces the use-time required, as the MP is pre-installed when a virtual machine (VM) or cluster is created. CMAQ can be run on a 4-core VM (US$0.31/hr) for a small urban domain over New York City, completing in 2.8 minutes of run-time per simulation day.

For more compute-intensive jobs, Amazon's HPC system ParallelCluster uses a job scheduler to deploy the compute nodes to run CMAQ. This workflow ensures that the HPC compute nodes are only in-use during the compute-intensive CMAQ model run, with the scheduler VM (US$.31/hr) used for set-up and analysis. Running CMAQ for a larger 12-km contiguous U.S. domain (12US1) costs US$2.50 per simulation day using 192-core (3 nodes) (US$1.6832/hr/node) and completes in 26 minutes of run-time. An annual simulation for the 12US1 domain is estimated to cost US$1,000 (compute: US$912, scheduler: $49, storage (Amazon FSx for Lustre): US$39), with a run-time of less than one week.

The CMAQ on Azure Tutorial[5] provides instructions to install and run CMAQ on a single VM and on Azure's HPC system Cycle Cloud, allowing modelers to compare the user experience, run-time, and cost of running CMAQ on AWS and Azure. The CMAS Center developed a CMAQ on AWS Workshop to run CMAQ and analyze the results within a web browser. This AWS-sponsored workshop was held at the 22nd Annual CMAS Conference, and it is now available as on-demand training on the AWS Workshop Studio website.[6]

The method of using AWS ParallelCluster with pre-installed MPs provides modelers with an open-source, reproducible workflow to create on-demand and scalable HPC clusters. This approach gives modeling groups the opportunity to rapidly run CMAQ and analyze results. While pre-installed, the MP software is obtained from EPA's public GitHub code repository,[7] and modelers can readily build and run for their choice of available CMAQ options.

CMAQ benchmark scaling and cost performance charts in the tutorials show that price performance is dependent on domain size, grid resolution, and the number and type of CPUs utilized. HPC compute nodes are compute-optimized CPUs with high core densities, located in close-proximity for optimal network performance, and heavily discounted to encourage use. An AWS blog describes how to accurately compare costs for in-house versus cloud HPC clusters.[8] Running AQMs on the cloud is an alternative to owning clusters, and reduces the risk that an in-house cluster will be either under-powered (limiting modeling domain and scenario length, or delaying time-to-solution) or under-utilized (used to meet a deadline, but not needed for the full lifetime of the cluster) after completing modeling in support of a State Implementation Plan or to meet other objectives.

EPA and the CMAS Center noticed significant synergies when HPC clusters are pre-installed with software and configured to use data hosted in the cloud, providing modelers a

