atmosphere

Article

NOAA's Global Forecast System Data in the Cloud for Community Air Quality Modeling

Patrick C. Campbell [1,2,*], Weifeng (Rick) Jiang [2], Zachary Moon [2,3], Sonny Zinn [2] and Youhua Tang [1,2]

[1] Cooperative Institute for Satellite Earth System Studies (CISESS), Center for Spatial Information Science and System (CSISS), George Mason University, 4400 University Drive, Fairfax, VA 22030, USA; youhua.tang@noaa.gov

[2] Air Resources Laboratory, NOAA, NCWCP, 5830 University Research Ct., College Park, MD 20740, USA; rick.jiang@noaa.gov (W.J.); zachary.moon@noaa.gov (Z.M.); sonny.zinn@noaa.gov (S.Z.)

[3] Earth Resources Technology (ERT), Inc., Laurel, MD 20707, USA

* Correspondence: patrick.c.campbell@noaa.gov

Abstract: Herein, we present the initial application of the NOAA-EPA Atmosphere-Chemistry Coupler (NACC) in the cloud ("NACC-Cloud", version 1), which processes NOAA's operational Global Forecast System version 16 (GFSv16) meteorology on-demand and produces model-ready meteorological files needed to drive U.S. EPA's Community Multiscale Air Quality (CMAQ) model. NACC is adapted from the U.S. EPA's Meteorology-Chemistry Interface Processor version 5 (MCIPv5) and is used as the primary model coupler in the current operational NWS/NOAA air quality forecasting model. The development and use of NACC-Cloud in this work are critical to provide the scientific community streamlined access to NOAA's operational GFSv16 data and user-defined processing and download of model-ready, meteorological input for any regional CMAQ domain worldwide. The NACC-Cloud system was implemented on the Amazon® Web Services High-Performance Computing platform, and results from this work show that the NACC-Cloud system is immediately beneficial to the air quality modeling community worldwide.

Keywords: cloud computing; Global Forecast System; air quality modeling; community

check for updates

Citation: Campbell, P.C.; Jiang, W.; Moon, Z.; Zinn, S.; Tang, Y. NOAA's Global Forecast System Data in the Cloud for Community Air Quality Modeling. *Atmosphere* **2023**, *14*, 1110. https://doi.org/10.3390/atmos14071110

Academic Editor: Philippe Thunis

Received: 18 May 2023
Revised: 26 June 2023
Accepted: 30 June 2023
Published: 4 July 2023

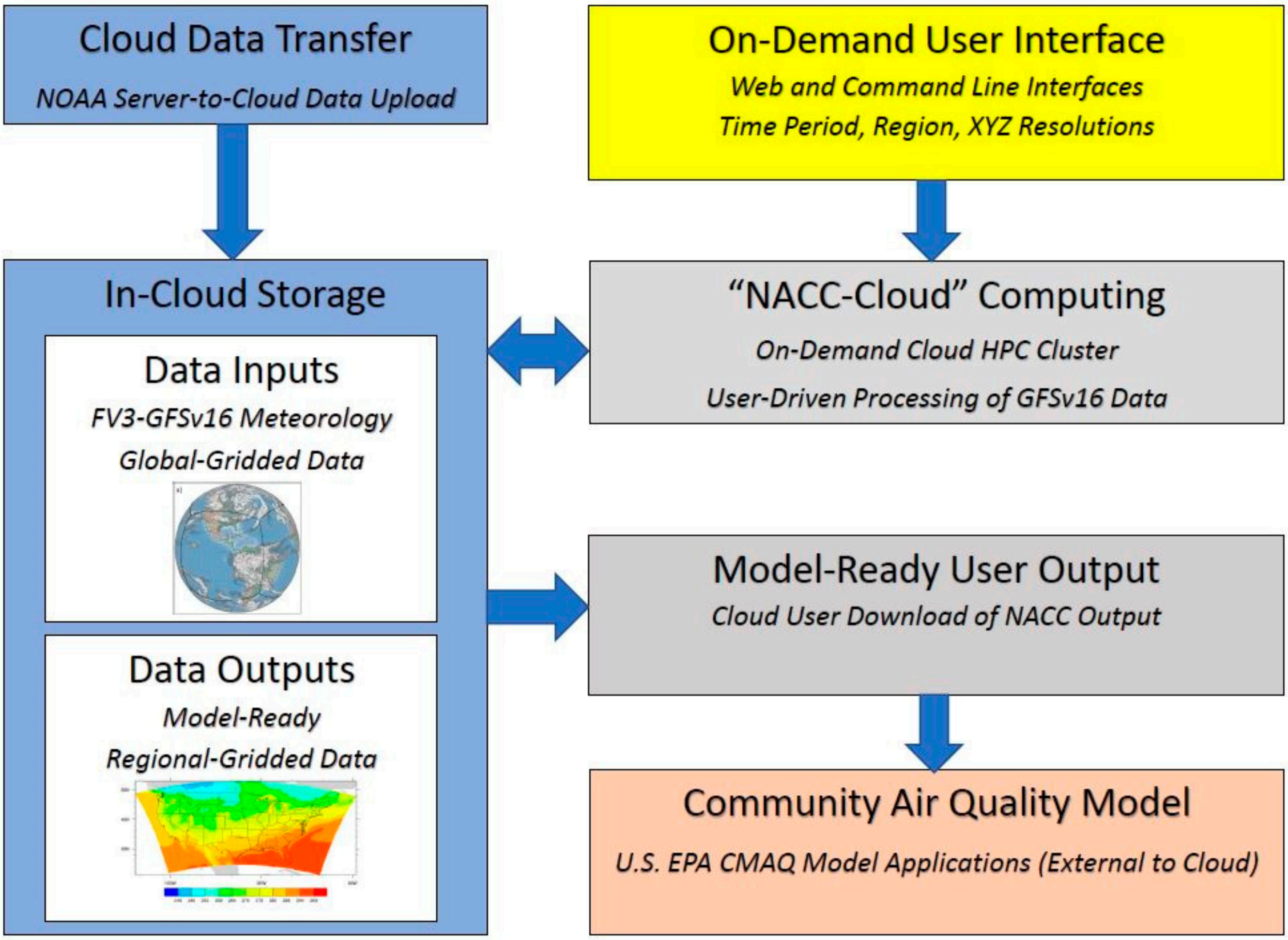
1. Introduction

In contrast to on-premises High-Performance Computing (HPC) systems, cloud computing platforms are more readily being used for a wide array of geophysical modeling applications worldwide. This is because on-premises HPCs require significant capital investment and have high long-term operational costs, while cloud computing has more flexibility, constantly refreshed hardware, high reliability, geo-distributed computing and networking, and a "pay for what you use" pricing model. Specific to atmospheric chemistry and composition (ACC) models, there are numerous applications of widely used models that have been recently moved into the cloud. Some examples include (1) the Weather Research and Forecasting (WRF) model [1] in the Google Cloud Platform (GCP, https://cloud.google.com/blog/topics/hpc/weather-forecasting-using-the-wrf-model-on-google-cloud, accessed on 16 June 2023), (2) the Goddard Earth Observing System (GEOS)-Chemistry (GEOS-Chem) model [2] (10.5281/zenodo.7383492) in the Amazon® Web Services (AWS) Cloud platform (https://cloud-gc.readthedocs.io/, access on 16 June 2023), and (3) NOAA's Unified Forecast System Weather Model (https://ufs-srweather-app.readthedocs.io/, accessed on 16 June 2023) on the GCP, AWS, and Microsoft Azure (https://azure.microsoft.com/en-us, accessed on 16 June 2023) cloud platforms (https://epic.noaa.gov/the-epic-team-has-successfully-run-the-ufs-weather-model-on-all-3-csps/, accessed on 16 June 2023). These coupled ACC models have readily been moved into the cloud, and there are studies that have applied them for scientific investigations, improved computational performance and scaling, and the development of novel cloud-based tools for model code testing and analysis [3,4]. Up to this point, however, there has been no demonstration of moving well-vetted, operationally based ACC model components into the cloud to better facilitate community air quality modeling practices.