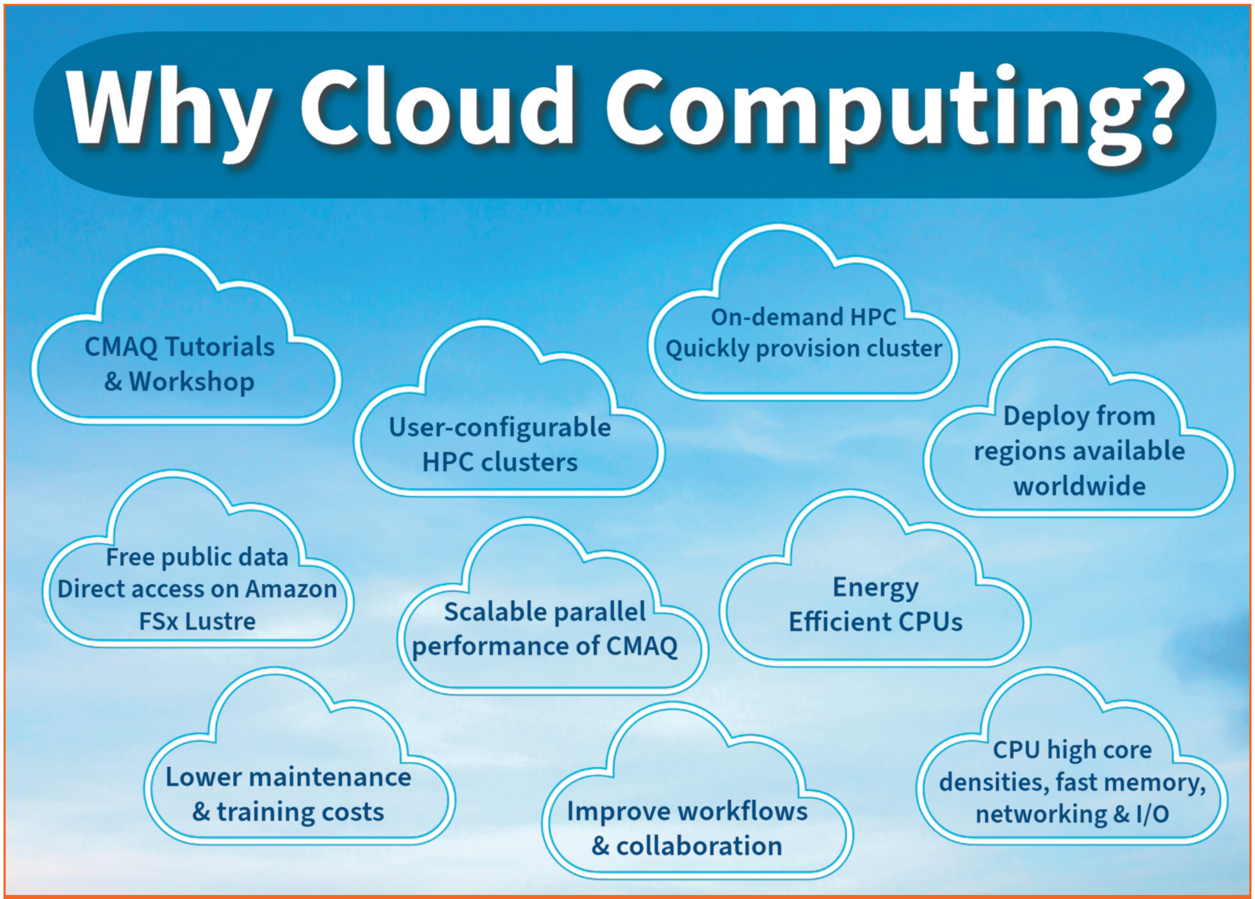
Why Cloud Computing?
CMAQ Tutorials & Workshop
User-configurable HPC clusters
On-demand HPC Quickly provision cluster
Deploy from regions available worldwide
Free public data Direct access on Amazon FSx Lustre
Scalable parallel performance of CMAQ
Energy Efficient CPUs
Lower maintenance & training costs
Improve workflows & collaboration
CPU high core densities, fast memory, networking & I/O

Figure 1. Benefits of Cloud Computing. Each cloud in the figure helps illuminate a benefit of using cloud computing for air quality modeling.

choice of available MPs. To facilitate collaboration, 50 TB of CMAQ datasets are freely accessible from the CMAS Data Warehouse on the AWS Open Data Registry.[9] Open-data datasets currently available include CMAQ-ready inputs from EPA's Air QUAlity TimE Series (EQUATES)[10] project for 2017, several recent National Emissions Inventory (NEI) MPs (2016, 2019, 2020),[11] and the CMAQv5.4 MP for 2018. The CMAS Center's Dataverse site[12] hosts an inventory of all available open-data datasets with metadata. Open-data datasets may be downloaded to in-house clusters free of charge. For cloud-based modeling, open-data datasets may be copied at no cost to Azure Cycle Cloud and AWS ParallelCluster or accessed directly using Amazon FSx for Lustre file system as demonstrated in the AWS Tutorial and Workshop. Typically, uploading data to the cloud is free, storage costs US$0.023 per GB for the first 50 TB on Amazon S3, and downloading costs US$0.09 per GB for the first 10 TB/month as listed in the S3 pricing guide.[13]

In addition to these success stories from EPA, LADCO, and the CMAS Center, AQMs are being explored in the cloud by federal and state governments, academia, private entities, and non-profit organizations. The New York State Department of Environmental Conservation downloaded the CMAQv5.4 MP for 2018 from the CMAS Data Warehouse for CMAQ modeling on their in-house HPC system, taking advantage of the cost-free AWS Open Data Registry and fast transfer speeds. Anecdotally, learning to run CMAQ with cloud computing has been eased by using the CMAS Tutorials. Verisk's Atmospheric and Environmental Research was an early adopter of a cloud-first strategy for AQMs, to shift capital cost to operational cost, outsource hardware maintenance, and avoid paying for idle resources.[14,15] The National Oceanic and Atmospheric Administration (NOAA) uses AWS ParallelCluster to process NOAA's operational Global Forecast System meteorology on-demand, allowing modelers to create CMAQ-ready meteorological input for their domain.[16] PSE Healthy Energy, a nonprofit research institute, relies on cloud-based implementations of AQMs to study how energy policy impacts air quality, public health and the environment. Lawrence Berkeley National Laboratory has secured multi-year cloud-vendor discounted, systemwide payer agreements to encourage modelers to try their workloads on the cloud.[17] Cost-effective cloud-based modeling systems such as WRF Cloud reduce idle computer server time while making weather forecasting accessible to new users.[18]

Innovations by CSPs continue to improve the price, compute, and energy performance of HPC clusters,[19] and development of sharable MPs with installed software, scripts and datasets, improves AQM workflows. The benefits of running

CMAQ in the cloud are illuminated in Figure 1. Cloud-based HPC enables new applications and workforce development, allowing the modeling community to experience designing, running, and analyzing state-of-the-science AQMs, beyond their current capabilities. Training and open-data opportunities for AQMs in the cloud are now readily accessible. The process of developing cloud-based HPC solutions will be unique to each organization, but the goal is to share pre-installed MPs and tutorials using open-data and compute-intensive cloud-based AQMs that can be widely adopted and adapted by a worldwide user base. Unique and tailored modeling solutions can be achieved by modifying the existing MPs. Promotional credits are available from AWS's Amazon Sustainability Data Initiative,[20] Microsoft for Startups Founders Hub,[21] or Azure's Government Free Trial.[22] Cloud computing makes state-of-the-science AQMs more accessible and shareable among non-profits, private entities, universities, and government organizations (U.S. and internationally), and serves to grow the community and accelerate exploration and discoveries. **em**

Elizabeth Adams (corresponding author) is a Research Associate with the Institute for the Environment, Community Modeling and Analysis System (CMAS) at The University of North Carolina at Chapel Hill; **Christos I. Efstathiou** is a Scientist with Physicians, Scientist, and Engineers for Healthy Energy, Oakland, CA; **Zac Adelman** is Executive Director of the Lake Michigan Air Directors Consortium, Chicago, IL; **Kristen Foley** and **Fahim Sidi** are both with the Center for Environmental Measurement and Modeling, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC; **Winston Hao** is a Research Scientist with the New York State Department of Environmental Conservation, Albany, NY; **Saravanan Arunachalam** is Deputy Director of the Institute for the Environment at The University of North Carolina at Chapel Hill, Chapel Hill. Email: **lizadams@email.unc.edu**.

Acknowledgment: The U.S. Environmental Protection Agency (EPA) partially funded and collaborated in the research described here under contract EP W 16 014 to the University of North Carolina at Chapel Hill. The authors also thank Amazon Web Services, Microsoft Azure, and Research Computing at Information Technology Services at the University of North Carolina at Chapel Hill for cloud engineering support. Microsoft Azure provided UNC with computing credits to help with running CMAQ on Azures HPC system Cycle Cloud.

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily represent the views or policies of EPA.