For almost 20 years, NOAA's National Weather Service (NWS) has been providing air quality forecasts to the public based on a foundation of modeling work provided by NOAA's Air Resources Laboratory. First implemented into operations in September 2004, the National Air Quality Forecasting Capability (NAQFC) (https://www.weather.gov/sti/stimodeling_airquality_predictions, accessed on 5 April 2022) has been regularly developed, updated, and applied across both the NOAA operational and greater scientific communities [5–10]. The primary goal of the NAQFC is to help protect the public against the harmful effects of air pollution and associated costly medical expenses. Most recently, the NAQFC modeling system has been substantially updated to include coupling of the latest NOAA/NWS operational Finite-Volume Cubed-Sphere (FV3) air quality modeling Version 16 (FV3GFSv16) [11] meteorology with the Community Multiscale Air Quality (CMAQ) model version 5.3.1 [12,13]. The major upgrade to GFSv16 largely improves the meteorological model forecast performance while also providing enhanced forecast products, and has extended the forecast outlook period from 48 to 72 h [10].

The meteorological–chemical coupling of the GFSv16 to the regional CMAQ v5.3.1 model is achieved via the development of the NOAA-EPA Atmosphere Chemistry Coupler (NACC) (NACC, i.e., "knack", meaning an acquired skill) (https://github.com/noaa-oar-arl/NACC, accessed on 1 July 2023). NACC has been adapted from the US EPA's Meteorology-Chemistry Interface Processor (MCIP) version 5 [14]. The NACC and CMAQ coupling (hereafter referred to as NACC-CMAQ) involves a number of structural and scientific advancements, where NACC delivers the capability to couple the operational GFSv16 meteorological forecasts to CMAQ for air quality modeling applications in the scientific research community [10].

The global 3D, gridded GFSv16 data sets are very large (on the order of TBs/month), and thus it can be very cumbersome to use basic transfer and processing tools for usage in the scientific modeling community. Thus, the use of the cloud in this work is critical to provide the scientific community streamlined access to the near-real-time GFS forecast output and to facilitate user-defined NACC processing of GFS data to generate model-ready, meteorological input for any U.S. EPA Community Multiscale Air Quality (CMAQ) domain and air quality application worldwide (i.e., "NACC-Cloud" version 1) (https://nacc.arl.noaa.gov/nacc/, accessed on 1 July 2023). Cloud computing and storage platforms are desirable as they are highly customizable, on-demand, and much more scalable than traditional local servers. Such a cloud interface for GFS-driven CMAQ applications was not currently available and is advantageous to the air quality scientific community and beyond (Figure 1).



Cloud Data Transfer

NOAA Server-to-Cloud Data Upload

On-Demand User Interface

Web and Command Line Interfaces

Time Period, Region, XYZ Resolutions

In-Cloud Storage

Data Inputs

FV3-GFSv16 Meteorology

Global-Gridded Data

Data Outputs

Model-Ready

Regional-Gridded Data

"NACC-Cloud" Computing

On-Demand Cloud HPC Cluster

User-Driven Processing of GFSv16 Data

Model-Ready User Output

Cloud User Download of NACC Output

Community Air Quality Model

U.S. EPA CMAQ Model Applications (External to Cloud)

Figure 1. Schematic representation of the "NACC-Cloud" version 1 product to facilitate GFS-driven CMAQ applications for the scientific air quality modeling community.

2. Data and Methodology

2.1. NOAA's Global Forecast System (GFS) Version 16 Data

The GFS model was updated (February 2021) from v15.3 to v16 by NOAA's Environmental Monitoring Center (EMC) and included numerous changes and upgrades such as the incorporation of the FV3-based dynamical core (Figure 2), enhanced horizontal and vertical model resolutions, improved physical parameterizations, data assimilation, and enhanced forecast products (e.g., hourly output out to 72 h (3 d) forecast) [11]. Ultimately, these upgrades and the availability of operational global meteorological data were a major impetus in the development of NACC to couple GFSv16 to the CMAQ model for regional air quality modeling applications.

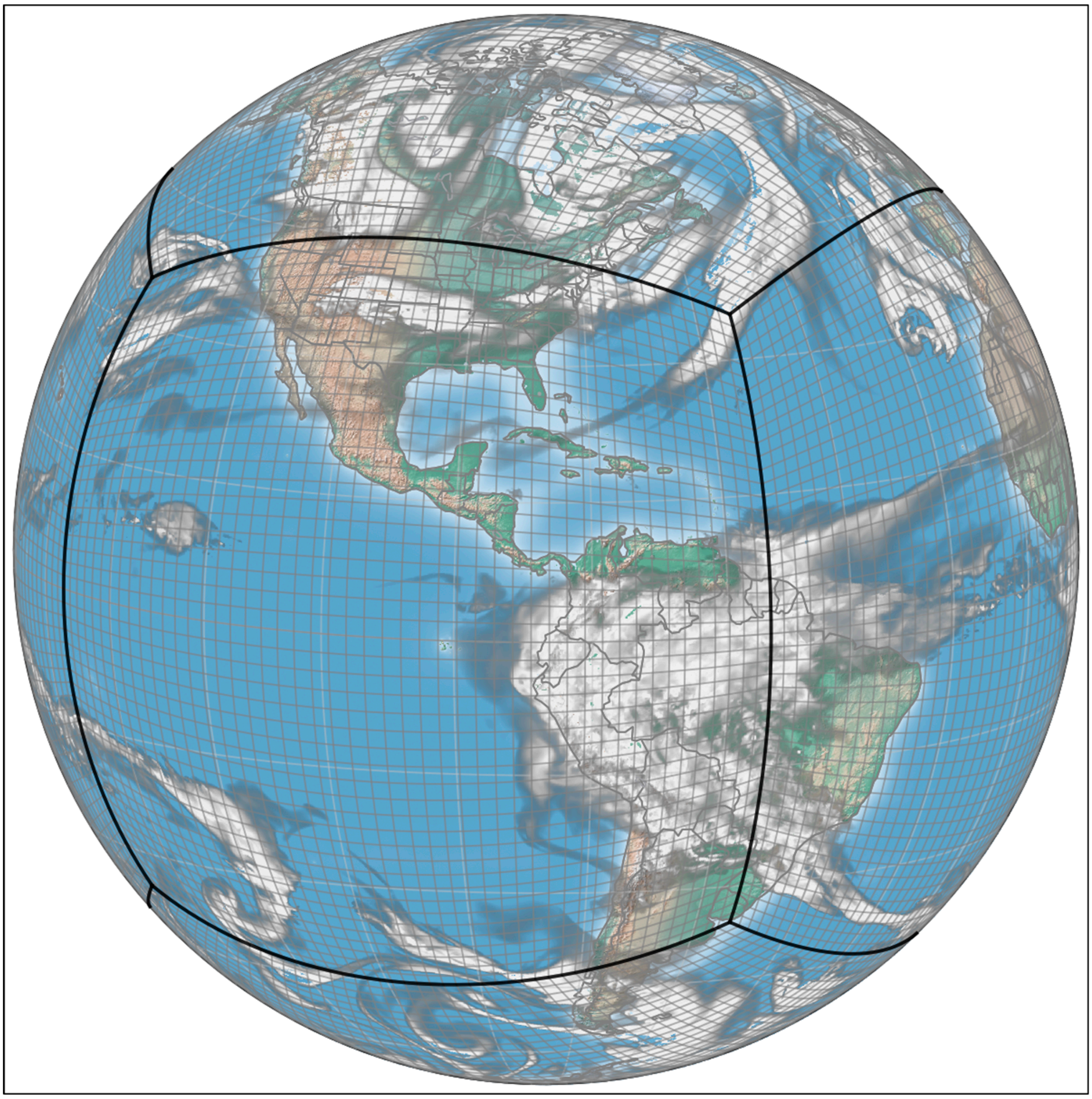


Figure 2. The native Finite Volume Cubed Sphere (FV3) gnomonic grid at 2° resolution for the GFSv16 weather model (image courtesy of Dusan Jovic, NOAA). Figure adapted from Campbell et al., 2022 [10].

2.2. The NOAA-EPA Atmosphere Chemistry Coupler (NACC)

The development of NACC stemmed from the need to update and use a coupled GFSv16 and recent CMAQv5.3.1 model as the latest NOAA/NWS operational NAQFC [10]. NACC is based on U.S. EPA's MCIPv5 [14] (https://github.com/USEPA/CMAQ, accessed on 25 January 2023); however, the major difference of NACC is that it includes a variable-dependent bilinear or nearest-neighbor horizontal interpolation of the GFSv16 Gaussian gridded (~13 × 13 km) fields (e.g., 2 m temperature, 2 m specific humidity, 10 m wind speed and direction, and sea level pressure) to a Lambert conformal conic (LCC) projection (i.e., "IOAPI GDTYP" option 2; see https://www.cmascenter.org/ioapi/documentation/all_versions/html/GRIDS.html for more details, accessed on 25 January 2023) at a user-defined output grid resolution via FORTRAN 90 name list controls. Conversely, MCIPv5 is only capable of handling native meteorological inputs from the WRF model, while NACC was adapted to also interpolate FV3-based GFSv16 inputs for CMAQ (Figure 3). Further details of the NACC processing and description may be found at https://github.com/noaa-oar-arl/NACC (accessed on 25 January 2023).

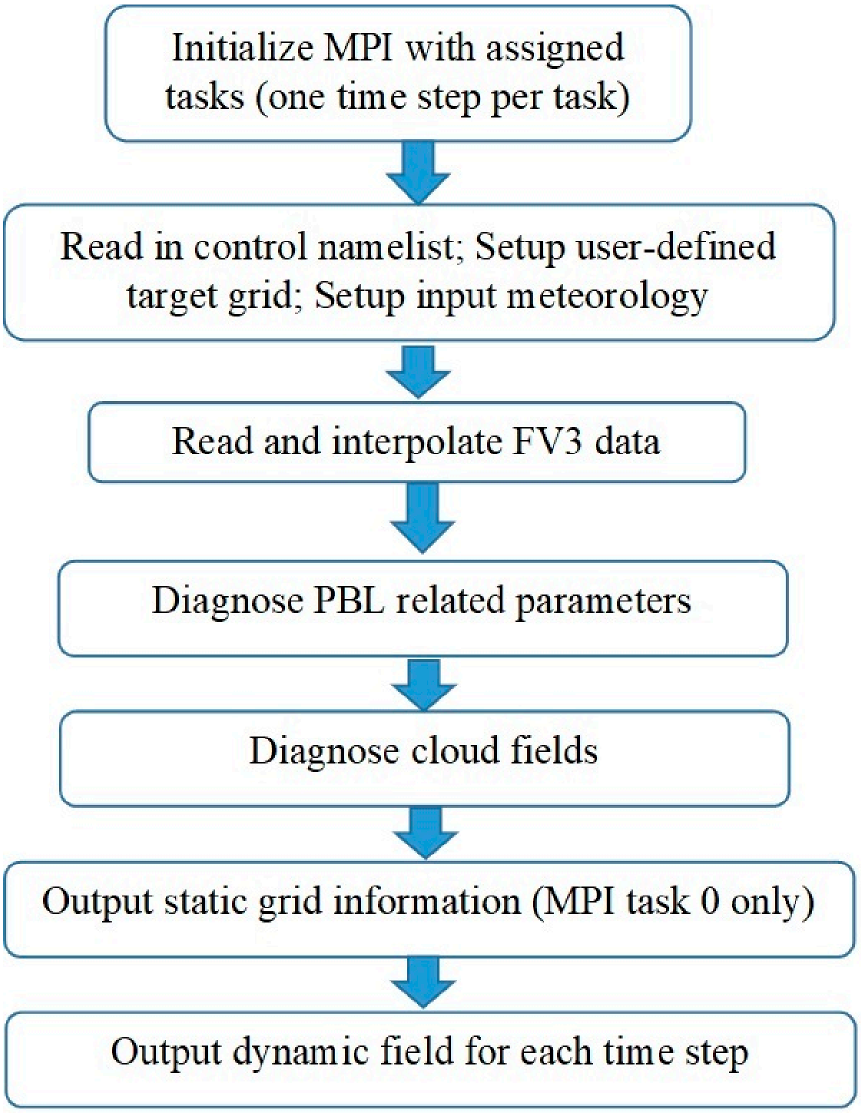

Initialize MPI with assigned tasks (one time step per task)
Read in control namelist; Setup user-defined target grid; Setup input meteorology
Read and interpolate FV3 data
Diagnose PBL related parameters
Diagnose cloud fields
Output static grid information (MPI task 0 only)
Output dynamic field for each time step

Figure 3. Flowchart of the main NACC processes.

In general, NACC defines the computational domain for CMAQ, subsequently extracts meteorological model output and interpolates both horizontally and vertically on the computational domain that is prescribed, and processes all required meteorological fields for CMAQ. Meteorological fields such as atmospheric temperature, pressure, humidity, and winds are acquired directly from the meteorological model (i.e., "passed through"), while NACC also uses the available meteorological fields to compute additional fields that are required by the CMAQ but are not part of the meteorological model's output stream (e.g., the Jacobian used for model coordinate transformations). The NACC code is parallelized using MPI and time-splitting techniques, where one time step is processed with each MPI task, which results in much improved computational efficiency. NACC outputs both static and dynamic files in I/O API format that contain geospatial and meteorological information used by CMAQ (Figure 3).

Recent updates to NACC also allow for processing meteorological LCC grid projections from NOAA's Unified Forecast System (UFS), Limited Area Model (LAM), Short-Range Weather App model (https://ufs-srweather-app.readthedocs.io/en/develop/#, accessed on 25 January 2023) for model-ready inputs to CMAQ. In other words, the interpolation-based NACC can use various meteorological outputs to drive the CMAQ model, even if they are on different grids. A recent comparison of the use of NACC-CMAQ vs. WRF-CMAQ showed that using global GFSv16 meteorology with NACC to directly drive CMAQ via interpolation is feasible and yields reasonable chemical predictions compared to the commonly used WRF approach [15]. At the time of writing this paper, the latest NACCv2.1.1 (available at: https://github.com/noaa-oar-arl/NACC/releases/tag/v2.1.1, accessed on 25 April 2023) further allows for interpolation of input global GFSv16 to I/O API GDTYP option 6 (Polar Stereographic) for Hemispheric-CMAQ applications [16–18]. The development of NACC thus serves to open a myriad of new regional and hemispheric CMAQ applications worldwide.

2.3. Amazon® Web Services HPC Cloud Platform and Configuration

Amazon® Web Services (AWS) is one of the world's most comprehensive and broadly adopted cloud computing systems. Millions of customers including government agencies are using AWS or other cloud services to lower costs, become more agile, and innovate faster. Among over 200 fully featured services, Amazon® Simple Storage Service (Amazon® S3), AWS ParallelCluster, and Amazon® FSx for Lustre are the three major ones used in the NACC-Cloud application development.

2.3.1. Amazon® S3 for NOAA-to-AWS GFSv16 NetCDF Transfer and Storage

Amazon® S3 is an object storage service created to store and retrieve an unlimited amount of data files from virtually any endpoints on the internet, offering high scalability, data availability, security, and performance. It is an ideal solution for storing large amounts of GFSv16 data (~200 GB/day, or ~6 TB/month) for the NACC-Cloud application's input when the right storage classes are chosen. Amazon® S3 offers a variety of storage classes that can be selected to meet different workload requirements, such as data access, resiliency, or cost. Among eight different S3 storage classes (https://aws.amazon.com/s3/storage-classes/, accessed on 16 June 2023) at the time of writing this paper, the S3 Glacier Instant Retrieval (IR) for archived data was chosen for storing the NACC-Cloud application's GFSv16 input data, as Glacier-IR allows for immediate data access. The AWS S3 Standard bucket was chosen for the NACC-Cloud output data, which delivers the required low latency and high throughput for user access needs due to other factors such as project cost controls.

An automated script containing AWS Command Line Interface (CLI) commands (CLI provides a consistent interface for interacting with all parts of AWS) is run daily (at 1700 UTC) on an on-premises NOAA server to automatically transfer operational GFSv16 data to an Amazon® S3 bucket: aws s3 cp $LOCAL_NOAA_SERVER_PATH/. s3://nacc-in-the-cloud/inputs/${yyyymmdd}-storage-class GLACIER_IR-recursive-exclude "*"—include "$GFSv16_FILENAMES".