References

1. Dancheva, T.; Alonso, U.; Barton, M. Cloud benchmarking and performance analysis of an HPC application in Amazon EC2; *Cluster Computing* (2023).
2. Modeling in the Cloud Workgroup. See https://github.com/CMASCenter/modeling-in-the-cloud/ (accessed 2023-11-28).
3. U.S. Environmental Protection Agency (EPA) Community Multiscale Air Quality Model (CMAQ) Website. See https://www.epa.gov/cmaq (accessed 2023-11-28).
4. CMAQ on AWS Tutorial. See https://pcluster-cmaq.readthedocs.io (accessed 2023-11-28).
5. CMAQ on Azure Tutorial. See https://cyclecloud-cmaq.readthedocs.io (accessed 2023-11-28).
6. CMAQ on AWS Workshop. See https://catalog.workshops.aws/cmaq-tutorial/en-US (accessed 2023-11-28).
7. U.S. Environmental Protection Agency (EPA) Community Multiscale Air Quality Model (CMAQ) GitHub Site. See https://github.com/USEPA/CMAQ/ (accessed 2023-11-28).
8. AWS Cloud Financial Management. See https://aws.amazon.com/blogs/aws-cloud-financial-management/five-things-you-should-do-to-create-an-accurate-on-premises-vs-cloud-comparison-model/ (accessed 2023-11-28).
9. CMAS OpenData Registry. See https://registry.opendata.aws/cmas-data-warehouse/ (accessed 2023-11-28).
10. U.S. Environmental Protection Agency (EPA) Air QUAlity TimE Series (EQUATES). See www.epa.gov/cmaq/equates (accessed 2023-11-28).
11. U.S. Environmental Protection Agency (EPA) Emissions Modeling Platforms. See https://www.epa.gov/air-emissions-modeling/emissions-modeling-platforms (accessed 2023-11-28).
12. AWS S3 Pricing. See https://aws.amazon.com/s3/pricing/ (accessed 2023-11-28).
13. CMAS Dataverse. See https://dataverse.unc.edu/dataverse/cmascenter (accessed 2023-11-28).
14. Alvarado, M. (2017). "Lessons Learned About Running Weather, Climate, and Air Quality Models on the Amazon Cloud." Presented at the NCAR Cloud Modeling Research in the Cloud Workshop, Boulder, CO, May 31, 2017; https://www.unidata.ucar.edu/events/2017CloudModelingWorkshop/presentations/01_1430_Alvarado_NCAR_2017_Cloud_v4.pdf.
15. Alvarado, M. (2021). "CMAQ inverse modeling, WRF downscaling, and MPAS simulations on the cloud: Challenges and best practices." Presented at the 20th Annual CMAS Conference, Chapel Hill, NC, October 20, 2021; https://cmascenter.org/conference/2021/slides/alvarado-cmaq-wrf-mpas-cloud-2021.pdf.
16. Campbell, P.C.; Jiang, W.; Moon, Z.; Zinn, S.; Tang, Y. (2023). NOAA's Global Forecast System Data in the Cloud for Community Air Quality Modeling; Atmosphere 2023, 14 (7), 1110; https://doi.org/10.3390/atmos14071110.
17. Cloud Computing for Science at LBL. See https://it.lbl.gov/cloud-computing-for-science-at-lbl/ (accessed 2023-11-28).
18. WRF Cloud: Powerful Forecasting Made Easy. See https://www.wrfcloud.com/ (accessed 2023-11-28).
19. What we can learn from the energy efficiency of supercomputers?; New Electronics.com; https://www.newelectronics.co.uk/content/features/what-we-can-learn-from-the-energy-efficiency-of-supercomputers/ (accessed 2023-11-28).
20. AWS Promotional Credits. See https://aws.amazon.com/government-education/sustainability-research-credits/ (accessed 2023-11-28).
21. Microsoft for Startups Founders Hub. See https://startups.microsoft.com/blog/credit-levels/ (accessed 2023-11-28).
22. Microsoft Azure Government Free Trial. See https://azure.microsoft.com/en-us/pricing/offers/ms-azr-usgov-0044p (accessed 2023-11-28).