2.3.2. Amazon® FSx for Lustre to Connect S3 Storage and NACC-Cloud Computing

Amazon® FSx for Lustre (hereafter referred to as FSx) provides fully managed shared storage built on the AWS high-performance file system (https://aws.amazon.com/fsx/lustre/, accessed on 16 June 2023). FSx accelerates compute workloads with shared storage that provides sub-millisecond latencies, millions of input/output operations per second, and up to hundreds of GB/s of throughput. GFSv16 input data files are accessed and processed by the NACC-Cloud HPC system through the FSx, which is linked to the specific S3 bucket where GFSv16 data is stored. To save space on the relatively expensive FSx, the space occupied by the input data files is released following NACC data processing. Meanwhile, output data files are written to this same FSx, and are then moved to a specific location within the S3 bucket, where end users can access and download output data through the NACC-Cloud application's web interface (see Section 2.4). All existing and newly uploaded GFSv16 data files are structurally visible to the NACC-Cloud system, and specific input data files will automatically be synchronized over to the system for use once the application is ready to read input for data processing. After the NACC-Cloud jobs are complete, the synced or archived output data files will be deleted from the FSx. The main commands for the NACC-Cloud FSx-to-S3 data workflow and cleanup are shown below:

1. Export output files from FSx to the S3 bucket;

```
TASK_ID = $($FSX_CLIENT create-data-repository-task --type EXPORT_TO_REPOSITORY --file-system-id $FSX_FILESYSTEM_ID --paths $LOCAL_PATH --report Enabled = true,Path = “$S3_EXPORT_PATH”,Format = “REPORT_CSV_20191124”,Scope = “FAILED_FILES_ONLY” | jq -r ‘.DataRepositoryTask.TaskId’).
```

2. Find and delete exported output files from FSx;

```
# adding archived output files to a file
find “$output_directory” -type f -print0 | xargs -0 -n 1 -P 8 sudo lfs hsm_state | grep archived | awk ‘{ print $1 }’ > “$file_list”
# replacing all colons with an empty string
sed -i ‘s/://g’ “$file_list”
# deleting all files on the list
while read file; do
rm “$file”
done < “$file_list”
```

3. Release space used by input data files on FSx;

```
sudo lfs hsm_release/fsx/inputs/{ymd}/*
```

4. Delete “old” archived output data files on the S3 bucket.

```
aws s3api delete-object --bucket $NACC_S3_BUCKET_NAME --key $NACC_OBJECT _NAME
```

2.3.3. AWS ParallelCluster-HPC Development for NACC-Cloud

The High-Performance Computing (HPC) cluster used for NACC-Cloud was created based on AWS ParallelCluster (Figure 4), which is a cluster management tool built on the open-source CfnCluster project (https://aws.amazon.com/hpc/parallelcluster/, accessed on 16 June 2023). AWS ParallelCluster uses a text file or simple graphical user interface (GUI) (introduced since version 3.5.0) to customize and configure all the necessary resources for one’s HPC applications in an automated way, and its source code is hosted on the AWS GitHub repository (https://github.com/aws/aws-parallelcluster, accessed on 16 June 2023). The AWS ParallelCluster 3.0 with Slurm as a job scheduler was used to create the cluster for NACC-Cloud v1.

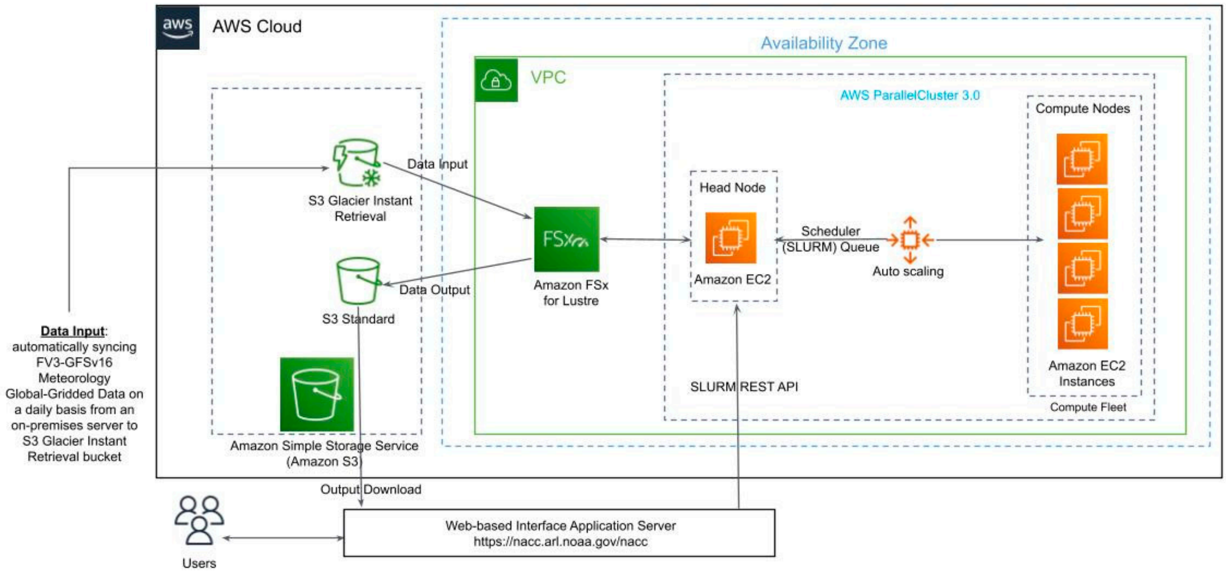

aws
AWS Cloud
Availability Zone
VPC
AWS ParallelCluster 3.0
Compute Nodes
Data Input
S3 Glacier Instant Retrieval
Head Node
FSx
Scheduler (SLURM) Queue
Auto scaling
Amazon FSx for Lustre
Amazon EC2
Data Output
S3 Standard
Data Input: automatically syncing FV3-GFSv16 Meteorology Global-Gridded Data on a daily basis from an on-premises server to S3 Glacier Instant Retrieval bucket
SLURM REST API
Amazon EC2 Instances
Compute Fleet
Amazon Simple Storage Service (Amazon S3)
Output Download
Web-based Interface Application Server
https://nacc.arl.noaa.gov/nacc
Users

Figure 4. Schematic representation of the streamlined AWS-based cloud development of Amazon® S3 storage, Amazon® FSx, and AWS ParallelCluster for the NACC-Cloud product.

The major elements used for the cluster are the head node's instance type—c5n.2xlarge; OS—alinux2; region—us-east-1; scheduler-slurm; compute node's instance type—r6i.32xlarge (MaxCount: 4 and MinCount: 0) and capacity type—ONDEMAND; and shared storage type—FsxLustre (ImportPath: s3://nacc-in-the-cloud/inputs and ExportPath: s3://nacc-in-the-cloud/outputs) (Figure 4).

2.3.4. NACC-Cloud Data Flow for User Download

GFSv16 meteorology global-gridded data files as input data are uploaded from an on-premises NOAA server to an Amazon® S3 Glacier Instant Retrieval (IR) bucket daily (Figure 4). When a user submits a processing request through the NACC-Cloud web interface, input data is synced automatically over to the Amazon® FSx mounted in the cluster's head node and accessible from all compute nodes. Once the processing request is finished, the application's output data files are synchronized over to the S3 bucket, from which the end user can download the results. Output data files are deleted from S3 after two weeks.

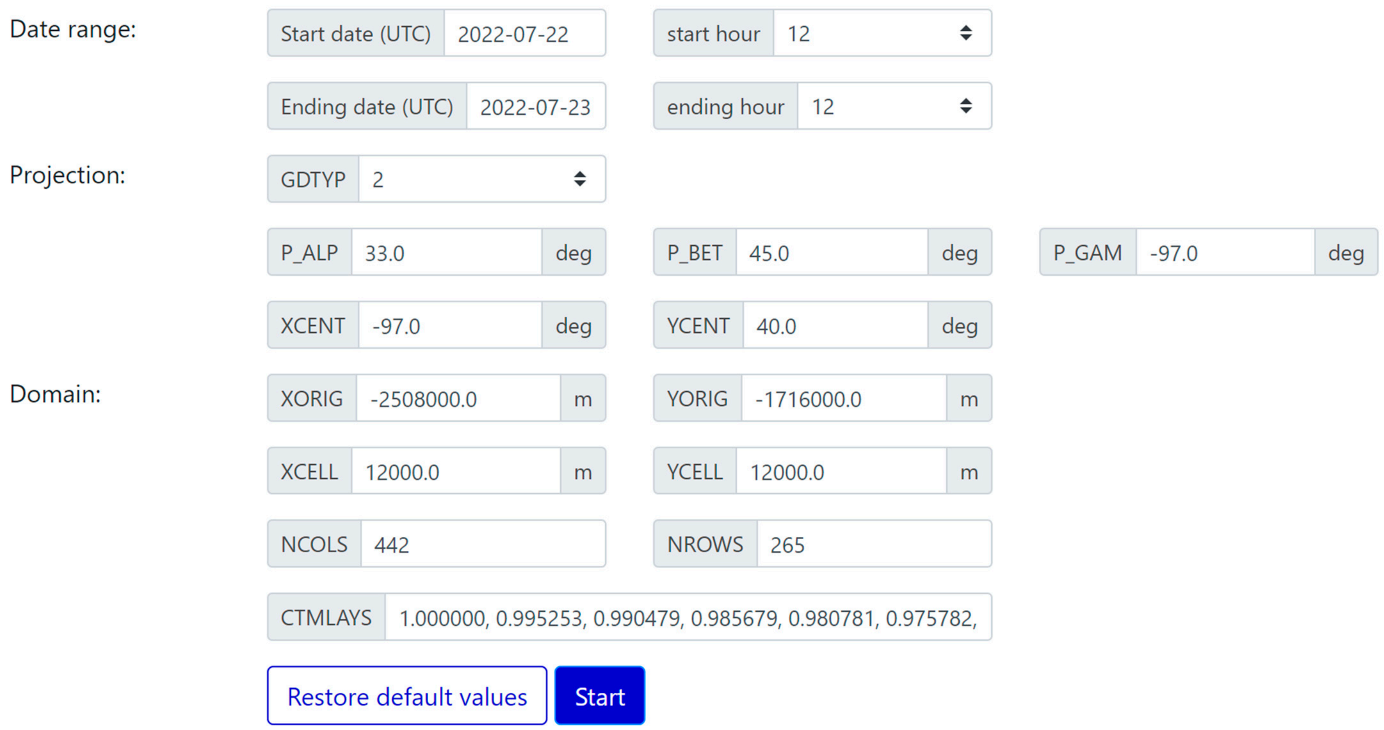
2.4. Development of a Web-Based User Interface for NACC-Cloud

The READY framework [19] (Real-time Environmental Applications and Display System; https://www.ready.noaa.gov/index.php, accessed on 16 June 2023) has successfully enabled users to run the NOAA HYSPLIT model and receive outputs via the web for many years. More recently, READY has been advanced using a next-generation Java Spring ecosystem (https://spring.io/, accessed on 16 June 2023), which was recently used to create the HYSPLIT Locusts web app (https://www.ready.noaa.gov/READYLocusts.php, https://locusts.arl.noaa.gov/, accessed on 16 June 2023). In this work, we adopt the READY framework for the NACC-Cloud web-based user interface (hereafter referred to as the "web app").

Upon authentication, users are able to access a form on the web app where they can specify NACC run parameters for CMAQ applications (Figure 5). Users can specify the time period, spatial (xyz) domains, and map projection. The map projection options are currently limited to Lambert conformal (GDTYP = 2) and polar stereographic (GDTYP = 6) for CMAQ applications, but they may be expanded in the future. The integer identifiers for the projections refer to those in I/O API (https://www.cmascenter.org/ioapi/documentation/all_versions/html/GRIDS.html, accessed on 16 June 2023), and this is documented on the form page below the form itself.

When the user submits the form ("Start" in Figure 5), the web app performs input validation, e.g., checking that the projection parameters (P_ALP, P_BET, P_GAM) are valid floats, central longitude (XCENT) is within −180–180 degrees, number of cells (NCOLS, NROWS) is not overly large (to avoid exceeding current memory constraints of NACC-Cloud), etc. If all checks pass, the web app invokes a Python script that generates a NACC name list, passing the necessary information to the Python script via a web API endpoint. The Python code does some additional validation, e.g., to estimate how much memory is needed and that the cluster will support the request (Section 3.1). If these checks pass, the Python code attempts to submit Slurm batch script job(s) (jobs get split at 12z, such that none have a time period more than 24 h) to run NACC for daily time periods (i.e., 24 h) using the Slurm REST API via a generated Python client library.

As part of the Slurm job, after NACC running is complete, data flow steps (Section 2.3.3) necessary to conserve the limited FSx space (necessary to run NACC) are employed. In the job information window, the web app presents S3 URLs to the output files that the user can easily download, e.g., with the wget Linux command.



Date range:
Start date (UTC) 2022-07-22 | start hour 12
Ending date (UTC) 2022-07-23 | ending hour 12

Projection:
GDTYP 2
P_ALP 33.0 deg | P_BET 45.0 deg | P_GAM -97.0 deg
XCENT -97.0 deg | YCENT 40.0 deg

Domain:
XORIG -2508000.0 m | YORIG -1716000.0 m
XCELL 12000.0 m | YCELL 12000.0 m
NCOLS 442 | NROWS 265
CTMLAYS 1.000000, 0.995253, 0.990479, 0.985679, 0.980781, 0.975782,

Figure 5. Example of the settings form within the NOAA-ARL READY web app for NACC-Cloud CMAQ applications, showing the default settings. https://nacc.arl.noaa.gov/nacc/setup, accessed on 16 June 2023.

3. Results and Analysis

3.1. AWS-HPC Components and Scalability for NACC-Cloud

Based on our experience in developing the NACC-Cloud version 1, we determined that cloud computing is an ideal solution for experimenting with new concepts and ideas in the environmental and geophysical modeling community. There was no huge upfront investment for NACC-Cloud, and it was possible to build an HPC environment within a short period of time, such as within days (even hours) or weeks. Furthermore, it is relatively easy to rebuild the environment required by the change of resources needed or through the limitation of funding.

Table 1 shows all major cloud components used by the NACC-Cloud application (as of April 2023). For the main HPC "head node", to accommodate changes to scripts or the NACC code, an Amazon® EC2 reserved instance (RI) is used to get a significant discount, which can be up to 72% compared to On-Demand pricing. For the "compute nodes", which are the major cost driver for the NACC-Cloud application, Amazon® EC2 On-Demand instances are chosen since this is a research application that does not run operationally (i.e., 24 × 7), where component costs are incurred only when end-users submit NACC-Cloud jobs through the web-app.

As described in Section 2.2, NACC has MPI-based parallelism, where each time (hour) runs in a separate MPI task. We ran a series of experiments with different time counts and domain sizes to examine the scalability of NACC-Cloud. We found that the maximum job memory, diagnosed using Slurm tools, scales linearly with domain size (Figure 6a). We use this linear relationship to estimate the memory needs for Slurm jobs and request the resources from the scheduler accordingly. The run time scaling (Figure 6b) is more variable. The "warmth" of the EC2 On-Demand instance/cluster when the job starts introduces variability. In general, the increases in run time are roughly linear with domain size and number of times. These results indicate that NACC scales well.

Table 1. AWS components necessary for NACC-Cloud version 1.0.

Component	Option	Notes
Head Node—c5n.2xlarge	Reserved	4 cores, 21 GB RAM, up to 25 Gbps network bandwidth
Compute Nodes—r6i.32xlarge × 4	On Demand	(1) each node: 96 cores, 1 TB RAM, 50 Gbps network bandwidth (2) assuming: 4 nodes, 4 h per day
File system—Amazon® FSx for Lustre	-	1200 GB
Data input storage—Amazon® S3 Glacier Instant Retrieval	-	(1) GFSv16 data files: increase—200 GB/day, 6 TB/month, 72 TB/year (2) Monthly cost—should be accumulated from previous months
Data output storage—Amazon® S3 Standard	-	Output for one run (one-month data as input): 3.5 GB, 35 GB/day, 1 TB/month
End User Data download	-	Output for one run (one-month data as input): 3.5 GB, 35 GB/day, 1 TB/month
Other resources	-	VPC, EBS, Elastic IP,
AWS Support	-	10% of monthly AWS usage.

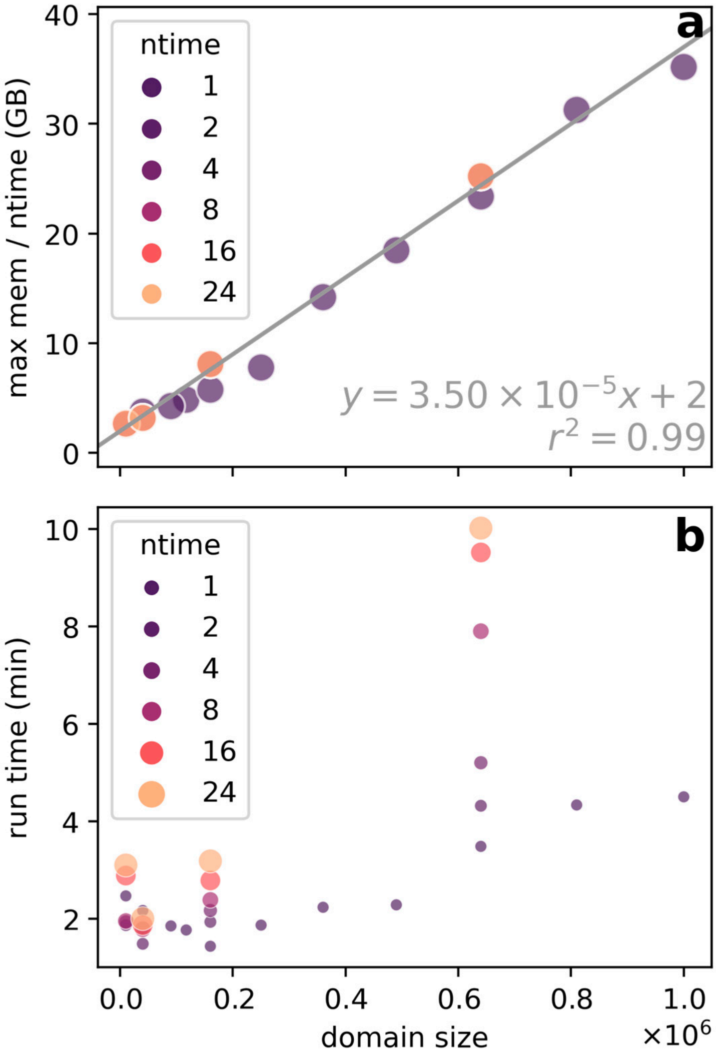


Figure 6. (**a**) NACC-Cloud memory usage scaling with domain size (number of grid cells, NCOLS × NROWS). (**b**) Variation of NACC-Cloud runtime with the number of times (MPI tasks) and domain size. Sample size n = 31.

3.2. Running NACC-Cloud for CMAQ Applications

It is relatively simple to run NACC-Cloud using the web app described in Section 2.4. The initial step to first gain access is to submit a request via Google Form to https://forms.gle/jUWFKLyY6WGKySkv6 (accessed on 25 January 2023), which is used to properly review any user access to the AWS HPC resources dedicated to NACC-Cloud. Following access approval, the user-set NACC configurations are established at run-time (rather than at compile time) via Fortran namelist variables that are controlled by the available web-app options. It is assumed that the user is familiar with the necessary CMAQ domain parameters (Figure 5; P_ALP, P_BET, P_GAM, XCENT, YCENT, XORIG, YORIG, XCELL, YCELL, NCOLS, and NROWS). The major inputs include 2D/3D GFSv16 meteorological NetCDF files stored in NRT on the AWS S3 (Figure 4), which are available from 23 March 2021 to the current day (uploaded daily at ~1700 UTC). The NACC-Cloud output files include grid information and 2D/3D meteorology land surface/soil data in IO API format needed for CMAQv5 applications (Table 2).

Table 2. NACC-Cloud generated output files needed for CMAQv5 *.

File Name	Format	Description
GRIDDESC	ASCII	Grid description file with coordinate and grid definition information
GRID_BDY_2D	I/O API	Time-independent 2-D boundary meteorology file
GRID_CRO_2D	I/O API	Time-independent 2-D cross-point meteorology file
GRID_CRO_3D	I/O API	Time-independent 3-D cross-point meteorology file
GRID_DOT_2D	I/O API	Time-independent 2-D dot-point meteorology file
LUFRAC_CRO	I/O API	Time-independent fractional land use by category
MET_BDY_3D	I/O API	Time-varying 3-D boundary meteorology file
MET_CRO_2D	I/O API	Time-varying 2-D cross-point meteorology file
MET_CRO_3D	I/O API	Time-varying 3-D cross-point meteorology file
MET_DOT_3D	I/O API	Time-varying 3-D dot-point meteorology file
SOI_CRO	I/O API	Time-varying soil properties in each soil layer

* See the I/O API grid documentation for more information (https://www.cmascenter.org/ioapi/documentation/all_versions/html/GRIDS.html, accessed on 16 June 2023).

Following completion of a NACC-Cloud run, the output geospatial and meteorological files (Table 2) are transferred from the FSx back to the AWS S3 standard bucket (Figure 4) for user download. The approximate end-to-end time from user web-based NACC-cloud job submission to file download is on the order of minutes; however, a longer processing time period may occur for large NACC domain/time requests.

3.3. Assessment of NACC-Cloud Output Meteorological Fields

Here, we compared the original GFSv16 input at a horizontal resolution of ~13 × 13 km to the interpolated NACC-Cloud output at 12 × 12 km for three LCC domains centered on the contiguous U.S, Alaska, and Hawaii (Figure 7).

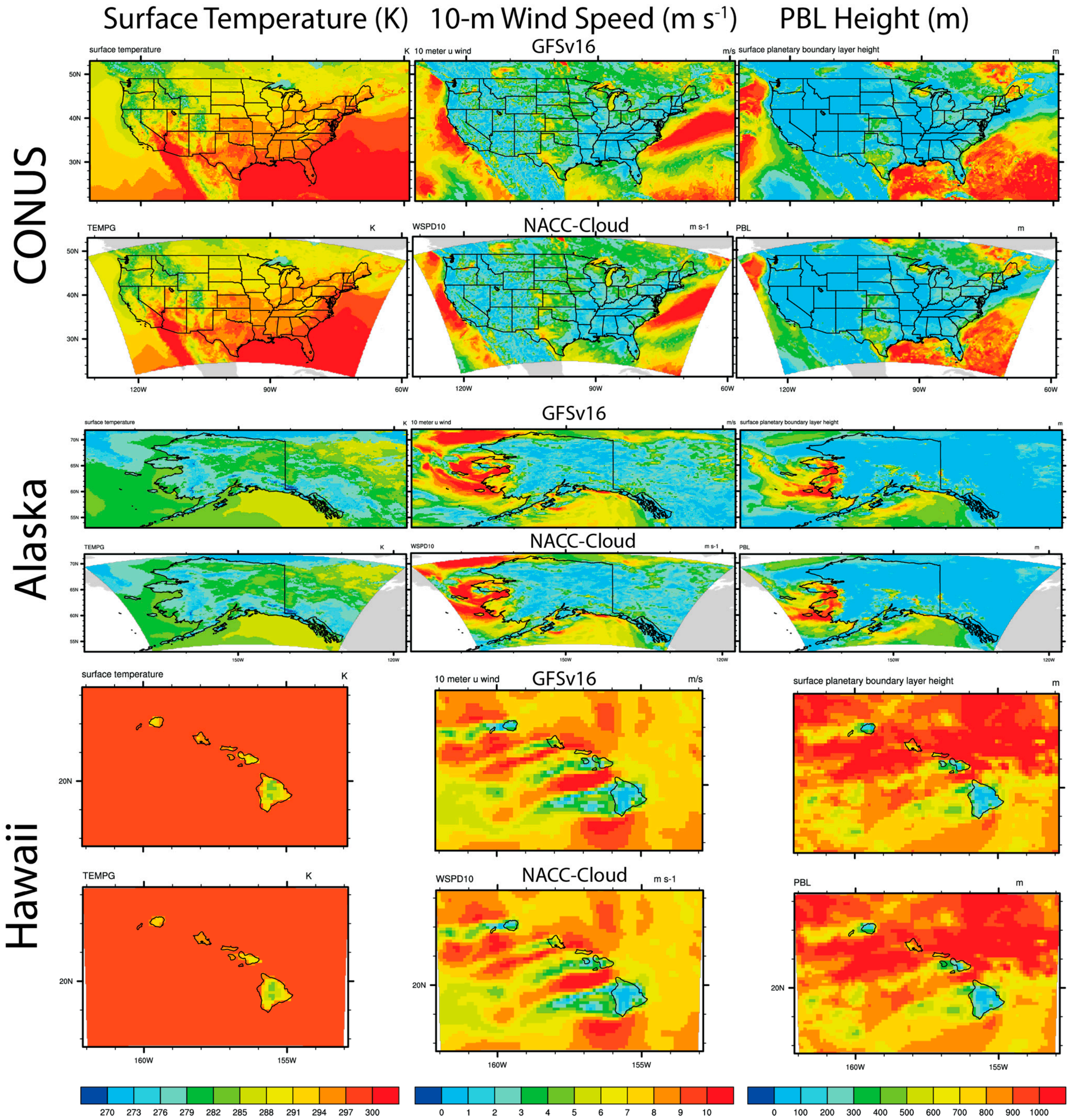


Figure 7. Comparison of input GFSv16 meteorology (i.e., surface temperature, 10-m wind speed, and planetary boundary layer height at ~13 × 13 km) and NACC-Cloud interpolated results (12 × 12 km) for example CMAQ LCC projection domains centered on the contiguous U.S. (**top rows**), Alaska (**middle rows**), and Hawaii (**bottom rows**). The time period shown is for 22 July 2022 at 1200 UTC.

The qualitative agreement between the original GFSv16 input and NACC-Cloud output is excellent, where the interpolated LCC domains capture both the relative large- and fine-scale temperature, wind, and PBLH gradients important for driving CMAQ. While not shown, there is a similarly excellent qualitative spatial agreement for all other 2D geospatial and meteorological variables output by NACC-Cloud.

To further quantify the performance of the interpolated NACC-Cloud (12 × 12 km) outputs vs. the original GFSv16 (~13 × 13 km) inputs, we perform a regression analysis over a month (1–30 September 2021) of hourly data for a large inner portion of the contiguous U.S. domain (Figure 8). Here we do a "nearest-neighbor" matching of the closest NACC-Cloud output and GFSv16 input grid cells and eliminate a conservative number of the outer NACC-Cloud domain grid cells from analysis (10 degrees latitude and longitude on each boundary). This helps avoid unfair errors due to global GFSv16 input grid cells that are separated by a significant distance from the NACC-Cloud outer domain during matching. Thus, the NACC-Cloud lat/lon boundary (lower left to upper right corner) analyzed in Figure 8 are from ~26.2° N/−126.7° W to 48.0° N/−63.8° W (see Figure 7 for reference). In the matching, the NACC-Cloud data are paired sequentially with each of the Gaussian grid GFSv16 "i-column" and "j-row" indices (e.g., 1,1; 2,2; 3,3; etc.), and thus the total number of paired sampled points (N) in this analysis is 130,152.

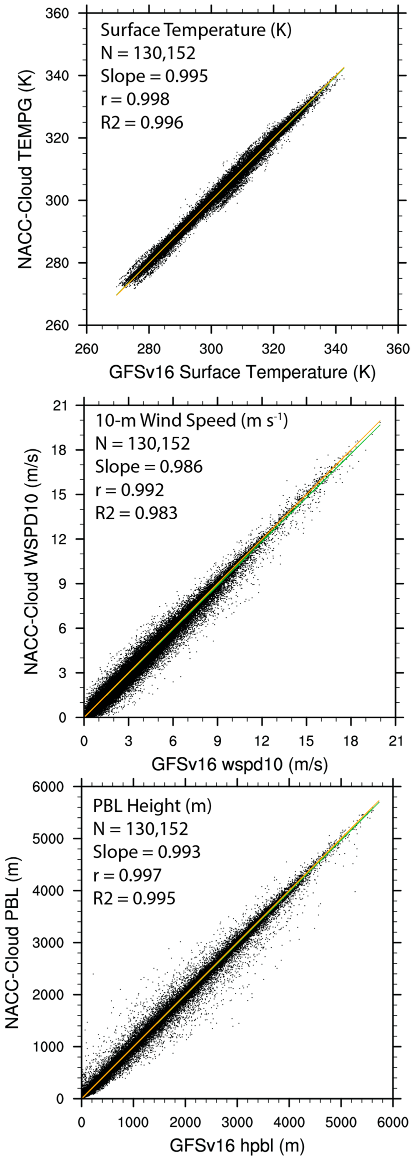


Figure 8. 1–30 September 2021 hourly X-Y regression plots between the original Gaussian GFSv16 (~13 × 13 km) inputs and interpolated NACC-Cloud (12 × 12 km) outputs (from top to bottom: surface temperature, 10-m wind speed, and planetary boundary layer height) for a subset (lat/lon corners ~26.2° N/−126.7° W to 48.0° N/−63.8° W) of the contiguous U.S. domain (see Figure 7 for reference). The calculated regression line is in green and the 1:1 line is orange. The insert statistics in each plot include the sample size (N), slope, Pearson correlation coefficient (r), and coefficient of determination (R2).

Results show an excellent agreement in hourly NACC-Cloud outputs and GFSv16 inputs for surface temperature, 10-m wind speed, and planetary boundary layer height, with a slope, Pearson correlation coefficient (r), and coefficient of determination (R2) very close to 1, and the calculated x-y regression line nearly overlapping the 1:1 line (Figure 8). Previous work also showed excellent agreement in the vertical structure of the GFSv16 input and NACC output 3D meteorological variables also necessary for driving CMAQ applications [10]. These results provide the utmost confidence in using the NACC-Cloud system for user GFS-driven CMAQ applications and facilitate a myriad of new research and science questions when coupling NOAA's GFSv16 to CMAQ for any regional domain globally.

4. Conclusions and Path Forward

Here we have described the vision and implementation of NOAA's GFS Data in the cloud for community air quality modeling (NACC-Cloud v1). This work has been made possible by the development of NACC at NOAA, which is used as the main atmosphere-chemistry coupler in the latest operational NAQFC. NACC-Cloud v1 (https://nacc.arl.noaa.gov/nacc/, accessed on 16 June 2023) has been implemented on the AWS HPC platform, and results from this work show that the NACC-Cloud system is feasible, worthwhile, relatively affordable and scalable, and the geospatial and meteorological outputs as I/O API format can be readily downloaded and used for any user-defined regional CMAQ application worldwide. Furthermore, the input global GFSv16 NetCDF files are now also publicly available on the NACC-Cloud AWS S3 Glacier IR location (see Data Availability Statement below).

While NACC-Cloud provides the user with a streamlined web app to generate operational GFSv16 meteorological inputs for CMAQ, we understand that there are numerous other inputs necessary to run CMAQ. Thus, there is a potential future goal of expanding NACC-Cloud to include more air quality modeling inputs (e.g., emissions, boundary conditions, external geospatial/land/soil inputs, etc.) and additional chemical transport model components (e.g., CMAQ, CAMx, etc.) to our AWS cloud platform, such that users can be able to run such applications more efficiently with NOAA operational products.

As the operational GFS evolves from version 16 at NOAA in the future (e.g., GFSv17 and GFSv18), there are plans to continue the NACC-Cloud project and further support such model upgrades. As noted above, NACC is also capable of processing outputs from the UFS Short Range Weather App for potential regional, high-resolution CMAQ applications. Hence there is potential for adding this UFS-SRW-App capability to NACC-Cloud for preset regional domains used in current NWS/NOAA applications (e.g., North America), or in a more robust manner using UFS capabilities and on-demand definition of the regional domain of interest in AWS. Finally, we can envision many upgrades that could be made to future versions of NACC-Cloud, which include user GUIs that provide an additional option for users (particularly those initially unfamiliar with CMAQ) to graphically select regional model domain projection parameters/configurations, rather than manually adding those in the web-app settings form (Figure 5). Any possible future developments of NACC-Cloud are contingent upon available funding.

Author Contributions: Conceptualization, P.C.C. and W.J.; methodology, P.C.C., W.J., S.Z. and Z.M.; software, P.C.C., W.J., Z.M., S.Z. and Y.T.; validation, P.C.C., W.J. and Z.M.; formal analysis, P.C.C., W.J. and Z.M.; investigation, P.C.C., W.J. and Z.M.; resources, P.C.C., W.J. and Z.M.; data curation, P.C.C., W.J. and Z.M.; writing—original draft preparation, P.C.C., W.J., Z.M. and Y.T.; writing—review and editing, P.C.C., W.J., Z.M., S.Z. and Y.T.; visualization, P.C.C., W.J., Z.M. and Y.T.; supervision, P.C.C. and W.J.; project administration, P.C.C. and W.J.; funding acquisition, P.C.C. and W.J. All authors have read and agreed to the published version of the manuscript.

Funding: This project: classified under OAR Cloud Strategic Theme 2: Cloud Compute Capabilities, was a 2021 OAR Cloud Technology Incubator (CTI) project, funded through the Weather Program Office (WPO) by the OAR Cloud Leadership Team, and co-funded by the Atmospheric Sciences and Modeling Division (ASMD) of Air Resources Laboratory (ARL). This study was also funded by the UMD/NOAA/CISESS: GMU Air Surface Exchange and Atmospheric Composition Research (Sponsor Number: 79785-Z7554202 Amend A).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All NACC-Cloud version 1.0 descriptions, data inputs, code bases, and user forms are found at the six following locations (all links last access: 16 June 2023): (1) NACC-Cloud: https://www.arl.noaa.gov/research/surface-atmosphere-exchange-home/tools-and-products/nacc-nacc-cloud/; (2) GFSv16 data on AWS: https://nacc-in-the-cloud.s3.amazonaws.com/inputs/\T1\textless{}yyyymmdd\T1\textgreater{}/ (note: use from command line only); (3) NACC codes on GitHub: https://github.com/noaa-oar-arl/NACC; (4) NACC-Cloud on AWS: https://nacc.arl.noaa.gov/nacc/; (5) NACC-Cloud Web-Based User Access: https://forms.gle/jUWFKLyY6WGKySkv6; (6) NACC-Cloud Google Form for User Feedback: https://forms.gle/ETwKS2qzoHndaruJ9.

Acknowledgments: We first acknowledge the entire NOAA-ARL research team that helped make NACC and the NAQFC a success, which was pivotal to the development of NACC-Cloud. This includes Pius Lee (retired), Barry Baker, and Rick Saylor. Furthermore, before and during the AWS HPC environment buildup used for the NACC-Cloud application, the Amazon® Web Services' NOAA Program team provided us with constructive suggestions, particularly, we would like to thank the following individuals for their help: Dev Jodhun, Senior Account Manager; Aaron Bucher, Principal Solutions Architect/HPC Specialist; Rayette Toles-Abdullah, Principal Solutions Architect; and John Kolman, Customer Solutions Manager. Meanwhile, the general AWS support team provided technical insights and solutions for some specific issues we encountered during the NACC-Cloud development. Tristan Dietz, Enterprise Architect for NOAA Research, provided coordination for hosting ARL's AWS environment within OAR's cloud system boundary and funding OAR's CTI projects including NACC-Cloud.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Skamarock, W.C.; Klemp, J.B.; Dudhia, J.; Gill, D.O.; Liu, Z.; Berner, J.; Wang, W.; Powers, J.G.; Duda, M.G.; Barker, D.; et al. *A Description of the Advanced Research WRF Version 4.3*; NCAR: Boulder, CO, USA, 2019; p. 145. [CrossRef]
2. The International GEOS-Chem User Community. Geoschem/GCClassic: GEOS-Chem Classic 14.0.2 (14.0.2). *Zenodo* **2022**. [CrossRef]
3. Zhuang, J.; Jacob, D.J.; Lin, H.; Lundgren, E.W.; Yantosca, R.M.; Gaya, J.F.; Sulprizio, M.P.; Eastham, S.D. Enabling high-performance cloud computing for Earth science modeling on over a thousand cores: Application to the GEOS-Chem atmospheric chemistry model. *J. Adv. Model. Earth Syst.* **2020**, *12*, e2020MS002064. [CrossRef]
4. Powers, J.G.; Werner, K.K.; Gill, D.O.; Lin, Y.; Schumacher, R.S. Cloud Computing Efforts for the Weather Research and Forecasting Model. *Bull. Am. Meteorol. Soc.* **2021**, *102*, E1261–E1274. [CrossRef]
5. Kang, D.; Eder, B.K.; Stein, A.F.; Grell, G.A.; Peckham, S.E.; McHenry, J. The New England Air Quality Forecasting Pilot Program: Development of an Evaluation Protocol and Performance Benchmark. *J. Air Waste Manag. Assoc.* **2005**, *55*, 1782–1796. [CrossRef] [PubMed]
6. Eder, B.; Kang, D.; Mathur, R.; Yu, S.; Schere, K. An operational evaluation of the Eta-CMAQ air quality forecast model. *Atmos. Environ.* **2006**, *40*, 4894–4905. [CrossRef]
7. Eder, B.; Kang, D.; Mathur, R.; Pleim, J.; Yu, S.; Otte, T.; Pouliot, G. A performance evaluation of the National Air Quality Forecast Capability for the summer of 2007. *Atmos. Environ.* **2009**, *43*, 2312–2320. [CrossRef]
8. Stajner, I.; Davidson, P.; Byun, D.; McQueen, J.; Draxler, R.; Dickerson, P.; Meagher, J. US National Air Quality Forecast Capability: Expanding Coverage to Include Particulate Matter. In *NATO Science for Peace and Security Series C: Environmental Security*; Springer: Dordrecht, The Netherlands, 2011; pp. 379–384. [CrossRef]
9. Lee, P.; McQueen, J.; Stajner, I.; Huang, J.; Pan, L.; Tong, D.; Kim, H.; Tang, Y.; Kondragunta, S.; Ruminski, M.; et al. NAQFC Developmental Forecast Guidance for Fine Particulate Matter (PM2.5). *Weather. Forecast.* **2017**, *32*, 343–360. [CrossRef]
10. Campbell, P.C.; Tang, Y.; Lee, P.; Baker, B.; Tong, D.; Saylor, R.; Stein, A.; Huang, J.; Huang, H.-C.; Strobach, E.; et al. Development and evaluation of an advanced National Air Quality Forecasting Capability using the NOAA Global Forecast System version 16. *Geosci. Model Dev.* **2022**, *15*, 3281–3313. [CrossRef] [PubMed]
11. Yang, F.; Tallapragada, V.; Kain, J.S.; Wei, H.; Yang, R.; Yudin, V.A.; Moorthi, S.; Han, J.; Hou, Y.T.; Wang, J.; et al. Model Upgrade Plan and Initial Results from a Prototype NCEP Global Forecast System Version 16. In Proceedings of the 2020 AMS Conference, Boston, MA, USA, 15 January 2020; Available online: https://ams.confex.com/ams/2020Annual/webprogram/Paper362797.html (accessed on 6 April 2022).
12. US EPA Office of Research and Development: CMAQ (Version 5.3.1). *Zenodo* **2019**. [CrossRef]

13. Appel, K.W.; Bash, J.O.; Fahey, K.M.; Foley, K.M.; Gilliam, R.C.; Hogrefe, C.; Hutzell, W.T.; Kang, D.; Mathur, R.; Murphy, B.N.; et al. The Community Multiscale Air Quality (CMAQ) model versions 5.3 and 5.3.1: System updates and evaluation. *Geosci. Model Dev.* **2021**, *14*, 2867–2897. [CrossRef] [PubMed]
14. Otte, T.L.; Pleim, J.E. The Meteorology-Chemistry Interface Processor (MCIP) for the CMAQ modeling system: Updates through MCIPv3.4.1. *Geosci. Model Dev.* **2010**, *3*, 243–256. [CrossRef]
15. Tang, Y.; Campbell, P.C.; Lee, P.; Saylor, R.; Yang, F.; Baker, B.; Tong, D.; Stein, A.; Huang, J.; Huang, H.-C.; et al. Evaluation of the NAQFC driven by the NOAA Global Forecast System (version 16): Comparison with the WRF-CMAQ during the summer 2019 FIREX-AQ campaign. *Geosci. Model Dev.* **2022**, *15*, 7977–7999. [CrossRef]
16. Mathur, R.; Xing, J.; Napelenok, S.; Pleim, J.; Hogrefe, C.; Wong, D.; Gan, C.-M.; Kang, D. Multiscale modeling of multi-decadal trends in ozone and precursor species across the northern hemisphere and the United States. In *Air Pollution Modeling and Its Application XXIV*; Steyn, D., Chaumerliac, N., Eds.; Springer: Cham, Switzerland, 2016; pp. 239–243. ISBN 978-3-319-24476-1.
17. Mathur, R.; Xing, J.; Gilliam, R.; Sarwar, G.; Hogrefe, C.; Pleim, J.; Pouliot, G.; Roselle, S.; Spero, T.L.; Wong, D.C.; et al. Extending the Community Multiscale Air Quality (CMAQ) modeling system to hemispheric scales: Overview of process considerations and initial applications. *Atmos. Chem. Phys.* **2017**, *17*, 12449–12474. [CrossRef] [PubMed]
18. Mathur, R.; Kang, D.; Napelenok, S.; Xing, J.; Hogrefe, C. A Modeling Study of the Influence of Hemispheric Transport on Trends in O3 Distributions over North America. In *Air Pollution Modeling and Its Application XXV*; Mensink, C., Kallos, G., Eds.; Springer Proceedings in Complexity; Springer: Cham, Switzerland, 2018; pp. 13–18. [CrossRef]
19. Rolph, G.; Stein, A.; Stunder, B. Real-time Environmental Applications and Display System: READY. *Environ. Model. Softw.* **2017**, *95*, 210–228. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.